REVISTA DE OBRAS PUBLICAS, julio 1982. Págs. 433 a 439

Criterios mínimo-cuadráticos de ajuste de distribuciones de probabilidad a datos experimentales (*)

Por E. CASTILLO
E. MORENO
J. PUIG-PEY

Dpto. de Matemáticas Aplicadas
ETS de Ingenieros de Caminos, Canales y Puertos
Universidad de Santander

En este trabajo se analiza la influencia que tienen las escalas de gradación de los ejes en los papeles probabilísticos, sobre los errores que se cometen al ajustar la recta de mínimos cuadrados a la nube de puntos experimental. Se presentan seguidamente diversos criterios mínimo-cuadráticos de ajuste de distribuciones de probabilidad a datos experimentales, estudiando su mayor o menor utilidad en función de la zona de la distribución en que interesa un mejor ajuste. Finalmente se aplican los diversos criterios al ajuste de datos de oleaje mediante una nueva familia de distribuciones de extremos, haciéndose el estudio comparativo y presentándose las conclusiones correspondientes.

1. INTRODUCCION

Un problema frecuente con el que se enfrenta el ingeniero o el físico en su vida práctica consiste en elegir, dentro de una familia, la distribución de probabilidad que mejor se ajusta a una muestra experimental. Ello exige precisar el sentido que se da a la palabra «mejor», es decir, exige definir una cierta relación de orden en el conjunto de las distribuciones de la familia elegida mediante un criterio acorde con sus deseos. Desgraciadamente no es corriente encontrar proyectistas que justifiquen los criterios elegidos y es muy frecuente la utilización de criterios inadecuados al objetivo que se persigue. Así muchos proyectistas utilizan ajustes gráficos «a ojo» y no son conscientes de que dichos ajustes son sensibles al tipo de representación que se elija (escala aritmética, papel doble logarítmico, etc...).

En la figura 1 se presentan los diagramas de frecuencia acumuladas de una muestra de alturas de ola significante máxima anual correspondientes a veinticuatro años, medidas en Myken-Skomvaer (Noruega) entre 1949 y 1976 —Houmb et al. (1978)—. Así mismo, se dibujan las rectas de mejor ajuste por el método de mínimos cuadrados realizadas «a ojo» para varios tipos de escala elegidos sobre los ejes coordenados. En la figura 1a) se utiliza papel aritmético, en la 1b) papel semilogarítmico, en la figura 1c) el doble logarítmico y en la figura 1d) el papel de Gumbel.

Por otro lado, la figura 2 muestra todos los ajustes juntos en escala aritmética, donde puede observarse no sólo que no son coincidentes (sólo uno de ellos es una recta), sino que se separan notablemente en algunas zonas y que algunos de ellos ajustan mucho mejor para valores grandes de la variable y otros lo hacen mejor para valores pequeños.

Ello es debido a que la función que se minimiza en el primer caso es

$$Q_1 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 \qquad (1)$$

en el segundo

$$Q_2 = \sum_{i=1}^{n} [Ly_i - (ax_i + b)]^2 \qquad (2)$$

en el tercero

$$Q_3 = \sum_{i=1}^{n} [Ly_i - (aLx_i + b)]^2 \qquad (3)$$

(*) Se admiten comentarios sobre el presente artículo, que podrán remitirse a la Redacción de esta Revista hasta el 31 de agosto de 1982.

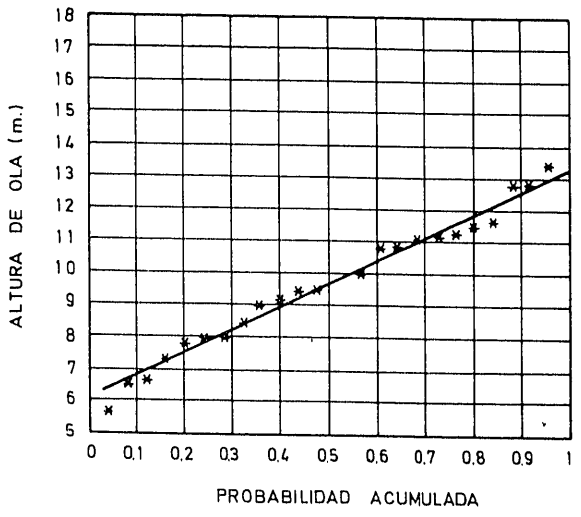


Fig. 1a

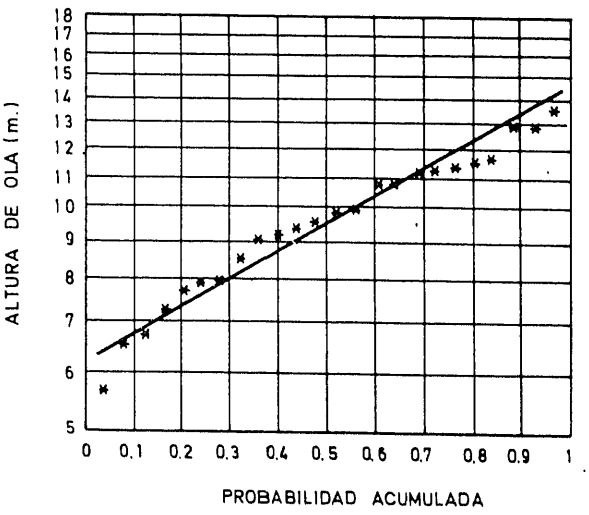


Fig. 1b

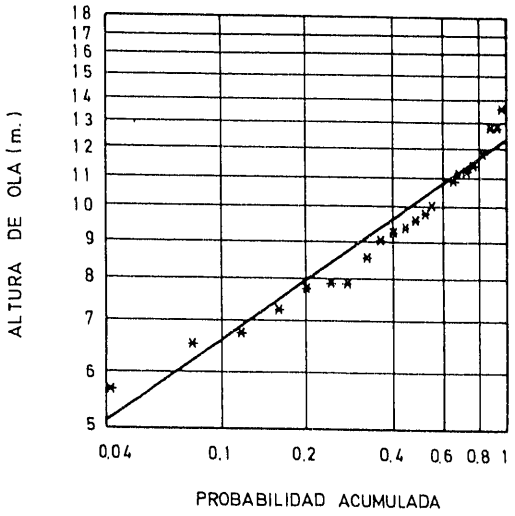


Fig. 1c

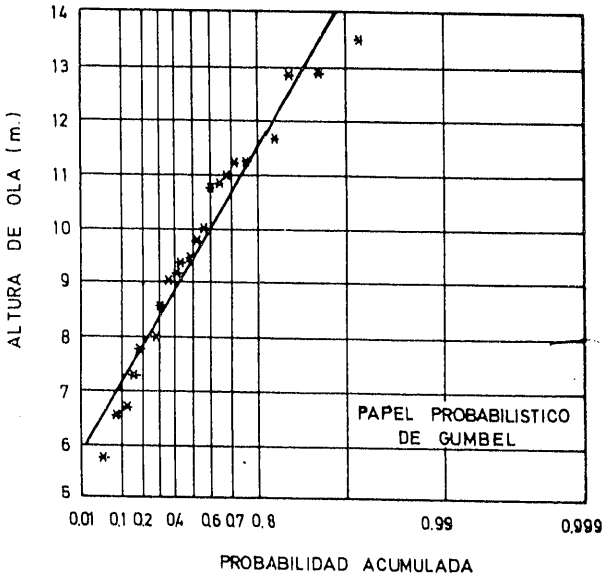


Fig. 1d

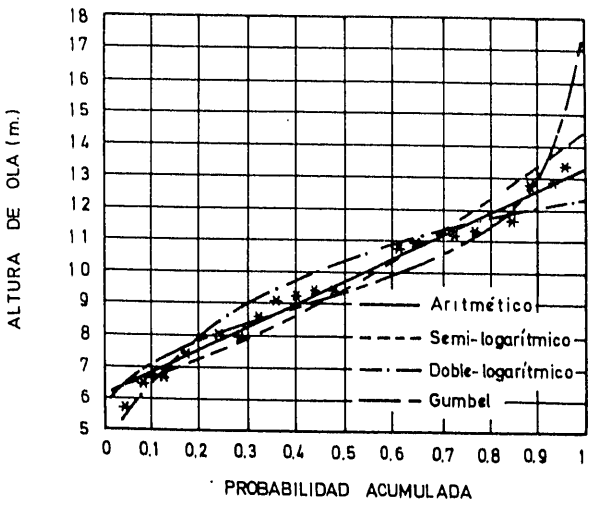


Fig. 2

y en el cuarto

$$Q_4 = \sum_{i=1}^{n} [y_i - (aLL\,(1/x_i) + b)]^2 \qquad (4)$$

Análogamente, el uso de otros papeles probabilísticos y ajuste por mínimos cuadrados lleva implícita la minimización de otras funciones.

Como puede observarse, los criterios (1), (2), (3) y (4) son completamente diferentes y deben corresponder a realidades físicas diferentes, siendo el ingeniero el que debe seleccionar de manera consciente cuál de ellos es el que debe utilizarse en cada caso.

Conviene señalar aquí que los térmicos que aparecen en los sumatorios de las expresiones (1), (2), (3) y (4) son las contribuciones de cada dato al error total. Por tanto, el valor y_i contribuye en (1) aditivamente y en (2), aditivamente pero como su logaritmo, lo que implica una reducción del peso relativo de los valores grandes frente a los pequeños y explica el por qué del peor ajuste para valores grandes que el método (1).

2. CRITERIOS DE AJUSTE

En lo que sigue se dan algunos métodos de ajuste, comentando en cada uno de ellos el significado físico, que puede justificar su uso. En la formulación analítica de cada criterio se representa por Z_i y $F(x_i)$ los valores de la función de distribución muestral y teórica correspondientes al dato x_i de la muestra de tamaño n, que se supone procede de una población con función de distribución $F(x,\Theta)$, donde $\Theta = (\Theta_1, \Theta_2, ...)$ es el vector paramétrico de la familia. Ante una muestra dada, hay que obtener los valores de los parámetros Θ_i que conducen al mejor ajuste con la información experimental, de acuerdo con el criterio que se considere.

Para el punteo de la función de distribución muestral, en las aplicaciones realizadas se ha utilizado la fórmula recomendada por Gumbel $Z_i = i/(n+1)$, donde i es la posición que ocupa el dato Z_i en la muestra ordenada de menor a mayor.

2.1. Método del error uniforme en probabilidad

Este método consiste en obtener la distribución aproximante dentro de la familia F(x) de modo que se minimice la expresión

$$G_1 = \sum_{i=k}^{n} [Z_i - F(x_i; \Theta)]^2 \qquad (5)$$

Este criterio da un mismo peso a un mismo error en la frecuencia acumulada en todo el rango de variación de la probabilidad, es decir, que según él es equivalente dar una frecuencia acumulada de 0,8 en vez de 0,9, que dar una de 0,1 en vez de 0,2.

Como se ve en la fórmula (5) se admite la posibilidad de despreciar la información de (k-1) valores de la muestra.

2.2. Método del error relativo en la probabilidad

Este método se basa en minimizar, dentro de la familia elegida, la expresión

$$G_2 = \sum_{i=k}^{n} \left[\frac{Z_i - F(x_i)}{F(x_i)} \right]^2 \qquad (6)$$

Este criterio da el mismo peso a un mismo error relativo en la frecuencia acumulada en todo el rango de variación de la probabilidad, es decir, que según él es equivalente dar una frecuencia acumulada de 0.99 en vez de 0.9, que dar una de 0.11 en vez de una de 0.1.

Este método sería correcto, por ejemplo, para ajustar en la cola izquierda de la distribucción, pues da más peso a los errores más a la izquierda que a los de la derecha. En efecto, la expresión (6) puede escribirse también como

$$G_2 = \sum_{i=k}^{n} \frac{1}{F^2(x_i)} [Z_i - F(x_i)]^2 \qquad (7)$$

donde se comprueba que es análoga a la (5), pero afectada de unos coeficientes, $1/F^2(x_i)$, que son más elevados para valores más bajos de x_i, lo cual lo hace insatisfactorio para ajustar en la cola derecha. Para este uso sería conveniente minimizar la expresión

$$G_3 = \sum_{i=k}^{n} \left[\frac{Z_i - F(x_i)}{1 - F(x_i)} \right]^2 \qquad (8)$$

que da el mismo peso a un mínimo error relativo de las frecuencias respecto de la probabilidad complementaria. Con este método sería equivalente dar una frecuencia acumulada de 0,85 en vez de 0,8 que dar una de 0,4 en vez de 0,2.

Como se verá este método va a surgir mediante otro criterio en el apartado 2.5

2.3. Método del error uniforme en el período de retorno

Este método se basa en minimizar, dentro de la familia elegida, la expresión

$$G_4 = \sum_{i=k}^{n} \left[\frac{1}{1 - Z_i} - \frac{1}{1 - F(x_i)} \right]^2 \qquad (9)$$

Como puede comprobarse, los sumandos del sumatorio de (9) son los cuadrados de los errores que se cometen en el período de retorno al utilizar como frecuencia acumulada Z_i en vez de $F(x_i)$. Obsérvese que idénticos errores absolutos en el período de retorno conducen a idénticas contribuciones al error total. Así, bajo este criterio, es equivalente dar como período de retorno un año en vez de dos años, que dar ciento un años en vez de ciento dos años.

La expresión (8) puede·excribirse también:

$$G_4 = \sum_{i=k}^{n} \frac{1}{(1 - Z_i)^2 \cdot [1 - F(x_i)]^2} [(Z_i - F(x_i)]^2 (10)$$

donde se observa que también es análoga a la (5), pero con coeficientes, $1/[(1-Z_i)^2 \cdot (1 - F(x_i))^2]$, mayores para los valores altos de x_i que para los valores bajos, lo cual lo hace válido para ajustar en la cola derecha aún tomando k igual a 1.

2.4. Método del error relativo en el período de retorno

Este método se basa en minimizar, dentro de la familia elegida, la expresión

$$G_5 = \sum_{i=k}^{n} \left[\frac{\frac{1}{1 - z_i} - \frac{1}{1 - F(x_i)}}{\frac{1}{1 - F(x_i)}} \right]^2 \qquad ,11)$$

Los sumandos del sumatorio de (11) son los cuadrados de los errores relativos que se cometen en el período de retorno al utilizar como frecuencia acumulada Z_i en vez de $F(x_i)$. Con este criterio es equivalente dar como período de retorno 1,1 años en vez de dar un año, que dar ciento diez años en vez de cien años.

La expresión (11) puede escribirse como:

$$G_5 = \sum_{i=k}^{n} \left[1 - \frac{1 - F(x_i)}{1 - Z_i} \right]^2 \qquad (12)$$

que será análoga a la dada en el apartado 2.5, o también como

$$G_5 = \sum_{i=k}^{n} \frac{1}{(1 - Z_i)^2} [Z_1 - F(x_i)]^2 \qquad (13)$$

lo que demuestra que da más peso a los valores altos de x_i que a los bajos, y por tanto, que es válida para ajustar en la cola derecha aún en el caso de tomar K igual a 1.

2.5. Método basado en el criterio de equivalencia en la cola derecha

Este método se basa en minimizar la expresión

$$G_3 = \sum_{i=k}^{n} \left[\frac{1 - Z_i}{1 - F(x_i)} - 1 \right]^2 =$$

$$= \sum_{i=k}^{n} \left[\frac{Z_i - F(x_i)}{1 - F(x_i)} \right]^2 \qquad (14)$$

que está fundamentado en el concepto de equivalencia en la cola derecha. Se dice que dos funciones de distribución F(x) y G(x), son equivalentes en la cola derecha si y sólo si

$$\lim_{x \to \infty} \frac{1 - F(x_i)}{1 - G(x_i)} = 1 \qquad (15)$$

y, por tanto, los sumandos del sumatorio de (14) para valores grandes de x_i deben ser muy pequeños en virtud de (15). La equivalencia en la cola derecha entre F (x) y G (x) significa que son equivalentes los infinitésimos que separan a F (x) y a G (x) del valor 1 en dicha cola.

Por otra parte, la expresión (14) puede ponerse como

$$G_3 = \sum_{i=k}^{n} \frac{1}{[1 - F(x_i)]^2} \cdot [Z_i - F(x_i)]^2 \qquad (16)$$

que muestra su validez para el ajuste en la cola derecha, pues los coeficientes, $1/[1 - F(x_i)]^2$, crecen con x_i.

Si lo que se desea es dar más importancia en el ajuste a la cola izquierda, el criterio de equivalencia es inmediato, sin más que cambiar $(1 - Z_i)$ y $[1 - F(x_i)]$ por Z_i y $F(x_i)$ en las expresiones anteriores.

2.6. Método de los errores ponderados

Las expresiones (5), (10), (13) y (16) sugieren un método que generaliza todos los anteriores y que consiste en minimizar la expresión

$$G_6 = \sum_{i=1}^{n} \alpha_i \cdot [(Z_i - F(x_i)]^2 \qquad (17)$$

donde los α_i ($i = 1,2,..., n$) son coeficientes de ponderación, cada uno de los cuales se asocia al dato experimental x_i y cuya magnitud dependerá de la zona de la distribución en la que se pretende un mejor ajuste.

3. EJEMPLOS DE APLICACION

En este apartado se aplican los métodos anteriores al ajuste de dos grupos de datos de altura de ola significante máxima anual. Uno, el ya citado en la introducción y obtenido por Houmb et al. (1978), y el otro, constituido por medidas realizadas durante veinte años en el cabo Machichaco y recogidas por Copeiro (1978). La familia de distribuciones ajustadas es

$$F(x; a, b, \gamma) = \exp[-\exp(-a \cdot |x|^{\gamma} \cdot \mathrm{sig}(x) + b)]; \; a<0, \gamma>0, x \in R \qquad (18)$$

que fue propuesta por Moreno (1980) para el ajuste de datos extremales. Los parámetros se obtienen mediante un proceso iterativo en ordenador que minimiza las expresiones asociadas a los diversos criterios.

Debido a la naturaleza de los datos, es claro que interesa obtener un buen ajuste en la cola derecha de la distribución, por ser los valores que ésta tome en dicha cola los que van a marcar la altura de cálculo, teniendo una importancia secundaria los datos experimentales que no pertenecen a dicha cola.

Las figuras 3a) y 3b) recogen ajustes referidos a los datos de Houmb, mientras que las 4a) y 4b) hacen lo propio con los de Copeiro.

En cuanto al método del error uniforme en probabilidad (UP), (Figs. 3a y 4a), obsérvese que el aspecto del ajuste con $K = 1$ es bueno para valores pequeños del período de retorno y malo para valores grandes, pues la curva se aleja de los datos en la cola derecha, sin embargo, los errores en probabilidad se mantienen pequeños (la escala de la probabilidad en dicha zona está muy distorsionada y grandes distancias implican pequeñas diferencias). Para el ajuste en la cola debe utilizarse un valor $k>1$ con lo que se mejoran notablemente los resultados. La figura 3a) también muestra el ajuste UP pero realizado con $K = 12$. De la comparación de ambas curvas se deduce la conveniencia de elegir valores altos de k.

En las figuras 3a) y 4a) también se recogen los ajustes por el método del error relativo en probabilidad (R.P.). Puede verse que cuando $k = 1$ el ajuste es bueno en la zona izquierda y completamente insatisfactorio en la cola derecha. No obstante, el ajuste mejora si se eligen valores de k mayores. obsérvese los ajustes R.P. con $k = 15$ y $k = 12$ para los datos de Houmb y Copeiro, respectivamente.

Observando en las figuras 3b) y 4b) los ajustes mediante el criterio del error uniforme en el período de retorno (UPR), con $k = 1$, se puede afirmar que en la cola derecha, dichos ajustes son notablemente mejores que con los métodos anteriores.

Los resualtados de emplear el método del error relativo en el período de retorno (RPR), con $k = 1$, recogidos asimismo en las figuras 3b) y 4b) corroboran su buen comportamiento para el ajuste en la cola derecha.

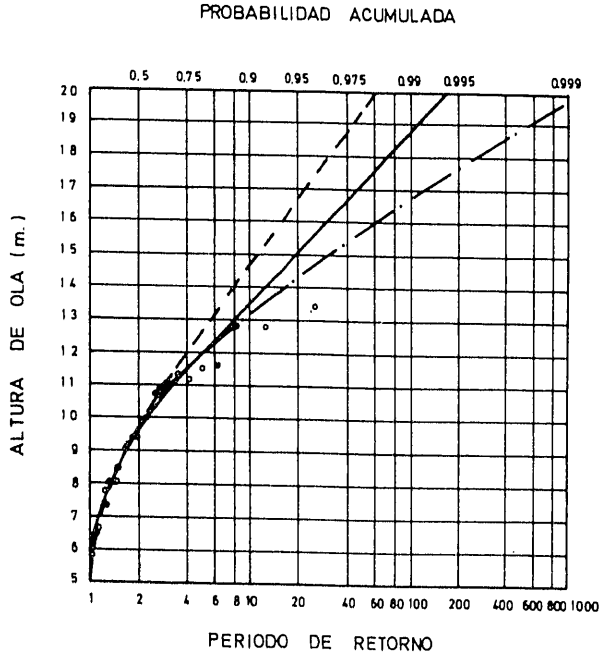
Finalmente, también en las figuras 3b) y 4b) se presentan los ajustes a los datos de Houmb y Copeiro respectivamente, empleando el método basado en el criterio de equivalencia en la cola derecha (E). Los resultados son muy satisfactorios sobre todo en la cola derecha, tomando $k = 1$. El aumento de k no mejora notablemente los resultados, como se ve en la curva de la figura 4b) obtenida con $k = 12$.

4. CONCLUSIONES

Las principales conclusiones de este trabajo son:

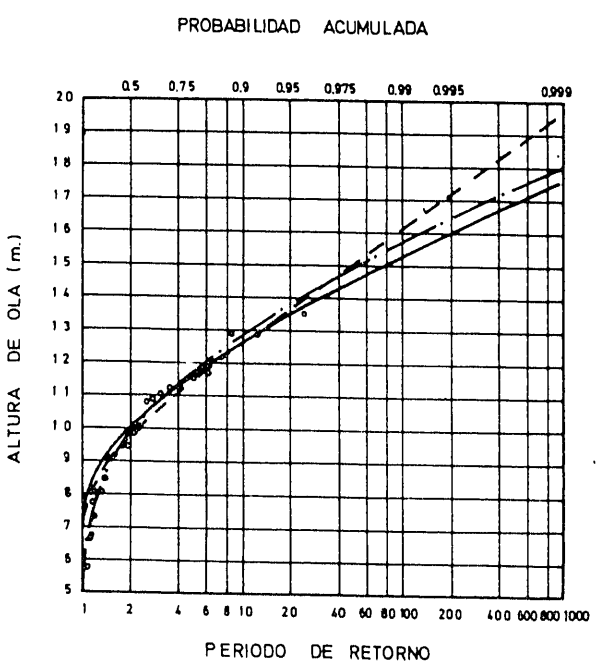
1. Es absolutamente necesario, al utilizar papeles probabilísticos, tener presente el peso que se está dando a los diversos datos, en su contribución al error total cometido al efectuar la regresión mediante una recta de los puntos representativos de la distribución acumulada experimental.

2. Las mayores diferencias de dispersión al utilizar diversos papeles, se dan para los



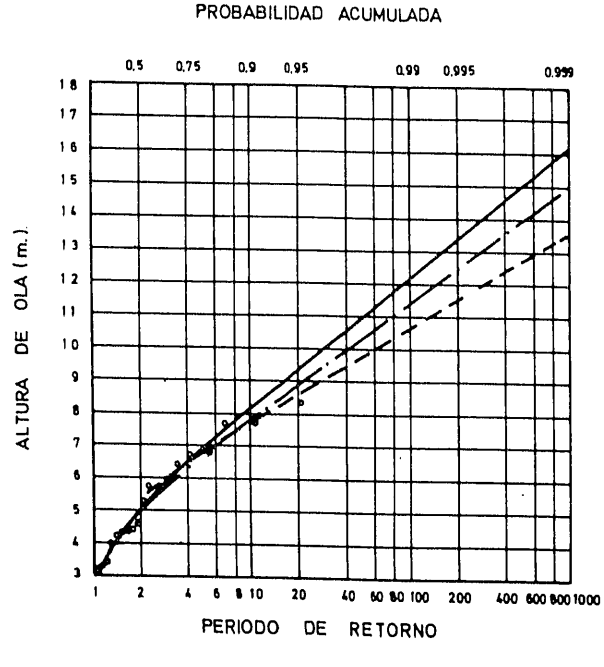
	k	a	b	γ
U.P.	1	0.80	5.11	0.85
R.P.	1	0.80	4.41	0.79
R.P.	15	0.10	1.54	3.03

Fig. 3 a



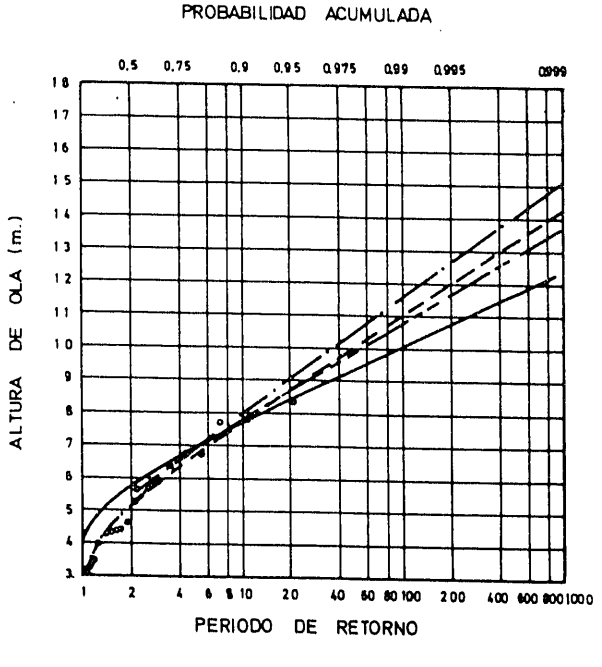
	k	a	b	γ
U.R.P.	1	0.1	3.9867	1.6351
R.P.R.	1	0.80	6.54	0.95
E	1	0.018	2.036	2.147

Fig. 3 b



	k	a	b	γ
U.P.	1	0.80	3.12	0.91
U.P.	12	0.80	3.94	1
R.P.	1	0.80	3.24	0.94
R.P.	12	0.80	3.89	1

Fig. 4 a



	k	a	b	γ
U.P.R.	1	0.80	4.82	1.07
R.P.R.	1	0.80	3.50	0.97
E	1	0.80	3.34	0.94
E	12	0.80	3.98	1

Fig. 4 b

valores extremos de las variables en juego, que son precisamente los que interesan en ingeniería, pues siempre se proyecta pensando en situaciones extremas.

3. De los criterios de ajuste cuadrático entre la distribución experimental y la teoría aquí representados, aquellos basados en el error referido al período de retorno dan mejores resultados para ajustar en la cola derecha que los que minimizan el error referido a la función de distribución.

4. De todos los criterios propuestos, el más adecuado para el ajuste en la cola derecha es el basado en el concepto de equivalencia en dicha cola.

5. Para la elección del valor de cálculo de la variable en estudio cuando interesan los valores extremos de ésta, puede convenir prescindir de datos registrados cuya probabilidad de presentación es grande pues no van a definir situaciones límites. Para ello, basta tomar $k>1$ en los criterios anteriores.

6. En los datos extremales analizados, la distribución (17) presentada manifiesta buenas cualidades para el ajuste.

5. REFERENCIAS

CASTILLO, E (1978): *Introducción a la Estadística Aplicada.* Ed. E. Castillo.

COPEIRO, E. (1978): *Análisis Extremal de Variables Geofísicas.* Tesis doctoral. ETS de Ingenieros de Caminos. Universidad de Santander.

GUMBEL, E. J. (1958): *Statistics of Extremes.* Columbia Univ. Press. New York.

HOUMB, O. G.; Mo, K., y OVERVIK, T. (1978): «Reliability Test of Visual Wave Data and Estimation of Extreme Sea States». Rapport No. 5 of *Port and Ocean Engineering.* The University of Trondheim. Norway.

MORENO, E. (1980): *Nuevos modelos de ajueste en las colas. Distribución de valores extremos. Aplicación a obras maritimas.* Tesis doctoral. ETS de Ingenieros de Caminos. Universidad de Santander.