Análisis Estadístico
de
Datos Climáticos

Análisis de correlación canónica

2015

Motivación

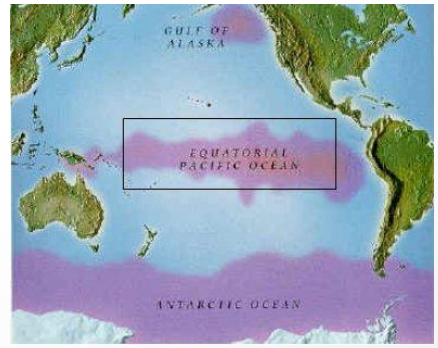
Si se tienen *dos* conjuntos de datos (A y B) que, por ejemplo, varían en el espacio y en el tiempo, (es decir 2 campos), nos podemos preguntar:

¿qué patrones tienden a ocurrir "conjuntamente" en ambos campos? o

¿cuál es el grado de conexión entre ambos campos?

Ejemplo

Los dos conjuntos de datos son:



- **X: TSM en la región 160E - 90W, 15S -15N del Océano Pacífico (grilla de 3.75º x 3.75º)**

- **Y: Precipitación en 15S - 45S, 75W – 40 W (grilla de 2.5º x 2.5º en región de América del Sur)**

Ambos en base mensual entre Ene 1979 y Dic 2006.

Motivación

El análisis de correlación canónica (ACC, o CCA en inglés) es útil para determinar los modos lineales dominantes de covariabilidad entre dos conjuntos de datos.

El método es útil para estudios de **diagnóstico** y también para construir modelos estadísticos de **pronóstico**.

Se parte de dos conjuntos de datos, o campos:

$$\tilde{X}_{n\,x\,p} \text{ e } \tilde{Y}_{n\,x\,q}$$

donde p y q se pueden interpretar como números de variables y n como número de realizaciones que, como se indica, deben coincidir para ambos campos.

Si n representa el número de observaciones en el tiempo y p y q son puntos en el espacio, cada columna es una serie temporal.

Las observaciones pueden ser simultáneas en el tiempo para ambos campos, o pueden estar desfasadas.

La primera opción se suele utilizar para estudios de diagnóstico y la segunda para pronóstico.

Se trabaja con anomalías. Llamamos X e Y a esos campos de anomalías.

Idea: Se trata de encontrar nuevas variables (U y V) que sean combinaciones lineales de las variables originales X e Y, y que tengan algunas propiedades (similar a ACP).
En lenguaje matricial:

$${}_nU_r = {}_nX_pA_r$$

$${}_nV_r = {}_nY_qB_r$$

siendo $r \leq \min(p, q)$

Las nuevas variables serán (U_1, U_2,…., U_r) y (V_1, V_2,….V_r) (las columnas de U y V, que tienen longitud n, en el "tiempo"), que se llaman **vectores canónicos.**

¿Qué propiedades queremos que cumplan estas nuevas variables?

Condiciones que deben cumplir U y V

Queremos que:

- la correlación entre U_1 y V_1 sea la máxima posible

- además:

corr $(U_1, V_1) \geq$ corr $(U_2, V_2) \geq \ldots \geq$ corr $(U_r, V_r) \geq 0$

$$\text{corr}\,(U_i\,, V_i) = \lambda_i$$ (coeficientes de correlación canónica)

$$\text{corr}\,(U_i\,, V_j) = 0 \;\; \text{si} \;\; i \neq j$$

$$\text{corr}\,(U_i\,, U_j) = 0 \;\; \text{si} \;\; i \neq j$$

$$\text{corr}\,(V_i\,, V_j) = 0 \;\; \text{si} \;\; i \neq j$$

o sea …

Si $i \neq j$, los U son *no correlacionados* entre sí,

" V " " " " "

U_i y V_j " " " " "

• además: var (U_i) = var(V_j) = 1. (o sea los Ui y Vj se estandarizan)

Se dice que se tienen r **modos canónicos.**

Se demuestra que los λ_i^2 y las matrices A y B son soluciones de un problema de autovalores y autovectores:

Si llamamos: $\mathbf{S_{XX} = \frac{1}{n-1} X^T X}$ $\mathbf{S_{YY} = \frac{1}{n-1} Y^T Y}$

$\mathbf{S_{XY} = S_{YX}^T = \frac{1}{n-1} X^T Y}$ (matrices de covarianza),

las ecuaciones para A y los λ_i^2 **son:**

$$\mathbf{det\,(S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX} - \lambda^2 I) = 0}$$

$$\mathbf{(S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX} - \lambda_i^2 I)\ A_i = 0 \quad (i = 1,2,...,r)}$$

Análogas ecuaciones se tienen para determinar **B,** o bien:

$\mathbf{B = S_{YY}^{-1} S_{YX} A\ \Lambda^{-1}}$ siendo $\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & . & \\ & 0 & & . \\ & & & \lambda_r \end{pmatrix}$

(si q < p, es preferible usar las ecuaciones análogas para B y luego calcular A)

Resumiendo: de X e Y, se obtienen A y B, y luego U y V:

$$\mathbf{U = XA \quad , \quad V = YB}$$

También se pueden estimar por mínimos cuadrados variables de un conjunto (p. ej. Y) a partir de las del otro (X):

$$\hat{\mathbf{Y}} = \mathbf{X\,A\,\Lambda\,B^{T}\,S_{YY}}$$

Ejemplo

Los dos conjuntos de datos son los dados antes:

- **X: TSM en la región 160E - 90W, 15S -15N del Océano Pacífico (grilla de 3.75º x 3.75º)**

- **Y: Precipitación en 15S - 45S, 75W – 40 W (grilla de 2.5º x 2.5º en región de América del Sur)**

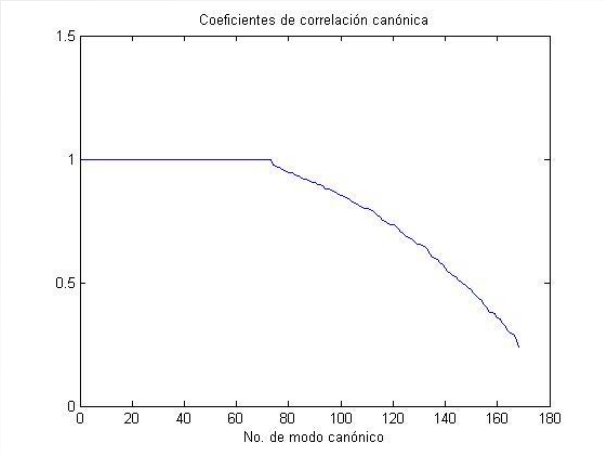
Ambos en base mensual entre Ene 79 y Dic 06.

n = 336, p = 8x30 = 240, q = 12x14 = 168.
r = min(p,q) = 168

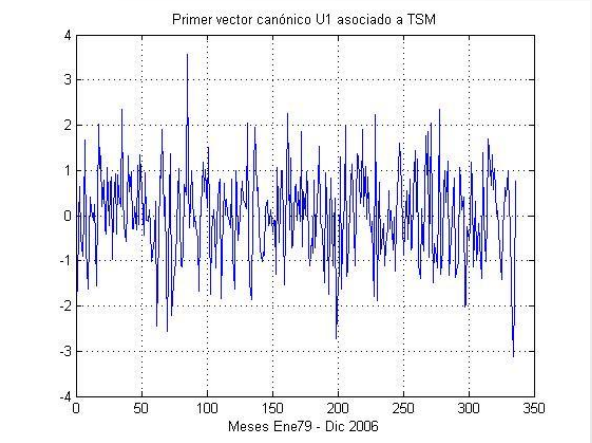
El comando canoncorr de Matlab calcula todos estos elementos.

[A,B,r,U,V] = canoncorr(X,Y);

Coef. de correl. canónica



Vector canónico U_1



En este caso, como p + q > n-1, hay muchas correlaciones iguales a 1, y hay muchos pares de vectores canónicos (U_i, V_i) que coinciden para ambos campos, en particular $U_1 = V_1$.

Suponiendo que X e Y son datos en espacio y tiempo, una forma relevante de extraer información de estos resultados es calcular la correlación de la serie temporal en cada punto en el espacio con los primeros vectores canónicos:

$$g(i) = corr(X_i, U_1) \quad (i = 1,.....p)$$

$$h(j) = corr(Y_j, V_1) \quad (j = 1,.....q)$$

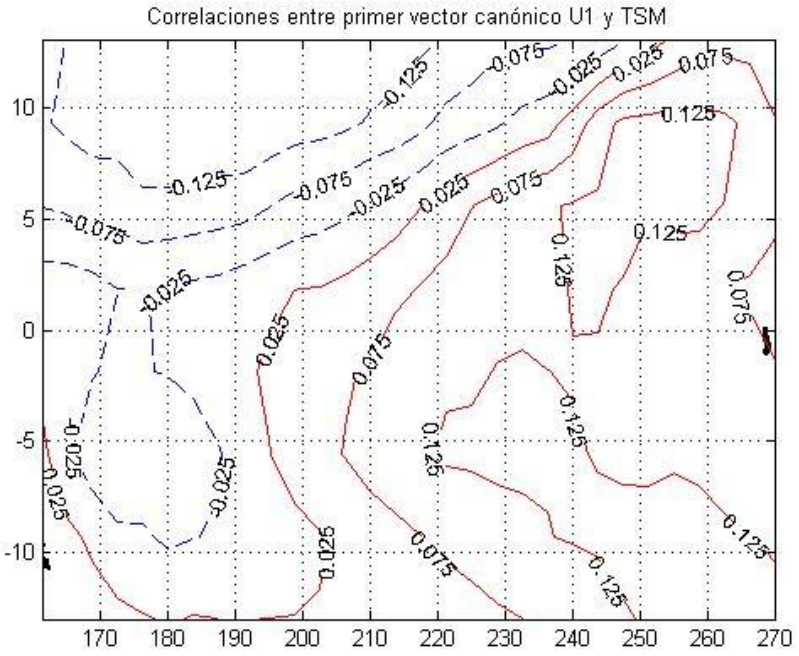
son **"mapas"**

(Idem con U_2 y V_2, etc)

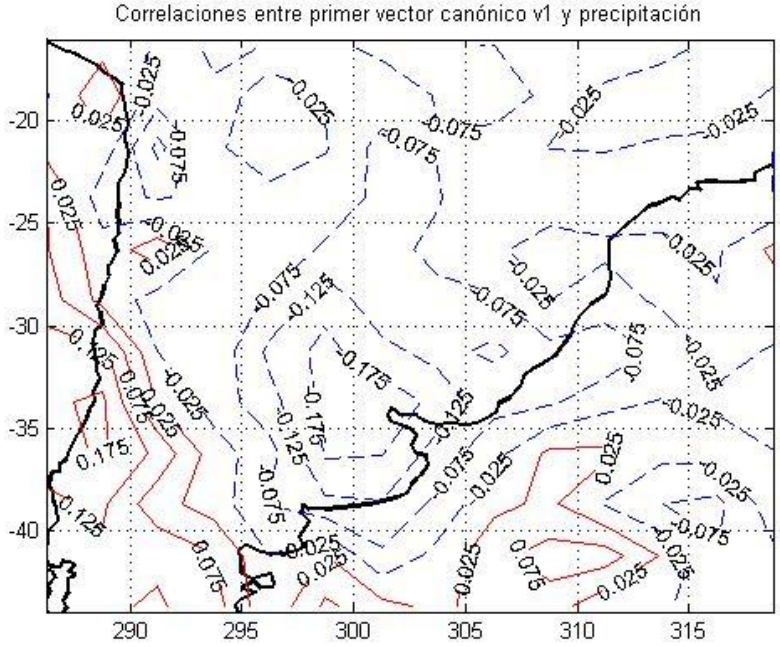
Se puede interpretar que los puntos (i y j) en que esas correlaciones sean significativas, determinan regiones en cada campo, que tienden a co-variar conjuntamente.

El conjunto U_k, V_k, g_k, h_k es un "representante" del k-ésimo modo canónico.

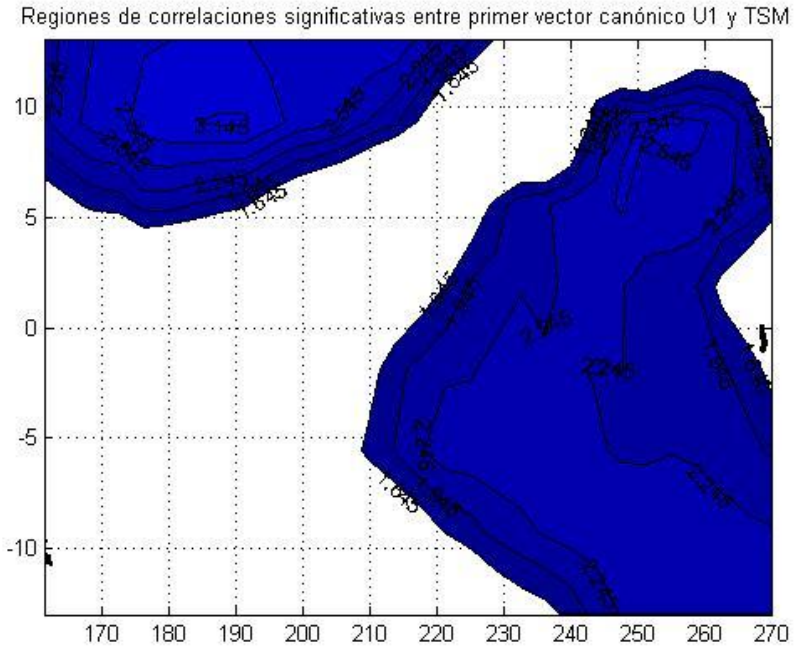
Correlaciones entre U1 y TSM



Correlaciones entre V1 y precip

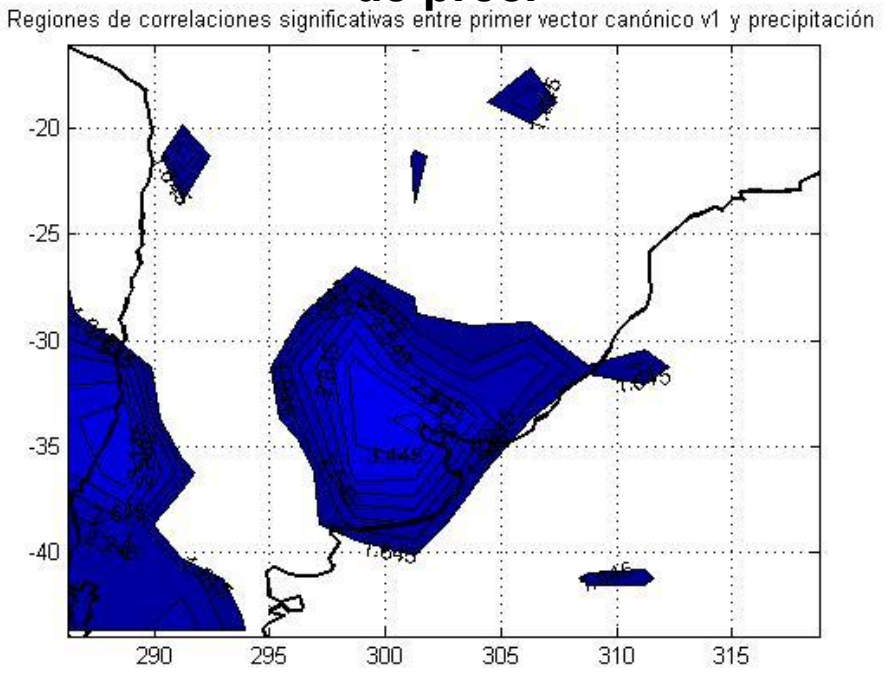


Regiones significativas al 5% en campo de TSM



El test para la significancia es el test de Student para correlaciones ya visto.

Regiones significativas al 5%en campo de prec.



Todos estos patrones dependen de las regiones que se tomen al comienzo del análisis.

Cabe preguntarse: ¿qué fracción de la varianza (p. ej., de Y, la precipitación) explica el primer modo canónico?

1) La **varianza total de Y** es:

$$\mathbf{Var(Y) = tr(S_{YY})}$$ (= 642.34 (mm^2 ?) en este ejemplo)

2) Se puede demostrar que **la varianza de la variable y_i explicada por el modo canónico j** es:

$$\mathbf{Var\ exp_{i,j} = \lambda_j^2\ [(S_{YY}\ B)_{i,j}]^2}$$

3) La **varianza total explicada por el modo canónico j** es:

$$\mathbf{Var\ exp_j = \lambda_j^2 \sum_{i=1}^{i=q} [(S_{YY}\ B)_{i,j}]^2}$$ = 3.94 mm^2 (sólo el 0.6% para j=1 !!!)

4) La **varianza explicada total** (por los 168 modos) es:

$$\textbf{Var exp total} = \textbf{tr}\,(\mathbf{S_{YY}\, B\, \Lambda^2\, B^t\, S_{YY}})$$

(= 549.47 mm^2 (85.5% de la varianza))

Existen expresiones similares para las fracciones de varianza de las variables X, explicada por cada modo canónico j.

Esas fracciones de varianza no son necesariamente decrecientes con los modos (como ocurría en el ACP).

Para el caso q = 1, tenemos como caso particular la regresión múltiple.

Resulta entonces que esos patrones hallados, representan una fracción muy escasa de la varianza total.

Se necesitan muchos modos para representar un % considerable de la varianza.

Entonces, ¿qué se puede hacer?

CCA con prefiltrado usando ACP

La idea es aplicar el CCA no a los campos originales sino a series formadas por algunos componentes principales de cada uno de los campos (los que expliquen más varianza).

¿Por qué?

- Para solucionar el problema anterior y además:
- Si p>n o q>n, en el procedimiento anterior pueden aparecer matrices no invertibles.
- Con el prefiltrado se puede obtener una reducción importante de la dimensión del problema.
- Se filtra el ruido de pequeña escala.
- De paso, el análisis de los EOF y CP puede permitir obtener un mayor conocimiento de los campos.

Desventajas:

•La formulación es (un poco) más complicada

•Problema del truncamiento (igual que en ACP, no es claro cuantos CP retener): cuantos menos retengo, más “ruido” elimino pero corro el riesgo de perder alguna “señal” físicamente relevante.
(Se pueden usar los criterios ya vistos para truncar.)

Con esta formulación, se “heredan” las ventajas y problemas del ACP.

Como primer paso, hallamos los PCs en ambos campos:
Primero X, las anomalías de TSM.

```
CX=cov(X); %  CX es de 240*240
[EX, LX]=eig(CX);
varianzaX=diag(LX)/trace(LX); % vector de fracciones de varianza
PCX=X*EX; % PCX es de 336*240
```

Usaremos el criterio (subjetivo) de retener los componentes necesarios para superar el 70% de la varianza en cada campo.

```
% con 2 componentes se supera el 70%  (87.8 %)
PCX1=PCX(:,end);
PCX2=PCX(:,end-1);
```

Idem para la precipitación (Y):
se llega a que son necesarios 5
PCs, que explican 70.9% de la varianza.

Luego el segundo paso es aplicar el CCA a los nuevos "campos" representados por las 2 matrices formadas por los CPs que retuvimos en cada campo, una de 336*2 (**X***, la nueva "**X**") y otra de 336*5 (**Y***, la nueva "**Y**"), donde cada una concentra gran parte de la varianza de los datos originales.

$$X^* = (PCX1 \quad PCX2)_{336*2} \qquad Y^* = (PCY1 \quad PCY2 \ldots\ldots PCY5)_{336*5}$$

Luego:

$$[A^*, B^*, r^*, U^*, V^*] = \text{canoncorr}(X^*, Y^*);$$

Valen las relaciones para varianzas explicadas ya vistas.

Ahora r = min (p,q) = 2, de modo que sólo habrá 2 modos canónicos.

Se tiene: $\lambda_1 = 0.61$ y $\lambda_2 = 0.38$

Con sólo 5 vectores por un lado y 2 por otro, no fue posible obtener una correlación máxima mayor que 0.61 (cuando antes era 1).

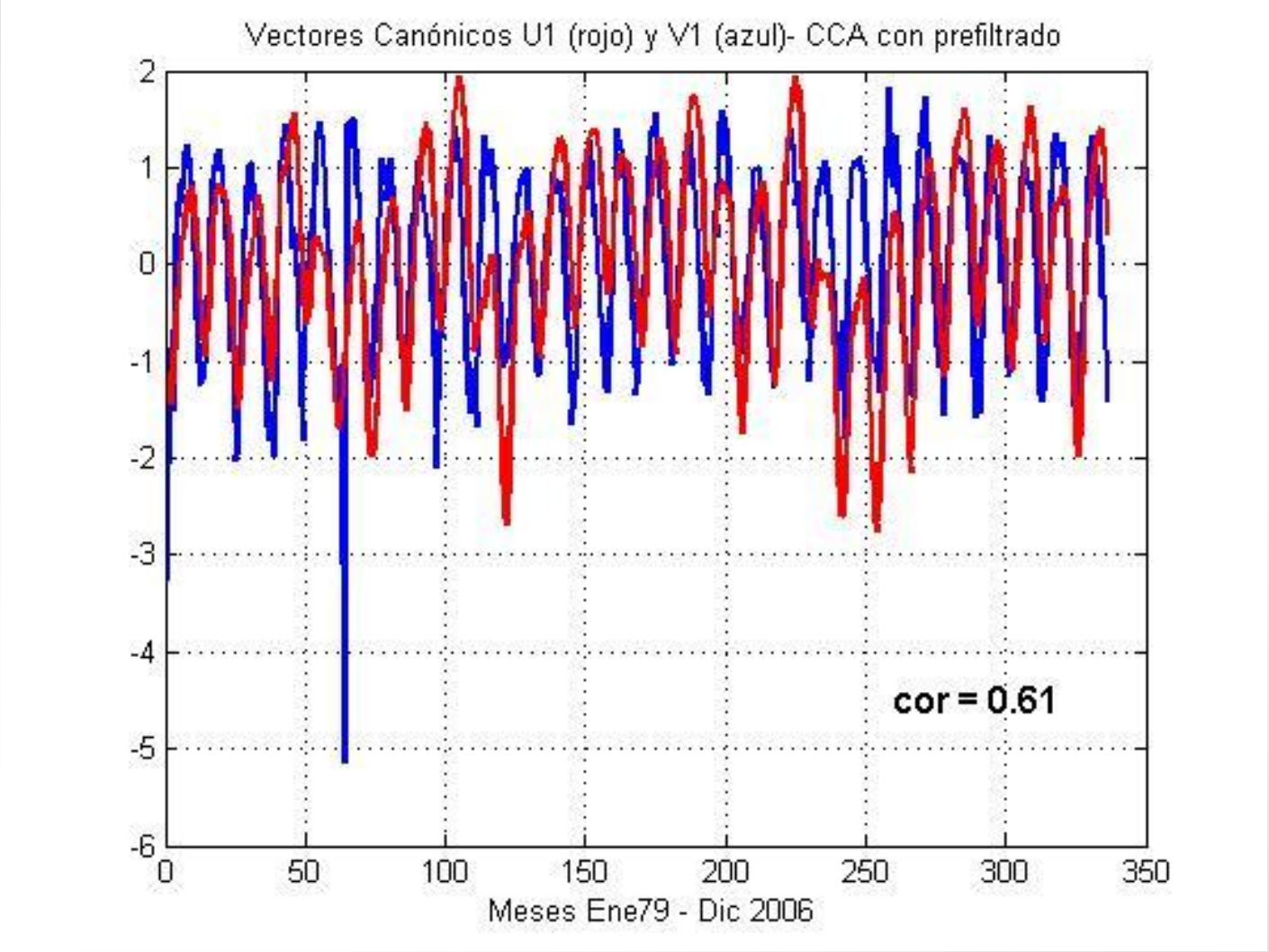
Ahora la varianza explicada por el primer modo es el 14.8% de la varianza total de las precipitaciones.

La varianza explicada por el segundo modo es aprox. el 3% de la total.

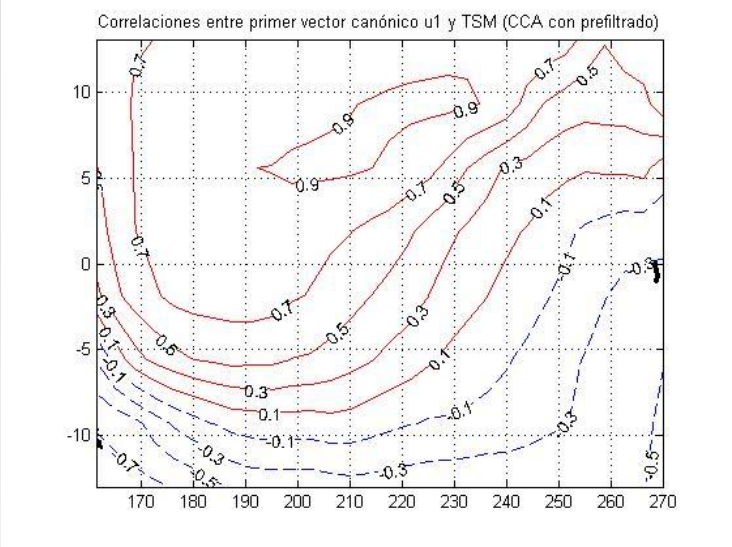
Tenemos no mucha varianza explicada pero está concentrada en 2 modos; antes teníamos más pero muy dispersa en muchos modos.

Si hubiéramos hecho el truncamiento, por ej. en 80% de la varianza, quizá tendríamos más varianza explicada, pero también más “ruido”.

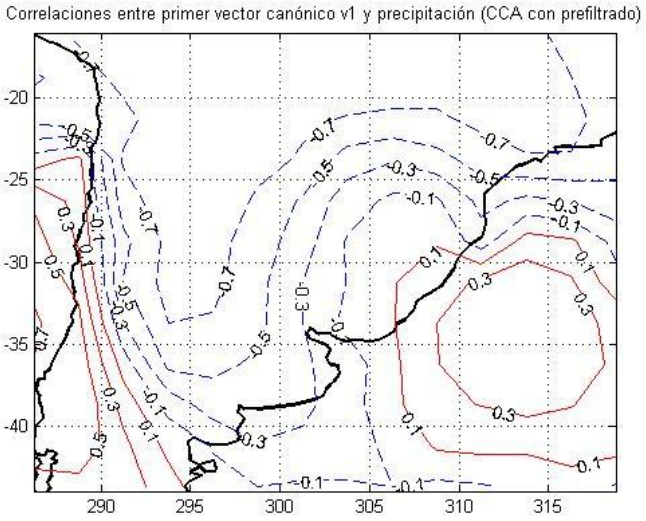
Vectores Canónicos U_1^* (rojo) y V_1^* (azul)- CCA con prefiltrado



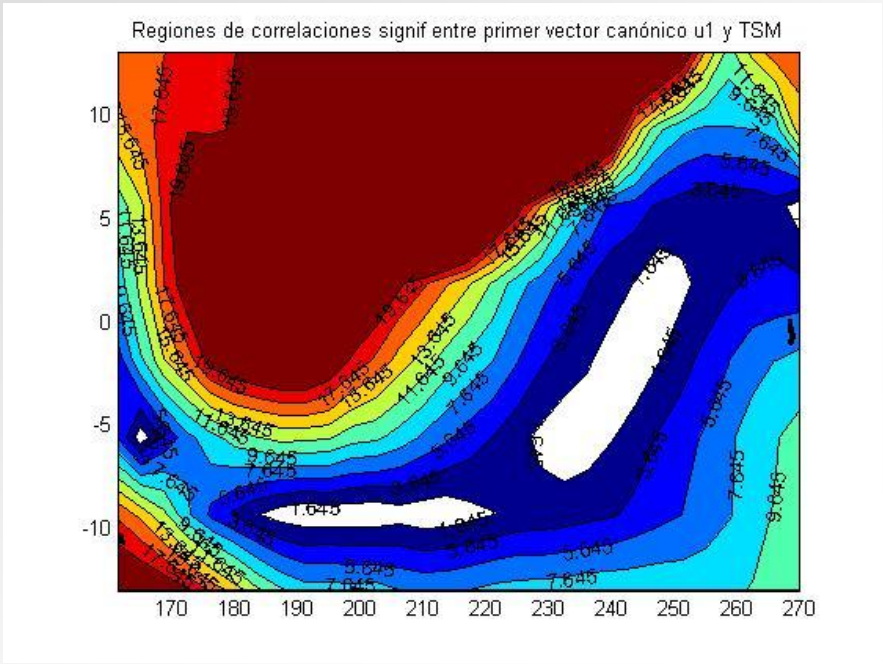
Correlaciones entre U_1 y TSM (CCA pref)



Correlaciones entre V_1 y precip (CCA pref)

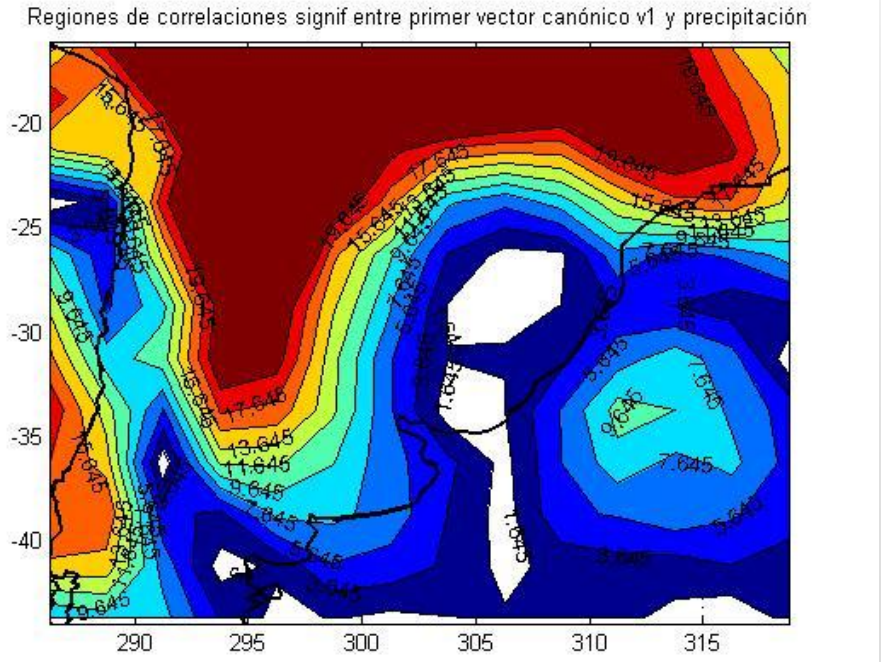


Regiones significativas al 5% en campo de TSM



Existen todos estos elementos también para el segundo modo.

Regiones significativas al 5% en campo de prec.



Aplicación a pronóstico

Una forma de aplicar este método a la predicción es considerando campos desfasados en el tiempo (por ejemplo Y hacia el futuro), y estimar por mínimos cuadrados variables de un conjunto (Y) a partir de las del otro (X).

También aparecen aquí los modos con los vectores canónicos, mapas de correlación y varianzas explicadas. Las regiones significativas serán indicativas de la predictibilidad.