Estadística Industrial – Ingeniero en organización industrial
Universidad Politécnica de Cartagena

Grupo 8

Javier Moreno Cortés
Ana Salmerón Baños
Antonio José Barcelona Vinadel

Problema 2 – Junio 2010

En el fichero **PFactorialEx.sav** se encuentran los datos correspondientes a la contaminación del aire en 80 ciudades de US para el año 1960. Las variables se definen a continuación.

TMR: Tasa de mortalidad.
SMEAN: Media aritmética de cantidades de sulfato.
SMAX: Cantidades máximas de sulfato cada dos semanas.
PMEAN: Media aritmética de partículas suspendidas.
PMAX: Cantidad máxima de partículas suspendidas.
PM2: Densidad de población por milla cuadrada.
GE65:Porcentaje de población con 65 años o más.
PERWH: Porcentaje de blancos en la población.
NONPOOR: Porcentaje de familias con ingresos sobre el nivel de la pobreza.
LPOP: Logaritmo (base 10) de la población.

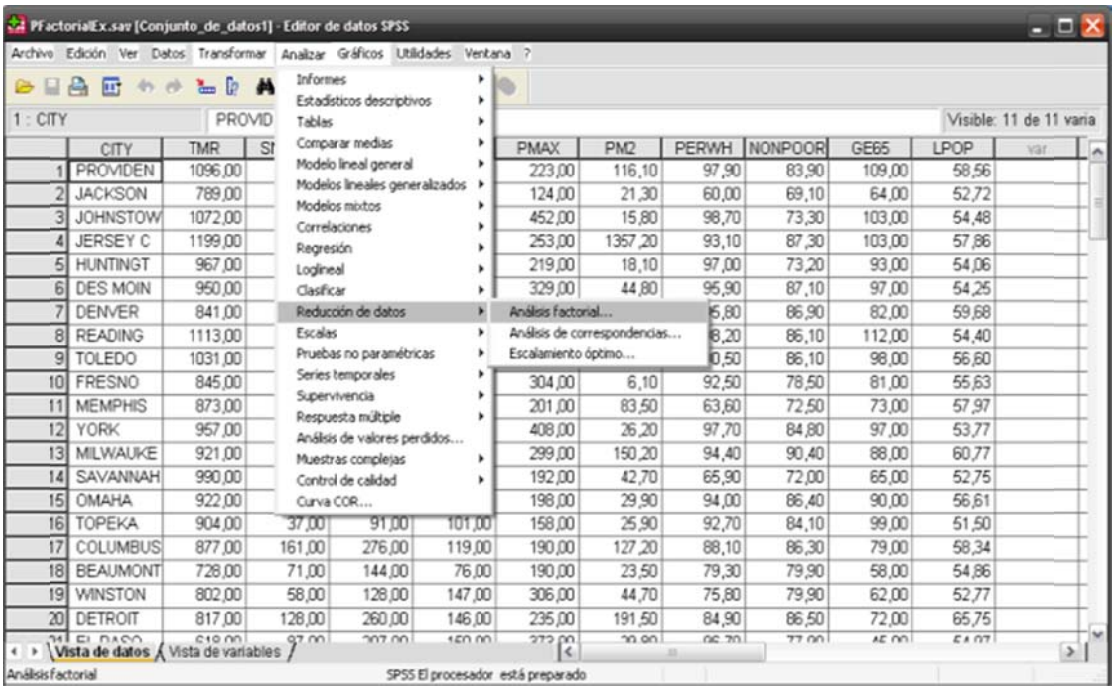
Se pretende aplicar una técnica de análisis multivariante que permita explicar la estructura de dependencia entre las variables del problema partir de unos pocos factores.

1. Explica razonadamente que tipo de análisis parece más adecuado para este estudio.

Como el objetivo de este problema es estudiar si las variables se pueden explicar con unos pocos factores comunes a todas estas variables, vamos a realizar un análisis factorial.

2. Determina si el análisis propuesto es adecuado en función de los datos muestrales. Indica qué procedimientos permiten responder a esta cuestión, explica en qué consisten y comenta los resultados obtenidos con SPSS.

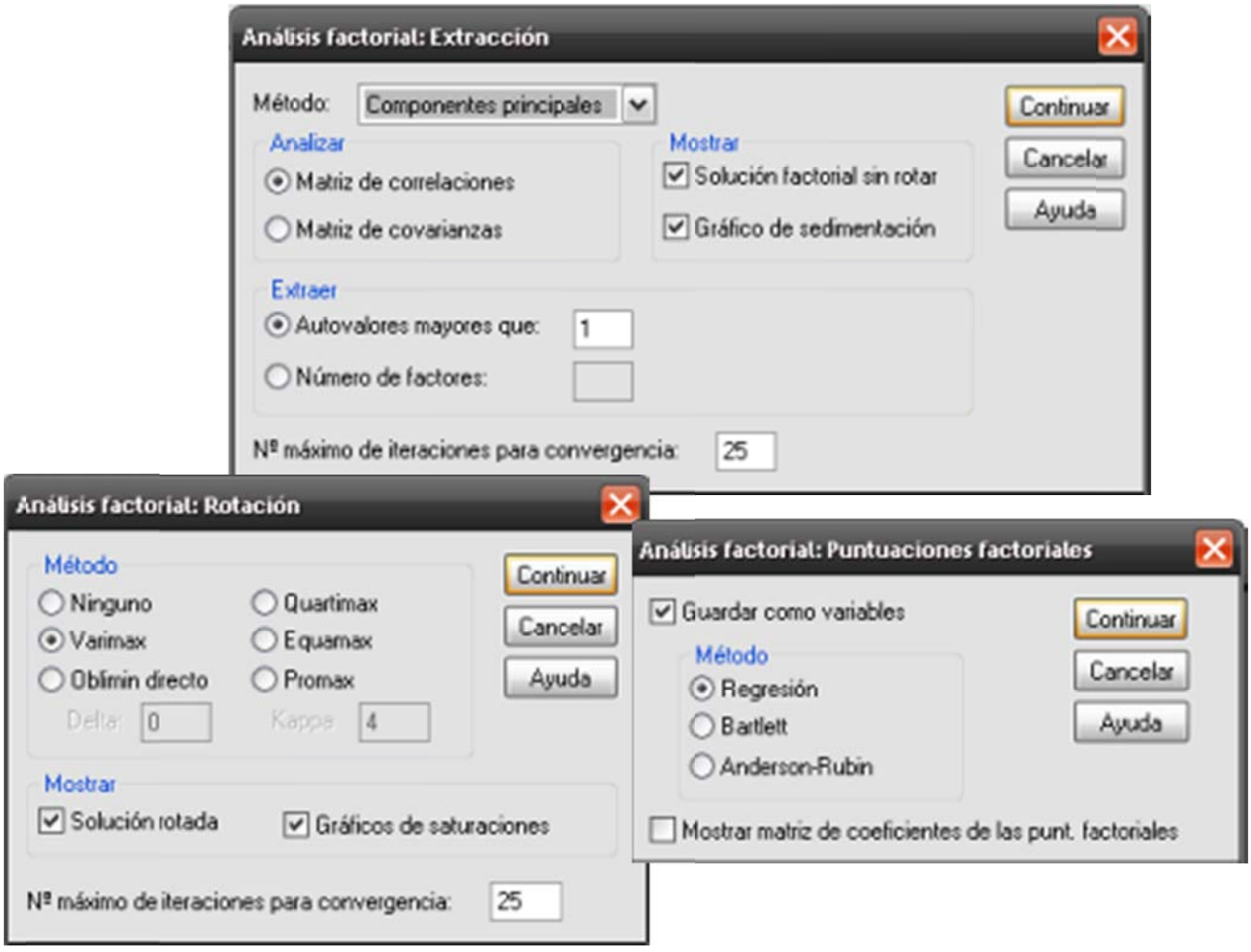
Para llevar a cabo el análisis, en la pantalla de datos, buscaremos pincharemos en Analizar >> Reducción de datos >> Análisis factorial.



Que nos llevará al siguiente cuadro, y en él deberemos mover las variables al cuadro de la derecha

En este menú marcaremos tantas opciones como nos sean necesarias para la solución de todas las cuestiones que nos propone el problema.

Las opciones marcadas en los submenús Descriptivos, Extracción, Rotación, Puntuaciones y Opciones quedan como los vemos en las siguientes ventanas.



El submenú opciones podríamos seleccionar que ordene las variables atendiendo a las cargas factoriales, lo que permite visualizar de forma más rápida qué variables están principalmente relacionadas con cada factor.

Con respecto a la pregunta sobre que procedimientos permiten dar respuesta a si el análisis propuesto es adecuado en función de los datos, nos sirve la matriz de correlación, la prueba de esfericidad de Bartlett, la prueba de Kaiser – Meyer – Olkin y la medida de adecuación muestral de cada variable.
Las dos primeras (correlaciones y test de Barlett) sirven para determinar si las variables de nuestro problema presentan algunas correlaciones significativas, mientras que las dos últimas (KMO y adecuación muestral)

sirven para cuantificar si los datos de nuestro problema permitirán un esquema factorial: todas las variables explicadas por unos pocos factores comunes y unas partes específicas de cada variable que presentan incorrelación con el resto. Para más detalles acerca de las expresiones numéricas (que contemplan las correlaciones simples y las parciales) véase los apuntes.

Para analizar si el análisis propuesto es adecuado en función de nuestros datos muestrales, la matriz de correlaciones nos dice que...

Matriz de correlaciones[a]

		TMR	SMEAN	SMAX	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
Correlación	TMR	1,000	,319	,120	-,068	-,135	,271	,299	,182	,860	,087
	SMEAN	,319	1,000	,832	,554	,339	,421	,208	,332	,192	,377
	SMAX	,120	,832	1,000	,560	,474	,196	,214	,250	,065	,256
	PMEAN	-,068	,554	,560	1,000	,657	,163	,179	,204	-,114	,304
	PMAX	-,135	,339	,474	,657	1,000	-,010	,099	,134	-,147	,118
	PM2	,271	,421	,196	,163	-,010	1,000	,057	,221	,115	,265
	PERWH	,299	,208	,214	,179	,099	,057	1,000	,637	,528	,064
	NONPOOR	,182	,332	,250	,204	,134	,221	,637	1,000	,256	,417
	GE65	,860	,192	,065	-,114	-,147	,115	,528	,256	1,000	,095
	LPOP	,087	,377	,256	,304	,118	,265	,064	,417	,095	1,000
Sig. (Unilateral)	TMR		,002	,144	,273	,116	,007	,004	,053	,000	,220
	SMEAN	,002		,000	,000	,001	,000	,032	,001	,044	,000
	SMAX	,144	,000		,000	,000	,041	,028	,013	,282	,011
	PMEAN	,273	,000	,000		,000	,074	,056	,035	,158	,003
	PMAX	,116	,001	,000	,000		,465	,191	,119	,096	,148
	PM2	,007	,000	,041	,074	,465		,307	,024	,155	,009
	PERWH	,004	,032	,028	,056	,191	,307		,000	,000	,288
	NONPOOR	,053	,001	,013	,035	,119	,024	,000		,011	,000
	GE65	,000	,044	,282	,158	,096	,155	,000	,011		,200
	LPOP	,220	,000	,011	,003	,148	,009	,288	,000	,200	

a. Determinante = ,002

Vemos que hay muchas variables que presentan correlación, es decir, su p-valor es bueno (círculo rojo), sin embargo, existen otras variables que no esta tan clara su correlación (círculo azul), llegando a valores de 0,465 en el p-valor.

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,563
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	450,343
	gl	45
	Sig.	,000

La prueba de esfericidad de Bartlett se corresponde con el siguiente contraste de hipótesis:

H_0 : Todas las correlaciones son nulas
H_1 : Alguna correlación es no nula

Como el p-valor vale 0, rechazamos la hipótesis nula H_0 y concluimos que existen correlaciones significativas entre las variables de nuestro problema.

Además la prueba de KMO, nos resulta favorable ya que nos interesa que el resultado sea mayor a 0,5, y se cumple

$$0{,}563 > 0{,}5$$

También debemos observar la medida de adecuación muestral para cada variable, que vemos en la siguiente tabla. La de Matriz anti-imagen.

Matrices anti-imagen

		TMR	SMEAN	SMAX	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
Covarianza anti-imagen	TMR	,168	-,044	,024	-,009	,007	-,072	,087	-,039	-,133	,075
	SMEAN	-,044	,179	-,161	-,073	,051	-,111	,019	-,034	,012	-,040
	SMAX	,024	-,161	,232	,014	-,090	,084	-,030	,027	-,003	,017
	PMEAN	-,009	-,073	,014	,387	-,244	-,002	-,092	,076	,041	-,117
	PMAX	,007	,051	-,090	-,244	,503	,041	,033	-,050	-,004	,057
	PM2	-,072	-,111	,084	-,002	,041	,696	-,016	-,026	,058	-,063
	PERWH	,087	,019	-,030	-,092	,033	-,016	,295	-,230	-,122	,170
	NONPOOR	-,039	-,034	,027	,076	-,050	-,026	-,230	,395	,060	-,233
	GE65	-,133	,012	-,003	,041	-,004	,058	-,122	,060	,139	-,084
	LPOP	,075	-,040	,017	-,117	,057	-,063	,170	-,233	-,084	,605
Correlación anti-imagen	TMR	,494[a]	-,255	,122	-,037	,023	-,209	,392	-,150	-,866	,236
	SMEAN	-,255	,649[a]	-,788	-,277	,170	-,313	,084	-,129	,075	-,120
	SMAX	,122	-,788	,654[a]	,048	-,264	,209	-,114	,090	-,015	,044
	PMEAN	-,037	-,277	,048	,682[a]	-,552	-,003	-,271	,195	,178	-,241
	PMAX	,023	,170	-,264	-,552	,659[a]	,069	,086	-,112	-,016	,103
	PM2	-,209	-,313	,209	-,003	,069	,655[a]	-,036	-,050	,185	-,097
	PERWH	,392	,084	-,114	-,271	,086	-,036	,424[a]	-,675	-,604	,402
	NONPOOR	-,150	-,129	,090	,195	-,112	-,050	-,675	,531[a]	,256	-,476
	GE65	-,866	,075	-,015	,178	-,016	,185	-,604	,256	,469[a]	-,290
	LPOP	,236	-,120	,044	-,241	,103	-,097	,402	-,476	-,290	,482[a]

a. Medida de adecuación muestral

Vemos que hay cuatro variables que tienen su medida de adecuación muestral, por debajo de 0,5. Por tanto, una posibilidad a valorar posteriormente consistirá en eliminar alguna de estas variables, empezando por la que tiene el valor más bajo, como porcentaje de blancos en la población que en principio es la variable que menos adecuación muestral presenta. De esta forma probablemente logremos mejorar la bondad del ajuste.

De todas formas, en primer lugar continuaremos el problema sin descartar ninguna variable de nuestro modelo factorial.

3. Lleva a cabo el análisis propuesto respondiendo a las siguientes cuestiones y comentando los procedimientos utilizados en cada caso:

   (a) Determina el número de factores a retener.

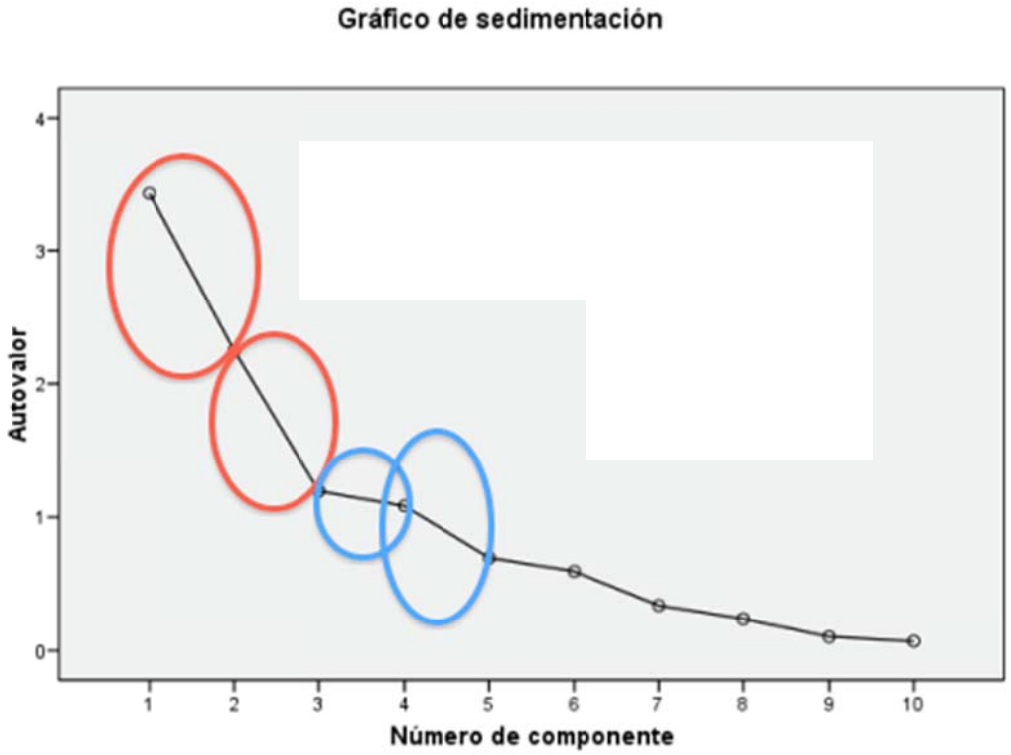
Atendiendo al criterio de Kaiser, vemos en la siguiente tabla que existen 4 valores propios mayores que "1" (incluso mayores de "0.7"), así que este criterio conduce a retener 4 factores.
Si nos basamos en el porcentaje de variabilidad explicado, estos 4 factores explicarían un 79,609 %, para explicar más del 90 % deberíamos coger 6 factores, lo cual no simplificaría el problema en gran medida. Además, en el caso de realizar un análisis factorial, sabemos que el porcentaje de variabilidad explicado es el criterio de menor importancia, mientras que para un análisis de componentes principales sería el de mayor peso.

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3,432	34,319	34,319	3,432	34,319	34,319	2,639	26,391	26,391
2	2,250	22,500	56,819	2,250	22,500	56,819	2,077	20,772	47,162
3	1,195	11,947	68,766	1,195	11,947	68,766	1,669	16,693	63,856
4	1,084	10,843	79,609	1,084	10,843	79,609	1,575	15,754	79,609
5	,695	6,947	86,557						
6	,594	5,940	92,497						
7	,338	3,376	95,873						
8	,238	2,381	98,254						
9	,104	1,042	99,296						
10	,070	,704	100,000						

Método de extracción: Análisis de Componentes principales.

Basándonos en el gráfico de sedimentación, también nos decantaríamos por retener 4 factores puesto que hasta el cuarto tramo del gráfico de sedimentación observamos que las pendientes son significativas (pronunciadas).



(b) Obtén la matriz de cargas factoriales e indica el Modelo de Análisis Factorial propuesto (indica al menos la primera y la última fila del modelo) ¿Qué métodos has utilizado para obtener las cargas factoriales? Comenta las ventajas e inconvenientes de este método frente a otros métodos que también permiten calcular las cargas factoriales.

Hemos obtenido la matriz de cargas factoriales por el método de componentes principales. Escoger este método nos asegura siempre la obtención de una solución, ya que no es un método con algoritmo iterativo.

Matriz de componentes[a]

	Componente			
	1	2	3	4
TMR	,426	,739	-,243	-,338
SMEAN	,851	-,161	-,293	-,146
SMAX	,766	-,336	-,081	-,268
PMEAN	,653	-,546	,110	-,118
PMAX	,475	-,583	,290	-,272
PM2	,442	,108	-,612	,278
PERWH	,533	,433	,611	,069
NONPOOR	,618	,230	,393	,527
GE65	,401	,819	,038	-,277
LPOP	,516	-,069	-,208	,594

Método de extracción: Análisis de componentes principales.

a. 4 componentes extraídos

El modelo de análisis factorial propuesto es:

$$TMR_{tip} = 0{,}426 * f_1 + 0{,}739 * f_2 - 0{,}243 * f_3 - 0{,}338 * f_4 + \varepsilon_1$$
$$SMEAN_{tip} = 0{,}851 * f_1 - 0{,}161 * f_2 - 0{,}293 * f_3 - 0{,}146 * f_4 + \varepsilon_2$$
$$SMAX_{tip} = 0{,}766 * f_1 - 0{,}336 * f_2 - 0{,}081 * f_3 - 0{,}268 * f_4 + \varepsilon_3$$
$$PMEAN_{tip} = 0{,}653 * f_1 - 0{,}546 * f_2 + 0{,}110 * f_3 - 0{,}118 * f_4 + \varepsilon_4$$
$$PMAX_{tip} = 0{,}475 * f_1 - 0{,}583 * f_2 + 0{,}290 * f_3 - 0{,}272 * f_4 + \varepsilon_5$$
$$PM2_{tip} = 0{,}442 * f_1 + 0{,}108 * f_2 - 0{,}612 * f_3 + 0{,}278 * f_4 + \varepsilon_6$$
$$PERWH_{tip} = 0{,}533 * f_1 + 0{,}433 * f_2 + 0{,}611 * f_3 + 0{,}069 * f_4 + \varepsilon_7$$
$$NONPOOR_{tip} = 0{,}618 * f_1 + 0{,}230 * f_2 + 0{,}393 * f_3 + 0{,}527 * f_4 + \varepsilon_8$$
$$GE65_{tip} = 0{,}401 * f_1 + 0{,}819 * f_2 - 0{,}038 * f_3 - 0{,}277 * f_4 + \varepsilon_9$$
$$LPOP_{tip} = 0{,}516 * f_1 - 0{,}069 * f_2 - 0{,}208 * f_3 + 0{,}594 * f_4 + \varepsilon_{10}$$

(c) ¿Cómo podemos facilitar la interpretación de los factores retenidos? Aplícalo e interpreta los factores resultantes.

Podemos facilitar la interpretación de los factores retenidos con la rotación del modelo factorial, nosotros, en primer lugar, hemos pedido que haga la rotación varimax, resultándonos la siguiente tabla:

Matriz de componentes rotados[a]

	Componente			
	1	2	3	4
TMR	-,003	,933	,057	,164
SMEAN	,729	,305	,044	,480
SMAX	,837	,163	,047	,223
PMEAN	,831	-,149	,146	,128
PMAX	,804	-,197	,110	-,163
PM2	,088	,215	-,073	,774
PERWH	,151	,377	,816	-,139
NONPOOR	,122	,064	,866	,313
GE65	-,069	,904	,296	-,001
LPOP	,145	-,126	,347	,713

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 9 iteraciones.

Según esta rotación, podemos asociar:

Factor1 → Media aritmética de cantidades de sulfato, cantidades máximas de sulfato cada dos semanas, media aritmética de partículas suspendidas y cantidad máxima de partículas suspendidas.

Factor2 → Tasa de mortalidad y porcentaje de población con 65 años o más.

Factor 3 → Porcentaje de blancos en la población y porcentaje de familias con ingresos sobre el nivel de la pobreza.

Factor 4 → Densidad de población por milla cuadrada y Logaritmo (base 10) de la población.

Podríamos interpretar que el factor 1 nos habla de la cantidad de componentes contaminantes contenidos en aire, que el factor 2 nos habla de muertes o riesgo de muertes en la población, al ser el estudio en los años

60 en Norteamérica podemos relacionar riqueza con blancos en la población, por lo que para nosotros el factor 3 es la riqueza de la población, y por último el factor 4 nos habla de la densidad de población.

Si hacemos esta vez, una rotación Quartimax, vemos que nos acaba saliendo la misma relación de los factores con las variables como vemos en la siguiente tabla.

Matriz de componentes rotados[a]

	Componente			
	1	2	3	4
TMR	,008	,935	,042	,159
SMEAN	,750	,306	,019	,447
SMAX	,847	,162	,023	,187
PMEAN	,839	-,149	,127	,095
PMAX	,798	-,199	,094	-,195
PM2	,119	,218	-,082	,768
PERWH	,166	,388	,807	-,144
NONPOOR	,157	,078	,860	,311
GE65	-,059	,909	,284	-,003
LPOP	,183	-,117	,342	,709

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Quartimax con Kaiser.

a. La rotación ha convergido en 7 iteraciones.

(d) Comenta la bondad del ajuste. Para ello, compara las correlaciones observadas con las reproducidas y determina si te parece adecuado el modelo factorial obtenido ¿Qué representan las comunalidades? ¿Te parecen adecuadas en nuestro modelo?

A esta cuestión daremos respuesta con la siguiente tabla:

Correlaciones reproducidas

		TMR	SMEAN	SMAX	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
Correlación reproducida	TMR	,901[b]	,364	,189	-,112	-,207	,323	,375	,160	,860	,019
	SMEAN	,364	,857[b]	,769	,628	,453	,498	,194	,297	,238	,425
	SMAX	,189	,769	,779[b]	,707	,609	,277	,195	,223	,103	,276
	PMEAN	-,112	,628	,707	,751[b]	,692	,130	,170	,258	-,149	,282
	PMAX	-,207	,453	,609	,692	,723[b]	-,106	,159	,130	-,200	,064
	PM2	,323	,498	,277	,130	-,106	,659[b]	-,073	,204	,166	,513
	PERWH	,375	,194	,195	,170	,159	-,073	,850[b]	,705	,572	,159
	NONPOOR	,160	,297	,223	,258	,130	,204	,705	,867[b]	,305	,534
	GE65	,860	,238	,103	-,149	-,200	,166	,572	,305	,910[b]	-,022
	LPOP	,019	,425	,276	,282	,064	,513	,159	,534	-,022	,667[b]
Residual[a]	TMR		-,045	-,068	,044	,071	-,052	-,076	,022	,000	,069
	SMEAN	-,045		,063	-,075	-,114	-,077	,014	,035	-,047	-,048
	SMAX	-,068	,063		-,146	-,136	-,082	,019	,027	-,038	-,020
	PMEAN	,044	-,075	-,146		-,036	,034	,009	-,055	,035	,022
	PMAX	,071	-,114	-,136	-,036		,096	-,060	,004	,053	,055
	PM2	-,052	-,077	-,082	,034	,096		,130	,017	-,051	-,248
	PERWH	-,076	,014	,019	,009	-,060	,130		-,068	-,044	-,095
	NONPOOR	,022	,035	,027	-,055	,004	,017	-,068		-,049	-,117
	GE65	,000	-,047	-,038	,035	,053	-,051	-,044	-,049		,118
	LPOP	,069	-,048	-,020	,022	,055	-,248	-,095	-,117	,118	

Método de extracción: Análisis de Componentes principales.

a. Los residuos se calculan entre las correlaciones observadas y reproducidas. Hay 24 (53,0%) residuales no redundantes con valores absolutos mayores que 0,05.

b. Comunalidades reproducidas

Vemos algunos valores residuales que son bajos, y podrían ser muy válidos (rojo), sin embargo, vemos otros bastante elevados, (azul) que hacen que nuestro modelo no sea tan bueno, de hecho, abajo vemos que un 53 % de los datos es mayor de 0.05.

En la diagonal de la parte superior de la tabla, vemos las comunalidades, las comunalidades, representan la parte común explicada por los factores, tomaremos como buenas comunalidades mayores de 0.8, en este modelo, la mitad, están por encima de este 0.8, (violeta), pero la otra mitad de las comunalidades se encuentran por debajo de este 0.8

Nuestra conclusión es que este modelo no es del todo bueno, si recordamos al principio, había 4 variables cuya medida de adecuación muestral estaba por debajo de 0.5. Vamos a volver a hacer este estudio eliminando una de las variables con esta medida más baja, que es el porcentaje de blancos en la población.

Análisis desde el principio eliminando PERWH.

Eliminando ahora este dato, y volviendo a realizar el análisis desde, vemos ahora que el análisis factorial resulta más adecuado para los datos muestrales:

Matriz de correlaciones[a]

		TMR	SMEAN	SMAX	PMEAN	PMAX	PM2
Correlación	TMR	1,000	,319	,120	-,068	-,135	,271
	SMEAN	,319	1,000	,832	,554	,339	,421
	SMAX	,120	,832	1,000	,560	,474	,196
	PMEAN	-,068	,554	,560	1,000	,657	,163
	PMAX	-,135	,339	,474	,657	1,000	-,010
	PM2	,271	,421	,196	,163	-,010	1,000
	NONPOOR	,182	,332	,250	,204	,134	,221
	GE65	,860	,192	,065	-,114	-,147	,115
	LPOP	,087	,377	,256	,304	,118	,265
Sig. (Unilateral)	TMR		,002	,144	,273	,116	,007
	SMEAN	,002		,000	,000	,001	,000
	SMAX	,144	,000		,000	,000	,041
	PMEAN	,273	,000	,000		,000	,074
	PMAX	,116	,001	,000	,000		,465
	PM2	,007	,000	,041	,074	,465	
	NONPOOR	,053	,001	,013	,035	,119	,024
	GE65	,000	,044	,282	,158	,096	,155
	LPOP	,220	,000	,011	,003	,148	,009

Vemos que la correlación de los datos es buena por norma general (rojo), pero que también encontramos valores malos en los datos de la correlación

		NONPOOR	GE65	LPOP
Correlación	TMR	,182	,860	,087
	SMEAN	,332	,192	,377
	SMAX	,250	,065	,256
	PMEAN	,204	-,114	,304
	PMAX	,134	-,147	,118
	PM2	,221	,115	,265
	NONPOOR	1,000	,256	,417
	GE65	,256	1,000	,095
	LPOP	,417	,095	1,000
Sig. (Unilateral)	TMR	,053	,000	,220
	SMEAN	,001	,044	,000
	SMAX	,013	,282	,011
	PMEAN	,035	,158	,003
	PMAX	,119	,096	,148
	PM2	,024	,155	,009
	NONPOOR		,011	,000
	GE65	,011		,200
	LPOP	,000	,200	

a. Determinante = ,008

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,631
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	360,626
	gl	36
	Sig.	,000

La prueba de esfericidad de Bartlett muestra la existencia de correlaciones entre las variables del problema y el KMO ha aumentado con respecto al caso inicial,

0,631 > 0,5

Por último la matriz anti-imagen, nos muestra una adecuación muestral de cada variable bastante mejor que nos salía anteriormente, en este caso, todas por encima de 0,5.

Matrices anti-imagen

		TMR	SMEAN	SMAX	PMEAN	PMAX
Covarianza anti-imagen	TMR	,199	-,060	,039	,023	-,004
	SMEAN	-,060	,181	-,162	-,073	,050
	SMAX	,039	-,162	,235	,006	-,089
	PMEAN	,023	-,073	,006	,418	-,254
	PMAX	-,004	,050	-,089	-,254	,507
	PM2	-,079	-,111	,083	-,007	,043
	NONPOOR	,064	-,035	,007	,010	-,044
	GE65	-,179	,032	-,024	,006	,015
	LPOP	,036	-,061	,041	-,082	,046
Correlación anti-imagen	TMR	,509[a]	-,314	,182	,079	-,012
	SMEAN	-,314	,633[a]	-,786	-,265	,164
	SMAX	,182	-,786	,644[a]	,018	-,257
	PMEAN	,079	-,265	,018	,753[a]	-,552
	PMAX	-,012	,164	-,257	-,552	,670[a]
	PM2	-,212	-,312	,206	-,014	,072
	NONPOOR	,169	-,098	,017	,017	-,073
	GE65	-,858	,159	-,106	,019	,045
	LPOP	,094	-,169	,099	-,150	,075

		PM2	NONPOOR	GE65	LPOP
Covarianza anti-imagen	TMR	-,079	,064	-,179	,036
	SMEAN	-,111	-,035	,032	-,061
	SMAX	,083	,007	-,024	,041
	PMEAN	-,007	,010	,006	-,082
	PMAX	,043	-,044	,015	,046
	PM2	,697	-,072	,080	-,064
	NONPOOR	-,072	,725	-,103	-,220
	GE65	,080	-,103	,219	-,026
	LPOP	-,064	-,220	-,026	,722
Correlación anti-imagen	TMR	-,212	,169	-,858	,094
	SMEAN	-,312	-,098	,159	-,169
	SMAX	,206	,017	-,106	,099
	PMEAN	-,014	,017	,019	-,150
	PMAX	,072	-,073	,045	,075
	PM2	,642[a]	-,101	,205	-,090
	NONPOOR	-,101	,723[a]	-,257	-,304
	GE65	,205	-,257	,504[a]	-,064
	LPOP	-,090	-,304	-,064	,762[a]

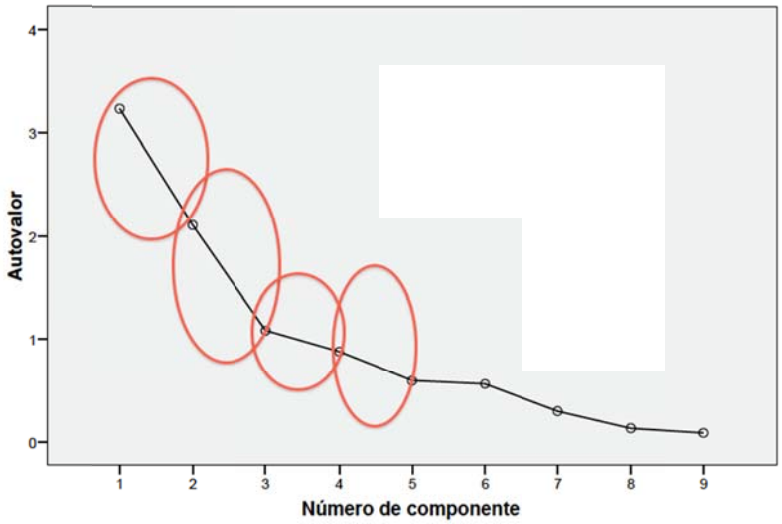
a. Medida de adecuación muestral

Atendiendo a los criterios de Kaiser, variabilidad explicada y gráfico de sedimentación, deberíamos valorar dos opciones: retener 3 factores o retener 4 factores.

Componente	Autovalores iniciales		
	Total	% de la varianza	% acumulado
1	3,233	35,926	35,926
2	2,105	23,391	59,316
3	1,086	12,069	71,385
4	,882	9,803	81,188
5	,598	6,640	87,828
6	,567	6,298	94,126
7	,301	3,350	97,476
8	,136	1,515	98,991
9	,091	1,009	100,000

Método de extracción: Análisis de Componentes principales.

Gráfico de sedimentación



Veamos primero la bondad del ajuste del modelo con tres factores, para ver si podemos seguir a delante, o debemos añadir algún factor más.

Correlaciones reproducidas

		TMR	SMEAN	SMAX	PMEAN	PMAX
Correlación reproducida	TMR	,923[b]	,350	,182	-,108	-,179
	SMEAN	,350	,804[b]	,747	,642	,501
	SMAX	,182	,747	,768[b]	,711	,629
	PMEAN	-,108	,642	,711	,745[b]	,680
	PMAX	-,179	,501	,629	,680	,685[b]
	PM2	,259	,378	,227	,160	-,010
	NONPOOR	,221	,428	,270	,217	,024
	GE65	,898	,283	,116	-,165	-,227
	LPOP	,044	,427	,280	,285	,073
Residual[a]	TMR		-,031	-,061	,039	,044
	SMEAN	-,031		,085	-,089	-,163
	SMAX	-,061	,085		-,151	-,156
	PMEAN	,039	-,089	-,151		-,023
	PMAX	,044	-,163	-,156	-,023	
	PM2	,013	,043	-,032	,003	,000
	NONPOOR	-,039	-,096	-,020	-,013	,110
	GE65	-,038	-,091	-,051	,052	,080
	LPOP	,043	-,050	-,024	,019	,045

Método de extracción: Análisis de Componentes principales.

		PM2	NONPOOR	GE65	LPOP
Correlación reproducida	TMR	,259	,221	,898	,044
	SMEAN	,378	,428	,283	,427
	SMAX	,227	,270	,116	,280
	PMEAN	,160	,217	-,165	,285
	PMAX	-,010	,024	-,227	,073
	PM2	,420[b]	,468	,233	,494
	NONPOOR	,468	,528[b]	,191	,575
	GE65	,233	,191	,879[b]	,015
	LPOP	,494	,575	,015	,672[b]
Residual[a]	TMR	,013	-,039	-,038	,043
	SMEAN	,043	-,096	-,091	-,050
	SMAX	-,032	-,020	-,051	-,024
	PMEAN	,003	-,013	,052	,019
	PMAX	,000	,110	,080	,045
	PM2		-,247	-,118	-,229
	NONPOOR	-,247		,064	-,158
	GE65	-,118	,064		,080
	LPOP	-,229	-,158	,080	

Método de extracción: Análisis de Componentes principales.

a. Los residuos se calculan entre las correlaciones observadas y reproducidas. Hay 19 (52,0%) residuales no redundantes con valores absolutos mayores que 0,05.

b. Comunalidades reproducidas

Los resultados que obtenemos para este modelo con tres factores, es que un 52 % de los datos son mayores de 0,05. Algunos con bastante residuo. Si miramos ahora las comunalidades, solo 3 de ellas están por encima de 0.8, que es el valor que marcamos para considerarlas buenas, el resto están por debajo, alguna de los valores, bastante por debajo. En principio no consideramos buena la bondad del ajuste. Vamos a proponer un modelo con 4 factores, a ver que resultados obtenemos.

Correlaciones reproducidas

		TMR	SMEAN	SMAX	PMEAN	PMAX
Correlación reproducida	TMR	,923[b]	,349	,181	-,108	-,179
	SMEAN	,349	,857[b]	,772	,628	,447
	SMAX	,181	,772	,780[b]	,704	,604
	PMEAN	-,108	,628	,704	,749[b]	,695
	PMAX	-,179	,447	,604	,695	,740[b]
	PM2	,258	,531	,300	,119	-,165
	NONPOOR	,221	,320	,219	,246	,134
	GE65	,898	,230	,091	-,151	-,174
	LPOP	,045	,376	,256	,299	,125
Residual[a]	TMR		-,030	-,061	,039	,044
	SMEAN	-,030		,060	-,074	-,109
	SMAX	-,061	,060		-,144	-,130
	PMEAN	,039	-,074	-,144		-,038
	PMAX	,044	-,109	-,130	-,038	
	PM2	,013	-,110	-,104	,045	,155
	NONPOOR	-,039	,012	,031	-,042	-1,56E-005
	GE65	-,038	-,038	-,026	,037	,027
	LPOP	,043	,001	,000	,005	-,006

Método de extracción: Análisis de Componentes principales.

		PM2	NONPOOR	GE65	LPOP
Correlación reproducida	TMR	,258	,221	,898	,045
	SMEAN	,531	,320	,230	,376
	SMAX	,300	,219	,091	,256
	PMEAN	,119	,246	-,151	,299
	PMAX	-,165	,134	-,174	,125
	PM2	,857[b]	,158	,081	,348
	NONPOOR	,158	,748[b]	,299	,678
	GE65	,081	,299	,932[b]	,066
	LPOP	,348	,678	,066	,720[b]
Residual[a]	TMR	,013	-,039	-,038	,043
	SMEAN	-,110	,012	-,038	,001
	SMAX	-,104	,031	-,026	,000
	PMEAN	,045	-,042	,037	,005
	PMAX	,155	-1,56E-005	,027	-,006
	PM2		,063	,034	-,083
	NONPOOR	,063		-,043	-,261
	GE65	,034	-,043		,030
	LPOP	-,083	-,261	,030	

Método de extracción: Análisis de Componentes principales.

a. Los residuos se calculan entre las correlaciones observadas y reproducidas. Hay 12 (33,0%) residuales no

En este caso vemos que la bondad del ajuste es mejor, solo hay un 33 % de residuos mayores que 0,05. Por otro lado 5 de los valores de las comunalidades son menores que 0.8, pero por otro lado son todas cercanas a este valor, mayores de 0.7 todas.

Este modelo que hemos comentado es el de 4 factores, el modelo será el siguiente:

Matriz de componentes[a]

	Componente			
	1	2	3	4
TMR	,340	,843	-,310	,001
SMEAN	,891	,019	-,101	-,231
SMAX	,814	-,206	-,253	-,109
PMEAN	,715	-,465	-,132	,062
PMAX	,538	-,545	-,314	,234
PM2	,461	,259	,374	-,662
NONPOOR	,527	,216	,452	,468
GE65	,267	,850	-,294	,230
LPOP	,547	,056	,608	,221

Método de extracción: Análisis de componentes principales.

a. 4 componentes extraídos

El modelo de análisis factorial propuesto es:

$$TMR_{tip} = 0{,}340 * f_1 + 0{,}843 * f_2 - 0{,}310 * f_3 + 0{,}001 * f_4 + \varepsilon_1$$
$$SMEAN_{tip} = 0{,}891 * f_1 + 0{,}019 * f_2 - 0{,}101 * f_3 - 0{,}231 * f_4 + \varepsilon_2$$
$$SMAX_{tip} = 0{,}814 * f_1 - 0{,}206 * f_2 - 0{,}253 * f_3 - 0{,}109 * f_4 + \varepsilon_3$$
$$PMEAN_{tip} = 0{,}715 * f_1 - 0{,}465 * f_2 - 0{,}132 * f_3 + 0{,}062 * f_4 + \varepsilon_4$$
$$PMAX_{tip} = 0{,}538 * f_1 - 0{,}545 * f_2 - 0{,}314 * f_3 + 0{,}234 * f_4 + \varepsilon_5$$
$$PM2_{tip} = 0{,}461 * f_1 + 0{,}259 * f_2 + 0{,}374 * f_3 - 0{,}662 * f_4 + \varepsilon_6$$
$$NONPOOR_{tip} = 0{,}527 * f_1 + 0{,}216 * f_2 + 0{,}452 * f_3 + 0{,}468 * f_4 + \varepsilon_7$$
$$GE65_{tip} = 0{,}267 * f_1 + 0{,}850 * f_2 - 0{,}294 * f_3 + 0{,}230 * f_4 + \varepsilon_8$$
$$LPOP_{tip} = 0{,}547 * f_1 + 0{,}056 * f_2 - 0{,}608 * f_3 + 0{,}221 * f_4 + \varepsilon_9$$

Los factores se relacionan con las variables de la siguiente manera:

Matriz de componentes rotados[a]

	Componente			
	1	2	3	4
TMR	,008	,942	,036	,187
SMEAN	,725	,263	,200	,472
SMAX	,834	,134	,091	,240
PMEAN	,831	-,139	,190	,050
PMAX	,810	-,152	,061	-,238
PM2	,067	,088	,160	,905
NONPOOR	,140	,202	,829	-,002
GE65	-,052	,954	,137	-,023
LPOP	,157	-,031	,800	,235

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 5 iteraciones.

Vamos a proponer la relación de los factores con cada variable:

f_1 = SMEAN, SMAX, PMEAN y PMAX
f_2 = TMR y GE65
f_3 = NONPOOR y LPOP
f_4 = PM2

El factor 1, nos habla de la cantidad de componentes contaminantes contenidos en el aire, el factor 2 nos habla de muertes o riesgo de muertes en la población, el factor 3 nos dice las familias con un buen nivel económico de la población y lo grande que es esta y por último el factor 4, se refiere a la densidad de población.

Estadística Industrial – Ingeniero en organización industrial
Universidad Politécnica de Cartagena

Grupo 8

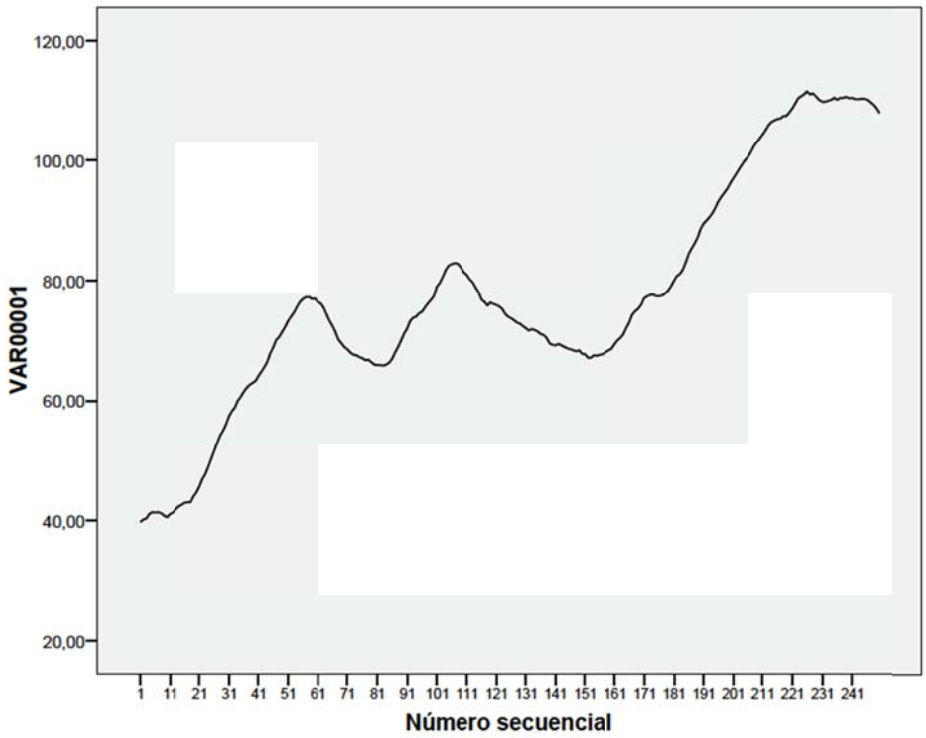
Javier Moreno Cortés
Ana Salmerón Baños
Antonio José Barcelona Vinadel

Problema 3 – Junio 2010

Usando los datos del fichero **PSeriesJun2010.sav** resuelve las siguientes cuestiones.

1. Representa la serie de datos en un gráfico temporal. ¿Crees que la serie proviene de un proceso estocástico estacionario? Justifica tu respuesta.

La serie representada es la siguiente:



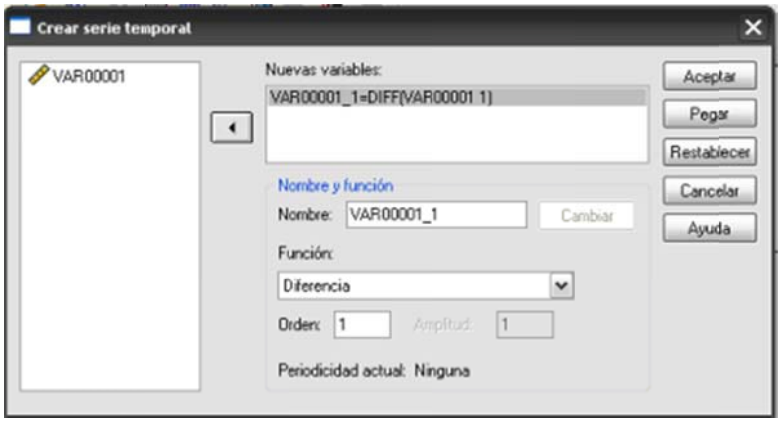
Observamos un crecimiento en la serie de datos, posiblemente lineal o cuadrático, podemos decir que el proceso no es estacionario al menos en media. Vemos que las fluctuaciones son constantes, por lo que diremos que la varianza si es constante.

2. En el caso de que la serie no sea estacionaria, realiza las transformaciones que estimes oportunas para convertirla en estacionaria, comentando por qué las realizas.

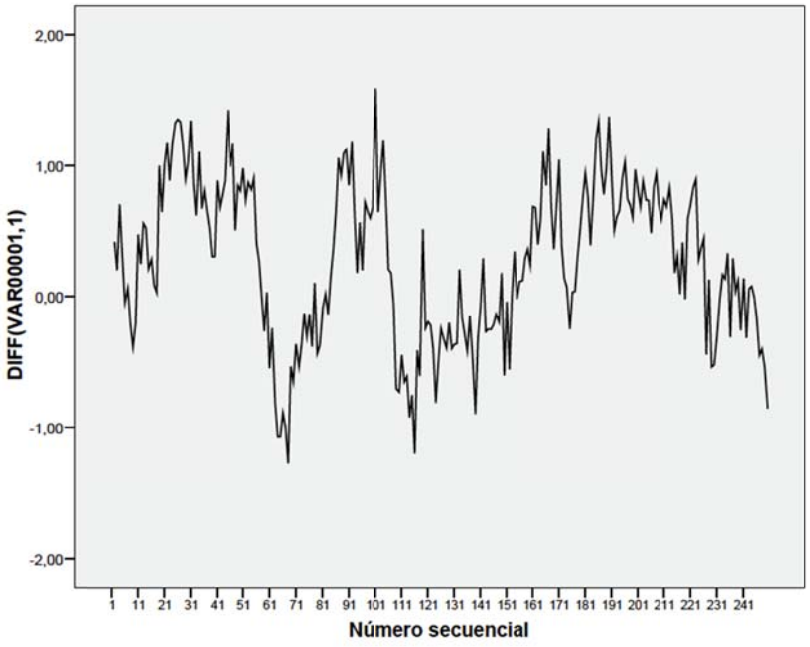
Como en este caso la serie no es estacionaria, no podemos hacer el análisis ARIMA directamente, debemos hacer una transformación que es hacer diferencias entre los valores de la serie, haremos tantas diferencias hasta que la serie quede de forma estacionaria.

En primer lugar haremos diferencias de orden uno, podemos simularlo directamente cuando hacemos el análisis de la serie, pero vamos a hacer las transformaciones.

Con Transformar > Crear serie temporal, y en la siguiente ventana, seleccionamos diferencia de orden 1.

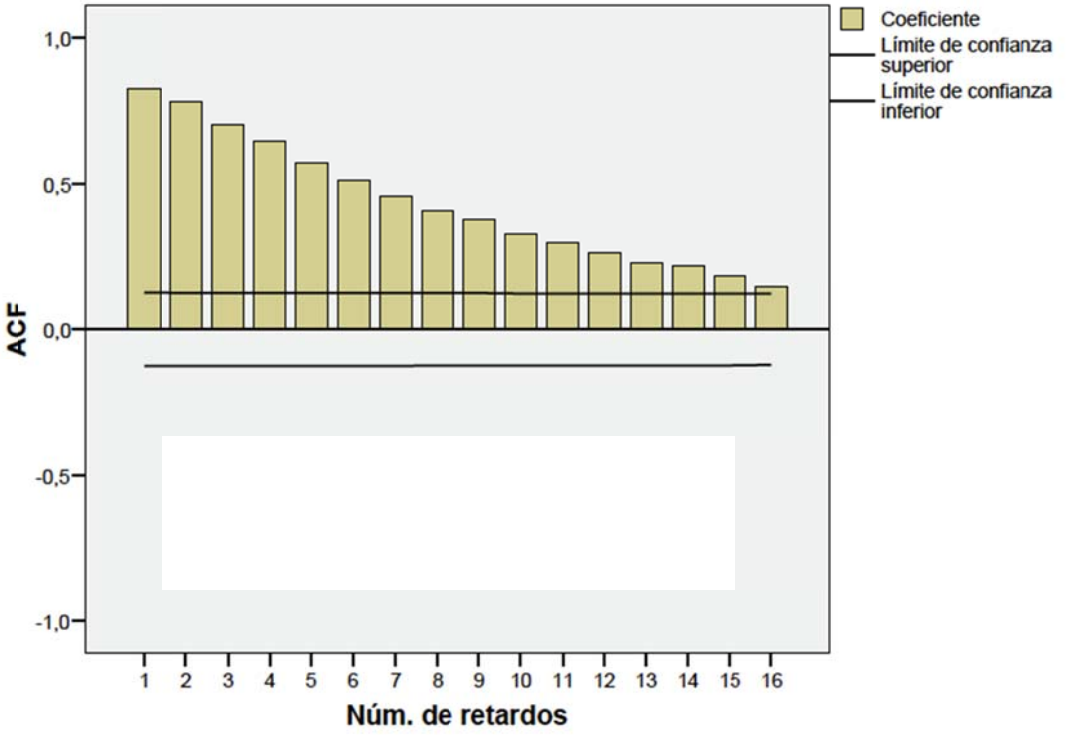


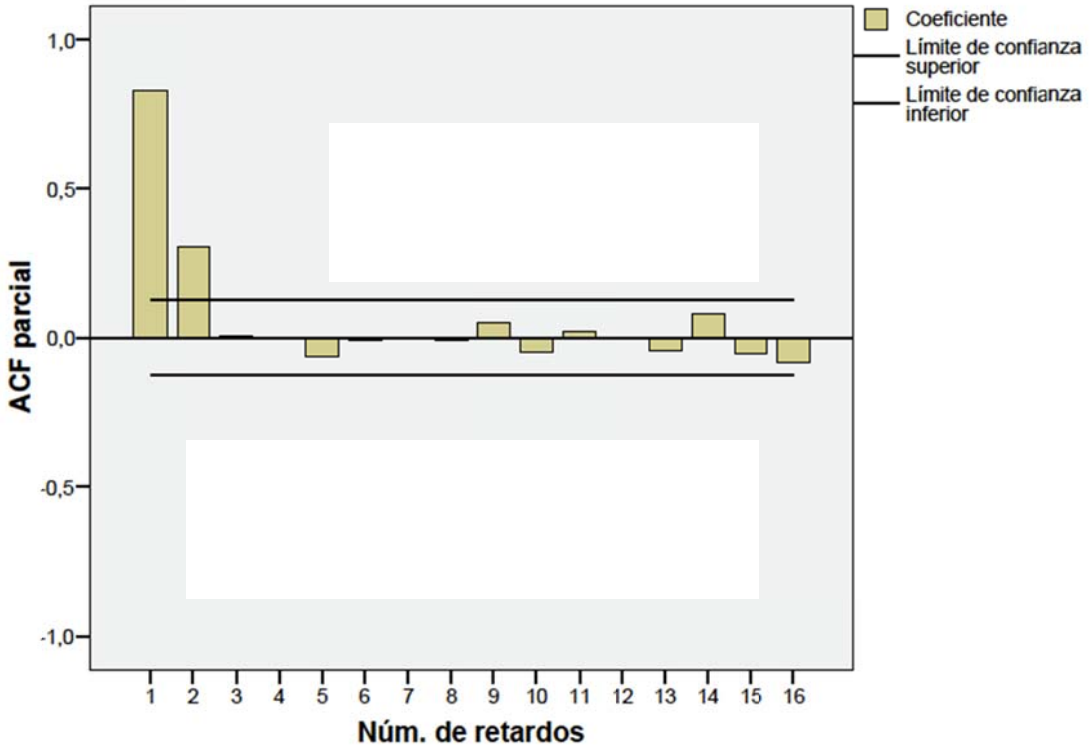
De esta manera tendremos la serie en el programa con diferencias de orden 1, y podremos trabajar ahora sobre esta, no olvidando que hemos hecho esta transformación. Haciendo diferencias de orden 1, es suficiente para obtener una serie estacionaria.



3. Obtén el autocorrelograma simple y parcial de la serie estacionaria. En función de los resultados, ¿Qué modelo(s) ARIMA propondrías como generador(es) de la serie en estudio? Justifica tu respuesta.

Los autocorrelogramas simples y parciales son los siguientes:





Viendo estos autocorrelogramas podemos decir que en ningún momento va a ser un modelo MA puro, ya que el correlograma simple tiene más coeficientes por encima del límite de confianza.

Estando 2 coeficientes en el autocorrelograma parcial por encima del límite de confianza, podríamos proponer un modelo AR (2), también, viendo que en el autocorrelograma simple hay muchos coeficientes por encima podríamos proponer un ARMA (2,X) siendo X tantos coeficientes como vemos por encima, pero la experiencia en el desarrollo de este tipo de ejercicios nos dice que esos coeficientes no serán significativos por los que vamos a empezar estudiando un modelo ARMA (2,2).

Por lo que nuestras propuestas de modelo son:

- AR (2)
- ARMA (2,2)

4. Para cada uno de los modelos teóricos propuestos en el apartado anterior, responde a las siguientes cuestiones:

    a. Determina los coeficientes del modelo, justificando si son o no significativos, y proporciona el modelo irreducible correspondiente.

Si proponemos el modelo AR (2), vemos en la siguiente tabla que los dos coeficientes son significativos, por lo que este modelo no es reducible, y será un posible candidato para la propuesta final

		Estimaciones	Error típico	t	Sig. aprox.
Retardos no estacionales	AR1	,576	,061	9,510	,000
	AR2	,312	,061	5,128	,000
Constante		,236	,173	1,365	,173

Se ha utilizado el algoritmo de Melard para la estimación.

Analizando el modelo ARMA (2,2) vemos que es reducible, que el coeficiente MA2, no es significativo

		Estimaciones	Error típico	t	Sig. aprox.
Retardos no estacionales	AR1	,443	,485	,913	,362
	AR2	,438	,456	,960	,338
	MA1	-,134	,493	-,273	,785
	MA2	,054	,198	,272	,786
Constante		,235	,176	1,337	,182

Se ha utilizado el algoritmo de Melard para la estimación.

El modelos siguiente ARMA (2,1) también es reducible, el coeficiente MA1 tampoco es significativo

		Estimaciones	Error típico	t	Sig. aprox.
Retardos no estacionales	AR1	,591	,196	3,014	,003
	AR2	,299	,168	1,786	,075
	MA1	,017	,206	,081	,935
Constante		,236	,174	1,354	,177

Se ha utilizado el algoritmo de Melard para la estimación.

Por lo que acabamos proponiendo el mismo modelo que proponíamos anteriormente, el AR (2).

b. Determina valores indicativos de la bondad del ajuste.

Los valores de bondad del ajuste son:

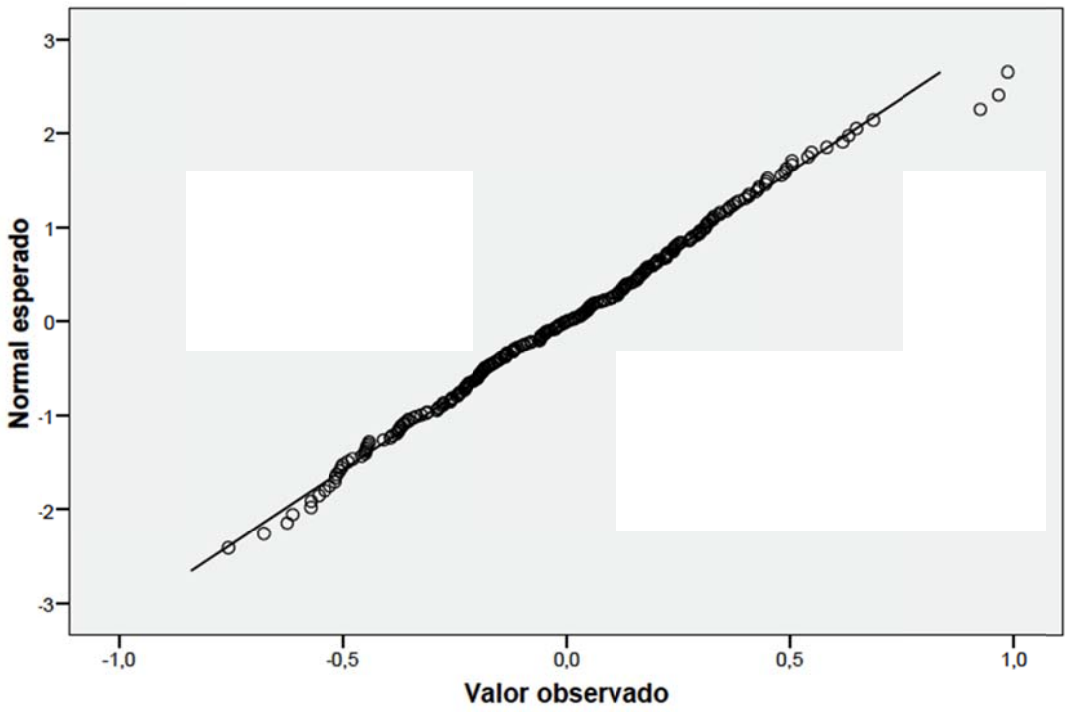
- AIC = 138,737
- BIC = 149,289

Como en principio solo tenemos una única propuesta, estos valores no son peores que otros, decimos que son buenos.

c. ¿El modelo es válido? Comprueba que se verifican las hipótesis sobre los residuos.

Para validar el modelo debemos validar las hipótesis de normalidad, homocedasticidad e independencia.

La hipótesis de Normalidad la validamos con un gráfico de probabilidad normal (Q-Q) y con las pruebas no paramétricas de Kolmogorov-smirnov,



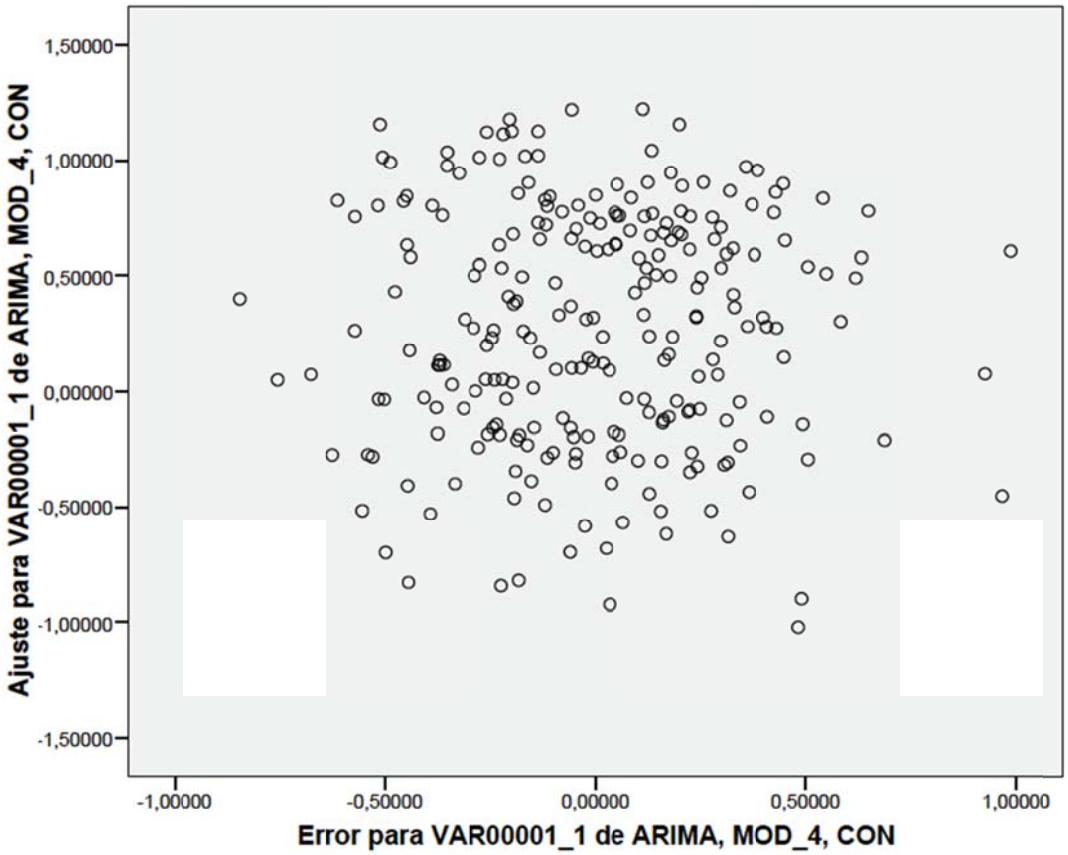
En el gráfico vemos que no muestra desviaciones significativas de la hipótesis de normalidad. Y en las pruebas no paramétricas de Kormogolov-Smirnov de a continuación, vemos que los p-valores de normalidad son altos, por lo que aceptamos la hipótesis de normalidad.

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Error para VAR00001_1 de ARIMA, MOD_4, CON	,035	249	,200*	,994	249	,442

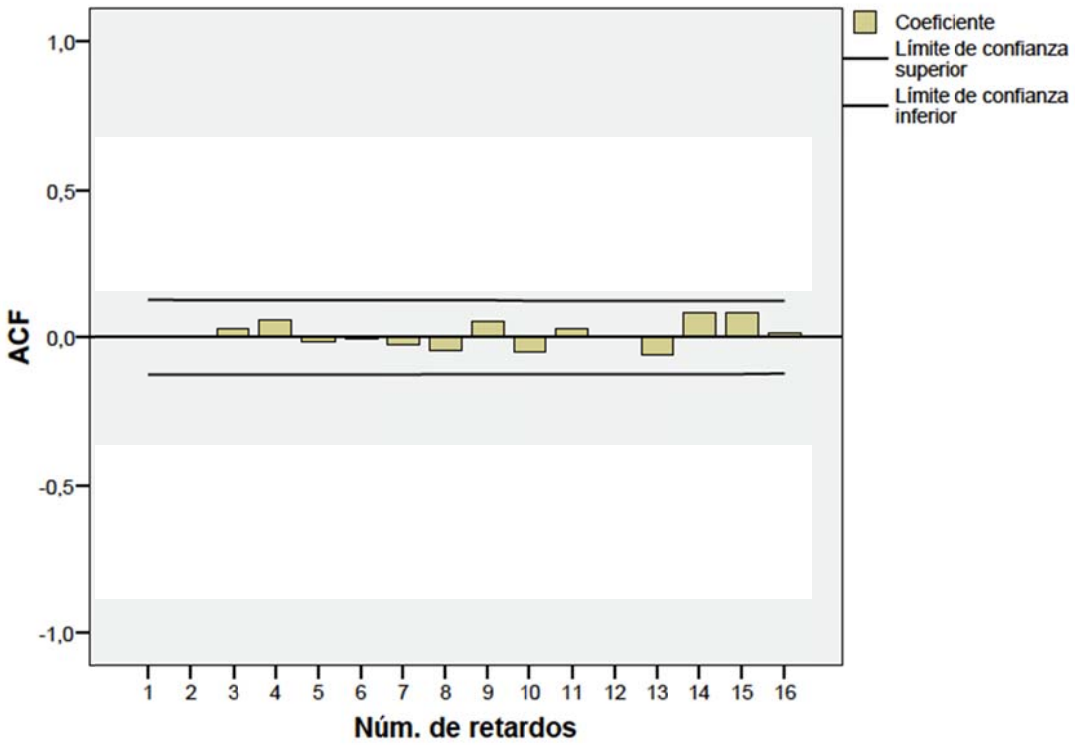
*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

La hipótesis de homocedasticidad podríamos aceptarla simplemente viendo la varianza constante en el gráfico de secuencia, pero para más seguridad, podemos hacer un gráfico de dispersión entre la predicción y su error, vemos que la nube de puntos tiene una dispersión más o menos constante, por lo que si aceptamos esta hipótesis.



La última hipótesis, que es la hipótesis de independencia, la vemos con el gráfico de autocorrelaciones del error de nuestro modelo, vemos que el gráfico no presenta ninguna correlación significativa, podemos suponer la independencia de los residuos.



Por lo que podemos decir que quedan validadas todas las hipótesis de nuestro modelo y que el modelo es válido

5. En función de los resultados del apartado anterior, indica qué modelo te parece más adecuado para representar la serie en estudio y exprésalo en la forma:

$$X_t = cte + a_1X_{t-1} + a_2X_{t-2} + \cdots + a_pX_{t-p} - b_1\varepsilon_{t-1} - b_2\varepsilon_{t-2} - \cdots - b_q\varepsilon_{t-q} + \varepsilon_t$$

donde X_t representa la serie original en estudio

Como ya hemos comentado antes, nos quedamos con el modelo AR (2), además si recordamos, habíamos realizado diferencias de orden 1, por lo que realmente el modelo propuesto será ARIMA (2,1,0)

Los coeficiente los vemos en la tabla de la predicciones del modelo:

		Estimaciones	Error típico	t	Sig. aprox.
Retardos no estacionales	AR1	,576	,061	9,510	,000
	AR2	,312	,061	5,128	,000
Constante		,236	,173	1,365	,173

Se ha utilizado el algoritmo de Melard para la estimación.

Sabemos que el SPSS no nos devuelve el valor real de la constante por la que la calcularemos con la siguiente ecuación.

$$Cte\ verd = Cte\ Spss * (1 - \sum_{i=1}^{p} a_i)$$

Siendo a_i los coeficientes de la parte autorregresiva, por lo que:

$$Cte\ verd = 0{,}236 * \big(1 - (0{,}576 + 0{,}312)\big) = 0{,}026$$

El modelo diferenciado es:

$$Y_t = 0{,}026 + 0{,}576 * Y_{t-1} + 0{,}312 * Y_{t-2} + \varepsilon_t$$

El modelo propuesto será:

$$Y_t = X_t - X_{t-1}$$

$$X_t - X_{t-1} = 0{,}026 + 0{,}576 * (X_{t-1} - X_{t-2}) + 0{,}312 * (X_{t-2} - X_{t-3}) + \varepsilon_t$$

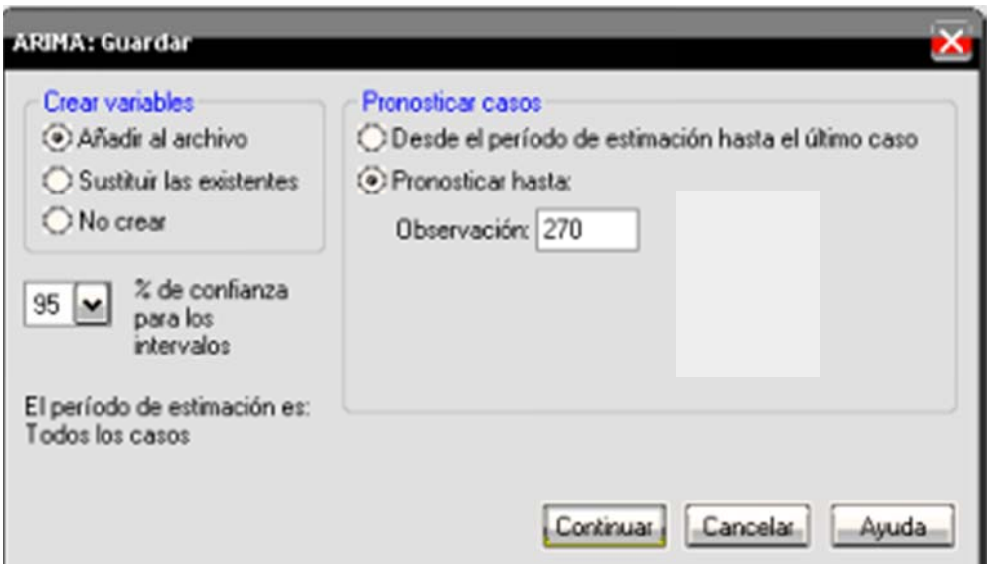
$$X_t = 0{,}026 + 1{,}576 * X_{t-1} - 0{,}576 * X_{t-2} + 0{,}312 * X_{t-2} - 0{,}312\, X_{t-3} + \varepsilon_t$$

$$X_t = 0{,}026 + 1{,}576 * X_{t-1} - 0{,}264 * X_{t-2} - 0{,}312\, X_{t-3} + \varepsilon_t$$

6. Realiza una predicción de la serie para los 20 instantes de tiempo siguientes (indica sólo las predicciones de los 3 primeros instantes). ¿Las 20 predicciones contienen el mismo error? Justifica tu respuesta.

Realizamos la predicción proponiendo el modelo ARIMA como lo habíamos hecho anteriormente pero dándole a la opción guardar hasta, y ponemos el número de datos que tenemos más el número de predicciones que queremos obtener.

Como tenemos 250 valores, y queremos 20 predicciones guardaremos hasta 270. Como vemos en la siguiente imagen.



Las 3 primeras predicciones son:

107,28529
106,67973
106,15928

El error no es constante en todas las predicciones, cada predicción más lejana tendrá más error, ya que nuestro modelo depende de los tres instantes anteriores al ser un modelo AR (3) no estacionario. Y cada vez que la predicción es más lejana depende de predicciones menos fiables. Este aspecto (crecimiento del error de predicción) se refleja de forma explícita cuando observamos el intervalo de confianza para las predicciones, el cuál aumenta su amplitud conforme solicitamos predicciones más lejanas.