

Modelos estadísticos para predecir la concentración de CO y NO₂



Ingeniería en Automática
y Electrónica Industrial
Autor: Raúl Bartolomé
Profesor: Eduard Llovet
Curso 2001/2

Objetivos del estudio

Análisis y predicción de la concentración de CO y NO₂ con los siguientes métodos estadísticos:

- **PCA:** Análisis de Componentes Principales
Supervisado, lineal, no paramétrico, clasificación/presentación
- **CA:** Análisis de Clusters
Euclidiano: Supervisado, lineal, no paramétrico, clasificación
No euclidiano: Supervisado, no lineal, no paramétrico, clasificación
- **PLS:** Regresión de Mínimos Cuadrados Parciales
Supervisado, no lineal, no paramétrico, clasificación

Matrices de datos

Matriz de respuesta X (n x p)

- 7 columnas = sensores (p)
S(1) ... S(p)
- 48 filas = muestras en ppm (n)
12 grupos de 4 muestras (i=1..4)
3 clases de gases Co, NO2 y Co+NO2
4 grupos de 4 muestras por clase

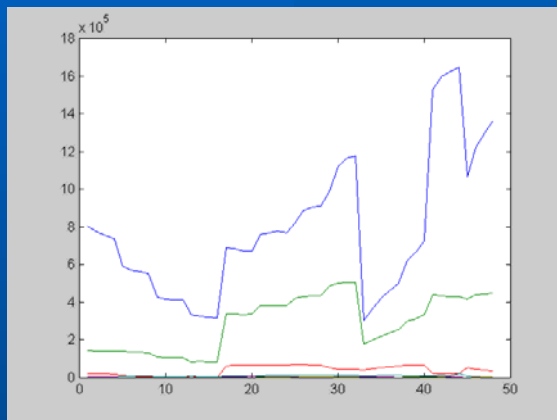
Matriz de concentraciones Y (n x 1)

- 1 columna = concentraciones en ppm
- 48 filas = valores de las concentraciones en ppm (n)

Gas	ppm	S(1)	S(2)	...	S(p-1)	S(p)
CO	20	1,1	1,2	...	1,p-1	1,p
	
	40	1i,1	1i,2	...	1i,p-1	1i,p
		2i,1	2i,2	...	2i,p-1	2i,p
		3i,1	3i,2	...	3i,p-1	3i,p
130	4i,1	4i,2	...	4i,p-1	4i,p	
NO2	10	5i,1	5i,2	...	5i,p-1	5i,p
	20	6i,1	6i,2	...	6i,p-1	6i,p
	40	7i,1	7i,2	...	7i,p-1	7i,p
	60	8i,1	8i,2	...	8i,p-1	8i,p
CO + NO2	20+10	9i,1	9i,2	...	9i,p-1	9i,p
	40+20	10i,1	10i,2	...	10i,p-1	10i,p
	80+40	11i,1	11i,2	...	11i,p-1	11i,p
		12i,1	12i,2	...	12i,p-1	12i,p
	130+60
		n,1	n,2	...	n,p-1	n,p

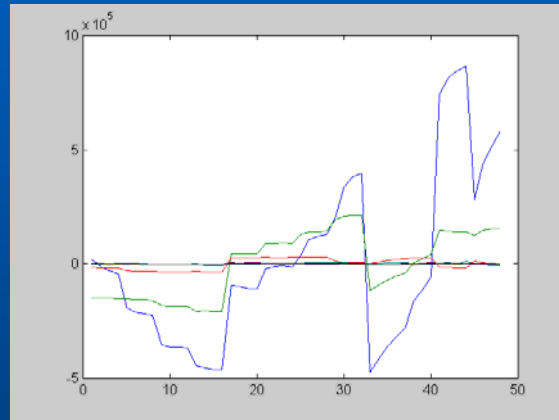
Preprocesamiento

Representación gráfica de la matriz X...



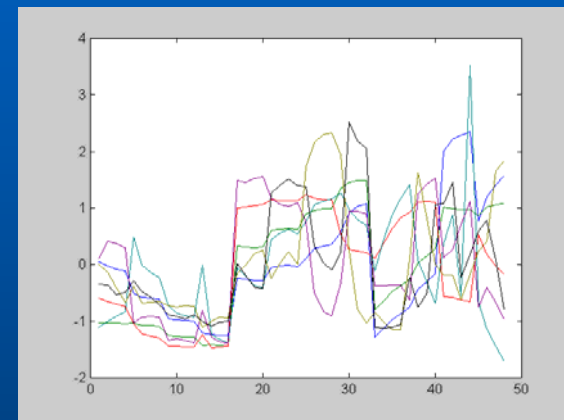
Sin preprocesamiento

Sólo se aprecian 3 sensores
En los modelos estadísticos se considerarían muchos sensores como irrelevantes



● Centrada

Eliminación del offset
Se resta la media a cada muestra



● Autoescalada

Todos los sensores poseen la misma varianza unidad
Se divide por la varianza cada muestra

Análisis de Componentes Principales PCA

Dada una matriz de respuesta X ($n \times p$) el PCA descompone X en:

$$X = t_1 p_1^t + t_2 p_2^t + \dots + t_k p_k^t + \varepsilon = TP^t + \varepsilon$$

$$k \leq \min(n, p)$$

Matriz de scores T ($n \times k$)

$$T = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1k} \\ t_{21} & t_{22} & \dots & t_{2k} \\ \dots & \dots & \dots & \dots \\ t_{n1} & t_{n2} & \dots & t_{nk} \end{bmatrix}$$

Matriz de loadings P ($p \times k$)

$$P = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1k} \\ p_{21} & p_{22} & \dots & p_{2k} \\ \dots & \dots & \dots & \dots \\ p_{p1} & p_{p2} & \dots & p_{pk} \end{bmatrix}$$

Matriz de residuos E ($n \times p$)

$$\varepsilon = X - TP^t$$

Es la varianza no contemplada por el modelo PCA

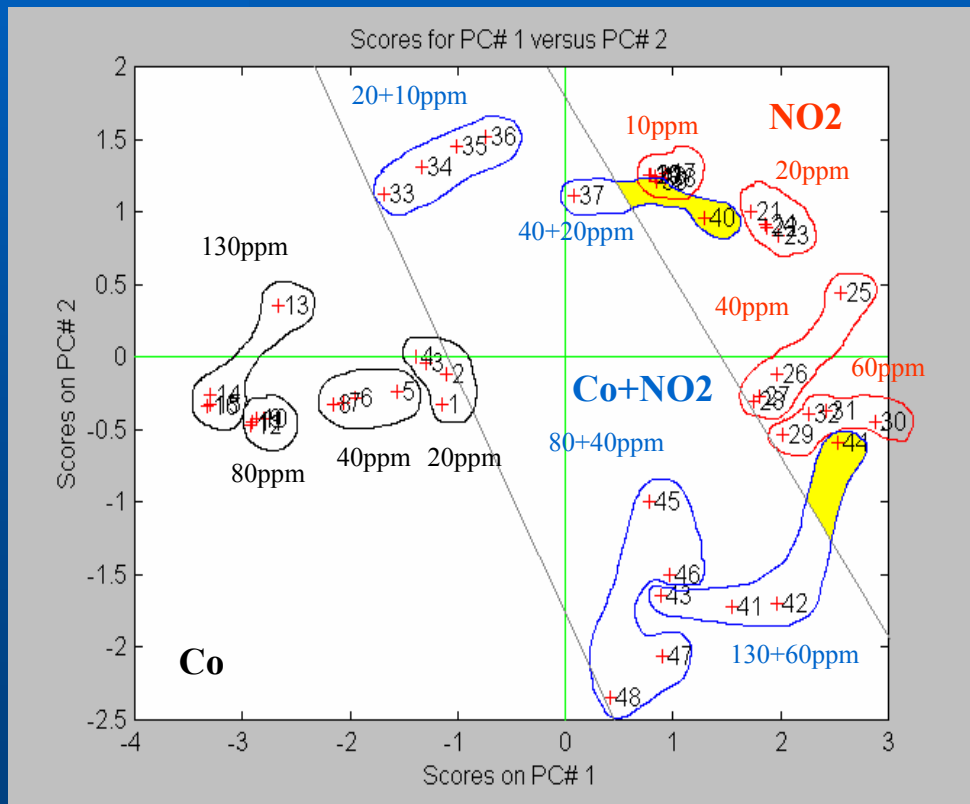
Scores para PC#2

Loading para PC#2

- Los PCs son ortogonales entre si
- El PC#1 sigue la dirección de la máxima varianza de X
- El PC#2 sigue la dirección de la segunda máxima varianza de X ...

Análisis de Componentes Principales PCA

Matriz X autoescalada

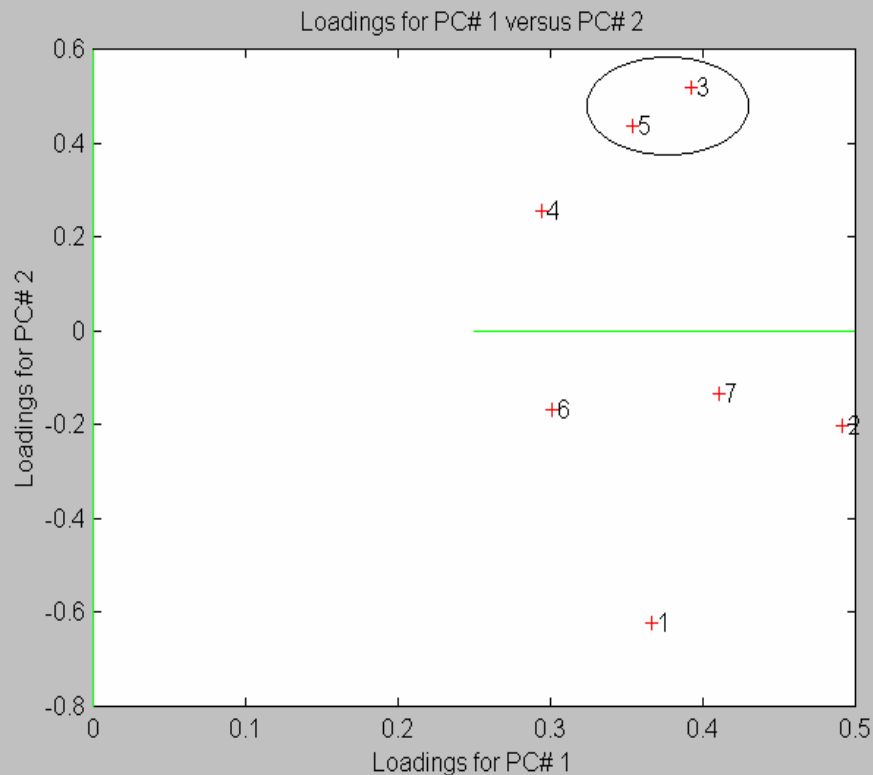


Porcentaje de varianza capturado por modelo PCA		
Número PC	Varianza PC actual %	Varianza acumulada %
1	53.60	53.60
2	14.22	67.82
3	13.95	81.78
4	10.57	92.35
5	4.85	97.20
6	2.37	99.57
7	0.43	100.00

Para una clasificación en ...

- 3 clases las muestras 38, 39 40 y 44 quedarían mal clasificadas
- 12 clases las muestras 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49 quedarían mal clasificadas

Análisis de Componentes Principales PCA



La distancias entre los loadings vecinos es parecida

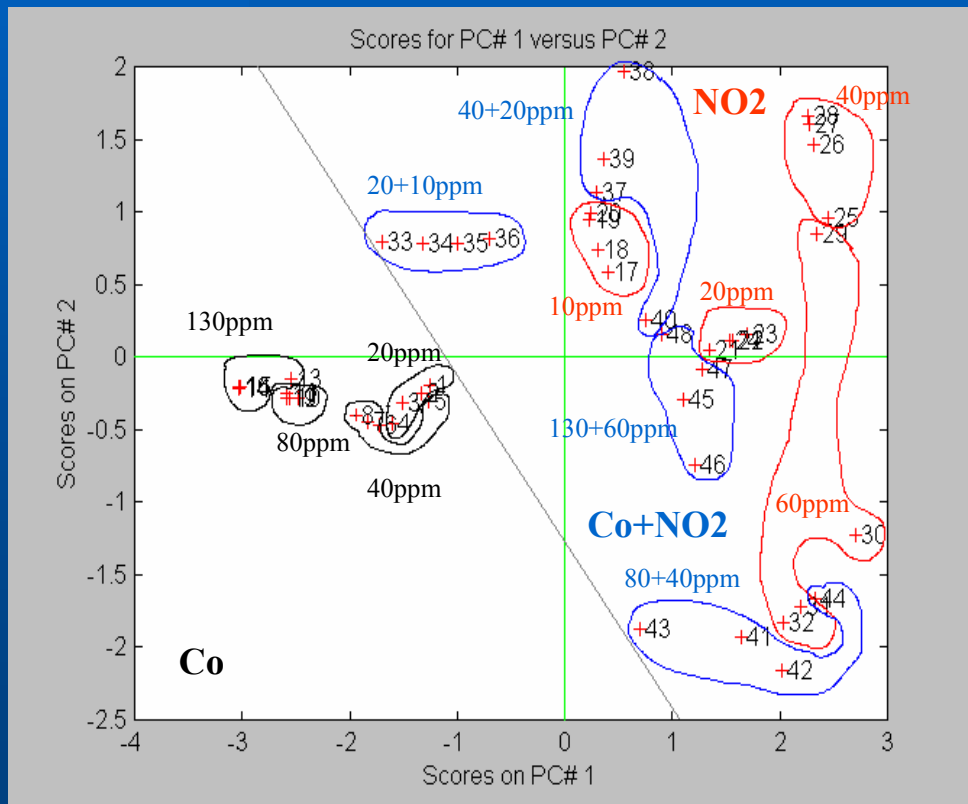


Por lo tanto no existe sensor alguno claramente superfluo

¿ Cómo afectaría la simplificación de los sensores 3 y 5?

Análisis de Componentes Principales PCA

Matriz X sin S5 y autoescalada



Porcentaje de varianza capturado por modelo PCA

Número PC	Varianza PC actual %	Varianza acumulada %
1	53.60	53.60
2	14.22	67.82
3	13.95	81.78
4	10.57	92.35
5	4.85	97.20
6	2.37	99.57
7	0.43	100.00

Para una clasificación en ...

- 3 clases sólo se puede clasificar bien el Co
- 12 clases la tarea de clasificación resulta muy poco efectiva o errónea

Análisis de Clusters CA

Dada una matriz de respuesta X ($n \times p$) el CA realiza los siguientes pasos:

- Generación de información de similitud entre medidas:
 - Distancia euclídea (lineal)
 - Distancia Mahalanobis (no lineal)
- Agrupación de medidas en un árbol jerárquico
- Determinación de un umbral para delimitar las agrupaciones

El CA las distancia entre grupos de puntos son definidos como ...

- Distancia entre centroides de los grupos (K-MNG)
- Distancia entre los vecinos más cercanos de cada grupo (K-NN)

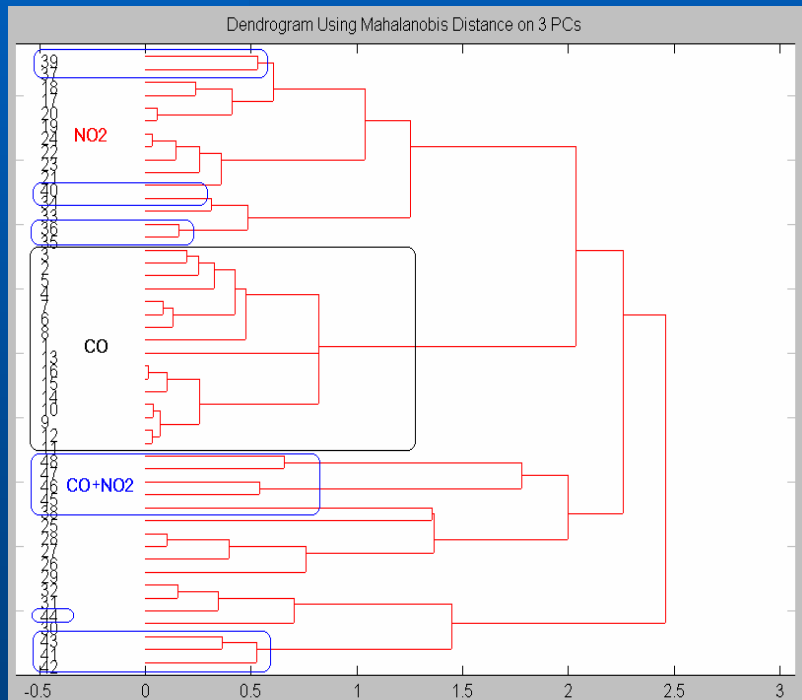
Para los siguientes análisis CA ...

- La matriz X esta autoescalada
- Se realiza un PCA previo con 3PCs

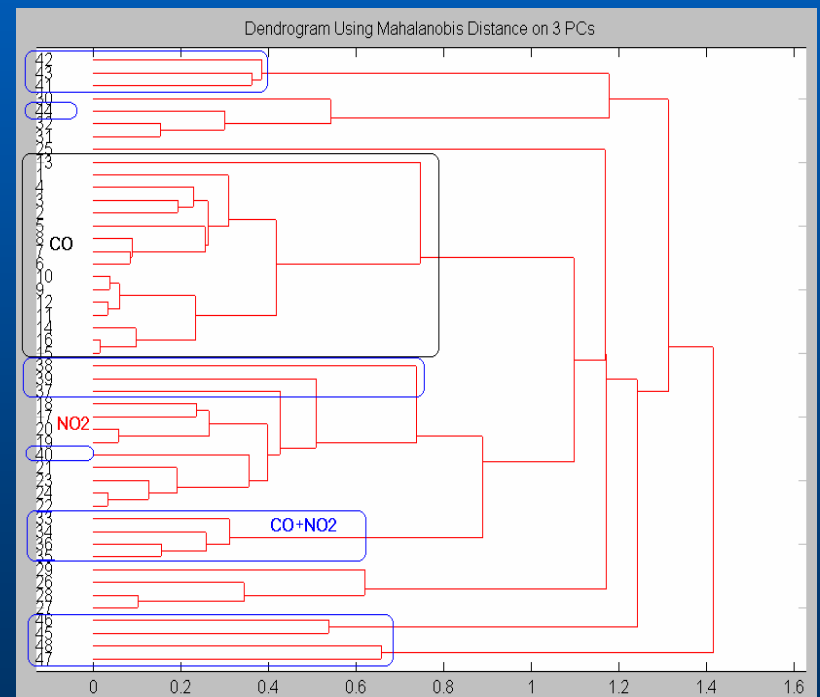
Análisis de Clusters CA

Dendogramas utilizando la distancia de Mahalanobis

● K-MNG



● K-NN

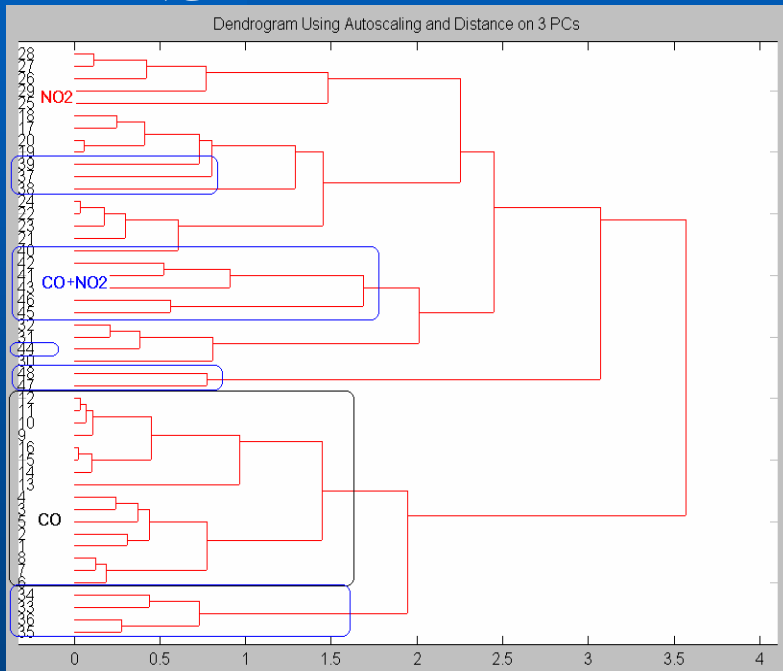


No existe un umbral de agrupación que satisfaga la clasificación en 3 clases y aún menos en 12 clases

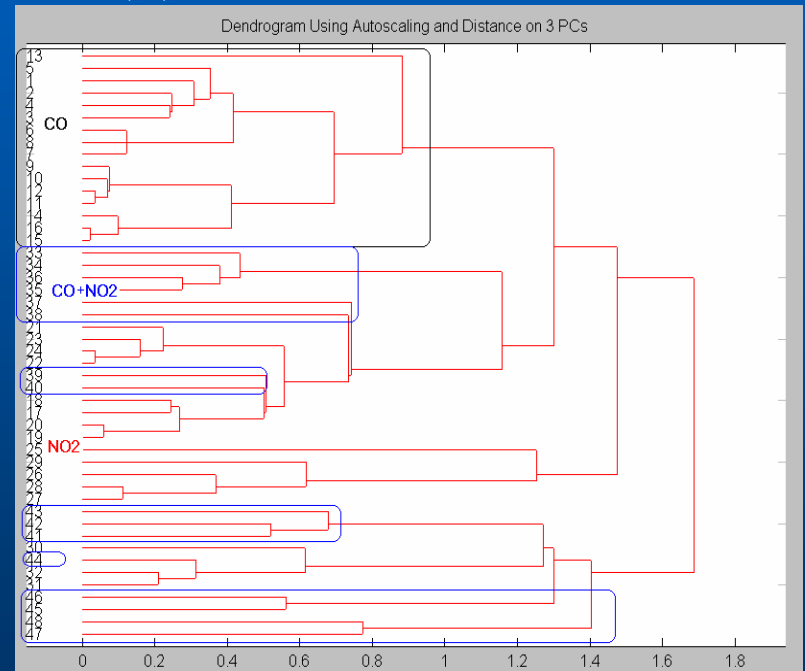
Análisis de Clusters CA

Dendogramas utilizando la distancia de Euclídea

● K-MNG



● K-NN



Ocurre nuevamente la misma situación

Obsérvese por ejemplo como la muestra 44 siempre se clasifica mal

Regresión de Mínimos Cuadrados Parciales PLS

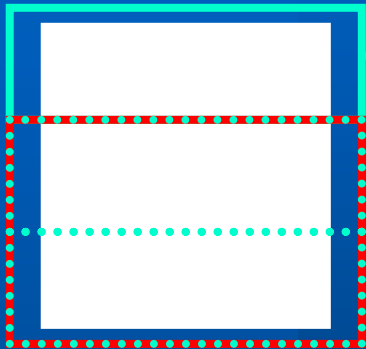
- El PLS busca factores que capturen varianza en X (predictor variables) y son correlacionados con Y (predicted variables)
- PLS intenta maximizar la covarianza

Antes del PLS es necesaria una validación cruzada PLS para determinar el número óptimo de factores LV (latent variables)

Regresión de Mínimos Cuadrados Parciales

Validación Cruzada PLS

Matriz X



Muestras utilizadas para la validación

Muestras utilizadas para construir el modelo PLS

Proyección de los scores sobre los LVs

PLS-1LV

PLS-2LV

...

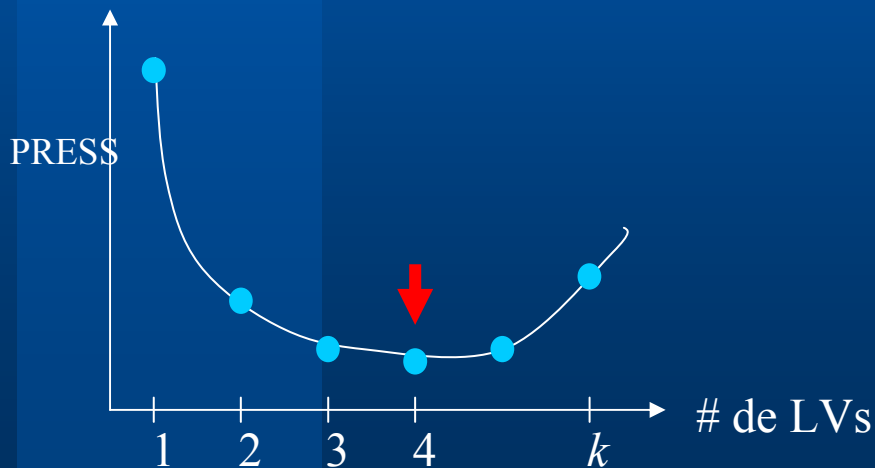
PLS-kLV

Predicción para las muestras de validación

PredErr
PLS-1LV

PredErr
PLS-2LV

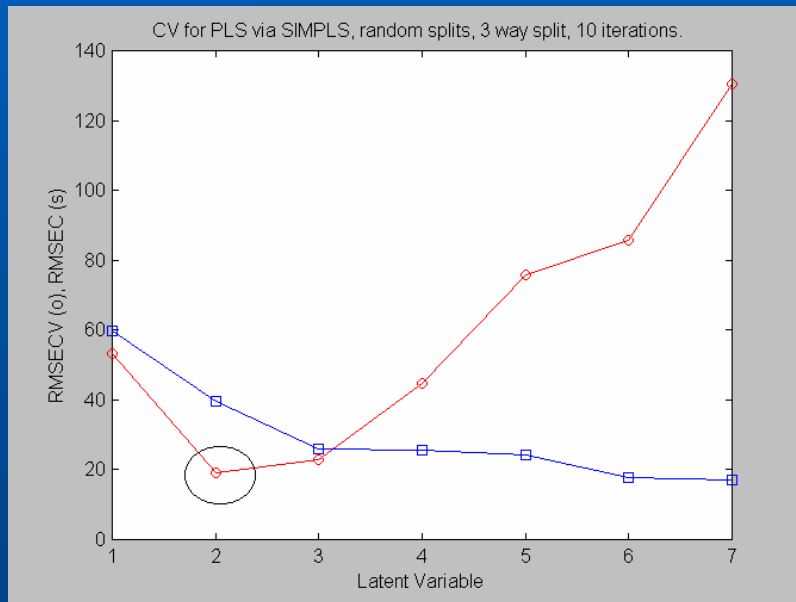
PredErr
PLS-kLV



Regresión de Mínimos Cuadrados Parciales

Validación Cruzada PLS para CO

Matriz de datos X y concentraciones Y sin preprocesar



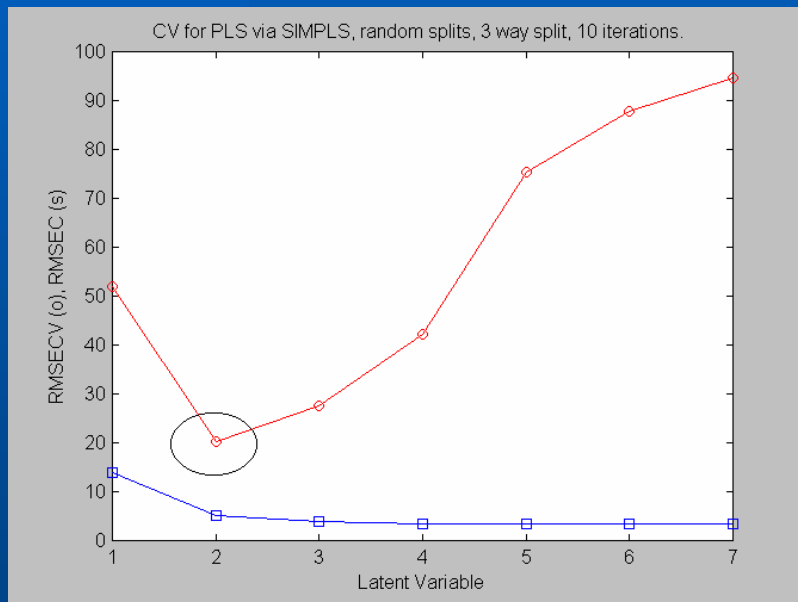
Porcentaje de varianza capturado por el modelo PLS				
LV	Bloque de entrada X		Bloque de salida Y	
	LV actual %	Acumulada %	LV actual %	Acumulada %
1	99.89	99.89	43.36	43.36
2	0.11	100.00	31.76	75.12
3	0.00	100.00	14.32	89.44
4	0.00	100.00	0.10	89.55
5	0.00	100.00	1.03	90.58
6	0.00	100.00	4.35	94.93
7	0.00	100.00	0.49	95.42

El porcentaje de varianza capturada para el bloque de salida es muy bajo !
 Esto se puede mejorar preprosanado las matrices....

Regresión de Mínimos Cuadrados Parciales

Validación Cruzada PLS para CO

Matriz de datos X y concentraciones Y centradas



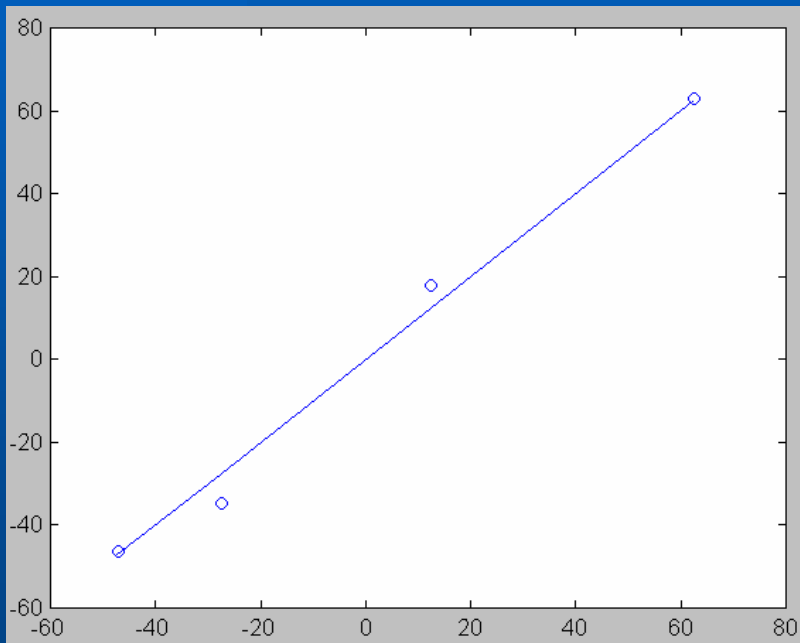
Porcentaje de varianza capturado por el modelo PLS				
LV	Bloque de entrada X		Bloque de salida Y	
	LV actual %	Acumulada %	LV actual %	Acumulada %
1	99.71	99.71	88.98	88.98
2	0.27	99.98	9.58	98.56
3	0.01	100.00	0.59	99.15
4	0.00	100.00	0.17	99.32
5	0.00	100.00	0.01	99.33
6	0.00	100.00	0.04	99.37
7	0.00	100.00	0.01	99.37

El porcentaje de varianza capturada para el bloque de salida se ha mejorado notablemente a costa de un insignificante decremento en el bloque de entrada

Regresión de Mínimos Cuadrados Parciales

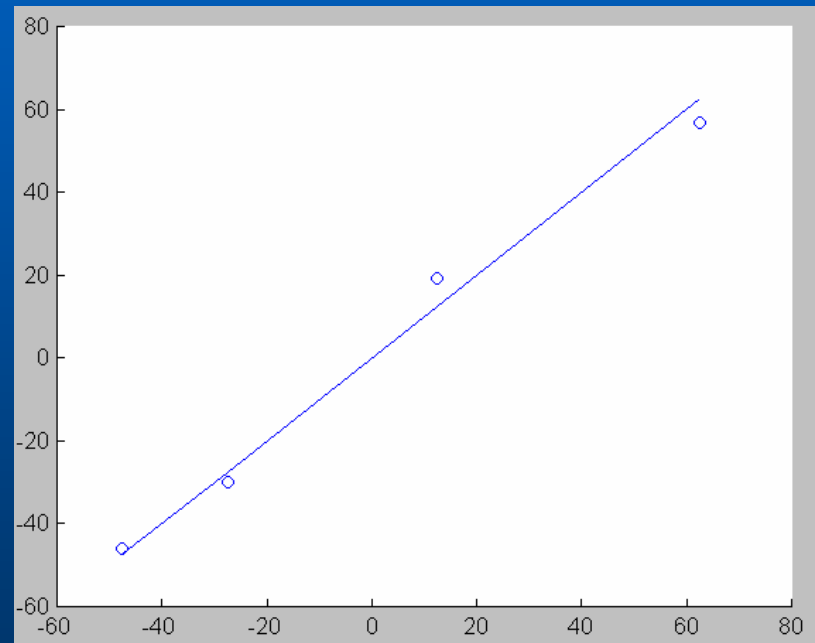
Modelo PLS para CO

Validación 1



Varianza capturada X =100.00% Y=98.81%
Error Relativo Acumulado = 14.43

Validación 2

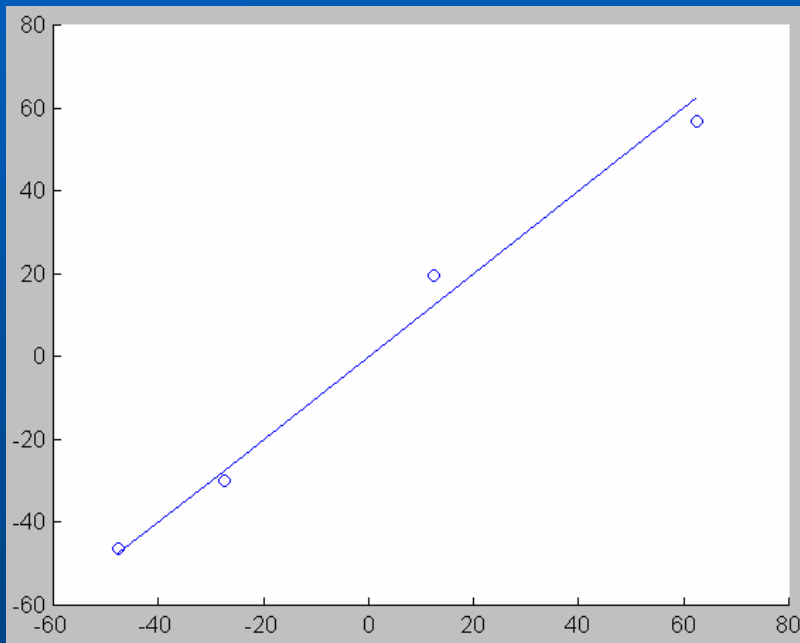


Varianza capturada X =99.98% Y=98.54%
Error Relativo Acumulado = 16.79

Regresión de Mínimos Cuadrados Parciales

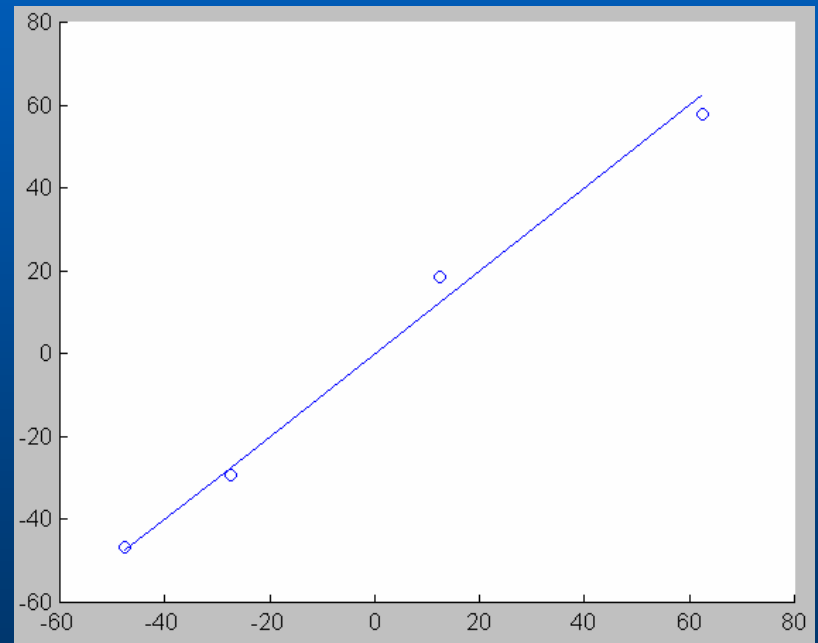
Modelo PLS para CO

Validación 3



Varianza capturada X =99.98% Y=98.46%
Error Relativo Acumulado = 16.44

Validación 4

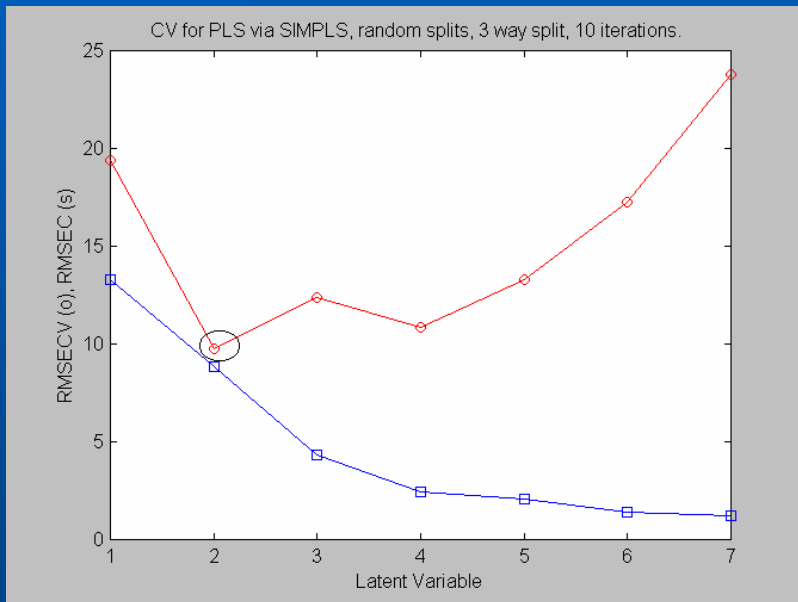


Varianza capturada X =99.98% Y=98.65%
Error Relativo Acumulado = 13.45

Regresión de Mínimos Cuadrados Parciales

Validación Cruzada PLS para NO2

Matriz de datos X y concentraciones Y sin preprocesar



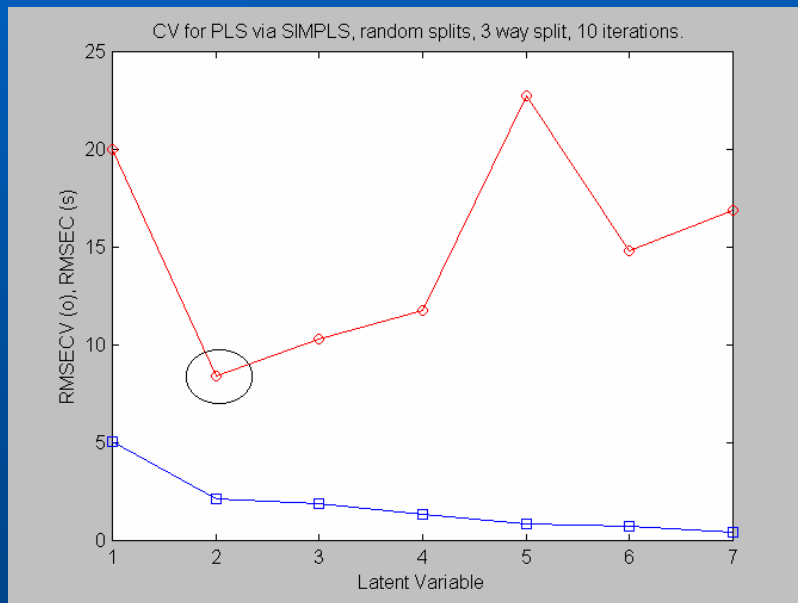
Porcentaje de varianza capturado por el modelo PLS				
LV	Bloque de entrada X		Bloque de salida Y	
	LV actual %	Acumulada %	LV actual %	Acumulada %
1	99.91	99.91	87.56	87.56
2	0.08	99.99	6.92	94.49
3	0.01	100.00	4.20	98.69
4	0.00	100.00	0.88	99.58
5	0.00	100.00	0.13	99.70
6	0.00	100.00	0.16	99.86
7	0.00	100.00	0.04	99.90

En este caso la varianza capturada por el bloque de salida podría ser suficiente, de todos modos se vuelven a autoescalar las matrices

Regresión de Mínimos Cuadrados Parciales

Validación Cruzada PLS para NO2

Matriz de datos X y concentraciones Y centradas

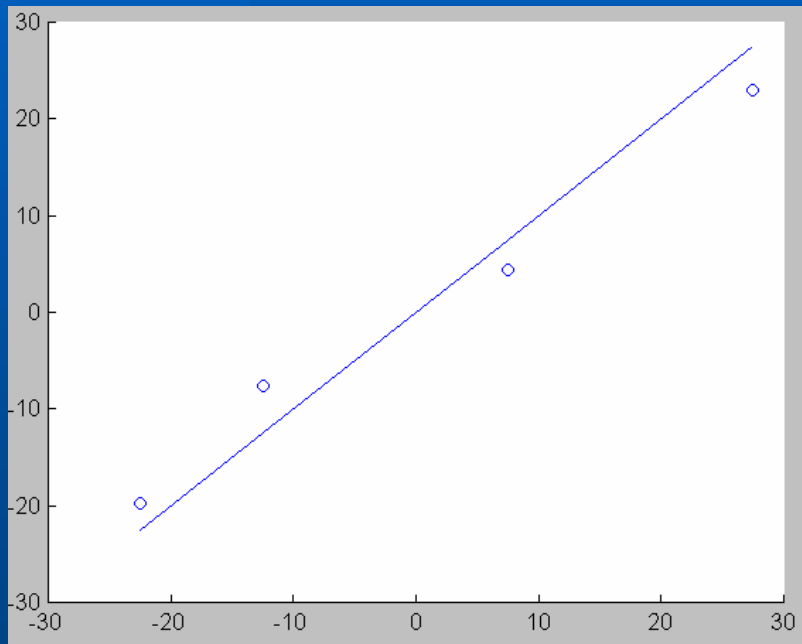


Porcentaje de varianza capturado por el modelo PLS				
LV	Bloque de entrada X		Bloque de salida Y	
	LV actual %	Acumulada %	LV actual %	Acumulada %
1	99.43	99.43	93.01	93.01
2	0.51	99.94	5.73	98.74
3	0.05	99.99	0.30	99.04
4	0.01	100.00	0.47	99.51
5	0.00	100.00	0.31	99.82
6	0.00	100.00	0.03	99.84
7	0.00	100.00	0.10	99.95

Regresión de Mínimos Cuadrados Parciales

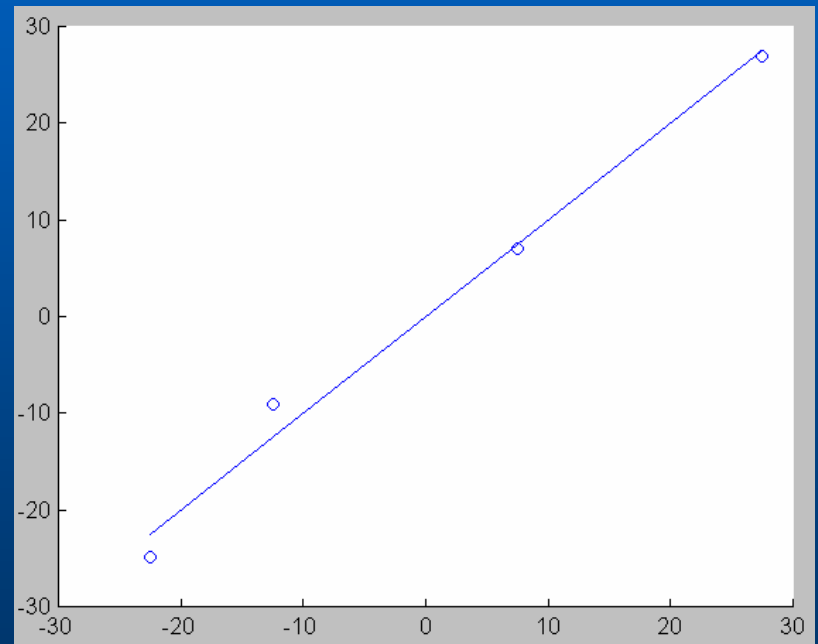
Modelo PLS para NO₂

Validación 1



Varianza capturada X =99.98% Y=99.15%
Error Relativo Acumulado = 15.55

Validación 2

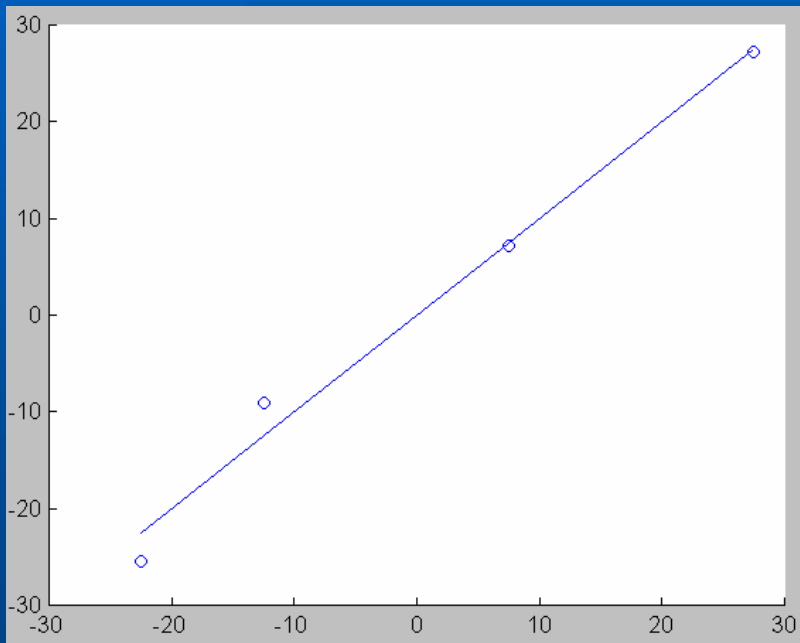


Varianza capturada X =99.93% Y=98.75%
Error Relativo Acumulado = 6.83

Regresión de Mínimos Cuadrados Parciales

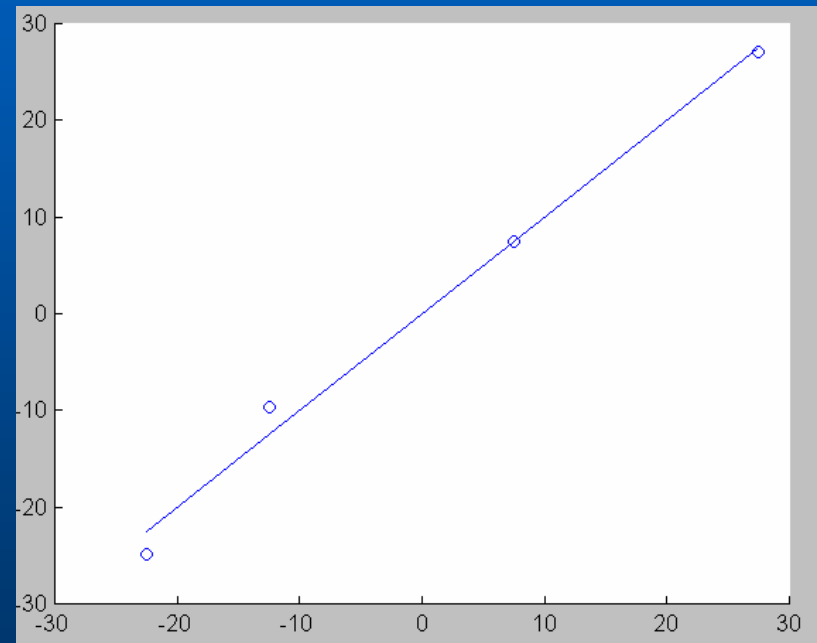
Modelo PLS para NO₂

Validación 3



Varianza capturada X =99.93% Y=98.81%
Error Relativo Acumulado = 6.93

Validación 4



Varianza capturada X =99.93% Y=98.67%
Error Relativo Acumulado = 5.76