

HISTOGRAMAS REPARTIDOS PARA EL ANÁLISIS DE MUESTRAS ESTADÍSTICAS ESCASAS II: COTAS RÍGIDAS, ROSAS Y MAPAS

F. Alejandro Nava

Depto. de Sismología, División de Ciencias de la Tierra, CICESE

A.P. 2732, Ensenada, 22800, B.C., México

E-mail: fnava@cicese.mx

RESUMEN

Se presenta la extensión del método de histogramas repartidos para el análisis de muestras escasas (Nava, 1998), a histogramas con cotas rígidas, histogramas de rosa e histogramas en 2D como son los mapas que presentan valores de la distribución de alguna observable en celdas (o clases) de 2D. Como para los histogramas convencionales, el método reduce la dependencia de los histogramas en el tamaño y la localización de las celdas; también se encuentra, empíricamente, que los histogramas producidos por este método pueden, en general, aproximar la distribución madre de procesos naturales mejor que los histogramas comunes.

INTRODUCCIÓN

La motivación y las bases del método de histogramas repartidos han sido expuestos en Nava (1998), pero aquí se resumirán brevemente de manera que este artículo pueda entenderse sin necesidad de consultar el anterior.

El problema en cuestión es tratar de identificar o estimar la distribución estadística $p(v)$ (*distribución madre*) de una población V que tiene M elementos (M puede ser infinito), a partir de una muestra $\{v_j\}$, donde cada v_j es un dato (producto de la observación) identificado por el índice $j=1,2,\dots,N$ y N es el número total de datos. Conocer la distribución $p(v)$ permite hacer inferencias probabilísticas respecto a la ocurrencia de eventos de V .

Las muestras se analizan muy a menudo mediante histogramas que consisten en la representación del número n_i de elementos de la muestra cuyos valores están contenidos dentro de un intervalo que abarca de $v_0 + i \Delta v$ a $v_0 + (i+1) \Delta v$, donde Δv es el tamaño de los intervalos de clase, o *celdas*, v_0 es un valor de referencia que indica el comienzo de la i ésima celda, e i es un entero que toma los valores necesarios para que las celdas cubran (por lo menos) toda la extensión de los N valores observados. Conforme crece el tamaño de la muestra y la razón $N/M \rightarrow 1$ el histograma se aproxima cada vez mejor a la distribución de la población. Una muestra abundante permite el uso de Δv pequeña y el histograma se aproxima a la distribución.

Desafortunadamente, no siempre es posible conseguir una muestra abundante y cuando la muestra es escasa la forma del histograma depende críticamente de la elección del tamaño y la posición de los intervalos y no se parecerá necesariamente a la distribución madre.

En Nava (1998) se propuso una variante del conocido método de convolucionar una distribución con una función de error (e.g. Silvermann, 1986; Fisher, 1989). El método consiste en repartir cada muestra (no, como se hace comúnmente, repartir el número de muestras en cada celda una vez hecho el histograma), considerada con área unitaria, sobre una extensión centrada en el valor medido y de acuerdo con alguna distribución de reparto dada, de manera que la contribución de la muestra a la celda en que cae su valor y a las celdas cercanas a ésta depende del área de la distribución de reparto que cae dentro de cada celda. En el artículo citado, se mostró que la dependencia de los histogramas repartidos en el punto de referencia y la dimensión de las celdas (siempre que éstas sean angostas con respecto a la función de repartición) es casi nula y se encontró empíricamente que los histogramas repartidos representan mejor a la distribución madre que los histogramas usuales.

COTAS RÍGIDAS

Es obvio que la convolución de los impulsos que representan a los datos con una distribución de error razonable (esto es, no impulsiva) aumenta la extensión que cubren dichos datos. Por ejemplo, si los datos están representados por las líneas discontinuas verticales de la figura 1, que cubren la extensión de $V_{\min} = 5^\circ$ a $V_{\max} = 90^\circ$, y consideramos $\Delta v = 1^\circ$ y una función de error normal con desviación estándar $\sigma = 2^\circ$, entonces, tras la convolución, tendremos celdas con valores distintos de cero (para cinco cifras decimales de aproximación) entre $V_n = -18^\circ$ a $V_x = 113^\circ$ (línea delgada Figura 1). Este resultado sería perfectamente aceptable si estuviéramos trabajando con datos de, digamos, azimuths o temperaturas; pero es absurdo si se trata, como en este ejemplo, de datos de echados de falla (e.g. Frez et al., 1999), cuyos valores aceptables son, por definición, de $V_1 = 0^\circ$ a $V_2 = 90^\circ$. Esto es, si tenemos una incertidumbre en una medición cercana a una cota rígida el error posible no puede

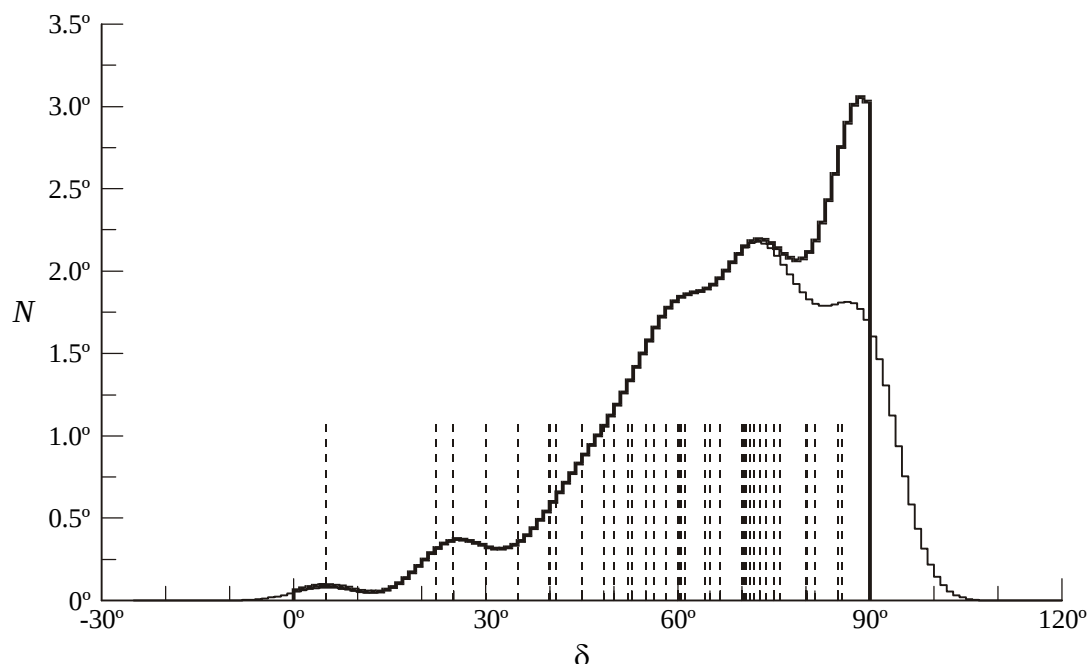


Figura 1. Histograma repartido de los datos de echado δ , representados por líneas discontinuas verticales, que cubren de $V_{min} = 5^\circ$ a $V_{max} = 90^\circ$. Para $\Delta v = 1^\circ$ y función de error normal con $\sigma = 2^\circ$, el histograma (línea delgada) abarca de $V_n = -18^\circ$ a $V_x = 113^\circ$. La línea gruesa muestra el efecto de considerar 0° y 90° como cotas rígidas.

distribuirse a través de la cota y es necesario tomar esto en cuenta al “repartir” los datos sobre las celdas. Naturalmente, si para cada valor se conoce la función de error apropiada que lo distribuya sin rebasar la cota rígida, función que será en general distinta para cada valor cercano a la cota, entonces se puede aplicar directamente la función apropiada a cada valor (lo cual puede ser complicado desde el punto de vista computacional). Pongamos como ejemplo el caso de mediciones de echados cercanos a los 90° por geólogos: si alguien hiciera un estudio estadístico para conocer la función de error y ésta fuera aplicable a las mediciones de cualquier geólogo, entonces se podría aplicar ampliamente esta función al estudio de fallas. Desafortunadamente, ésta y muchas otras funciones no son conocidas y sin embargo es necesario tomar en cuenta de alguna manera la incertidumbre en ellas.

Consideramos aquí dos métodos para tratar el caso de barreras rígidas mediante aplicación de la misma función de error utilizada para los demás datos. El primer método consiste en considerar un plegamiento (semejante al *faltung* que ocurre alrededor de la frecuencia de Nyquist en el análisis de Fourier) de la distribución de error alrededor de la cota rígida, según se ilustra en la figura 2a, donde la flecha indica la posición de un dato, la línea delgada es la repartición original con la parte que sobrepasa a la cota plegada y la línea gruesa es la resultante del plegamiento. Este tratamiento sesga los posibles errores, dando mayor probabilidad a valores entre el valor medido y la cota que al propio valor medido, lo cual no parece razonable.

El método elegido en este trabajo consiste simplemente en truncar la repartición en la cota rígida y renormalizar la repartición truncada a área unitaria, de manera que los valores no permitidos se redistribuyen proporcionalmente a la distribución ori-

ginal. Este tratamiento se ilustra en la figura 2b, donde la línea delgada es la distribución original y la línea gruesa es la distribución truncada y renormalizada. El resultado de la aplicación de este método al análisis de los datos del ejemplo se muestra como línea gruesa en la figura 1.

ROSAS

El histograma de rosa es una herramienta bien conocida para el análisis de datos con extensiones cíclicas (e.g. Fisher, 1989, 1993), como, por ejemplo, orientaciones de trazos de falla, azimuths de orientaciones de magnetismo en las rocas, horas del día, etc., donde las cotas inferior y superior coinciden en una única cota que no es rígida. Los histogramas de rosa presentan los mismos problemas que los histogramas comunes (no cíclicos), por lo que puede aplicárseles la técnica de repartición.

El único problema surge cuando hay datos cercanos a las cotas, de manera que dichos datos son repartidos entre celdas que sobrepasan las cotas. Por ejemplo, en el histograma de rosa con extensión de 0° a 360° de la figura 3, se muestra la repartición de dos datos indicados por flechas: uno con valor 165° que se reparte entre 142° y 188° sin problema alguno; otro con valor 5° que se reparte entre -18° y 28° , y las celdas del intervalo -18° a 0° (sombreadas) corresponden naturalmente a las de 342° a 360° . De la misma manera, la repartición de datos con valores levemente inferiores a 360° podrá afectar celdas para valores cercanos a 0° , pero asignar la repartición a las celdas apropiadas es sólo cuestión de teneduría de libros y no es problema en realidad.

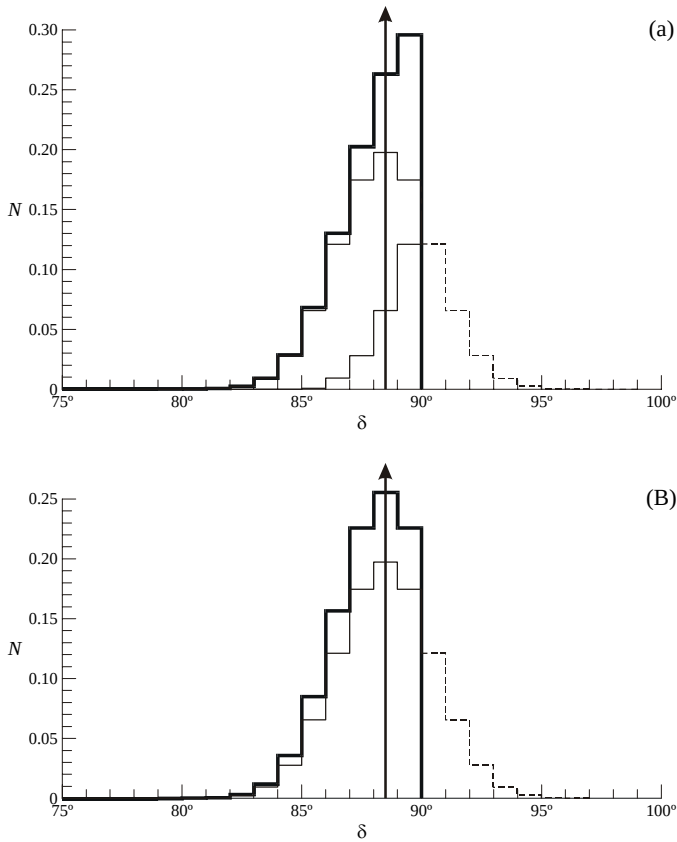


Figura 2. Dos diferentes maneras de hacer la repartición de un dato, indicado por una flecha y repartido originalmente según la línea delgada, en presencia de una cota rígida en $\delta=90^\circ$. La parte de la repartición original que sobrepasa la cota se indica por la línea punteada. En (a) esta sobrante es plegada sobre ésta y el resultado del plegamiento se muestra como la línea gruesa. En (b) la sobrante es truncada y el restante ha sido normalizado a área unitaria según muestra la línea gruesa.

Podemos considerar dos tipos de valor cíclico, a los que llamaremos uni- y bidireccional. El tipo unidireccional corresponde a datos cuyos valores están definidos unívocamente dentro de la extensión considerada; por ejemplo: azimuts de orientaciones magnéticas (0° a 360°), hora de ocurrencia dentro de un día (0 a 24 hrs), etc. El histograma de rosa unidireccional se ilustra en la figura 4 que muestra la distribución, sobre cada día, de los tiempos de origen t_0 de deslizamientos verticales en el extremo sur de la falla Imperial (Nava y Glowacka, 1999). Los datos son indicados por líneas discontinuas. Las celdas tienen $\Delta t = 0.05$ días y el largo de cada una corresponde al número repartido (mediante una gaussiana con $\sigma = 4\Delta t$) de eventos en ella, según la escala radial N , de manera que el área total contenida por el histograma es el número total de eventos.

El tipo bidireccional corresponde a datos cuyos valores tienen una ambivalencia de medio ciclo. Como ejemplo tenemos a las direcciones de falla expresadas como azimuts en la extensión de 0° a 360° ; como no se asocia un sentido determinado a las fallas, azimuts de 45° y de 225° describen al mismo rumbo NE. Como ilustración del ejemplo presentaremos los azimuts de plano de falla, obtenidos de los mecanismos focales, para

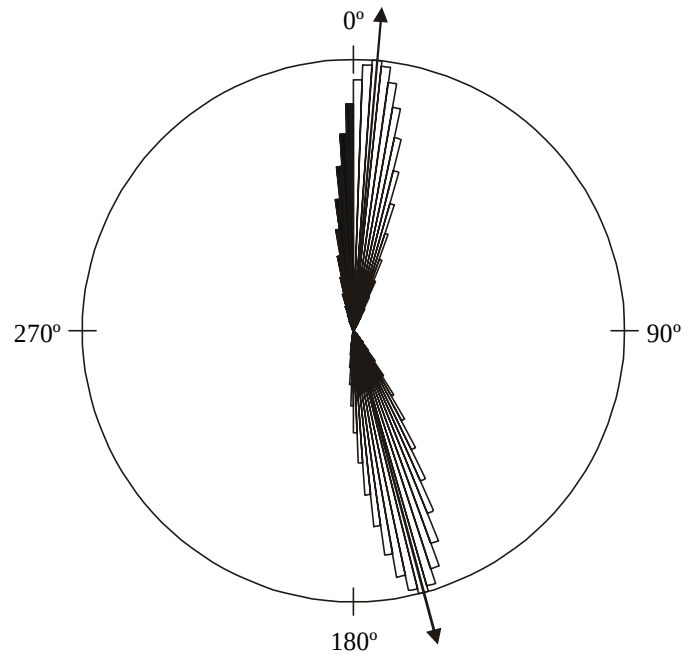


Figura 3. Repartición de dos datos indicados por flechas. El dato con valor 5° se reparte entre -18° y 28° , y las celdas del intervalo -18° a 0° (sombreadas) corresponden naturalmente a las de 342° a 360° .

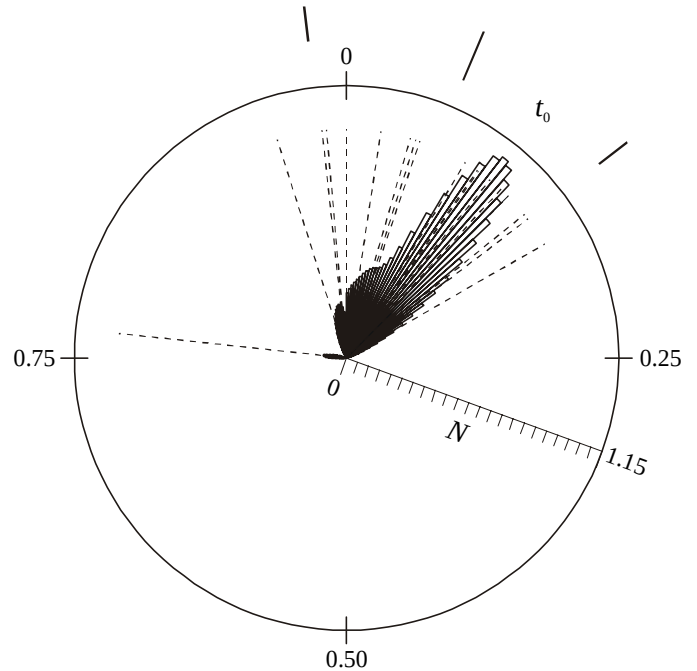


Figura 4. Histograma de rosa unidireccional que muestra la distribución, sobre cada día, de tiempos de origen t_0 ; los datos son indicados por líneas discontinuas, las celdas tienen $\Delta t = 0.05$ d y el largo de cada una corresponde al número repartido (mediante una gaussiana con $\sigma = 4\Delta t$) de eventos en ella según la escala radial N . El área total dentro de la curva corresponde al número de eventos. La raya mas gruesa extránea al círculo y con dirección radial indica el valor $\langle t_0 \rangle$ y las dos rayas a sus lados indican $\langle t_0 \rangle \pm S_{t_0}$.

sismos en el valle de Ojos Negros en el norte de Baja California (Frez *et al.*, 1999). Naturalmente, el tipo bidireccional puede reducirse al unidireccional mediante la reducción de la extensión a la mitad restando medio ciclo a los datos que superan este valor; en el ejemplo los azimuths son reducidos a la extensión de 0° a 180° , como se muestra en la figura 5. Sin embargo, es más fácil interpretar o visualizar la información del histograma si está expresada sobre la extensión natural de los datos; en el caso de azimuths del ejemplo el histograma de 0° a 360° puede sobreponerse o integrarse a mapas. Por lo tanto, los histogramas bidireccionales pueden presentarse sobre medio círculo, según se ilustra en la figura 6a para los mismos datos de la figura 5, o, mejor aún, con la información de medio círculo reflejada sobre la otra mitad (figura 6b), aunque en este caso el área total dentro de la curva es el doble del número de eventos.

En las figuras 5 y 6 se ha graficado, con línea gruesa, sólo la frontera externa de las celdas (sin las rayas que las separan), pues así se evita que, para celdas angostas, la concentración de rayas cerca del origen oscurezca los detalles de la distribución.

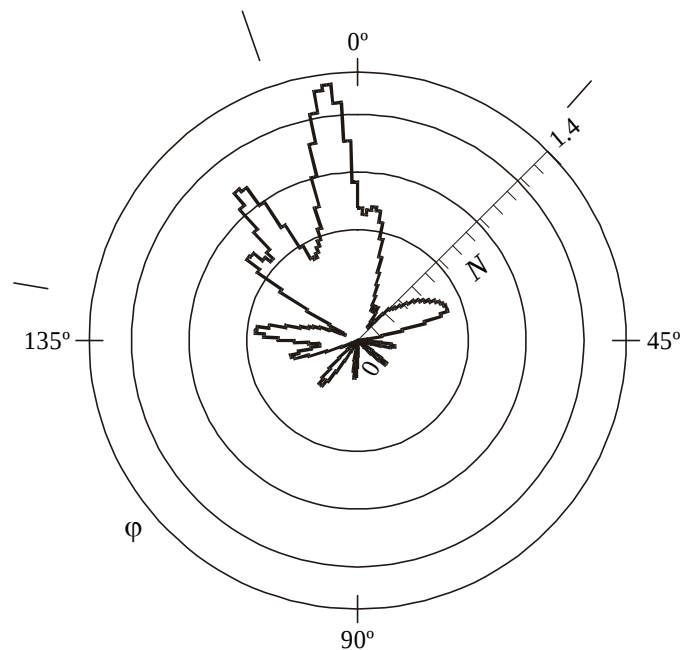


Figura 5. Histograma repartido de azimuths ϕ de plano de falla, distribuidos unidireccionalmente de 0° a 180° . Las rayas radiales extrañas al círculo indican $\langle \phi \rangle$ (centro) y $\langle \phi \rangle \pm S_\phi$.

Un aspecto en que difieren los histogramas de rosa de los no-cíclicos es el del valor medio, desviación standard y otros momentos de la distribución. Para cualquier tipo de distribución los parámetros mencionados están perfectamente definidos en un histograma no-cíclico, pero en el caso de una distribución uniforme, o aproximadamente uniforme, o multimodal con modos situados en puntos del ciclo lejanos entre sí, estos parámetros pueden no tener sentido pues sus valores dependen de la elección del origen, que es arbitraria. Por ejemplo, supóngase una distribución uniforme de azimuths, medida de 0°

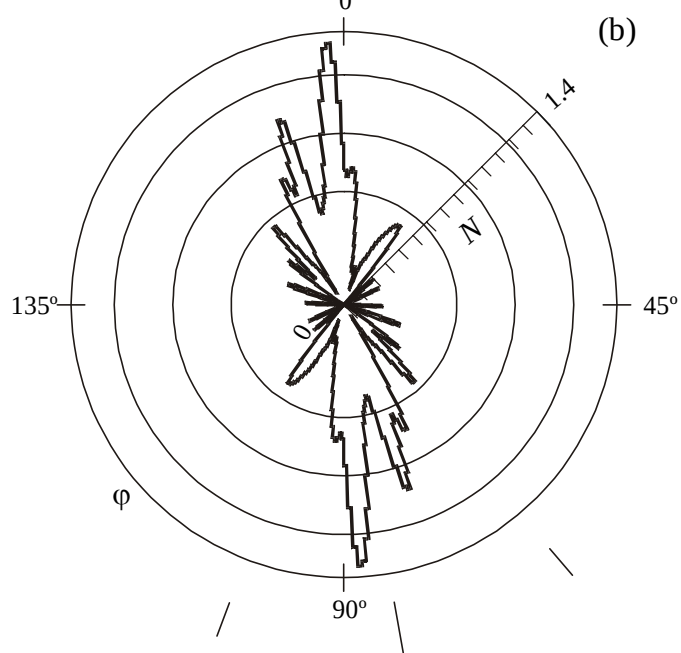
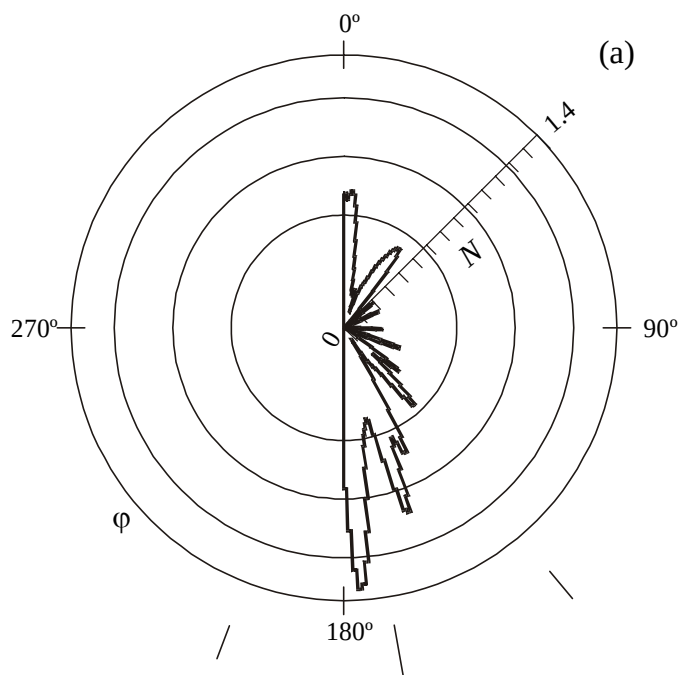


Figura 6. Histograma repartido de azimuths ϕ de plano de falla, distribuidos bidireccionalmente de 0° a 360° . En (a) el histograma se grafica de 0° a 180° . En (b) el histograma se refleja alrededor del origen, de manera que el área total bajo la curva es 2 veces el número de eventos. Rayas radiales como en la figura 5.

a 360° cuya media es, inmediatamente, 180° ; si medimos los azimuths de 45° a 405° , la media es 225° y obtendremos medias distintas para cada elección de origen.

Sin embargo, para distribuciones unimodales o, al menos, mayormente concentradas sobre una extensión que cubre una parte relativamente pequeña del ciclo, sí tiene sentido hablar de

valor medio y de desviación estándar. Por ejemplo, la distribución de tiempos origen t_0 de la figura 4 contiene eventos sólo en la extensión de 0.767d a 1.000d y de 0.000d a 0.167d y tiene un claro máximo $t_0^{\max} = 0.108d$; en este caso es perfectamente válido preguntarse cuál es el valor promedio de los tiempos origen.

Si la extensión de los valores observados no abarca el origen, la media, la desviación estándar, etc., pueden ser calculadas directamente; pero si, como en la figura 4, la extensión sí abarca el origen, entonces es necesario tener cuidado de corregir por el brinco que significa el paso por el origen. Una manera de hacer esto es utilizar el valor $t_0^{\max}$ como referencia y, al evaluar los parámetros, cada vez que $|t_0^{\max} - t_{0i}| > \frac{1}{2}T$, donde t_{0i} es el valor del i ésimo dato y T es el valor de un ciclo, en vez de t_{0i} utilizar $t_{0i} + T \operatorname{sgn}(t_0^{\max} - t_{0i})$. En la figura 4 la raya más gruesa con dirección radial extránea al círculo indica el valor medio $\langle t_0 \rangle = 0.063d$ y las dos rayas a sus lados indican $\langle t_0 \rangle \pm S_{t_0}$, donde $S_{t_0} \langle t_0 \rangle = 0.083d$ es la desviación estándar. En las figuras 5 y 6 rayas similares indican los valores de $\langle \phi \rangle$ y $\langle \phi \rangle \pm S_\phi$.

MAPAS

Para muestras cuyos valores cubren una superficie en 2D, un histograma considera celdas rectangulares de tamaño $\Delta x \Delta y$ y el número n_{ij} de elementos de la muestra cuyos valores están contenidos dentro de la celda h_{ij} cuyos lados son los intervalos $x_0 + i \Delta x$ a $x_0 + (i+1) \Delta x$ y $y_0 + j \Delta y$ a $y_0 + (j+1) \Delta y$. Por cada evento que cae dentro de una celda dada, un histograma convencional incrementa por 1 el número de elementos en esa celda y no hace cambio alguno en los contenidos de las demás celdas. En cambio, para cada evento, un histograma repartido incrementa el contenido de todas las celdas por el volumen que queda dentro de cada celda bajo alguna distribución bidimensional de error o incertidumbre (unitaria), alrededor del evento. De particular utilidad es la función gaussiana bidimensional:

$$g(x - x_k, y - y_k) = \frac{1}{\sigma_x \sigma_y 2\pi} e^{-\frac{(x-x_k)^2}{2\sigma_x^2} - \frac{(y-y_k)^2}{2\sigma_y^2}}, \quad (1)$$

con desviaciones estándar σ_x y σ_y , en general distintas, centrada sobre el k ésimo evento localizado en (x_k, y_k) . Esta distribución puede ser fácilmente rotada en caso de que la orientación de la incertidumbre en los datos no corresponda con los ejes X y Y.

La figura 7 ilustra el efecto de distribuir los eventos en 2D mediante una gaussiana simétrica con $\sigma_x = \Delta x/4 = \sigma_y = \Delta y/4$. En la parte superior a la izquierda se muestra el caso en que el evento cae en el centro de una celda; la gaussiana (cuya escala se ha exagerado) queda contenida (para el nivel de precisión utilizado) dentro de la celda y ésta tiene amplitud unitaria. En la parte superior a la derecha, se ilustra cómo se reparte un evento que cae en el centro de la frontera entre dos celdas; la mitad de la gaussiana es contenida en cada celda de manera que ambas tienen amplitud 0.5. Al frente a la derecha se muestra la repartición de un evento localizado en la frontera entre cuatro

celdas, que resulta en el valor de 0.25 para cada una. Al frente a la izquierda se muestra un evento ocurrido cerca de una frontera y cómo se reparte de manera desigual entre las celdas adyacentes. En un histograma usual los eventos de los tres últimos casos habrían sido asignados completamente a una celda u otra y, para eventos situados en fronteras, la celda habría sido escogida según algún criterio usualmente arbitrario. Lo anterior ilustra una de las razones por las que, al menos para muestras escasas, es más razonable repartir que simplemente asignar ocurrencias. Naturalmente, si la distribución de reparto es más ancha que las celdas, entonces un evento estará siempre repartido entre varias celdas.

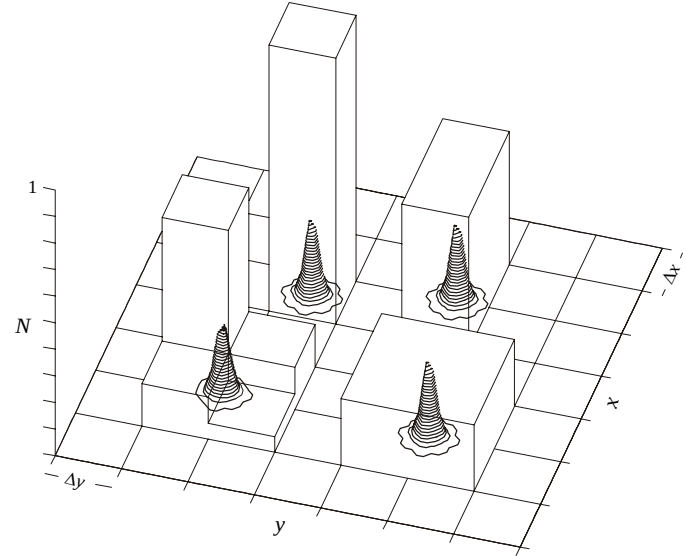


Figura 7. Ejemplos del efecto de la posición de un evento con respecto a las celdas en 2D sobre su distribución mediante una gaussiana simétrica con $\sigma_x = \Delta x/4 = \sigma_y = \Delta y/4$.

En Nava (1998) se discutieron los fuertes efectos que sobre un histograma convencional pueden tener diferentes valores del origen de las variables y del ancho de clases; los mismos efectos resultan para histogramas en 2D y, como en el caso unidimensional pueden ser en gran parte corregidos mediante la repartición de los datos. En particular, el uso de celdas muy pequeñas con relación a la extensión y/o incertidumbre en los datos hace que el histograma repartido sea esencialmente independiente de la elección del origen de coordenadas y del tamaño exacto de las celdas.

La figura 8 ilustra el histograma repartido de los sismos con $M \geq 4$ ocurridos en el norte de Baja California de 01/07/1973 a 29/05/1998, en el área que abarca de 114.8°W a 116.5°W y de 31.5°N a 32.7°N, con $\Delta x = \Delta y = 0.02^\circ$ y $\sigma_x = \sigma_y = 0.04^\circ$; sobre la figura se muestra la frontera internacional y algunas de las principales fallas de la región. Si se compara esta figura con la figura 4 de Frez y Frías-Camacho (1998) que ilustra la energía sísmica liberada en la región, puede verse que un simple histograma repartido (con ventana de magnitud que incluye solamente sismos relativamente grandes) puede dar una muy buena idea de la distribución de la sismicidad significativa en la

región. Naturalmente, si se hubiera asignado pesos a los sismos según su energía (o, más bien, según el logaritmo de ésta) el resultado hubiera sido muy parecido al de la obra citada, aunque no exactamente igual, porque en ésta se obtuvo primero un histograma normal y se repartió después.

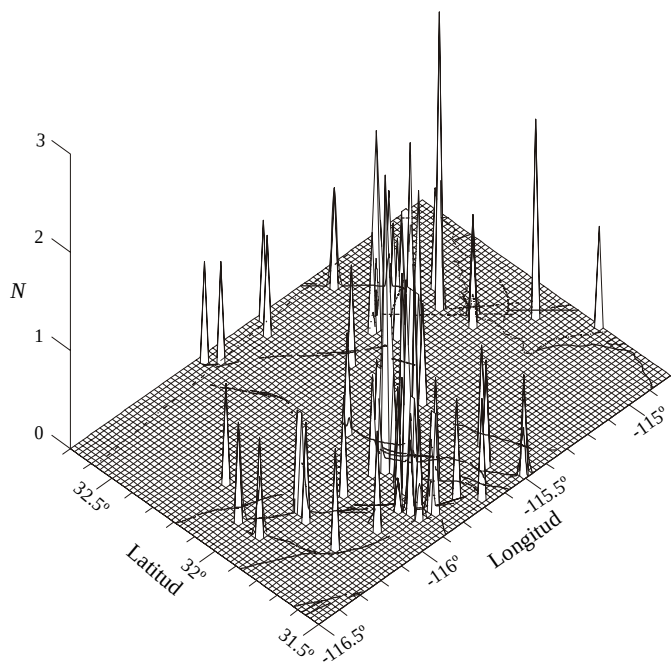


Figura 8. Histograma repartido en 2D de los sismos del norte de Baja California con $M \geq 4$ 01/07/1973 a 29/05/1998, de 114.8°W a 116.5°W y de 31.5°N a 32.7°N , con $\Delta x = \Delta y = 0.02^{\circ}$ y $\sigma_x = \sigma_y = 0.04^{\circ}$; sobre la figura se muestra la frontera internacional y algunas de las principales fallas de la región.

CONCLUSIONES Y DISCUSIÓN

La técnica de histogramas repartidos es perfectamente aplicable a distribuciones con cotas rígidas, a histogramas de rosa para el análisis de variables cíclicas y para histogramas en 2D. Como en los otros casos discutidos en Nava (1998) los histogramas repartidos son menos dependientes del tamaño y la localización de las celdas y casi independientes de estos parámetros cuando las celdas son bastante menores que la desviación estándar de la distribución de reparto.

Se encuentra empíricamente que para muestras escasas los histogramas repartidos se ajustan mejor a las distribuciones madre que los histogramas comunes. Por lo tanto, estimar o identificar una distribución madre desconocida a partir de una muestra escasa puede facilitarse mediante el uso de histogramas repartidos.

Tres programas, escritos en BASIC para PC, que elaboran los histogramas descritos en este artículo y en Nava (1998), se encuentran a disposición de los interesados (fnava@cicese.mx).

AGRADECIMIENTOS

Muchas gracias a José Frez por el uso de su catálogo Baja73PR, a los editores de GEOS por su apoyo y entusiasmo y a los árbitros que fueron un factor para la mejoración del presente trabajo.

REFERENCIAS

- Fisher, N., 1989, Smoothing a sample of circular data. *J. Struc. Geol.*, 11, p. 775-778.
- Fisher, N., 1993, Statistical analysis of circular data. Cambridge Univ. Press, UK, 277 pp.
- Frez, J. y Frías-Camacho, V., 1998, Distribución de energía sísmica en la región fronteriza de ambas Californias. *GEOS*, Vol. 18, No. 3, p. 189-196.
- Frez, J., González, J., Acosta, J., Nava, F., Méndez, I., Carlos, J., García-Arthur, R. and Alvarez, M., 1999, A detailed microseismicity study and current tectonics in the Peninsular Ranges of northern Baja California, Mexico: The Ojos Negros valley. *Bulletin of the Seismological Society of America*, (sometido).
- Nava, F., 1998, Histogramas repartidos para el análisis de muestras estadísticas escasas. *GEOS*, Vol. 18, No. 3, p. 174-179.
- Nava, F. and Glowacka, E., 1999, Fault-slip triggering, healing, and viscoelastic afterworking in sediments in the Mexicali-Imperial valley. *Pageoph* 1, p. 615-629.
- Silverman, B., 1986, Density estimation for statistics and data analysis. Chapman & Hall, UK, 326 pp.