Divulgaciones Matemáticas Vol. 16 No. 1(2008), pp. 73–86

Detección de mezclas Poisson por p-Partición

Detection of Poisson mixtures by p-Partition

Mauricio A. Bermúdez-Cella
(mbermude@euler.ciens.ucv.ve,maberce@yahoo.com.)
Laboratorios de Termocronología y Geomatemáticas, Escuela de Geología, Minas y Geofísica.Facultad de Ingeniería, Universidad Central de Venezuela, 1053. Ciudad Universitaria, Caracas-Venezuela.

Resumen

En este artículo se exponen los resultados de aplicar el proceso de p-Partición a mezclas simuladas de Poisson. Se explora la p-Partición como estimador del número de subpoblaciones presentes en la mezcla y como test para decidir si una muestra tiene una o más poblaciones.
Palabras y frases clave: distribución de Poisson, mezclas, p-partición, cubrimiento por intervalos, cubrimiento por conjuntos discretos, pruebas de hipótesis, huellas de fisión, edad.

Abstract

In this paper we apply a p-partitions methodology to simulated Poisson mixtures. The p-partitions are explored as a mean of estimating the number of suppopulations in the mixture and as a test for deciding whether the sample has one or more subpopulations.
Key words and phrases: Poisson distribution, mixtures, p-partition, coverage by intervals, coverage by discrete sets, hypothesis tests, fission tracks, age.

1 Introducción

Una huella de fisión [3] es una traza que se produce al ocurrir una fisión de los átomos de uranio-238 presentes en minerales accesorios constituyentes de rocas como el apatito, circón, etc. El número o contaje de huellas de fisión

Recibido 2006/02/28. Revisado 2006/07/03. Aceptado 2006/07/15.
MSC (2000): Primary 62H30; Secondary 62H12.

de un grano de roca, es considerado una variable aleatoria de Poisson cuya media depende de la densidad de uranio del grano en que se hace el contaje. El método del detector externo para la determinación de la antiguedad de una roca, utiliza el contaje de las huellas de fisión, sobre varios granos de la roca ([6]). El hecho de que la densidad de uranio en los granos de una misma roca no sea constante hace que sea poco realista considerar siempre, los contajes hechos como provenientes de una sola población de Poisson. Además, el hecho de que provengan de diversas poblaciones de Poisson puede afectar la calidad del estimado del tiempo. Por ello se han desarrollado métodos estadísticos que toman en cuenta la posible variación de densidades de uranio ([1]). Esto ha llevado, de manera natural, a considerar el problema de separación de poblaciones en mezclas de Poisson. Al menos en la literatura matemática de geología durante el desarrollo de esta investigación no se encontró trabajos en esa dirección. En este artículo se presentan algunos resultados iniciales de separación de mezclas. El concepto clave es el de p-Partición. Se explora dicho concepto como estimador del número de poblaciones de una mezcla y como test para discriminar entre muestras unipoblacionales y muestras bipoblacionales.

2 Definición de una p-Partición

Considérese p un número entre 0 y 1, y sea:

$$cuantil(p,k) = \inf_q \{q \in \mathbb{N}/P(x \leq q) \geq p\},$$

donde x es una variable aleatoria Poisson de media k. Dado $0 \leq p \leq 1$, la p-Partición de un conjunto finito de números enteros positivos D se hace en dos etapas. En primer lugar se define un p-Cubrimiento, luego se "disjuntiza.^{es}e p-Cubrimiento para obtener la p-Partición.

Ahora bien, para construir el p-Cubrimiento, se considera:

$$k_i = \text{mín}\,\{D_i\},$$

y se define el intervalo o clase C_i como:

$$C_i = [k_i; max(k_i, cuantil(p,k_i))]$$

Tomando:

$$D_{i+1} = D_i - C_i,$$

y:

$$D_1 = D.$$

Se obtiene el p-Cubrimiento $\{H_i\}$ de D al considerar:

$$H_i = C_i \bigcap D.$$

Se designa con m_i el promedio empírico de las observaciones en C_i y $P_{m_i}(x)$ la probabilidad de que la variable de Poisson X de media m_i valga x.

La sucesión dada por los m_i es creciente entre $min(D)$ y $max(D)$. En virtud de que los $\{H_i\}$ no necesariamente son disjuntos, se establece la siguiente regla de decisión:

$$\begin{aligned} S_i &= \{x \in H_i / (x \notin H_j, i \neq j)\} \\ &\quad \cup \left\{x \in H_{i+1} \cap H_i / P_{m_i}(x) > P_{m_{i+1}}(x)\right\} \\ &\quad \cup \left\{x \in H_{i-1} \cap H_i / P_{m_i}(x) > P_{m_{i-1}}(x)\right\} \\ &\quad \cup \left\{x \in H_{i+1} - H_i / P_{m_i}(x) = P_{m_{i+1}}(x)\right\} \end{aligned}$$

3 Estudio de la p-Partición como estimación del número de subpoblaciones de una mezcla de Poisson

Utilizando el paquete estadístico S^+ se realizaron diversas simulaciones con la finalidad de estudiar la capacidad de la p-Partición para estimar el número de subpoblaciones presentes en una mezcla Poisson. Se simularon N poblaciones Poisson $\{G_i\}_{i=1,...,N}$, cada una con medias m_i respectivamente para $N = 2; 3; ...; 7; 9$. Los resultados que se exponen a continuación son el producto de más de 500 corridas, este número resultó ser estable después de haber probado con 20, 100, 500 y 1000 corridas. La cercanía de las medias de las subpoblaciones afecta, en general, la calidad de los algoritmos de separación de muestras. En el caso de las distribuciones de Poisson la incidencia de este factor se complica por el hecho de que la varianza de la distribución está siempre relacionada con la media.

Las simulaciones fueron realizadas asumiendo que la separación entre dos medias consecutivas es constante. Esto reduce considerablemente el espectro de simulaciones. Se hicieron algunas simulaciones con variantes de esta condición y los resultados que se obtienen no se diferencian de aquellos obtenidos con esta condición restrictiva. Considérese delta (δ) la separación entre dos medias consecutivas, a continuación se presentan los resultados de las simulaciones para dos casos: medias separadas y cercanas.

El objetivo de estas simulaciones al igual que el de las dos secciones siguientes es establecer un rango de probabilidades p o una tabla de valores de p ideales que pueda ser usada por los futuros usuarios de este método con el fin de hacer una discriminación efectiva de las diferentes subpoblaciones existentes en un conjunto de datos provenientes de una mezcla de varias distribuciones Poisson.

3.1 Primer Caso: Medias separadas ($\delta \geq 20$).

Medias separadas significa distancia entre dos medias consecutivas ≥ 20. Los parámetros para la primera simulación son los siguientes:

- Media inicial: $m_1 = 10$.
- Número de subpoblaciones consideradas a priori: $N = 3$.
- Número de elementos en cada grupo: $n_i = 10$.
- El valor de probabilidad a priori: $p = 0{,}90$.

Al considerar valores a priori de probabilidad $p < 0{,}90$ se observó que no existían coincidencias entre la cantidad de subpoblaciones supuestas a priori ($N = 3$) y las estimadas por la p-Partición, es decir, al fijar un valor $p < 0{,}90$ la p-Partición estimaba más de 3 subpoblaciones. Los resultados para este caso en particular y para valores de p de 0.95, 0.99 y 0.9999 son mostrados en la tabla 1. En esta última se aprecia que a medida que el nivel p es incrementado aumenta el porcentaje de coincidencias entre el número de subpoblaciones consideradas a priori y las discriminadas por la p-Partición. Para $p = 0{,}95, 0{,}99$ y $0{,}9999$, los porcentajes de coincidencias son: 6.2, 51 y 98.2 % respectivamente.

p	**%Coinc.**
0.90	0.8
0.95	6.2
0.99	51
0.9999	98.2

Cuadro 1: Porcentaje de coincidencia entre las subpoblaciones supuestas ($N = 3$) y el número de grupos de la p-Partición para diferentes valores de p en el caso de medias separadas.

Al incrementar el número de subpoblaciones a $N = 4{,}5{,}6$, y 7 los resultados son similares a los expuestos en la tabla 1 para N=3. Estos resultados

son resumidos en la figura 1 considerando valores de p entre 0.90 y 0.9999. Mientras mayor es el número de poblaciones que conforman la mezcla, en el caso de que la media de cada población esté bastante alejada con respecto a la otra población se observa que el valor máximo de coincidencia es cuando se selecciona una probabilidad de 0.9999.

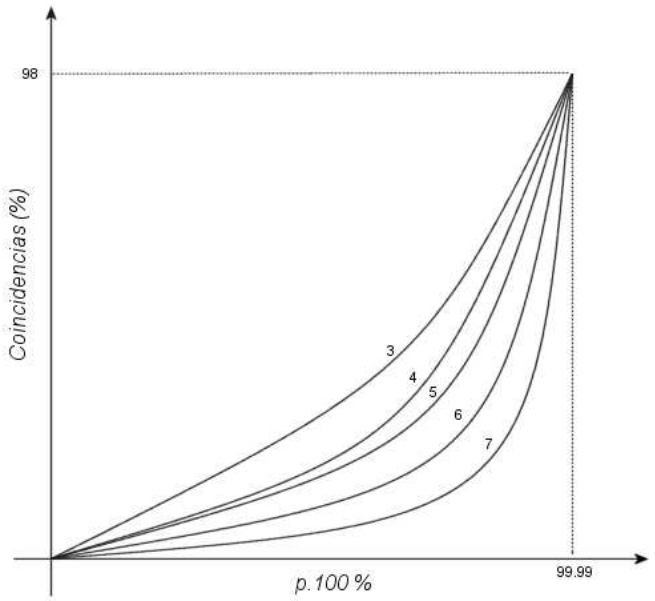

Figura 1: Porcentaje de coincidencias entre el número de subconjuntos generados por simulación y el número de poblaciones separadas por la p-Partición en el caso de medias distantes.

3.2 Segundo Caso: Medias cercanas ($\delta < 20$).

Dos subgrupos se dicen que tienen medias cercanas cuando $\delta \leq 20$. En este caso particular se tomó $\delta = 10$. En analogía con el caso anterior se asumió que m_1=10 y n_i=10. Obteniéndose para el caso en que el número de grupos considerados a priori es $N = 3$, los resultados mostrados en la tabla 2. En esta última se aprecia que cuando las medias de los 3 subgrupos están muy cercanas la probabilidad necesaria para separar "correctamente" es menor, situándose su valor entre 0.95 y 0.98.

p	**%Coinc.**
0.90	21
0.95	74.8
0.96	80.4
0.97	81.6
0.98	69.4
0.99	43
0.9999	0

Cuadro 2: Porcentaje de coincidencia entre las subpoblaciones supuestas ($N = 3$) y el número de grupos de la p-Partición para diferentes valores de p en el caso de medias cercanas.

Cuando la población está conformada por N=5 se obtuvieron los resultados mostrados en la tabla 3.

p	**%Coinc.**
0.90	41.4
0.91	54.8
0.92	64.6
0.93	72.8
0.94	71.2
0.95	51.8
0.96	31.8
0.97	15
0.98	5.6
0.99	0.4
0.9999	0

Cuadro 3: Porcentaje de coincidencia entre las subpoblaciones supuestas ($N = 5$) y el número de grupos de la p-Partición para diferentes valores de p en el caso de medias cercanas.

Como se puede apreciar en la tabla 3, en el caso de que existan 5 subgrupos, el valor de p ha descendido al rango comprendido entre 0.93 y 0.95. Este rango se mantuvo para un número de simulaciones de $20, 50, 100$ y 1000.

En el caso que existan 7 subpoblaciones en la tabla 4 se muestran los resultados obtenidos, para un número de 1000 corridas. Puede observarse que cuando el número de subgrupos que conforman la mezcla es 7, el rango óptimo de separación es alcanzado para p entre 0.90 y 0.93.

p	**%Coinc.**
0.90	60.6
0.91	66.2
0.92	62.3
0.93	47.4
0.94	27.1
0.95	10
0.96	2.7
0.97	0.3
0.98	0
0.99	0
0.9999	0

Cuadro 4: Porcentaje de coincidencia entre el número de subpoblaciones supuestas ($N = 7$) y el número de grupos de la p-Partición para diferentes valores de p en el caso de medias cercanas.

Los resultados para todas estas simulaciones 20,50,100,500 y 1000 corridas cuando las medias de cada una de las poblaciones $N = 3, 4, 5, 6$, y 7 están poco separadas pueden ser resumidas en la figura 2.

Observando la figura 2, se concluye que cuando las poblaciones tienen medias muy cercanas entre sí, para que coincidan el número de poblaciones separadas por la p-Partición con el número de subpoblaciones utilizadas para las simulaciones es necesario disminuir el valor de p. También se aprecia que a medida que la cantidad de poblaciones aumenta el porcentaje de coincidencia disminuye. Por ejemplo, cuando $N = 4$, el máximo de coincidencias entre las poblaciones supuestas a priori y las discriminadas por la p-Partición es 86 % y se alcanza en $p = 0{,}948$ mientras que si $N = 7$, el máximo de coincidencias es 61 % y se alcanza en $p = 0{,}905$.

Todas las simulaciones realizadas considerando diferentes números de subpoblaciones: $N = 3, 4, 5, 6$, y 7, junto con las figuras 1 y 2 permiten tener una idea del intervalo donde se encuentra el valor p a utilizar, dependiendo únicamente del grado de separación entre las medias de las subpoblaciones. Esta intervalo donde se encuentra p se denominará el rango de probabilidades p. A partir de la gran cantidad de simulaciones realizadas en esta investigación, puede establecerse un posible rango de probabilidades p de manera que el porcentaje de coincidencia entre el número de subpoblaciones supuestas a priori y las separadas por la p-Partición sea máximo, estos rangos son mostrados en la tabla 5. En la tabla 6 se resumen las observaciones antes realizadas, esta tabla

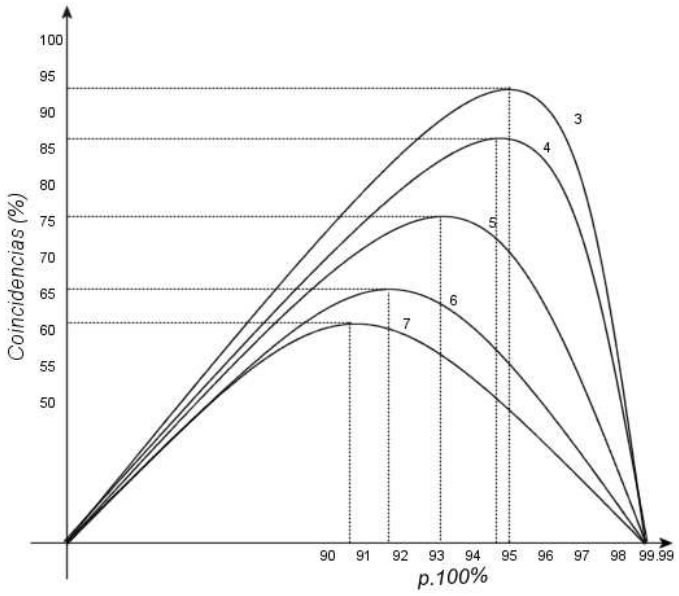

Figura 2: Porcentaje de coincidencias entre el número de subconjuntos generados por simulación y el número de poblaciones separadas por la p-Partición en el caso de medias cercanas.

proporciona una orientación precisa al momento de realizar las aplicaciones, pero es necesario tener una "idea.º conocimiento a priori acerca del grado de se-paración entre las medias de los subgrupos.

4 Pruebas de hipótesis aplicadas a la p-Partición de una población

Al aplicar la p-Partición a una población se puede no separar una muestra en dos subpoblaciones a pesar de que esta esté compuesta por observaciones que provienen de dos poblaciones poissonianas diferentes. O bien se puede separar una muestra en dos subpoblaciones a pesar de que la muestra provenga de una sola distribución de Poisson. Esta sección muestra estimados de las probabilidades de cada uno de los dos posibles errores señalados, obtenidos

# de Pob.	**Int.Prob.p**
3	$\approx (0{,}95, 0{,}97)$
5	$\approx (0{,}93, 0{,}95)$
7	$\approx (0{,}90, 0{,}93)$

Cuadro 5: Rango de confiabilidad para mezclas formadas por 3, 5 y 7 subpoblaciones con medias cercanas.

Distancia entre poblaciones: δ	**Niveles óptimos de** p
$\delta > 20$ (lejanas)	$0{,}99 \leq p < 1$
$15 \leq \delta \leq 20$ (intermedias)	$0{,}96 < p < 0{,}99$
$10 \leq \delta < 15$ (cercanas)	$0{,}94 < p \leq 0{,}96$
$5 < \delta < 10$ (muy cercanas)	$0{,}9 < p \leq 0{,}94$

Cuadro 6: Resumen de los valores óptimos a usar en la p-Partición según la distancia entre medias de subpoblaciones para diferentes valores de m_1.

mediante simulaciones.

La significancia de este test guarda relación con el p de la p-Partición. Antes de definir el test estadístico construido para esta situación, lo primero que se realizó fue buscar el valor óptimo de p, para el cual al introducir una población proveniente de una sola Distribución de Poisson no fuese dividido por en dos poblaciones, siendo este $p = 0.999999$, muy cercano a 1.

4.1 La p-Partición utilizada como test de hipótesis de que la muestra tiene una o más de una población.

Al aplicar la p-Partición como un test de hipótesis se conservan las características de las pruebas de hipótesis ([5]), con la diferencia que en lugar de discriminar parámetros de una población como la media y la varianza, discrimina hasta que punto al introducir la unión de dos grupos provenientes de dos distribuciones de Poisson distintas, la p-Partición en realidad separe estas poblaciones en una o dos. El buen ajuste del test depende del nivel de significancia p.

Para definir los parámetros que la p-Partición utilizada como test de hipótesis debe discriminar, se considera la siguiente matríz:

Sea:

$$M = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$$

donde:

$x_{11} + x_{12}$ es el número de veces que se genera una muestra de Poisson con dos poblaciones diferentes .

$x_{21} + x_{22}$ es el número de veces que se genera una segunda población de Poisson.

En esta matriz:

x_{11} indica el número de veces que al aplicar la p-Partición esta constó de un sólo grupo (el número de veces que no coincide con el número de poblaciones de la muestra).

x_{12} indica el número de veces que la p-Partición indica que hay más de un grupo.

x_{21} indica el número de veces que la p-Partición indica que hay un sólo grupo y

x_{22} indica el número de veces que la p-Partición indica que hay dos grupos (el número de veces que no coincide con el número de poblaciones de la muestra).

Lo que se desea al utilizar la p-Partición como test de hipótesis es que los valores de x_{12} y x_{21} sean máximos y al compararlos entre sí se pueda tomar una decisión sobre si las muestras están formadas por una sola población o por más de una población de acuerdo al significado de las componentes de la matriz M dado anteriormente. Por ejemplo, supóngase que se generan 100 muestras Poisson (es decir: $x_{11} + x_{12}$=100 y $x_{21} + x_{22}$=100) y que la matríz M está dada por:

$$M = \begin{pmatrix} 10 & 90 \\ 90 & 10 \end{pmatrix}$$

Como x_{12}=90 esto significa que con un 90 % de probabilidad cuando se introdujo una muestra formada por varios subgrupos Poisson, la p-Partición detectó una mezcla de poblaciones. Ahora bien, la otra componente x_{21}=90, indica que cuando se introdujo la muestra como una sóla población la p-Partición identifica en un 90 % que no hay mezclas de poblaciones. Para formalizar esto, es necesario pensar el número de subgrupos dados por una p-Partición, en términos de test de hipótesis y asumiendo que:

H_0: la muestra está constituida por dos subpoblaciones.

H_1: la muestra está constituida por una subpoblación.

Así:

$$\frac{x_{11}}{x_{11} + x_{12}}$$

es un estimado de la probabilidad del error de Tipo I del test y

$$\frac{x_{22}}{x_{21} + x_{22}}$$

es un estimado de la probabilidad del error de Tipo II del test.

Un modo de optimizar el test es maximizando la función:

$$f(M) = (x_{12} + x_{21}) - (x_{11} + x_{22})$$

Se probó optimizar con otras funciones y los resultados varían muy poco.

4.2 Resultados de las simulaciones

A lo largo de las simulaciones, se procedió generando siempre muestras de 20 elementos. En el caso de que la muestra conste de dos subpoblaciones, cada una de las subpoblaciones generadas tenía 10 elementos. Además se consideró:

$$x_{11} + x_{12} = x_{21} + x_{22} = 100.$$

Los resultados de este test, son mostrados según el grado de separación entre las distancias de las medias de las subpoblaciones. En el caso de que las medias (Poisson) de las subpoblaciones están muy cercanas, es decir, medias de 8 y 10 respectivamente, el intervalo donde se encuentra el nivel p-óptimo es el (0.995,0.996), la función de maximización establece un p-óptimo de 0.995611, obteniéndose que el 58 % de las veces en que hay dos grupos, el test discrimina 2 grupos, mientras que el 67 % de las veces en que hay un sólo grupo el test discrimina un sólo grupo. En este caso el error de tipo II es 0.33 y el error de tipo I es 0.42. Todas estas observaciones son resumidas en la tabla 7. En estas situaciones de separación de medias tan cercanas el test tiene un desempeño pobre.

Media de G_1	m_1=8
Media de G_2	m_2=10
Intervalo donde se encuentra el p-óptimo	[0,995, 0,996]
p-óptimo	0.995611
Matriz M	$\begin{pmatrix} 42 & 58 \\ 67 & 33 \end{pmatrix}$
Valor de la función $f(M)$	50

Cuadro 7: Resultados de la p-Partición como test de hipótesis en el caso de m_1=8 y m_2=10

Fijando m_1=8 y tomando m_2 =10;12;...;20, se estudiaron 6 casos, el último de ellos es resumido en la tabla 8. En este caso, el intervalo donde se alcanza el máximo está entre [0.998, 0.99999], el nivel p-óptimo es de 0.999777. Para

este caso particular, si se observa los valores estadísticos de los errores de tipo I y II se aprecia que la probabilidad de que ocurran estos errores es pequeña: 1 %, en ambos casos. El valor de la función donde el máximo es alcanzado es de 196, lo cual representa una buena aproximación.

Media de G_1	m_1=8
Media de G_2	m_2=20
Intervalo donde se encuentra el p-óptimo	[0,998, 0,9999]
p-óptimo	0.999777
Matriz M	$\begin{pmatrix} 1 & 99 \\ 99 & 1 \end{pmatrix}$
Valor de la función $f(x)$	196

Cuadro 8: Resultados de la p-Partición como test de hipótesis en el caso de m_1=8 y m_2=20

Análogamente, se hicieron otras pruebas tomando medias mayores, con los siguientes criterios: se tomaba como media del primer subgrupo m_1, y $m_2 = m_1 + 2i$ con $i = 1 \ldots n$ hasta que m_2 fuera el doble de m_1. En la figura 3, se resumen los 8 casos considerados, con medias m_1 de 8, 14, 18, 20, 30, 40, 50 y 60. El eje x representa la distancias entre las dos medias de las dos poblaciones consideradas, el eje y representa la altura del valor de $f(M)$ obtenido con el p-óptimo. Mientras más cerca estén las medias el valor del p-óptimo será más cercano a 0.99, mientras que a mayor distancia entre las medias el valor de p se acerca mucho a 1, es decir, p=0.999999. La noción de cercanía y lejanía para medias grandes cambia con respecto a medias pequeñas.

En la mayoría de las simulaciones se observó que al aumentar considerablemente m_1, debido a que la varianza de una Distribución de Poisson esta relacionada con su media $\overline{m}$ mediante:

$$\theta^2(m) = \overline{m}$$

esta varianza también se incrementaba, y la p-Partición tiende a no separar eficazmente poblaciones con medias çercanas". Esto es cierto para el caso en que $m_1 > 100$.

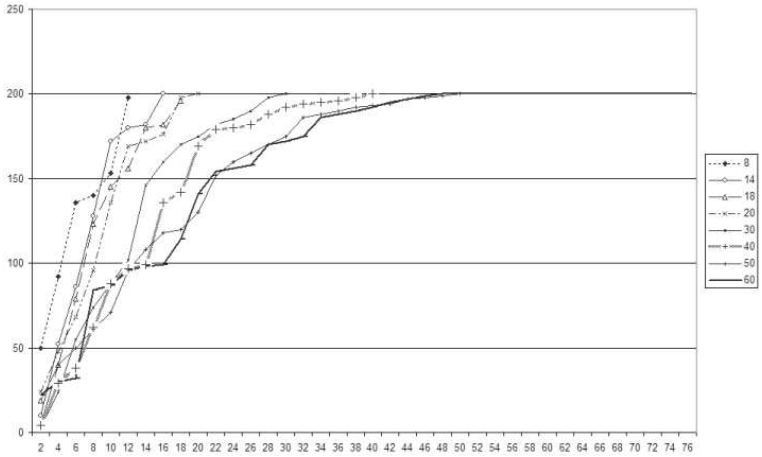

Figura 3: Resultados de la p-Partición para diferentes distancias entre medias.

5 Agradecimientos

Este artículo es parte de [1] y [2], trabajos dirigidos por Pedro Alson a quien le agradezco sus comentarios y sugerencias para la versión final de este artículo.

6 Conclusiones

El algoritmo descrito en este artículo fue aplicado a varios de los conjuntos de datos que motivaron este estudio. Esto permitió:

1) Hacer un estudio más detallado sobre la estructura de ciertos datos y compararlos con los resultados que se obtienen con el test de Galbraith ([4]). Se pudo constatar ([1], [2]), que en numerosos casos el test de Galbraith es inefectivo, mientras que el test, basado en la p-Partición, sí detecta distanciamiento de la hipótesis nula.

2) Mostrar que el problema físico de tener varias concentraciones de uranio en muestras si afecta la calidad del estimado en la edad proporcionada por el método de huellas de fisión y que la influencia de una distribución heterogénea de uranio no había sido resuelto por la técnica del detector externo como se pensó desde 1981 (ver [6]).

3) La aplicación de la p-Partición a los datos de huellas de fisión lleva a la

noción de p-Patrones ([2]) que permite mejorar la calidad de la estimación de la edad de las rocas para los casos de datos en donde la distribución heterogénea del uranio afecta la calidad de la estimación de la edad.

4) Permitió definir un nuevo estimador ([1], [2]) de la edad proporcionada por el método de huellas de fisión que es mucho más eficiente que los estimadores convencionales para este fin.

Las rutinas de S^+ que sirvieron para este estudio pueden ser suministradas por el autor si le son solicitadas.

Referencias

[1] Bermúdez C., Mauricio A. *Estudio de métodos estadísticos para la datación de material rocoso utilizando huellas de Fisión.* Tesis de Maestría, Escuela de Matemáticas, Universidad Central de Venezuela, 92 p., 2002.

[2] Bermúdez C., Mauricio A. *Aspectos estadísticos de la datación de eventos tectotérmicos por el método de huellas de Fisión.* Trabajo de Ascenso, Escuela de Geología, Minas y Geofísica, Universidad Central de Venezuela, 124 p., 2005.

[3] Bermúdez, M., Alson, P., y Mora, J. *Ecuación Fundamental de la Edad para la datación de minerales y su adaptación a la ecuación práctica para el método de huellas de fisión.* Revista de la Facultad de Ingeniería, 2005, Volumen 20, N° 2. Caracas, Venezuela.

[4] Galbraith, R.F. *On Statistical Models for Fission Track Counts.* Math. Geol., 1981, Vol. 13, No. 6, USA.

[5] Mendenhall, W., Scheaffer, R., and Wackerly, D. *Mathematical Statistics with Applications.* Second Edition, Duxbury Press, Boston-Massachusetts, 686 p., 1981.

[6] Wagner G. and Van Den Haute P. *Fission Track Dating.* Solid Earth Sciences Library, Kluwer Academic Publishers. Netherlands, 285 p., 1992.