Análisis de Frecuencia

CI41C

Primavera 2008

Introducción

- Para efectos de análisis hidrológico, muchas veces es necesario estimar caudales cuyo período de retorno T es superior al período de registro.
- Afortunadamente, se encuentra que las variables hidrológicas de interés se "distribuyen" de acuerdo a funciones conocidas de densidad de frecuencia

Introducción

- Al comienzo del curso repasamos algunas distribuciones probabilísticas contínuas que se usan en hidrología:
  - Normal
  - Log-normal
  - Gumbel
  - Extrema tipo I, II y III
  - Pearson

Método del factor de frecuencia

- Uno sabe que:

$$F(x_T) = \frac{T}{T-1}$$

- Dado T, si F es invertible, uno podría encontrar x_T
- Pero F rara vez es invertible! entonces,

$$x_T = \mu + K_T \sigma$$

factor de frecuencia, varia dependiendo de la distribución usada

Métodos de ajuste a una distribución de probabilidad

Momentos

Momentos de fdp son iguales a los de la muestra (ej: distribución exponencial)

$$f(x)=\lambda e^{-\lambda x} \qquad x>0$$

$$\mu = E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

$$= \int_{0}^{\infty} x\lambda e^{-\lambda x} dx$$

$\mu=1/\lambda$

$\lambda=1/\bar{x}$

Máxima Verosimilitud

Mejor valor de los parámetros son aquellos que maximizan la probabilidad conjunta de ocurrencia de la muestra

función de verosimilitud

$$L = \prod_{i=1}^{n} f(x_i)$$

$$Ln(L) = \sum_{i=1}^{n} Ln(f(x_i))$$

$f(x)=\lambda e^{-\lambda x}$

$$Ln(L) = \sum Ln(\lambda e^{-\lambda x})$$
$$Ln(L) = \sum [Ln\lambda - \lambda x_i]$$
$$= nLn\lambda - \lambda \sum x_i$$
$$= nLn\lambda - \lambda n\bar{x}$$

$$\frac{\partial(LnL)}{\partial\lambda} = 0 = \frac{n}{\lambda} - n\bar{x}$$

$\lambda=1/\bar{x}$

Pruebas de Bondad de ajuste

Kolmogorov Smirnov

Chi cuadrado χ^2: compara frecuencia relativa de muestra ($f_s(x_i)$) con valor teórico de distribución $p(x_i)$

$$\chi_c^2 = \sum_{i=1}^{m} \frac{n(f_s(x_i) - p(x_i))}{p(x_i)}$$

Se compara con fdp χ^2 con n grados de libertad y nivel de confianza 1-α

Número de intervalos

$n = m - p - 1$

Número de parámetros de la fdp

Se rechaza si

$\chi^2_c > \chi^2_{n,1-\alpha}$

Análisis de confiabilidad

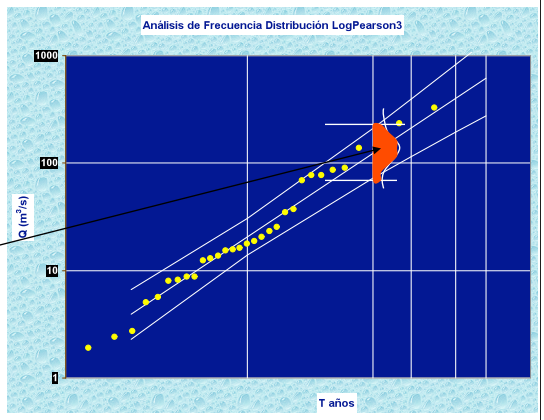
- OK, logramos ajustar una fdp, obtuvimos el factor de frecuencia, y calculamos x_T en función de los datos observados; cuál es el grado de confiabilidad que merecen nuestras estimaciones?

La precisión de resultados depende de la longitud de los registros disponibles

Puede suponerse que la estimación tiene distribución normal

$$x_T \pm z_{1-\alpha}\sqrt{Var(x_T)}$$

Nivel de Confianza $\beta = 1-2\alpha$



Valor aproximado fdp normal

$$Var(x_T) = S_e^2 = \frac{S_x^2}{n}\left(1 + \frac{K_T^2}{2}\right)$$

Valor aproximado fdp Gumbel

$$S_e = \left[\frac{1}{n}(1 + 1{,}1396K_T + 1{,}1K_T^2)\right]^{1/2} S_x$$

Valor aproximado para fdp Pearson3 y LP3

$P(x^*_T \le B_{T,\alpha}(x))=\alpha$

$P(x^*_T \ge A_{T,\alpha}(x))=\alpha$

Nivel de Confianza $\beta=1-2\alpha$

$$A_{T,\alpha}(x) = \bar{x} + K^A_{T,\alpha} S_x$$
$$B_{T,\alpha}(x) = \bar{x} + K^B_{T,\alpha} S_x$$

$$a = 1 - \frac{z^2_{1-\alpha}}{2(n-1)}$$
$$b = K_T^2 - \frac{z^2_{1-\alpha}}{n}$$

$$K^A_{T,\alpha} = \frac{K_T + \sqrt{K_T^2 - ab}}{a}$$

$$K^B_{T,\alpha} = \frac{K_T - \sqrt{K_T^2 - ab}}{a}$$

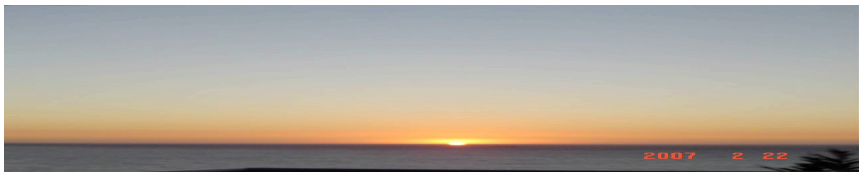
Ejemplo

T=100 años y β=90%

B=297,3 y A=1563 m³/s

Distribución normal y log normal

Intervalos exactos con distribución student t no centrada



Construcción de gráficos de probabilidad

Distribución	Procedimiento
Normal	Graficar x(i) ordenados vs. z(pi), donde z(pi) es la variable normal estándar para pi = 1-qi y qi = probabilidad de excedencia según fórmula general qi = (i – a)/(n + 1 – 2a). Usar a = 3/8
Lognormal	Graficar log[x(i)] ordenados versus z(pi). Usar a = 3/8 para estimar probabilidad de excedencia
Gumbel y Weibull	Para Gumbel graficar observaciones ordenadas x(i) versus variable reducida yi = -ln[-ln(1-qi)]. Usar posiciones de gráfico de Gringorten. Para Weibull graficar ln[x(i)] versus ln[-ln(qi)]
Pearson 3	Graficar observaciones ordenadas x(i) versus Kpi(Cw) donde pi = 1 – qi. Usar a = 3/8. Cw es el coeficiente de asimetría
Log Pearson 3	Graficar log[x(i)] versus Kpi(Cw) donde pi = 1 – qi. Usar a = 3/8. Cw es el coeficiente de asimetría. Papel lognormal también sirve.

Posiciones de gráfico

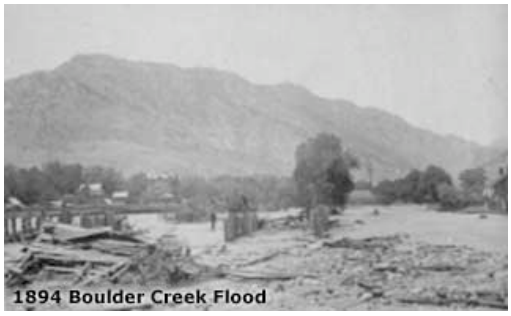
- General: $$p_i = \frac{i - a}{n + 1 - 2a}$$

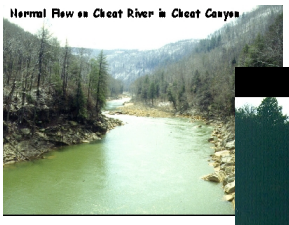
Nombre	a	T_1
Hazen	0.50	2n
Chegodayev	0.30	1.43n + 0.6
Gringorten	0.44	1.79n + 0.2
Weibull	0.00	n + 1

* California: pi = i/n

Incorporación de información histórica

- Registro histórico
  - Periódicos
  - Cartas
  - Marcadores de crecidas
- Paleocrecidas
  - Anillos de árboles
  - Depósitos de sedimentos

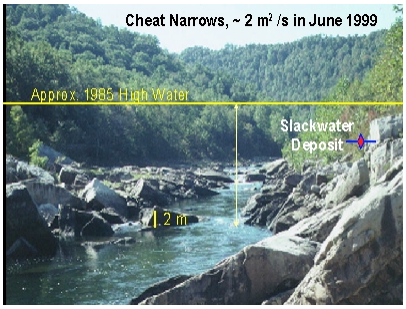


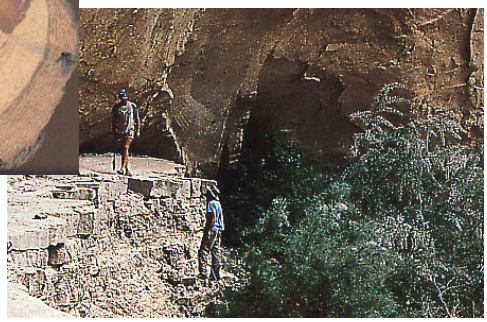



Causeway Area Floods

Dike Threatened, Saved; Water Down

River Reaches 15.3 Feet To Equal Record





Incorporación info. histórica

- registro histórico de h años de duración
- registro medido de s años de duración
- período total n = s + h
- Un total de r crecidas superan “umbral de percepción” Q_0
- e es el número de crecidas medidas que superan el umbral de percepción

$$\hat{p}_i = p_e \frac{i-a}{r+1-2a}; \quad i = 1,\ldots,r$$

$$p_e = r/n$$

$$\hat{p}_j = p_e + (1-p_e)\left(\frac{j-a}{s-e+1-2a}\right); \quad j = 1,\ldots,(s-e)$$

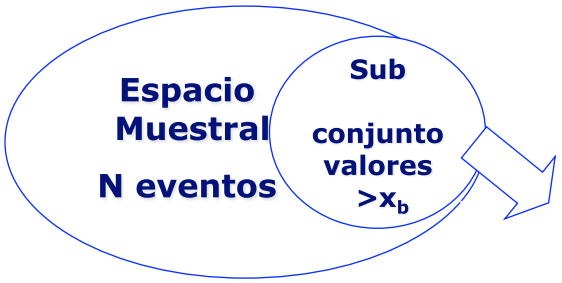
Si dato registro sistemático de longitud n es el mayor (m=1) en un periodo de r años

$P(X \geq Q^*)=1/(r+1)$

Si se determina un caudal máximo por las trazas dejadas en la ribera

- **si no se sabe lo sucedido en período intermedio (antes de iniciarse el registro sistemático), agregar como valor adicional en muestra. Se tiene n+1 valores.**
- **Si se sabe que es el mayor valor del período, agregar a la muestra insertándolo en el orden que corresponde. Todos los valores de mayor orden cambian su probabilidad.**

Análisis de Frecuencias de muestras con Valores Nulos



Análisis de frecuencia

$$P(y) \equiv \frac{m}{N} P(evento > x_k / eventos > 0)$$

Análisis de frecuencia series con valores nulos

- Típico de regiones con clima semi-arido
- Se aplica mismo criterio que en incorporación de info. histórica
- Total N observaciones, r son mayores que 0 (o algún otro límite de detección)

$$P(y) \equiv \frac{m}{N}\left(\frac{i-a}{m+1-2a}\right)$$

P(hay flujo)

P(excedencia dado que hubo flujo)

Periodos de Retorno para el Diseño de Obras Hidráulicas

Tipo de Obra	T (años)
Puentes	
Carretera Principal	50-100
Carretera Secundaria	10-50
Alcantarillas	
Carretera Alto Tráfico	50-100
Carretera Tráfico Intermedio	10-25
Carretera Bajo Tráfico	5-10

Tipo de Obra	T (años)
Drenaje Urbano	
Ciudades Pequeñas	2-25
Ciudades Grandes	25-50
Aeropuertos	
Alto Tráfico	50-100
Tráfico Intermedio	10-25
Bajo Tráfico	5-10
Vertederos de Presas sin Riesgo de Pérdida de Vidas Humanas	
Presas Pequeñas	50-100
Presas Intermedias	>100

Embalse Pequeño 7 a 12 m de altura

Embalse Intermedio 12 a 30 m de altura

Grandes Presas >30 m de altura

Si existe riesgo de pérdida de vidas humanas se estima la crecida máxima probable (VLE) y el vertedero se diseña con:

50 a 100% VLE

100% VLE

Riesgo: estructura falla si Caudal de diseño se excede al menos 1 vez, durante la vida útil de la obra.

$$R=1-[1-P(x \quad x_T)]^n$$

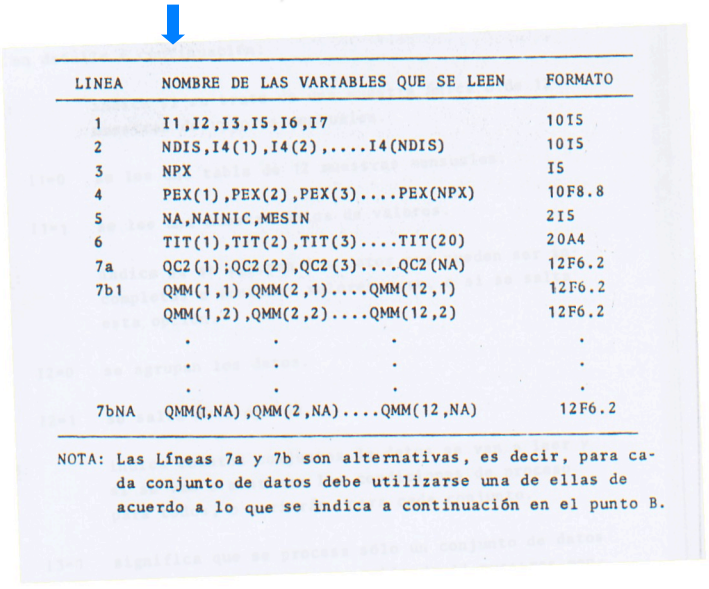
Capacidad adoptada

FS=C/L o MS=C-L

Capacidad de diseño

PROGRAMA FRECU

ARCHIVOS: ENTRADA.DAT, SALIDA.DAT



LINEA	NOMBRE DE LAS VARIABLES QUE SE LEEN	FORMATO
1	I1,I2,I3,I5,I6,I7	10I5
2	NDIS,I4(1),I4(2),....I4(NDIS)	10I5
3	NPX	I5
4	PEX(1),PEX(2),PEX(3)....PEX(NPX)	10F8.8
5	NA,NAINIC,MESIN	2I5
6	TIT(1),TIT(2),TIT(3)....TIT(20)	20A4
7a	QC2(1);QC2(2),QC2(3)....QC2(NA)	12F6.2
7b1	QMM(1,1),QMM(2,1)....QMM(12,1)	12F6.2
	QMM(1,2),QMM(2,2)....QMM(12,2)	12F6.2
	. . .	.
	. . .	.
	. . .	.
7bNA	QMM(1,NA),QMM(2,NA)....QMM(12,NA)	12F6.2

NOTA: Las Líneas 7a y 7b son alternativas, es decir, para cada conjunto de datos debe utilizarse una de ellas de acuerdo a lo que se indica a continuación en el punto B.

I1: indica si se trata de una muestra única o de 12 muestras de valores mensuales.

I1=0 se lee una tabla de 12 muestras mensuales.

I1=1 se lee una muestra única de valores.

I2: indica si se agrupan los datos que pueden ser incompletos o contener valores cero, o si se salta esta opción.

I2=0 se agrupan los datos.

I2=1 se salta esta opción.

I3: indica cuántos conjuntos de datos se van a leer y si se van a mantener las condiciones de proceso para todos, o variarán para cada conjunto.

I3=0 significa que se procesa sólo un conjunto de datos (una muestra única o una tabla de 12 muestras mensuales).

I3=1 significa que se procesa a continuación del primer conjunto de datos, todos los conjuntos que vienen inmediatamente después. Estos conjuntos se procesarán con las mismas condiciones del primero, es decir, con las mismas distribuciones, la misma prueba de bondad de ajuste, con datos de iguales características, con las mismas probabilidades de excedencia, etc.

Cuando I3 toma el valor 1, el programa una vez terminado de procesar el primer conjunto de valores, sigue con el segundo conjunto, pero a partir de la línea 5. Ello significa que desde este segundo conjunto en adelante, sólo deben incluírse las líneas 5 a la 7 a o b.

El último conjunto debe terminar en la última línea con un 1 para indicar que no se procesan más datos.

Una vez que se ha especificado I3=1, no pueden utilizarse las otras opciones de este índice, ya que no lee más la línea 1.

I5: indica si se elige o no la distribución de mejor ajuste, de acuerdo al o los test estadísticos utilizados.

I5=0 significa que se hace el o los tests estadísticos eligiendo la distribución de mejor ajuste y realizando el análisis de frecuencias para su distribución.

I5=1 significa que se hace el o los tests estadísticos, pero no eligiendo la distribución de mejor ajuste. En este caso el análisis de frecuencias se hace para todas las distribuciones especificadas.

I6: indica si se desea obtener información para el análisis de frecuencias gráfico con fines de comparación.

I6=0 significa que se salta esta opción.

I6=1 significa que se desea obtener una tabla de valores ordenados de mayor a menor versus las probabilidades de excedencia y períodos de retorno asignados según fórmula de Weibull.

I7: indica qué tests estadísticos se usarán en la prueba de bondad de ajuste.

I7=0 significa que se utiliza el test Chi-cuadrado.

I7=1 significa que se utiliza el test de Kolmogorov-Smirnov.

I7=2 significa que se utilizan ambos tests.

ii) Línea 2.

En esta línea se especifica el número y tipo de distribuciones a usar en el análisis de frecuencias.

Ello se especifica según los siguientes índices enteros:

NDIS: indica el número de distribuciones a usar en el análisis de frecuencias.

NDIS=9 significa que se utilizan todas las distribuciones disponibles en el programa. En este caso, no es necesario especificarlas una por una. Basta con dejar en blanco todos los campos a la derecha de NDIS.

NDIS<9 significa que se utiliza el número de distribuciones especificada en NDIS. En este caso, a continuación del valor de NDIS debe especificarse cada una de las distribuciones usando el índice I4 que se explica a continuación.

I4(1),I4(2)...I4(NDIS) indican el tipo de distribución a utilizar, según la regla siguiente:

(I4(I),I=1,NDIS) =

1 distribución Normal
2 distribución Lognormal 2
3 distribución Lognormal 3
4 distribución Extrema Tipo I o Gumbel
5 distribución Gamma 2
6 distribución Pearson Tipo 3
7 distribución Loggamma 2
8 distribución Log Pearson Tipo 3 (Est. de parám. por met. de máxima verosimilidad).
9 distribución Log Pearson Tipo 3 (Est. de parám. por met. de los momentos mixtos).

iii) Línea 3.

En esta línea se especifica el N° de valores de la probabilidad de excedencia que se desean utilizar en el análisis de frecuencias.

NPX: indica el número de valores de la probabilidad de excedencia para los cuales se desea calcular el valor de la variable. NPX puede tomar un valor máximo igual a 20.

iv) Línea 4.

En esta línea se especifican los valores de la probabilidad de excedencia para los cuales se desea calcular los valores de la variable. Las probabilidades deben expresarse en tanto por uno.

PEX(1),PEX(2)...PEX(NPX) indica cada uno de los valores de la probabilidad de excedencia.

v) Línea 5.

En esta línea se especifica el número de años de estadística de la o las muestras y el año de comienzo de la estadística.

NA: indica el número de años de estadística de una muestra única o de 12 muestras de valores mensuales. El valor máximo que puede tener NA es 100.

NAINIC: indica el año de comienzo de la estadística.

MESIN: indica el mes de inicio del año hidrológico (muestra de valores mensuales)

vi) Línea 6.

En esta línea se escribe un título literal que sirve al usuario para identificar cada conjunto de datos en la impresión de resultados.

El título puede ubicarse en cualquier parte de las 80 columnas de esta línea, pero se recomienda centrarlo.

vii) Línea 7a.

En esta línea se incluyen los valores de una muestra única. Deben utilizarse tantas líneas de este tipo, como sea necesario para incluir todos los valores de la muestra.

QC2(1),QC2(2)...QC2(NA) indican cada uno de los valores de la muestra única.

viii) Línea 7b.

En esta línea se incluyen los valores de una tabla de 12 muestras mensuales, a ser procesadas paralelamente. Se utiliza una línea de este tipo por cada año de estadística, es decir, en ella se incluyen 12 valores de un año.

Las líneas 7a y 7b deben rellenarse con valores negativos cada vez que la estadística a procesar sea incompleta. En caso que dicha estadística contenga valores nulos, deberá rellenarse con ceros.

Finalmente, a continuación de la línea 7a o 7b viene cualquiera de las tres alternativas siguientes:

a) Otro conjunto de datos incluidos en las líneas 1 a la 7a o 7b si I3=2.

b) Otro conjunto de datos incluidos en las líneas 5 a la 7a y 7b si I3=1.

c) Ningún dato más si I3=0.

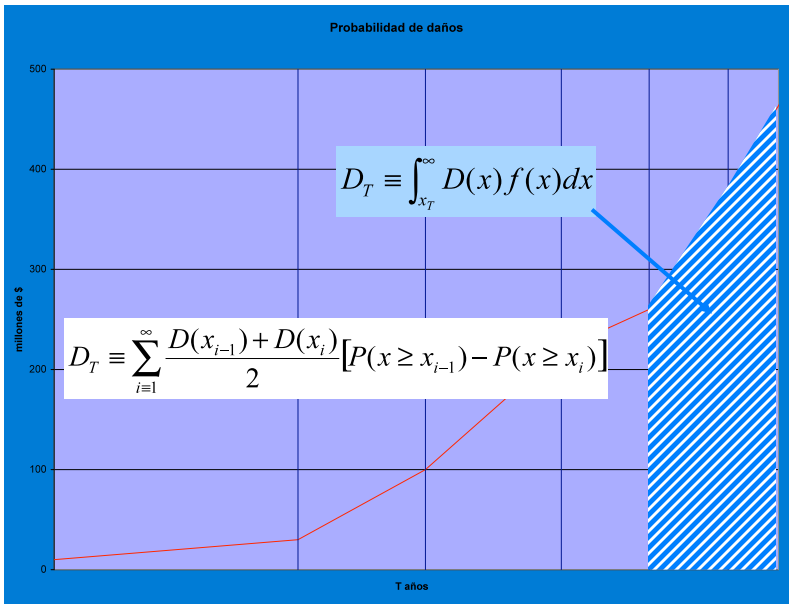
```
-M-E-N-S-A-J-E-S-


 DATOS DE ENTRADA PROPORCIONADOS POR EL USUARIO :
I1   I2   I3   I5   I6   I7
 1    1    0    1    0    0
NDIS
5
 1 2 3 4 5
NPX
 5
 PEX
.001    .010    .050    .200    .500
 33 1916   1
MAXIMOS ANUALES
63.3 59.5 49.2 29.9 22 20.4 9.0 7.31 7.0 5.85 5.3 5.14 3.22 2.48 2.41 1.87 1.62 1.5 1.49 1.46 1.45 1.44 1.41
 1.36 1.28 1.21 1.15 1.14 1.01 0.975 0.83 0.809 0.778
```

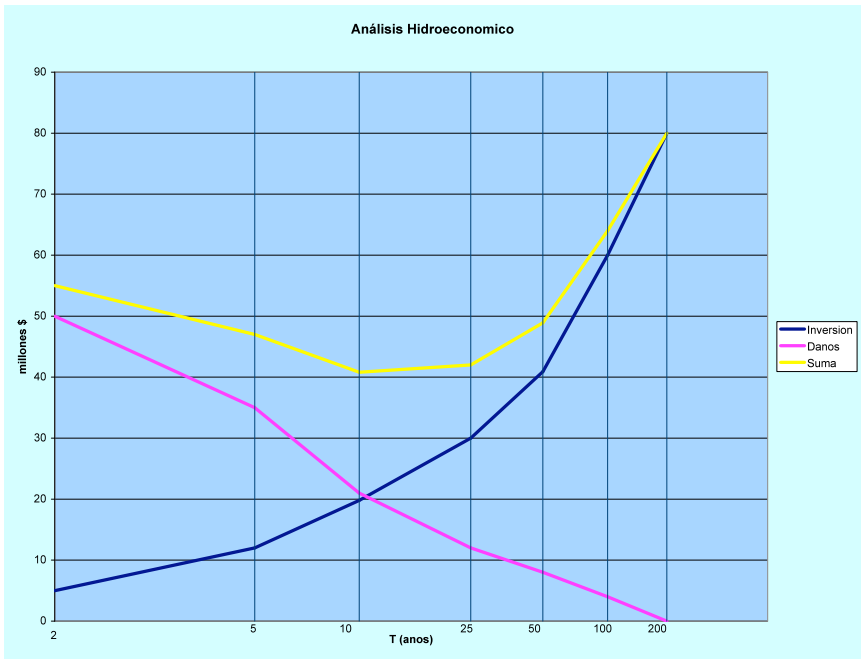
Análisis Hidroeconómico

Se determina el periodo de retorno óptimo que es aquel que minimiza los costos conjuntos de la inversión y los daños

Si se diseña para x_T, la estructura prevendrá de los daños para x x_T.

El costo esperado anual de daños se determina como el producto de la probabilidad f(x)dx de que ese evento ocurra en cualquier año y el daño D(x) que resultaria de ese evento




Análisis Hidroeconomico
millones $
90
80
70
60
50
40
30
20
10
0
2
5
10
25
50
100
200
T (anos)
Inversion
Danos
Suma