Análisis Estadístico de Datos Climáticos

Temas: Análisis de Varianza (ANOVA)
Regresión Múltiple

M.Barreiro, A. Diaz
2013

- El t-test compara dos grupos y determina si las medias de los dos grupos difieren. Si hay mas de dos grupos se podria proceder por pares, pero esto no es ni practico ni asegura resultados estadisticamente correctos.

- Cuando queremos comparar las medias de mas de dos grupos debemos usar ANOVA.

ANOVA 1 via

- Supongamos que se repite un experimento tal que resulta en N muestras de tamaño K, representado por variables aleatorias x_{nk},

  x_{nk}, n=1...N, k=1...K

  n identifica la muestra (estacion EFM en cada año)

  k identifica el elemento de la muestra (dado por corridas del modelo)

- Asumamos que x_{nk} son independientes, normales, y tienen igual desviacion standard.

- Asumamos ademas que para cada muestra n, la media es independiente de k, o sea

  $E(x_{nk})=\mu_n$

- Podemos escribir

$E(x_{nk}) = \mu_n = \mu + a_n$ donde $\mu = \frac{1}{N}\sum_{n=1}^{N} \mu_n$

Los coeficientes ($a_n = \mu_n - \mu$) son la diferencia entre la esperanza de x_{nk} y la media de todas las muestras.

a_n se denominan efectos de tratamiento (señal forzada por la TSM)

Un modelo estadístico apropiado para este tipo de datos es:

$x_{nk} = \mu + a_n + \varepsilon_{nk}$, donde $\varepsilon_{nk} \sim N(0,\sigma_N^2)$

y $\sum a_n = 0$

Como aplicamos ANOVA?

El efecto de las TSM sobre las variables atmosfericas da predictabilidad a la evolucion de la atmosfera pues “a priori” se puede pronosticar la futura evolucion de los oceanos con varios meses de antelacion.

Un primer paso es determinar cuales son las regiones del planeta influenciadas por la TSM

- Supongamos que quiero determinar la influencia de la TSM (1 factor=tratamiento) sobre la temperatura en S. America **(Predictabilidad Potencial).**
- Una posibilidad es usando Analisis de Varianza de 1 via.

Como disenamos el experimento?

Experimento: usando un Modelo de Circulacion General de la Atmosfera

- Hago evolucionar a la atmósfera desde el 1 de enero de 1950 hasta 31 de diciembre de 2007.
- Repito este experimento K veces, pero comenzando con condiciones iniciales el 1/1/1950 un poquito diferentes.
- Cada experimento k=1...K da una evolución diferente de la variable atmosférica X a lo largo de los 58 años: X_k

ANOVA

- El resultado del experimento es un ensemble de corridas de K-miembros (cada uno de 1/1/1950 hasta 31/12/2007).
- El modelo ANOVA asume que la evolucion de una variable atmosferica x en cada punto de grilla se puede aproximar como la suma de dos variables aleatorias independientes:

$$x_{nk} = \mu_n + \epsilon_{nk}, \quad k = 1...K, \quad n = 1...N,$$

Debido al forzante oceanico

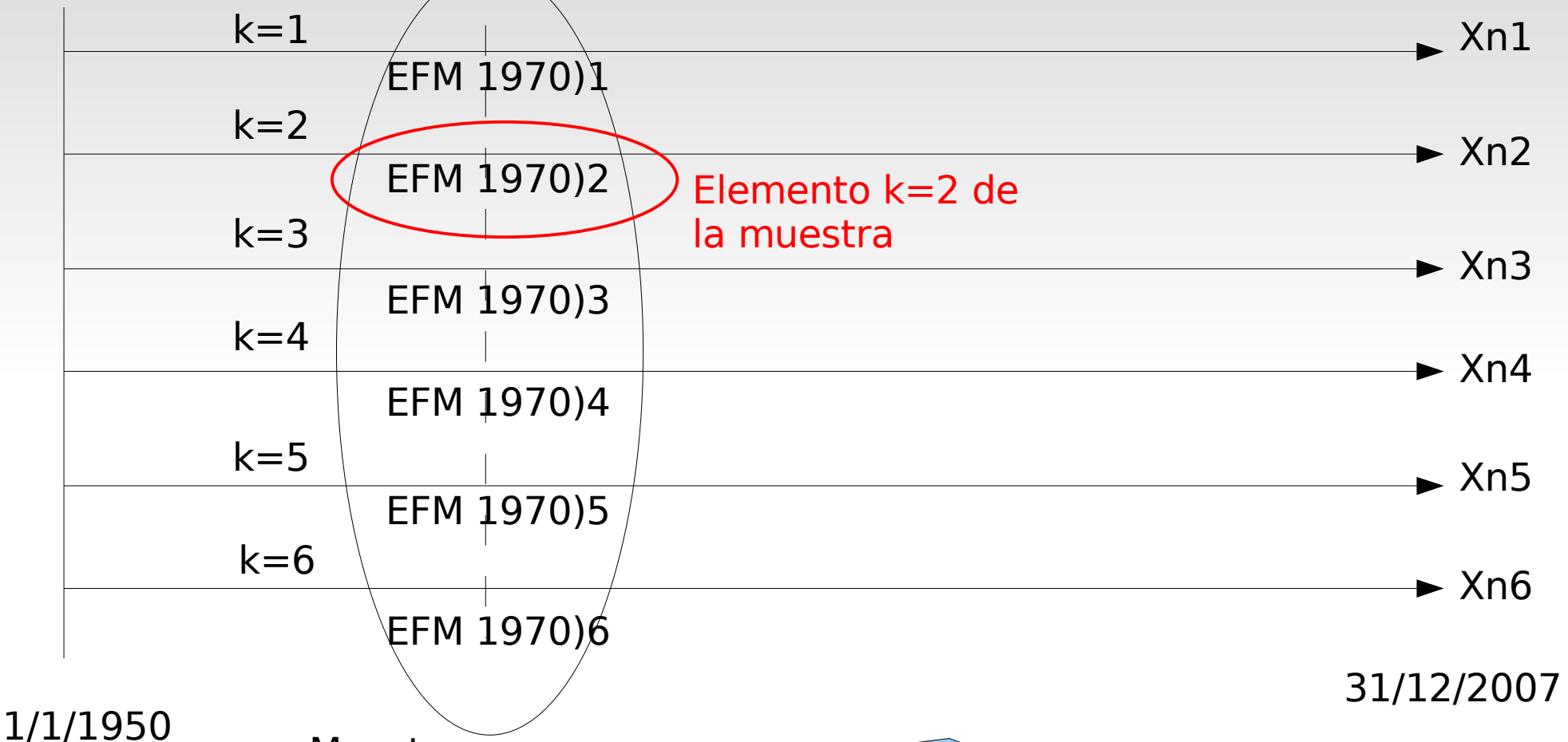
Debido a la dinamica atmosferica interna.

Por ej: consideramos solamente la estacion EFM -> 58 casos.

Diferentes condiciones iniciales en la atmosfera

$$x_{nk} = \mu_n + \epsilon_{nk}, \quad k = 1...K, \; n = 1...N.$$

$(EFM)_{ano,k} = \mu_{ano} + \varepsilon_{ano,k}$



Muestra correspondiente al ano n=1970

Condiciones de borde, o sea evolucion de las TSM, es la misma en cada corrida.

- Se asume que:
  - ϵ_{nk} son variables aleatorias independientes e identicamente distribuidas $N(0,\sigma_N^2)$
  - μ_n son variables aleatorias independientes e identicamente distribuidas $N(0,\sigma_F^2)$

- La predictabilidad potencial la definimos como la razon senal/ruido dada por

$$PP = \frac{variabilidad\ forzada}{variabilidad\ total} = \frac{\sigma_F^2}{\sigma_T^2} = \frac{\sigma_F^2}{\sigma_F^2 + \sigma_N^2}$$

- Con el modelo ANOVA podemos estimar las varianzas forzadas y totales.
- Definimos <u>media del ensemble</u>

$$\bar{x}_n = \frac{1}{K}\sum_{k=1}^{K} x_{nk}$$

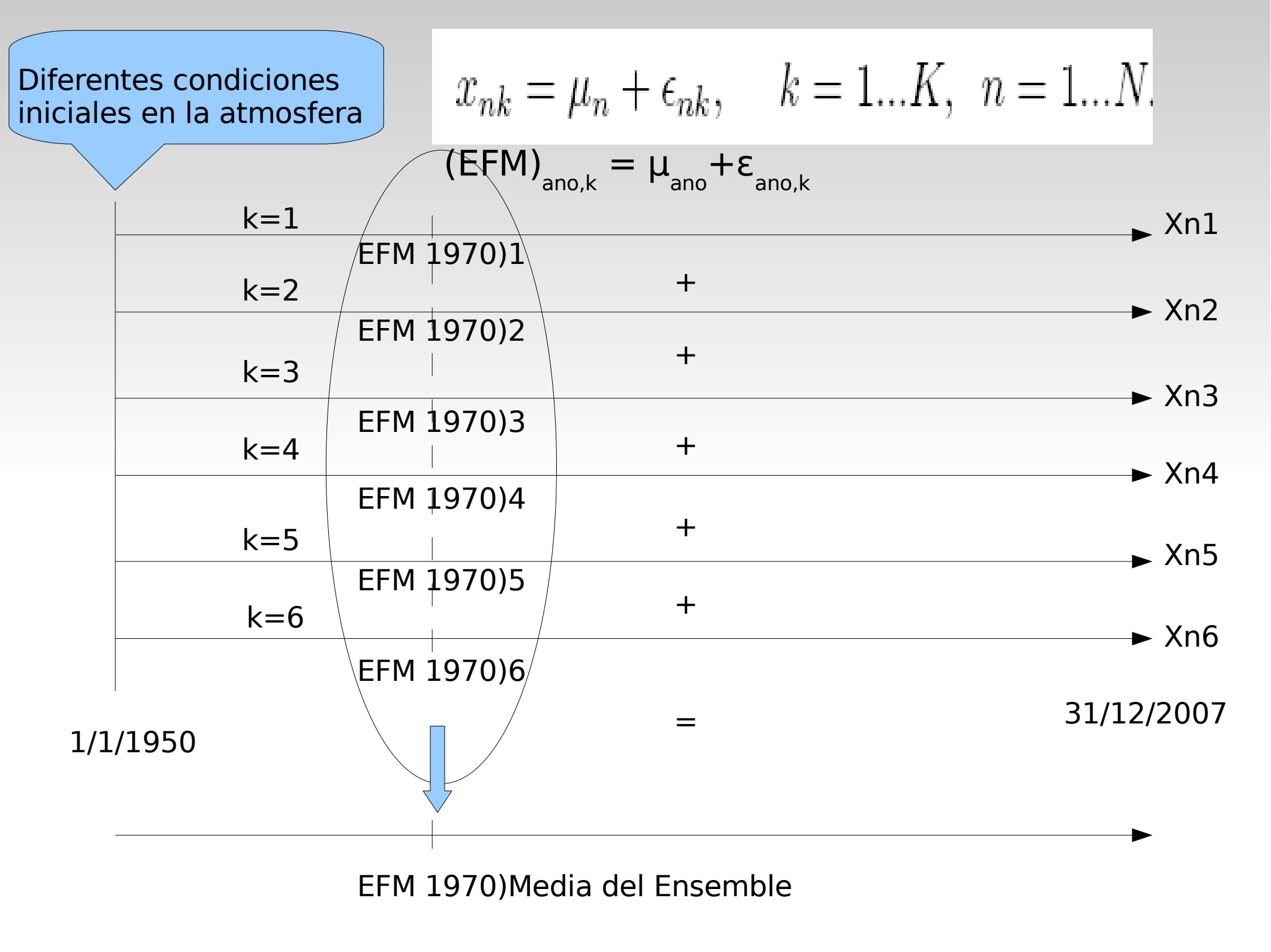
Entonces

$$\bar{x}_n = \mu_n + \frac{1}{K}\sum_{k=1}^{K} \epsilon_{nk}$$

Asi,

$$\sigma^2_{ME} = \sigma^2_F + \frac{1}{K}\sigma^2_N$$

y la varianza de la media del ensemble es un estimador sesgado de la varianza de la senal forzada.


Diferentes condiciones iniciales en la atmosfera
$x_{nk} = \mu_n + \epsilon_{nk}, \quad k = 1...K, \; n = 1...N.$
(EFM)ano,k = μano+εano,k
k=1
k=2
k=3
k=4
k=5
k=6
Xn1
Xn2
Xn3
Xn4
Xn5
Xn6
EFM 1970)1
EFM 1970)2
EFM 1970)3
EFM 1970)4
EFM 1970)5
EFM 1970)6
+
+
+
+
+
=
1/1/1950
31/12/2007
EFM 1970)Media del Ensemble

- Entonces la varianza forzada la estimo como

$$\sigma_F^2 = \sigma_{ME}^2 - \frac{1}{K}\sigma_N^2$$

donde la varianza de la media del ensemble esta dada por:

$$\hat{\sigma}_{EM}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\bar{x}_n - x)^2, \qquad = \text{SSA/n(J-1)}$$

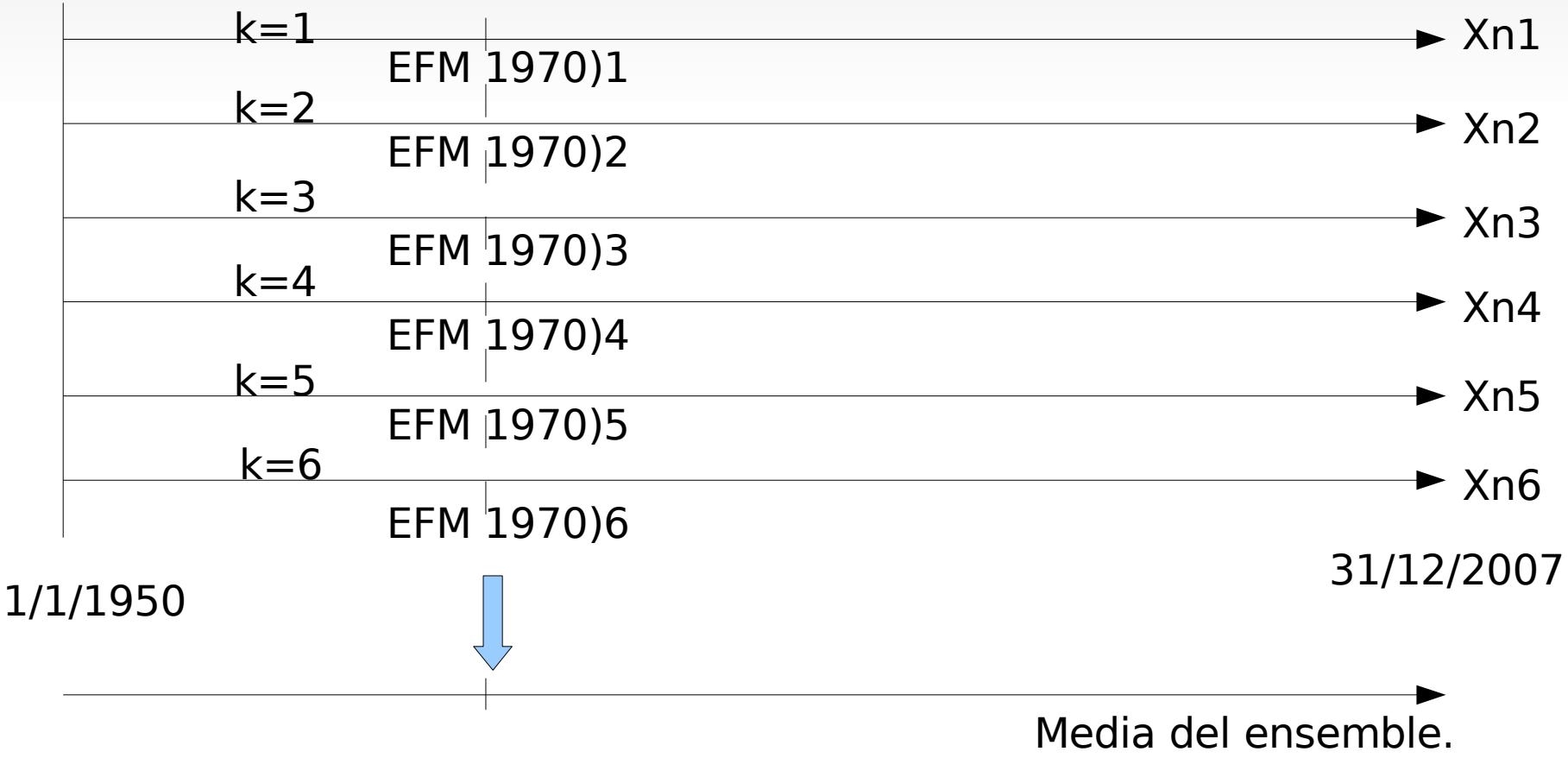
$$x = \frac{1}{NK}\sum_{k=1}^{K}\sum_{n=1}^{N} x_{nk},$$

x es la media de la variable de todos los datos.
Por ej: EFM promediado en el ensemble y en todos los anos.

- La varianza de la variabilidad interna la estimo como la desviacion respecto a la media del ensemble

$$\hat{\sigma}_N^2 = \frac{1}{N(K-1)} \sum_{n=1}^{N} \sum_{k=1}^{K} (x_{nk} - \bar{x}_n)^2$$

=SSE/J(n-1)



- El test que mide la significancia de la predictabilidad potencial estimada usando ANOVA es:
  - H0: $\sigma_F^2 = 0$
  - H1: $\sigma_F^2 > 0$
  - De la ecuacion $\sigma_{ME}^2 = \sigma_F^2 + \frac{1}{K}\sigma_N^2$

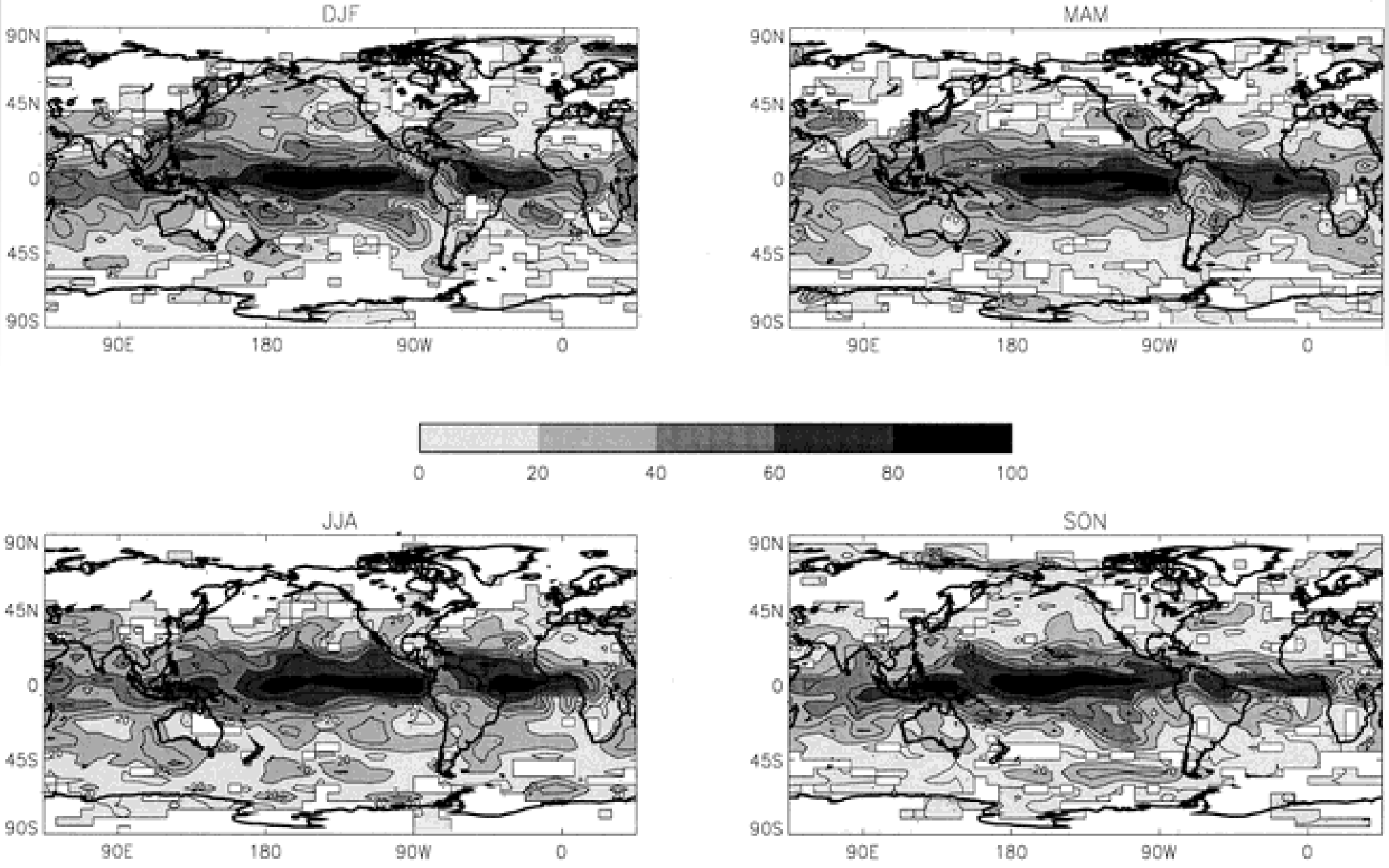
    vemos que la estadistica $F = \sigma_{ME}^2 / \frac{1}{K}\sigma_N^2$ =SSA/(J-1) / SSE/J(n-1)

    sirve de test. F sigue una distribucion-F con (N-1, N(K-1)) grados de libertad.

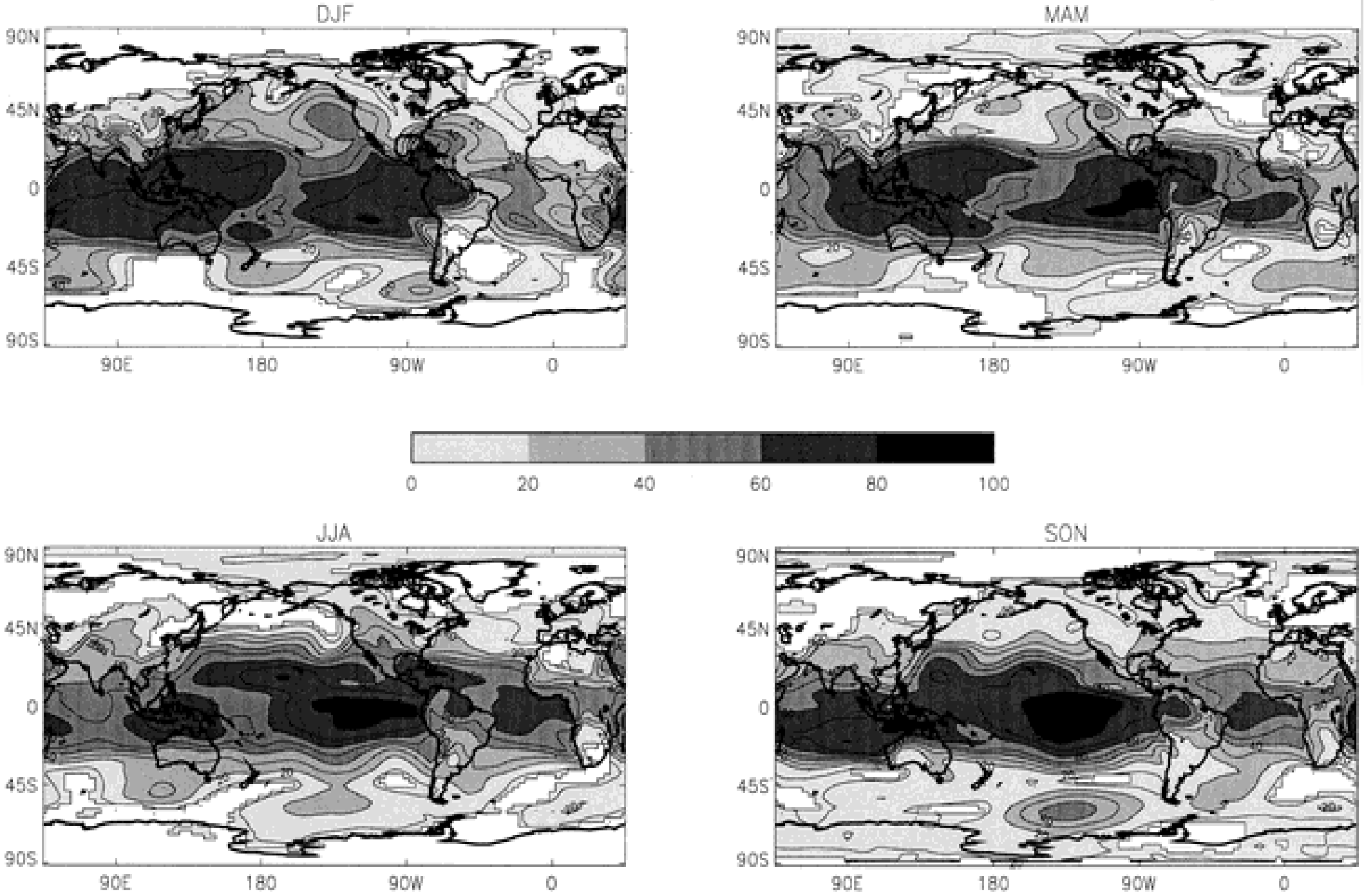
  - Usando F podemos escribir $PP = \frac{F-1}{F+(K-1)}$

    y calculamos los valores criticos de significancia de PP.

% de Variabilidad de precipitacion forzada por TSM (PP)



% de variabilidad de PS forzada por TSM (PP)



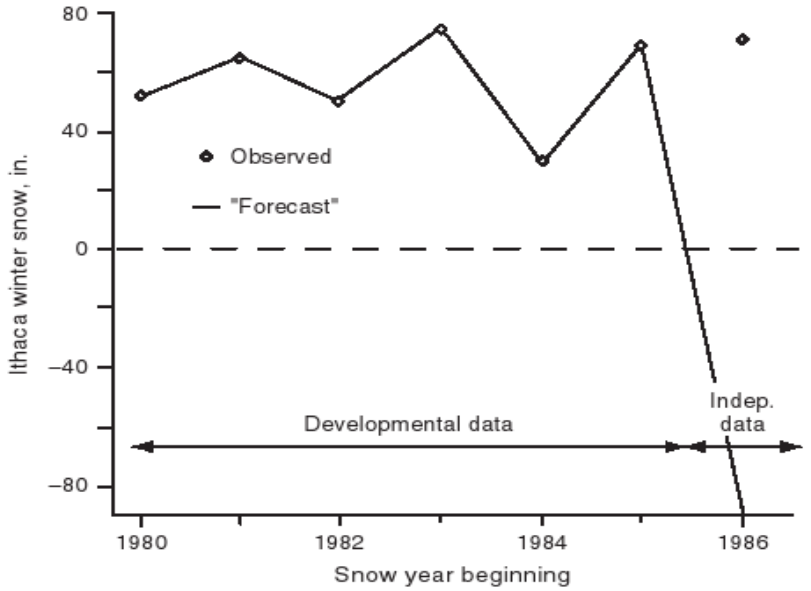
Regresion lineal multiple

- Un predictando y muchos (K) predictores

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K$$

- Al igual que en la regresion lineal simple, los coeficientes b se calculan por el metodo de minimos cuadrados resolviendo (K+1) ecuaciones.
- El resultado se presenta en la tabla de ANOVA.

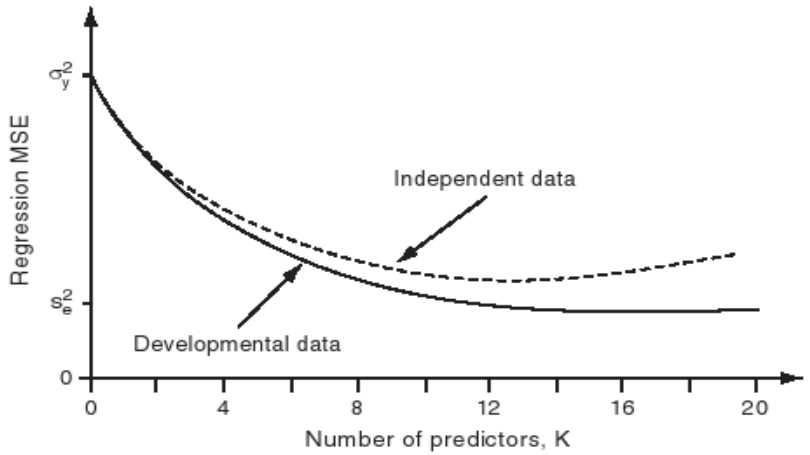
- La eleccion de predictores es crucial
  - Demasiados predictores usualmente es una mala idea; preferible tener pocos que muchos.
  - El uso de muchos predictores es bueno para explicar los datos con los cuales ajusto el modelo, pero las predicciones del modelo seran muy malas pues predictores que en realidad no estan casi correlacionados con el predictando tienen mucho peso relativo.



- Buenas costumbres:
  - Usar predictores que tengan sentido en el problema.
  - Separar el conjunto de datos en dos partes. Entrenar el modelo con una parte y validarlo con la otra parte del conjunto de datos disponibles. Si la performance del modelo es “mucho” mejor considerando los datos que se usaron para entrenar el modelo, seguramente se usaron demasiados predictores.
  - Usar conjuntos de datos grandes para asegurar la estabilidad del modelo.
  - Evitar usar predictores correlacionados entre si pues introduce informacion redundante en el modelo.

Como elegimos predictores?

- Seleccion hacia adelante:
  - Si hay M predictores posibles, se va eligiendo de a pasos el predictor que tenga mayor relacion lineal con el predictando.
    - En el primer paso se elige el predictor que tenga mas correlacion con el predictando: $\hat{y} = b_0 + b_1 x_1$
    - Segundo paso: se elige el predictor que de mejor regresion $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ de acuerdo a la tabla ANOVA (menor MSE)
    - Y asi sucesivamente hasta que MSE ya casi no disminuya.



- Por eliminacion
  - Se comienza ajustando el modelo con todos los M posibles predictores $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_M x_M$

y luego se van eliminando de a uno de acuerdo a su importancia en la regresion.

Este metodo no tiene por que dar el mismo resultado que el metodo anterior.