Documentos de Trabajo

09/2011

Metodología de estimación de Diplomados en Estadística del Estado en las delegaciones provinciales del INE

Julio César Hernández Sánchez

Cristobal Rojas Montoya

Primera versión: septiembre 2011

Versión actual: septiembre 2011

Metodología de estimación de Diplomados en Estadística del Estado en las delegaciones provinciales del INE

Resumen
El objetivo es estimar los diplomados en estadística del estado necesarios en cada una de las delegaciones provinciales del INE. La metodología sugiere identificar diferentes componentes en la estimación. Se estimarán cuatro modelos para los bloques de operaciones económicas, demográficas, bienales y censo electoral y padrón, que se acumulan para obtener la primera predicción. Además, se creará un modelo global y ambas predicciones se combinarán.

Palabras clave

DEE, censo electoral, padrón, cargas de trabajo, estimación, delegaciones provinciales del INE

Autores y Afiliaciones

Julio César Hernández Sánchez

Delegado provincial del INE en Zamora

Cristobal Rojas Montoya

Delegado provincial del INE en Salamanca

Metodología de estimación de Diplomados en Estadística del Estado en las delegaciones provinciales del INE

Autores:
Julio César Hernández Sánchez
Cristobal Rojas Montoya

Instituto Nacional de Estadística
Madrid, septiembre de 2011

Índice

1. Introducción

La estimación del número de diplomados en estadística del estado en las delegaciones del INE ha sido un tema estudiado durante años, con una filosofía que ha constituído el punto de partida de este trabajo, el cual propondrá un enfoque diferente al que se ha utilizado hasta ahora.

Se ha considerado necesario utilizar datos de cospro[1] de todo un año, y la elección del mismo condiciona completamente los resultados obtenidos, siendo conveniente repetir todo el trabajo que se presenta en este informe una vez transcurridos varios años, aunque nuestra opinión es que los parámetros de los modelos no van a cambiar significativamente.

Se han utilizado datos completos del año 2008, y para las operaciones que no existían en este año, como es el caso de la estadística de bibliotecas, empleo del tiempo y salud se ha considerado conveniente estudiar los datos de cospro de 2009. Hasta ahora sólo se había trabajado con datos de algunos meses para la construcción de los modelos.

Se ha realizado una revisión exhaustiva del libro de cargas[2] en las delegaciones para dicho año 2008. Este trabajo es imprescindible para que la construcción de regresiones no esté sustentada en información incorrecta, que pudiera condicionar los valores de los coeficientes en las mismas, de modo que se puedan obtener unos coeficientes lo más estables y coherentes posibles. Partiendo de ellas se ha generado un resumen de cargas por operaciones y por delegaciones, que será la variable independiente que se utilizará en el proceso econométrico, la cuál tratará de explicar la variabilidad de los datos de cospro.

Del mismo modo se han clasificado los códigos de cada operación de cospro en distintas agrupaciones que serán las que conformen las distintas regresiones, correspondientes a las operaciones que determinarán el número de DEE estimados (Ver Figura 1). Mediante una macro de visual basic se ha leído cospro para cada delegación y se han resumido los datos de cospro por operaciones en un fichero, que en el análisis posterior corresponderá a la variable dependiente.
Hay que tener en cuenta que los datos de cospro recibidos son de A2, y que no sólo los DEE pertenecen a esta categoría, sino que también hay jefes de gestión y analistas que son A2. Este inconveniente se solventa puesto que en la clasificación citada de los códigos de cospro en las operaciones, se consideran códigos que corresponden a tareas estadísticas que son responsabilidad de los DEE, por lo que el cómputo de cospro para estos es correcto.

[1] Cospro es el nombre con el que se conoce a la herramienta de gestión del tiempo que utiliza el INE y sobre la que los empleados graban el tiempo dedicado a cada tarea en su jornada laboral

[2] El libro de cargas es un documento interno del instituto que contiene información de la organización de la recogida de datos en las delegaciones provinciales a nivel de operación estadística

	COSPRO 2008								
			Piloto Censo				ETCL y otras		
TOTAL A2	TAREA	CODIGO	CENSO	CIS	COY.EMP	EB	ECL	ECV	ED
2,75	Índices de Producción Industrial	3005000000			3005000000				
2,22	Índices de Precios Industriales	3005100000			3005100000				
1,79	Índices de Cifras de Negocios	3005200000			3005200000				
1,60	Índices de Entradas de Pedidos	3005300000			3005300000				
0,08	Estadística sobre Actividades de I+D	3005800000							
1,63	Encuesta sobre Innovación en las Empresas	3006100000							
0,08	Índices de Precios de Materiales e Índices Nacionales de la Mano de Obra	3006200000			3006200000				
0,00	Consumos Intermedios e Inversión	3006400000							
0,21	Encuesta de Consumos Energéticos	3007000000							
2,32	Otros Indicadores Coyunturales de la Industria	3007100000			3007100000				
0,05	Indicadores del Sector de TIC	3008100000							
5,45	Índices de Comercio al por Menor	3010300000							
0,55	Taxis	3010900000							
0,01	Encuesta de Coste Laboral	3013200000					3013200000		
22,06	Índice de Precios de Consumo (IPC)	3013800000							
0,08	Hipotecas	3014900000							
9,58	Encuesta de Ocupación en Alojamientos Turísticos	3016200000							
0,53	Estadística de Transporte de Viajeros	3016300000			3016300000				

Figura 1: Resumen de códigos de cospro por operaciones

Paralelamente a estos trabajos se han corregido las cargas de trabajo para ese periodo con las cargas efectivas que hubo realmente para conseguir unos coeficientes lo más estables y coherentes posibles. Partiendo de ellas se ha generado un resumen de cargas por operaciones y por delegaciones, que será la variable independiente que se utilizará en el proceso econométrico.

2. Descripción de operaciones objeto de estudio

Se han realizado regresiones individuales para cada una de las siguientes operaciones:

1. CIS(Comercio Internacional de Servicios), Estadísticas coyunturales, ECL(Encuestas de Coste Laboral), Encuestas estructurales, Hoteles, IASS (Indicadores de Actividad del Sector Servicios), ICPM(Índice de Comercio al Por Menor), IPC(Índice de Precios de Consumo)

2. ECV(Encuesta de Condiciones de Vida), EDAD(Encuesta sobre discapaci-

dades), EET(Encuesta de Empleo del Tiempo), SALUD(Encuestas sobre salud), EPAP(Encuesta de Población Activa con entrevistas personales), EPAT(Encuesta de Población Activa CATI), EPF(Encuesta de Presupuestos Familiares), MH(Encuesta de Morbilidad Hospitalaria), MNP(Movimiento Natural de Población), TICHP(Encuesta de Tecnologías de la Información y las Comunicaciones en los Hogares), EJ(Estadísticas Judiciales) y EEEA(Encuesta de Estructuras Agrarias)

3. Padrón y censo electoral

En estas regresiones se han estudiado los outliers o puntos críticos, así como contrastes para detectar heteroscedasticidad, normalidad de residuos y especificación de los modelos.

A la vista de los resultados obtenidos en estas pruebas, en las cuales se podía detectar falta de normalidad y heteroscedasticidad en algunos casos , se ha intentado buscar alguna solución, que con la filosofía consensuada en grupos de trabajo anteriores, pudiera mejorar esta situación.

3. Transformaciones Box-Cox

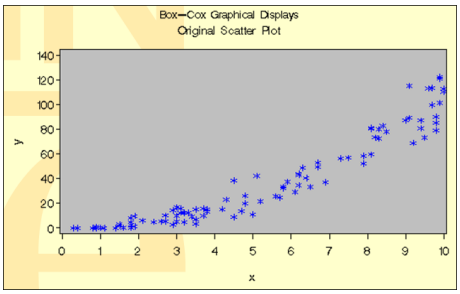
La familia de transformaciones más utilizada para resolver los problemas de falta de normalidad y de heteroscedasticidad es la familia de Box-Cox.
Dicha familia también se utiliza a veces para corregir la falta de linealidad en determinados modelos econométricos. En la Figura 2 se puede apreciar esta situación.

Se desea transformar la variable Y, cuyos valores muestrales se suponen positivos (en caso contrario se suma una cantidad fija M tal que $Y + M > 0$). La transformación Box-Cox depende de un parámetro λ por determinar y viene dada por

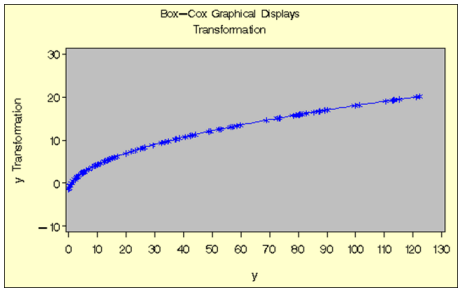
$$Z(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(y) & \text{si } \lambda = 0 \end{cases}$$

Si se quieren transformar los datos para conseguir normalidad, hay varias formas de hacerlo. Siguiendo a Pengfei Li [1] los dos caminos viables serían utilizar métodos bayesianos o bien mediante el método de máxima verosimilitud. Éste último método es bastante utilizado puesto que es sencillo conceptualmente y la función de verosimulitud es fácil de calcular. A su vez se puede determinar un intervalo de confianza para el parámetro λ por las propiedades asintóticas de los estimadores máximo verosímiles.

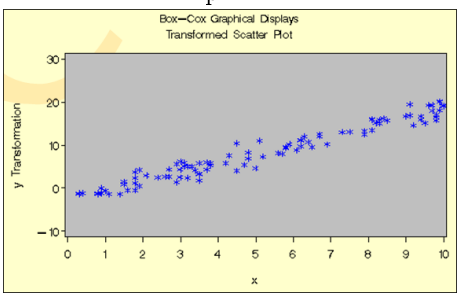
Asumiendo que la variable transformada $y(\lambda) \sim N(X\beta, \sigma^2 I_n)$ tenemos que la función de densidad para la variable transformada $y(\lambda)$ es:



Nube de puntos inicial



Transformación box-cox



Variable transformada vs variable indepte

Figura 2: Resultados de una transformación Box-Cox

$$f(y(\lambda)) = \frac{exp(\frac{-1}{2\sigma^2}(y(\lambda) - X\beta)'(y(\lambda) - X\beta))}{(2\pi\sigma^2)^{\frac{n}{2}}}$$

Si llamamos $J(\lambda, y)$ al Jacobiano de la transformación consistente en pasar de y a $y(\lambda)$, entonces la función de densidad para y es:

$$L(\lambda, \beta, \sigma^2 \mid y, X)) = f(y) = \frac{exp(\frac{-1}{2\sigma^2}(y(\lambda) - X\beta)'(y(\lambda) - X\beta))}{(2\pi\sigma^2)^{\frac{n}{2}}} J(\lambda, y)$$

Para obtener el estimador máximo verosimil de esta ecuación, se puede observar que para cada valor fijo de λ esta expresión es proporcional a la ecuación de verosimilitud para estimar (β, σ^2) para los valores $y(\lambda)$. Por tanto los estimadores de máxima verosimilitud para el par (β, σ^2) son:

$$\tilde{\beta}(\lambda) = (X'X) - Xy(\lambda),$$
$$\hat{\sigma^2}(\lambda) = \frac{y(\lambda)'(I_n - G)y(\lambda)}{n}$$

siendo $G = X(X'X) - X'$.

Sustituyendo ahora $\tilde{\beta}(\lambda)$ y $\hat{\sigma^2}(\lambda)$ en la ecuación de verosimilitud, y teniendo en cuenta la transformación de Box-Cox, $J(\lambda, y) = \prod_{i=1}^{n} y_i^{\lambda-1}$, obtenemos el

perfil de la función de verosimilitud maximizada en (β, σ^2) para un λ dado:

$$l_P(\lambda) = l(\lambda \mid y, X, \tilde{\beta}(\lambda), \hat{\sigma^2}(\lambda)) = C - \frac{n}{2} log(\hat{\sigma^2}(\lambda)) + (\lambda - 1) \sum_{i=1}^{n} log(y_i)$$

Sea g la media geométrica del vector respuesta, es decir, $g = (\prod_{i=1}^{n} y_i)^{\frac{1}{n}}$ y sea $y(\lambda, g) = \frac{y(\lambda)}{g^{\lambda-1}}$. Entonces es sencillo deducir que:

$$l_P(\lambda) = C - \frac{n}{2} log(s_\lambda^2)$$

donde s_λ^2 es la suma de cuadrados residual entre n resultante de ajustar el modelo lineal $y(\lambda) \sim N(X\beta, \sigma^2 I_n)$.

Luego para maximizar el perfil, o mejor dicho, la función de verosimilitud, sólo tenemos que encontrar aquel valor de λ que haga mínimo

$$s_\lambda^2 = \frac{y(\lambda, g)'(I_n - G)y(\lambda, g)}{n}$$

Podemos establecer, utilizando los métodos tradicionales de verosimilitud, un contraste de razón de verosimilitudes para contrastar:
$\begin{cases} H_0 : \lambda = \lambda_0 \\ H_1 : \lambda \neq \lambda_0 \end{cases}$ mediante el estadístico $W = 2[l_P(\hat{\lambda}) - l_P(\lambda_0)]$.

Asintóticamente, W se distribuye como una χ_1^2, aunque hay que tener en cuenta que W es una función que depende de los datos muestrales, a través de λ, y de λ_0.

Para muestras grandes se puede obtener un Intervalo de Confianza de forma sencilla invirtiendo la expresión del contraste de razón de verosimilitudes.
Sea $\hat{\lambda}$ el estimador máximo verosimil de λ. Entonces un intervalo de confianza para λ con un nivel de confianza de $(1 - \alpha)100\,\%$ sería:

$$\{\lambda \mid n * log(\frac{SSE(\lambda)}{SSE(\hat{\lambda})}) \leq \chi_1^2(1 - \alpha)\}$$

donde $SSE(\lambda) = y(\lambda, g)'(I_n - G)y(\lambda, g)$.

La precisión de la aproximación viene dada por la expresión:

$$P(W \leq \chi_1^2(1 - \alpha)) = 1 - \alpha + O(n^{\frac{-1}{2}})$$

En la práctica, se calcula $l_P(\lambda)$ en un enrejado (grid) de valores de λ que permite dibujar aproximadamente dicha función y se obtiene el máximo de la misma:

$$\hat{\lambda}_{MV} = \lambda_0 / l(\lambda_0) \geq l(\lambda), \forall \lambda$$

Mediante el paquete estadístico SAS 9.2 , utilizando el procedimiento proc transreg (ver Figura 3) se han obtenido los valores óptimos de λ para cada

operación estadística, figurando los mismos en la Figura 4.

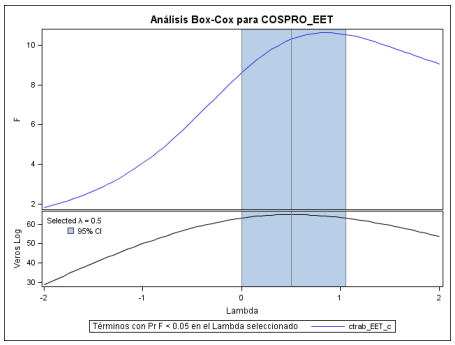
```
ods graphics on;
proc transreg details data=datos_&vbledep ss2  rsquare outtest=par_boxcox_&vbledep plots=all;
   model BoxCox(&vbledep / lambda=-&lambdaini to &lambdafin by 0.05) = identity(&vbleindep);
   output out = boxcox_&vbledep  predicted residuals;
run;
ods graphics off;
```

Figura 3: Procedimiento Proc transreg

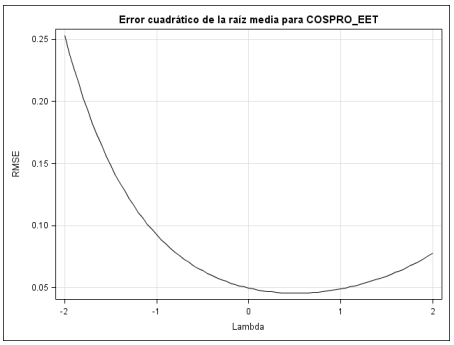
Variable	Lambda
cospro_cis	1,2
cospro_coy	1,35
cospro_eb	0,25
cospro_ecl	1,55
cospro_ecv	0,55
cospro_edad	0,3
cospro_eeea	0,1
cospro_eet	0,5
cospro_ej	-0,3
cospro_epap	0,05
cospro_epat	4,45
cospro_epf	0,15
cospro_est	-0,35
cospro_hotel	0,25
cospro_iass	0,9
cospro_icpm	0,1
cospro_ipc	0,55
cospro_mh	0,4
cospro_mnp	0,15
cospro_pce	0,3
cospro_salud	1,25
cospro_tichp	0,55

Figura 4: Estimadores máximo verosímiles de λ para cada una de las regresiones estudiadas

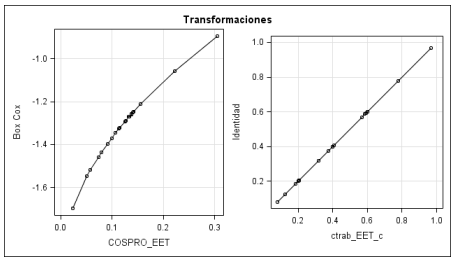
Dicho procedimiento ofrece además varios gráficos que soportan el componente teórico descrito anteriormente, como son por ejemplo, para la operación EET (Encuesta de Empleo del Tiempo) , los mostrados en la Figura 5.



Función de verosimilitud



Error cuadrático



Transf. variable independiente

Figura 5: Salidas del prodecimiento "proc transreg"de SAS

4. Proceso de estimación de coeficientes para cada operación

Como hemos descrito, para mejorar la explicación de cada uno de los modelos realizaremos transformaciones Box-Cox en cada una de las regresiones de estudio. Antes de considerar cada una de ellas, y en base a las enseñanzas de Wooldridge [2] es necesario hacer una serie de consideraciones al respecto.

4.1. Predicción de y cuando $log(y)$ es la variable dependiente

En muchas ocasiones se realiza esta transformación para estudiar un modelo más estable (que como sabemos es un caso particular de transformación box-cox):

$$logy = log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{1}$$

En esta ecuación, cada x_i podrían ser transformaciones de otras variables a su vez.

Calculando los estimadores de los parámetros por el método de mínimos cuadrados ordinarios, sabemos que para predecir $logy$, para cada valor de las

variables independientes se obtiene de la forma:

$$\hat{logy} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \tag{2}$$

Puesto que la función exponencial deshace la transformación logarítmica, la primera idea que uno tiene para predecir y es símplemente calcular la exponencial de $log(y) : \hat{y} = exp(\hat{logy})$.

La cuestión clave es que esto no es correcto, puesto que de hecho este cálculo infraestima el valor esperado de y. De hecho, si el modelo (1) cumple con las hipótesis del modelo lineal general, se puede demostrar que

$$E(y \mid x) = exp(\frac{\sigma^2}{2}) * exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

donde x denota el conjunto de variables independientes y σ^2 es la varianza de u. Si los residuos $u \sim Normal(0, \sigma^2)$, entonces el valor esperado de $exp(u)$ es $exp(\frac{\sigma^2}{2})$. Por tanto, esta ecuación denota que es necesario un ajuste para calcular la predicción de y:

$$\hat{y} = exp(\frac{\hat{\sigma^2}}{2}) * exp(\hat{logy}) \tag{3}$$

donde $\hat{\sigma^2}$ es un estimador insesgado de σ^2. Como $\hat{\sigma}$, que es el error estandar de la regresión, es conocido, es sencillo obtener los valores predichos de y. Además, al ser $\hat{\sigma^2} > 0$, $exp(\frac{\sigma^2}{2}) > 1$. Luego para valores altos de $\hat{\sigma^2}$, el factor de ajuste puede llegar a ser sustancialmente mayor que la unidad.

La predicción en (3) no es insesgada, pero es consistente. No existen predicciones insesgadas de y, y en muchos casos, (3) funciona bien. De todos modos, esta reflexión se basa en la normalidad del término de error, u. Por tanto, es útil disponer de una predicción que no dependa de este supuesto. Sabemos que el método de mínimos cuadrados ordinarios tiene propiedades aceptables, incluso cuando la distribución del término de error no es normal. Si suponemos que u es independiente de las variables explicativas, entonces tenemos que:

$$E(y \mid x) = \alpha_0 * exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \tag{4}$$

donde α_0 es el valor esperado de $exp(u)$, el cuál debe ser mayor que 1.

Dada una estimación $\hat{\alpha_0}$, podemos predecir y como :

$$\hat{y} = \hat{\alpha_0} * exp(\hat{logy}) \tag{5}$$

que lo que nos muestra es que es necesario calcular la exponencial del valor predicho para el modelo logarítmico y multiplicar el resultado por un valor que llamamos $\hat{\alpha_0}$.

Luego resulta que hemos obtenido un estimador consistente de $\hat{\alpha_0}$ fácilmente.

4.2. Proceso práctico de predicción de y cuando $log(y)$ es la variable dependiente:

La primera parte del proceso es calcular la estimación de α_0, para lo cuál seguimos los siguientes pasos:

(i) Obtenemos los valores ajustados $\hat{logy_i}$ de la regresión de $logy$ sobre $x_1, \ldots, x_k$.

(ii) Para cada observación i, calculamos $\hat{m}_i = exp(\hat{logy})$.

(ii) Efectuamos una regresión de y sobre la variable $\hat{m}$ sin término independiente, es decir, hacemos que pase por el origen. El coeficiente de $\hat{m}$ es la estimación de α_0

Una vez que hemos obtenido $\hat{\alpha_0}$, éste se usa, junto con los coeficientes estimados de $logy$ para predecir y de la siguiente forma:

(i) Para cada conjunto dado de valores de $x_1, \ldots, x_k$, obtenemos $\hat{logy}$ de (2)

(ii) Obtenemos la predicción $\hat{y}$ de (5)

4.3. Algunas consideraciones sobre este procedimiento

Podemos usar el anterior algoritmo consistente en obtener predicciones para determinar cómo de bien es capaz de explicar y el modelo con $log(y)$. Ya existen medidas de bondad en la determinación de modelos en los que y es la variable dependiente, como son el R^2 y el R^2 ajustado. Nuestro objetivo es encontrar una medida de bondad de ajuste en el modelo que tiene $log(y)$ como variable dependiente que pueda ser comparado con el R^2 habitual en los modelos cuya variable dependiente es y.

Hay varias formas de encontrar una medida así, aunque describiremos una que es sencilla de implementar:

Tras efectuar la regresión de y sobre $\hat{m}$ sin término independiente (descrita en el paso (iii)), obtenemos los valores ajustados para la regresión, $\hat{y_i} = \hat{\alpha_0} * \hat{m}_i$. Entonces, calculamos la correlación simple entre $\hat{y_i}$ y el valor y_i de la muestra. El cuadrado de este valor se puede comparar directamente con el R^2 que obtuvimos usando y como variable dependiente en una regresión lineal ordinaria. Recordemos también que el valor R^2 en la ecuación:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

es justamente el cuadrado de la correlación entre los valores y_i e $\hat{y_i}$.

4.4. Predicción de y cuando la variable dependiente no se obtiene como transformación logarítmica

Como se ha comentado anteriormente, para intentar solucionar el problema de la heteroscedasticidad se aplican las transformaciones Box-Cox, calculando

el valor de λ óptimo.

En el caso de que este valor sea distinto de cero, que correspondería a la transformación logarítmica, habrá que ajustar un modelo de regresión lineal a esta variable transformada. Si se cumplen las hipótesis básicas, posteriormente habrá que deshacer la transformación efectuada para obtener las estimaciones de la variable y.

Pero es bien conocido (e.g. Granger y Newbold, 1976) que las estimaciones obtenidas mediante este tipo de transformaciones no mantienen sus propiedades óptimas cuando deshacemos la transformación. Esto ocurre porque la estimación de la media de una distribución simétrica en la escala transformada se convierte en la mediana al calcular la inversa de esa transformación. Hay diversos métodos que intentar corregir ese sesgo, como los de Neyman y Scott(1960), Miller (1984), y Taylor(1986).

Consideremos nuestro modelo:

$$\sum_{j=1}^{p} x_{lj}\beta_j + \sum_{i=1}^{q+1} \xi_i = \begin{cases} \frac{(y_l^{\lambda}-1)}{\lambda}; & \lambda \neq 0 \\ logy_l; & \lambda = 0 \end{cases} \tag{6}$$

$$\text{l = 1,2,...,n}$$

con y_l la variable dependiente, $\xi_i \sim Normal(0, \sigma_i^2)$ para todo i, β_j es un parámetro fijo desconocido, y x_{lj} son valores conocidos.

Este modelo se puede reescribir de la siguiente forma:

$$\sum_{j=1}^{p} x_{lj}\beta_j + \sum_{i=1}^{q+1} \sigma_i e_i = \begin{cases} \frac{(y_l^{\lambda}-1)}{\lambda}; & \lambda \neq 0 \\ logy_l; & \lambda = 0 \end{cases} \tag{7}$$

donde $e_i \sim N(0,1)$ para todo i.

Despejando del modelo anterior la variable y_l tenemos:

$$y_l = \begin{cases} (1 + \lambda \sum_{j=1}^{p} x_{lj}\beta_j + \lambda \sum_{i=1}^{q+1} \sigma_i e_i)^{\frac{1}{\lambda}}; & \lambda \neq 0 \\ e^{\sum_{j=1}^{p} x_{lj}\beta_j + \sum_{i=1}^{q+1} \sigma_i e_i}; & \lambda = 0 \end{cases} \tag{8}$$

Dado $x = x_0$ y $\lambda \neq 0$ se verifica:

$$E(y_l|x_0) = \int (1 + \lambda \sum_{j=1}^{p} x_{0lj}\beta_j + \lambda \sum_{i=1}^{q+1} \sigma_i e_i)^{\frac{1}{\lambda}} dF(e_1, e_2, ..., e_{q+1})$$

Para asegurar que todas las observaciones son positivas con una probabilidad suficientemente alta, se debe cumplir para los valores de $\lambda \neq 0$ la siguiente condición debida a Draper y Cox(1969):

$$|c_i| = |\frac{\lambda\sigma_i}{1 + \lambda \sum_{j=1}^{p} x_{0lj}\beta_j}| << 1$$

Desarrollando en serie de Taylor de segundo orden la expresión anterior de la esperanza queda:

$$E(y_l|x_0) = (1+\lambda\sum_{j=1}^{p} x_{0lj}\beta_j)^{\frac{1}{\lambda}} \int (1+\sum_{i=1}^{q+1} c_i e_i)^{\frac{1}{\lambda}} dF(e_1, e_2, ..., e_{q+1})$$

$$\cong (1+\lambda\sum_{j=1}^{p} x_{0lj}\beta_j)^{\frac{1}{\lambda}} \int (1+\frac{\sum_{i=1}^{q+1} c_i e_i}{\lambda}+\frac{(1-\lambda)(\sum_{i=1}^{q+1} c_i e_i)^2}{2\lambda^2})dF(e_1, e_2, ..., e_{q+1})$$

Puesto que $e_i, e_{i'}$ son independientes para todo $i \neq i'$ entonces:

$$E(y_l|x_0) \cong (1+\lambda\sum_{j=1}^{p} x_{0lj}\beta_j)^{\frac{1}{\lambda}} \int ... \int (1+\frac{\sum_{i=1}^{q+1} c_i e_i}{\lambda}+\frac{(1-\lambda)(\sum_{i=1}^{q+1} c_i e_i)^2}{2\lambda^2})dF(e_1)...dF(e_{q+1})$$

$$= (1+\lambda\sum_{j=1}^{p} x_{0lj}\beta_j)^{\frac{1}{\lambda}}(1+(1-\lambda)\sum_{i=1}^{q+1}\frac{c_i^2}{2\lambda^2})$$

$$= (1+\lambda\sum_{j=1}^{p} x_{0lj}\beta_j)^{\frac{1}{\lambda}}(1+(1-\lambda)\sum_{i=1}^{q+1}\frac{\sigma_i^2}{2(1+\lambda\sum_{j=1}^{p} x_{0lj}\beta_j)^2})$$

Por tanto, un estimador de la media condicionada para un valor futuro será:

$$\hat{E}(y_l|x_0) = (1+\hat{\lambda}\sum_{j=1}^{p} x_{0lj}\hat{\beta}_j)^{\frac{1}{\hat{\lambda}}}(1+(1-\hat{\lambda})\sum_{i=1}^{q+1}\frac{\hat{\sigma_i}^2}{2(1+\hat{\lambda}\sum_{j=1}^{p} x_{0lj}\hat{\beta}_j)^2}) \quad (9)$$

Si $\lambda = 0$, y_l se distribuye como una variable log-normal cuya media condicionada se puede estimar por:

$$\hat{E}(y_l|x_0) = exp\{\sum_{j=1}^{p} x_{0lj}\hat{\beta}_j + \frac{1}{2}\sum_{i=1}^{q+1}\hat{\sigma_i}^2\}$$

Con estas expresiones es sencillo calcular el sesgo debido a la transformación Box-Cox para predecir las medias condicionadas. Hay estudios al respecto que constatan que para este tipo de modelos lineales este problema del sesgo que se presenta al transformar la variable dependiente no es significativo. Por este motivo, en los cálculos que se han realizado en este trabajo y para simplificar todo el proceso de estimación se ha considerado la inversa de la transformación Box-Cox sin el factor teórico descrito en la expresión de $\hat{E}(y_l|x_0)$ para un λ cualquiera. Como se puede observar en la expresión (9) el primer factor es la inversa de la transformación Box-Cox, constituyendo el segundo el factor de corrección del sesgo.

5. Implementación del proceso de estimación de coeficientes

En el caso de la estimación de los DEE necesarios por delegación se han analizado las operaciones descritas en el punto tercero, y salvo las referentes a Estadísticas Judiciales , Estadística de Bibliotecas y la Encuesta de estructuras agrarias, en las cuales las regresiones iniciales no eran significativas y se ha decidido estimar la media, se han obtenido un conjunto de coeficientes para las respectivas estimaciones, los cuales aparecen en la Figura 6.

VARIABLE	termindep	ptereg	coef_v1_reg	coef_v2_reg
COSPRO_CIS	-0,749776619	0,008674548		
COSPRO_COY	-0,445400127	0,023766011		
COSPRO_EB	0,052397025	0		
COSPRO_ECL	-0,69718493	0,041961751		
COSPRO_ECV	-1,36913944	0,144439138		
COSPRO_EDAD	-1,828147673	0,084618829		
COSPRO_EET	-1,485403224	0,422940546		
COSPRO_EEEA	0,026471653	0		
COSPRO_EJ	0,031842105	0		
cospro_salud	-0,700403991	0,026128511		
COSPRO_EPAP	-1,134592148	0,097300546		
COSPRO_EPAT	-0,32336756	0,015130405		
COSPRO_EPF	-1,055708707	0,085246463		
COSPRO_EST	-0,844361669	0,02823651		
COSPRO_HOTEL	-1,152230418	0,082956807		
COSPRO_IASS	-0,773913617	0,014606261		
COSPRO_ICPM	-1,783775752	0,117997774		
COSPRO_IPC	-0,94323424	0,063175579		
COSPRO_MH	-1,893651294	0,44816292		
COSPRO_MNP	-1,517326752	0,095175095		
COSPRO_PCE	-0,324612732		0,077253884	-1,213802255
COSPRO_TICHP	-1,489507612	0,178433818		

Figura 6: Tabla final de coeficientes para la estimación de los DEE necesarios por operación, una vez realizada la transformación box-cox

Hay que tener en cuenta que antes de realizar este proceso se calcularon regresiones sin la variable dependiente transformada. Siempre se filtran aquellos casos en los que tanto la variable cospro como la variable cargas de trabajo tenían valor en cada delegación provincial. Esta consideración parte del hecho de que había ocasiones que teniendo carga de trabajo cospro no lo reflejaba y viceversa, probablemente debido a errores en la cumplimentación de la herramienta de

control horario.

Por otra parte hay que especificar que las regresiones se han hecho eliminando los outliers o puntos de alta influencia, y para ello se han utlizado tanto las salidas de SAS de varios estadísticos destinados al efecto (estadístico de Cook, dfbetas, ...) como el conocimiento de las características de cada delegación provincial. Estas dos aclaraciones pueden verse en las Figuras 7 y 8 .

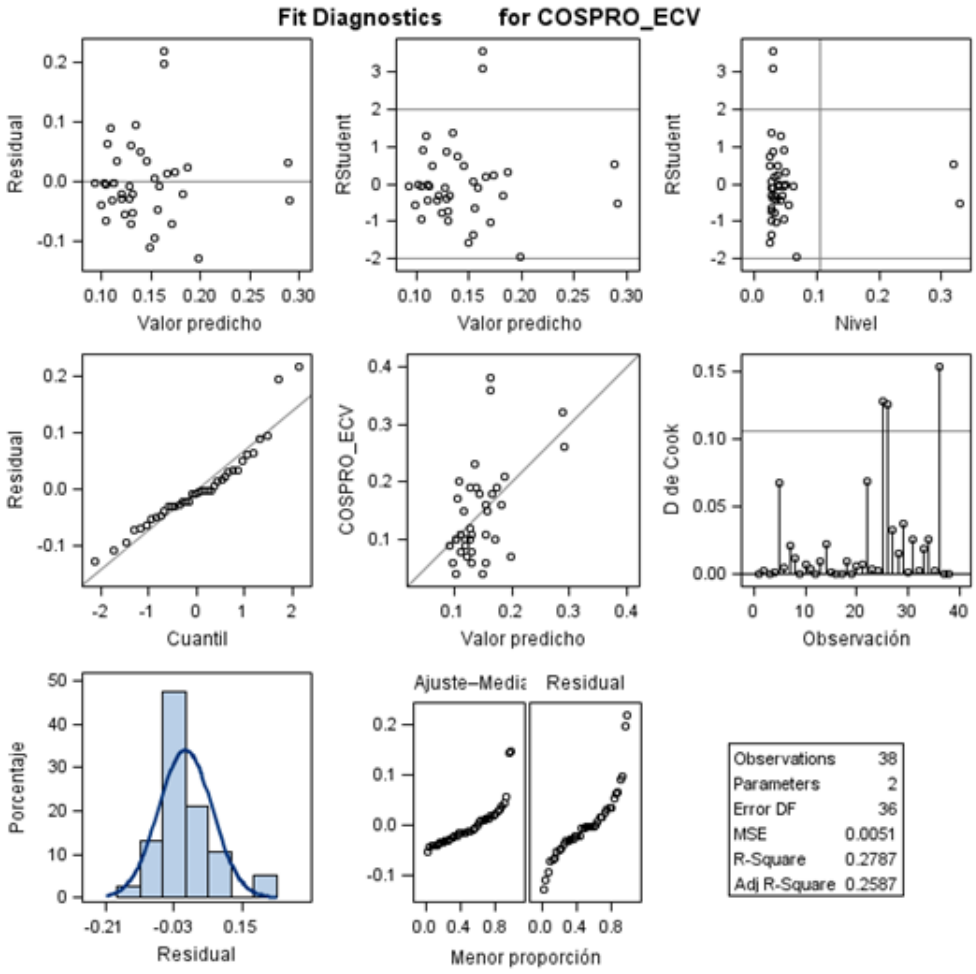


Figura 7: Algunas salidas analizadas en cada regresión

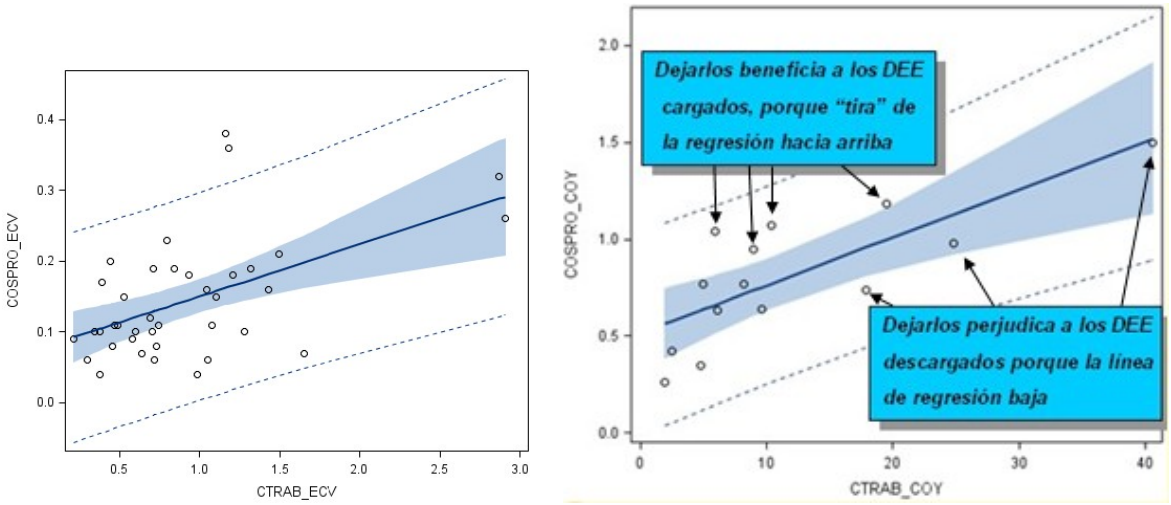


Figura 8: Consideraciones sobre outliers

```
PROC REG data=&oper._bc_sinoutliersmcero outest=&fich_params tableout alpha=0.1;
    MODEL &vbleDep=&vblesIndeps /spec;
    output out=WORK.&fichSAS_predic p=pred_&vbledep r=resid_&vbledep;
run;
quit;

PROC AUTOREG data=&oper._bc_sinoutliersmcero;
    MODEL &vbleDep=&vblesIndeps /reset;
run;
quit;
```

Figura 9: Procedimiento que efectúa cada regresión

Para obtener la tabla descrita se ha utilizado el procedimiento "proc reg"de SAS, mostrado en la Figura 9. En ese procedimiento se ha efectuado un contraste de heteroscedasticidad, mediante la opción "spec", que lo que calcula es el Test de White, siendo éste en todos los casos negativo con p-valores no significativos. También se ha estudiado en cada caso un Test Reset de especificación del modelo, consistente en determinar si éste está bien definido o existen variables relevantes que habría que incluir en él. Este test de conoce como Test de Ramsey[3].

[3]La idea del Test RESET es sencilla. Si el modelo original dado por

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{10}$$

verifica que $E(u \mid x_1, x_2, \ldots, x_k) = 0$, entonces si añadimos funciones no lineales de las variables independientes a (10) serán significativas en el modelo.
Para poder implementar este proceso, debemos decidir cuantas funciones de los valores ajustados incluimos en una regresión ampliada. Normalmente, términos al cuadrado y al cubo se

Las salidas de SAS para este test, en cada operación estadística, al igual que en el caso del estudio de la heteroscedasticidad, han arrojado valores no significativos, por lo que los modelos estarían bien definidos.
Igualmente se han realizado contrastes de normalidad de cada estadística con el prodedimiento "proc univariate"de SAS (Figura 10) entre los cuales se encuentra por ejemplo el de Shapiro-Wilks, resultando no significativos.

```
%macro NormalityResiduals(datosSAS=,nombreVarResid=,oper=);
/*
Below we use proc kde to produce a kernel density plot. kde stands for kernel
density estimate. It can be thought as a histogram with narrow bins and a moving average
*/
proc kde data=&datosSAS  out=&oper._proc_kde;
  var &nombreVarResid;
run;

proc sort data=&oper._proc_kde;
  by &nombreVarResid;
run;

goptions reset=all;
symbol1 c=blue i=join v=none height=1;
proc gplot data=&oper._proc_kde;
  plot density*&nombreVarResid=1;
run;
quit;
/*
Proc univariate will produce a normal quantile graph. qqplot plots the quantiles of a
variable against the quantiles of a normal distribution. qqplot is most sensitive to
non-normality near two tails. and probplot As you see below, the qqplot command shows a
slight deviation from normal at the upper tail, as can be seen in the kde above
*/
goptions reset=all;
proc univariate data=&datosSAS normal;
 var &nombreVarResid;
 qqplot &nombreVarResid / normal(mu=est sigma=est);
run;

/*
Severe outliers consist of those points that are either 3 inter-quartile-ranges below
the first quartile or 3 inter-quartile-ranges above the third quartile. The presence of
any severe outliers should be sufficient evidence to reject normality at a 5% significance
level. Mild outliers are common in samples of any size.
```

Figura 10: Macro de SAS para contrastar normalidad de residuos

Por tanto se han comprobado las hipótesis básicas del modelo lineal general

---

han probado que son útiles en la mayoría de la aplicaciones. Sea pues $\hat{y}$ los valores estimados por el método de mínimos cuadrados ordinarios. Consideremos el modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 \tag{11}$$

No estaremos interesados en los parámetros estimados de (11); sólo utilizaremos esta ecuación para contrastar si (10) tiene términos no lineales que no están especificados en el modelo original. Recordemos que $\hat{y}^2$ y $\hat{y}^3$ son justamente funciones no lineales de los x_j.
La hipótesis nula es que (10) está especificado correctamente. Entonces, RESET es el estadístico F para contrasta la hipótesis nula $H_0 : \delta_1 = 0, \delta_2 = 0$ en el modelo extendido.
Por tanto un valor del estadístico F que sea significativo sugeriría que existe algún problema en la especificación de la forma funcional del modelo inicial propuesto. La distribución del estadístico F es aproximadamente una $F_{2,n-k-3}$ en muestras grandes bajo la hipótesis nula (así como las hipótesis de Gauss-Markov)

necesarias para poder realizar los estudios econométricos necesarios y descritos en apartados anteriores.

PROCEDIMIENTO ACTUAL	
VARIABLE	R cuadrado
COSPRO_CIS	0,35026
COSPRO_COY	0,60840
COSPRO_ECL	0,89310
COSPRO_ECV	0,38602
COSPRO_EDAD	0,31013
COSPRO_EET	0,37674
cospro_salud	0,22849
COSPRO_EPAP	0,14990
COSPRO_EPAT	0,60041
COSPRO_EPF	0,46295
COSPRO_EST	0,72380
COSPRO_HOTEL	0,49299
COSPRO_IASS	0,10713
COSPRO_ICPM	0,38028
COSPRO_IPC	0,35998
COSPRO_MH	0,05877
COSPRO_MNP	0,53154
COSPRO_PCE	0,48712
COSPRO_TICHP	0,62653

Figura 11: R^2 equivalente por operación con el procedimiento descrito

Como resultado de estos procesos se han mejorado ligeramente algunos de los R^2 . La tabla final se muestra en la Figura 11.

6. Consideraciones sobre el método inicial y ajustes necesarios

Analizando las nubes de puntos, y realizadas las transformaciones pertinentes, se observan R^2 todavía bajos, por lo que con la filosofía actual será dificil aumentar la explicación de la variabilidad de cospro en cada delegación.

En la estimación final de los DEE hay que reconsiderar la agrupación actual de las operaciones estudiadas. Debido a los resultados poco satisfactorios de algunas regresiones es necesario efectuar una agregación mayor en los datos para minimizar el ruido que puede suponer la cumplimentación de cospro con los problemas que ello conlleva y los diferentes criterios que parecen existir respecto a cada operación estadística. Esta apreciación se constata claramente en el siguiente gráfico en el que están representadas las cargas de trabajo totales y el cospro de los DEE, y en el cual se aprecia una compactación de la nube, que facilitaría su modelización.

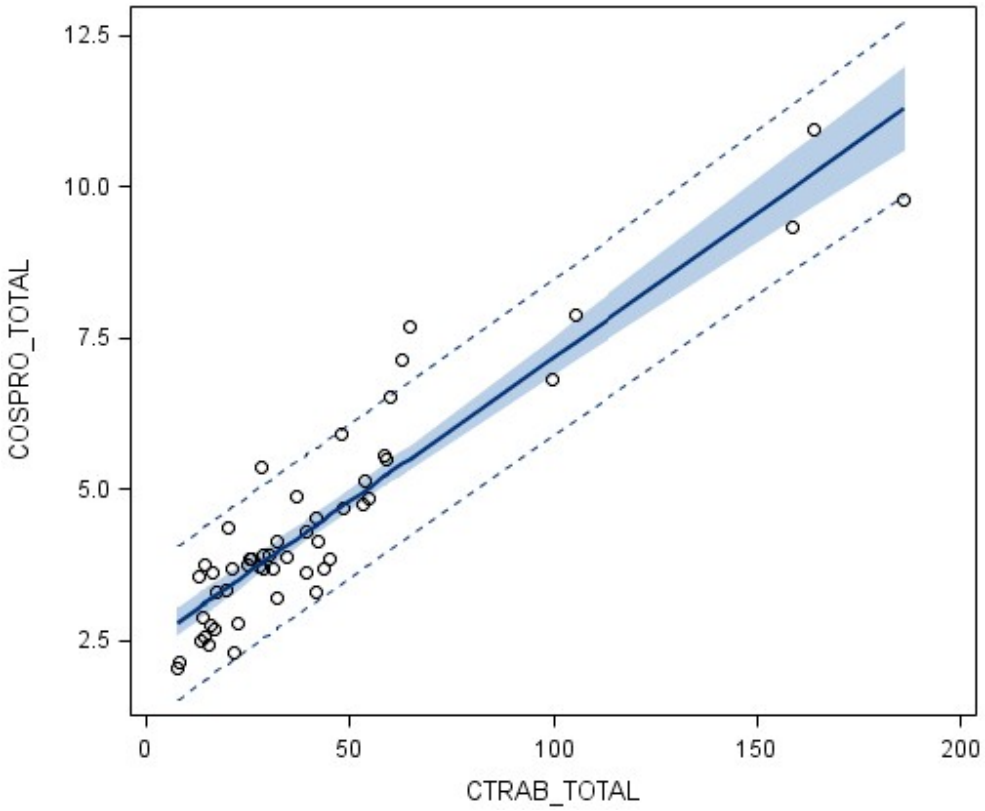


Figura 12: Agregación de cargas y cospro

Por otra parte, el análisis de Censo Electoral y Padrón arrojará nuevos resultados en la estimación de esta parte en cada delegación, por lo que cuando esté finalizado los resultados cambiarán. Este análisis consistirá básicamente en analizar las variables de censo y padrón que repercuten en la carga de trabajo de las delegaciones, para mediante un análisis de componentes principales tratar de reducirlas y construir un modelo econométrico que intente explicar los datos de cospro.

Por tanto, ha sido necesaria una reflexión sobre la filosofía de estimación para cambiarla e intentar introducir nuevas variables al estudio que aporten información a la estimaciones, como pudieran ser el número de operaciones que lleva cada DEE y cuáles son, el número de operaciones estadísticas en las que participa cada delegación, si una delegación tiene o no cati o urce, etc...

7. Nueva metodología de estimación de DEE en las DDPP del Ine

El grupo de trabajo del año 2010 se formó en el mes de Abril tras la reunión anual de delegados provinciales que tuvo lugar en Madrid. Dicho grupo trabajó en la misma línea que lo había hecho el anterior y con las mismas directrices y metodología, con el objetivo de proporcionar al consejo de dirección del Ine en el mes de Octubre una estimación del número de DEE necesarios en las delegaciones provinciales en el año en curso.

Dicha metodología fué actualizada en los términos descritos en los puntos anteriores, fundamentalmente en cuanto a la consideración de transformaciones box-cox, a una revisión exhaustiva de los libros de cargas, y a la utilización de datos de cospro de dos años completos, que han sido 2008 y 2009. Por otra parte, se mejoró en esa primera aproximación el planteamiento de los diferentes componentes que deben intervenir en la estimación. Hasta Abril de 2010 se utilizaban datos de cospro de algunos meses y los resultados estimados se elevaban a los datos de DEE que figuraban en la RPT. El esquema de estimación propuesto en una primera fase revisión hasta Octubre de 2010 ha sido el siguiente:

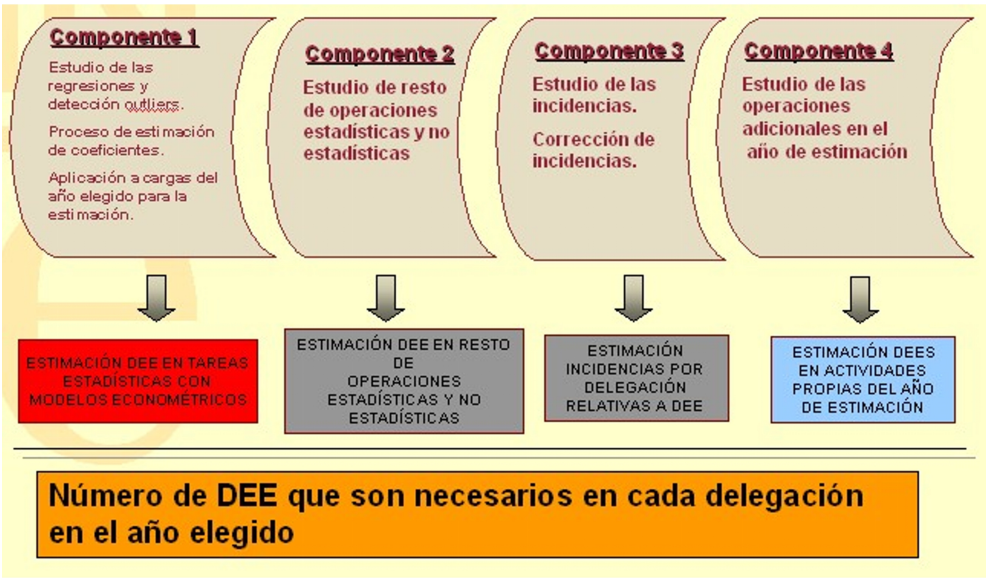


Figura 13: Esquema de estimación de DEE por el grupo de trabajo en 2010

Es decir, que la estimación total se compone de la parte de trabajos estadísticos que realizan los técnicos, más una parte de resto de operaciones estadísticas y no estadísticas, incidencias ajustadas y operaciones adicionales ejecutadas en el año en el cual se pretenden estimar los DEE.

Respecto a las incidencias, y para mantener la coherencia en la estimación, se han considerado los datos de 2008, puesto que es este periodo el que se ha utilizado para el cálculo de los coeficientes. Puesto que los datos de cospro son de A2 se han ajustado las incidencias por la relación existente en cada delegación de DEE respecto de A2. Pero como había delegaciones con valores altos en este apartado el grupo de trabajo se puso en contacto con aquellas con valores superiores a 0.5 para corroborar si la incidencia fue de un técnico en estadística o no, modificándose los valores de ese ajuste cuando correspondió. La importancia de incluir las incidencias en el esquema de estimación radica en el hecho de que esas posibles ausencias de técnicos, de larga o corta duración, afectan en la cumplimentación de cospro tanto de la persona que sufre la incidencia como de los que se hacen cargo de su trabajo.

Por ejemplo, si el DEE que sufre la incidencia la incluye en su cospro, pero ningún compañero refleja en el suyo el trabajo que ha asumido por esa ausencia (correspondería seguramente a bajas de corta duración), en el peor de los casos (si a todos los técnicos que llevan una tarea común les ocurriera esto) todas las nubes de puntos, de las tareas que lleva el técnico con la incidencia, estarían más bajas porque la dedicación reflejada en cospro sería menor al estar incluida la incidencia. Por tanto las estimaciones de esas tareas serían más bajas de lo que correspondería. Consecuencia: es necesario incluir las incidencias en la estimación.

7.1. Nuevo esquema de estimación de DEE en las DDPP

Uno de los aspectos que se ha corregido en la segunda fase que ha realizado el grupo, entre los meses de Noviembre de 2010 y Febrero de 2011 ha sido las estimaciones de cada uno de los componentes descritos en el esquema propuesto en la Figura 13.

Respecto del primer componente, es decir, la parte de estimación de dee en tareas estadísticas, se propone el esquema mostrado en la Figura 14. En el mismo se describe una combinación de dos predicciones: la primera está formada por la utilización de 4 modelos que agrupan los 22 modelos utilizados hasta ahora y la segunda se obtiene de la construcción de un único modelo que agrupa todas las variables y estima el número de DEE. El resultado de estas dos estimaciones se combina para obtener los DEE necesarios en cada delegación mediante una combinación lineal con un coeficiente α, que habrá que calcular de manera que minimice la varianza de los errores de ambas estimaciones.

El motivo de proceder de esta forma es minimizar el ruido que presenta cospro en las regresiones individuales y que se constata en unos bajos coeficientes de determinación, incluso introduciendo la mejora de las transformaciones box-cox. Mediante las agrupaciones conseguiremos que los datos de cospro sean más compactos y será más sencillo buscar nuevas variables que ayuden a explicar la

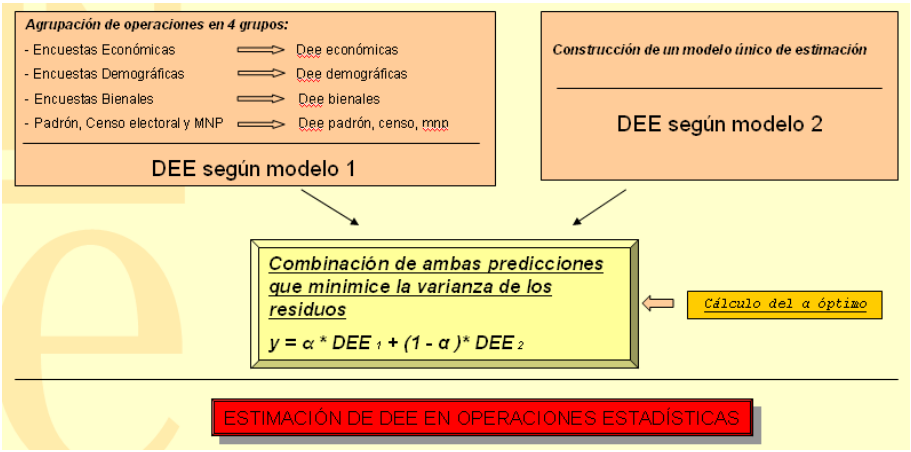
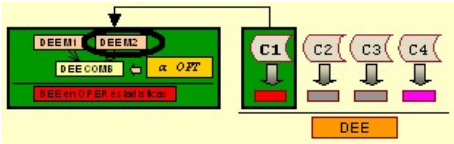


Figura 14: Estimación de DEE en tareas estadísticas

variabilidad de cospro.

7.1.1. Submodelo del total de operaciones



Comenzamos explicando el modelo 2, que agrupa al máximo la información de cospro y cargas e intenta utilizar en la estimación global otras variables que pudieran ser de interés. Entre ellas se ha intentado utilizar variables como si una delegación tiene cati o no , si tiene urce o no, el número de operaciones en las que la delegación participa (siguiendo el criterio anterior), el número de operaciones por técnico, los DEE reales en 2008 y numerosas interacciones entre ellas, resultando el siguiente modelo:

$$cospro_{total} = \beta_0 + \beta_1 * ctrab_{total} + \beta_2 * cati_{DDPPpeques} + \beta_2 * numoper_{totales} \quad (12)$$

Este modelo tiene un R^2 de 0.8834 y se comporta bien frente a la carga de trabajo global de las delegaciones. La variable cati toma el valor 1 para las delegaciones que son "pequeñas"[4], y cero para 2 delegaciones "grandes", puesto que de esta forma, además de ser significativa es capaz de explicar algunas cuestiones de la estimación final. La variable cati sin más no es significativa, entendida como si

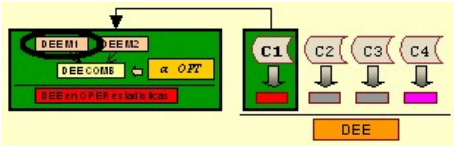
[4]La consideración de grandes y pequeñas se ha hecho según la carga de trabajo total, siendo esta muy diferente entre un grupo y otro.

la delegación tiene cati o no, y por eso ha hecho falta esta matización.
Los parámetros de este modelo, sobre el que no se ha hecho transformación alguna, son los siguientes:

Intercept	CTRAB_TOTAL	CATIP	num_oper_totales
1.1256093691	0.0409435199	0.8074916054	0.1147980506

Es necesario en este punto hacer una matización a la forma de contar el número de operaciones en las que una delegación participa (sea en el área que sea), que el grupo de trabajo, en función de la duración anual de cada operación, ha establecido de la siguiente manera: EJ:$\frac{2}{12}$, ECV:$\frac{4}{12}$, MNP:3, TICHCAPI:$\frac{2}{12}$, PCE:1, TICHCATI:$\frac{4}{12}$, PADRÓN:1, EPAP, EPAT, EPF, HOTEL, IASS, ICPM, EPC, MH:1, CIS:1, EET:$\frac{6}{12}$, COYUNTURALES:1 (cada una de ellas cuenta por una), SALUD:$\frac{6}{12}$, EB:$\frac{2}{12}$, EEEA:$\frac{3}{12}$, ETCL:1, EIAP:$\frac{3}{12}$, EACL:$\frac{3}{12}$, EIAE:$\frac{3}{12}$, EIT:$\frac{4}{12}$, TICCE:$\frac{3}{12}$, EAS:$\frac{4}{12}$, ECI:$\frac{3}{12}$, EAES:$\frac{3}{12}$.

7.1.2. Modelos del primer componente de la combinación



El modelo 1, por otra parte, desglosa la primera estimación en 4 partes, se compone de 4 modelos que son los siguientes:

(1) La primera parte utiliza el siguiente modelo:

$$cospro_{economicas} = \beta_0 + \beta_1 * ctrab_{economicas} \tag{13}$$

en el cual se han agregado las siguientes operaciones: CIS, Coyunturales, ECL, Estructurales, Hoteles, IASS, ICPM e IPC, en el sentido de considerar las variables de cospro y cargas de forma que se sumen estos componentes. Para este modelo se han hecho los contrastes necesarios para verificar que se cumplen las hipótesis del modelo lineal, como el reset, el test de white, contrastes de normalidad de residuos, etc...El R^2 del mismo es 0.7689, lo cual empieza a mostrar las bondades de este nuevo enfoque metodológico. Además, en este caso no es necesaria ninguna transformación de los datos, siendo los beta estimados 0.6961880437 y 0.0381020374.

(2) La segunda parte utiliza el siguiente modelo:

$$cospro_{demograficas}(\lambda) = \beta_0 * ctrab_{demograficas} + \beta_1 * numoper_{demograficas} \tag{14}$$

que incluye dos variables, que son las cargas de trabajo en demográficas y el número de operaciones en las que una delegación participa en el área de demográficas. Cabe destacar que el modelo no tiene término independiente.[5]
El grupo de estadísticas demográficas incluye ECV, Judiciales, EPAP, EPAT, EPF, MH y TICHP.
En este modelo ha sido necesario realizar una transformación box-cox para asegurar las hipótesis del modelo lineal, resultando los siguientes coeficientes: $\lambda = 0{,}25$, y los beta gorros del modelo transformado son respectivamente 0.0407968177 y -0.070510281. Se ha calculado el R^2 resultando ser 0.5588, que sigue siendo bajo, aunque aceptable.

(3) La tercera parte utiliza el siguiente modelo:

$$cospro_{bienales}(\lambda) = \beta_0 + \beta_1 * ctrab_{bienales} + \beta_2 * numoper_{bienales} \quad (15)$$

con el que se consigue un R^2 de 0.7152. Aplicando una transformación box-cox los parámetros resultantes son: $\lambda = 0{,}3$, y los beta gorros del modelo transformado son respectivamente -2.391093531, 0.1067716707 y 0.8679925732.
Este último modelo engloba EET, EB, Salud y EEEA.

(4) La cuarta parte, referente a censo electoral, padrón y mnp, la describimos en la siguiente sección en detalle, puesto que es uno de los puntos críticos y objetivos del estudio.

Por tanto, para obtener la estimación del modelo 1 se aplicarían las cargas del año actual a cada uno de estos modelos para obtener las 4 partes y al sumarlas se llegaría a los dee necesarios en operaciones estadísticas.

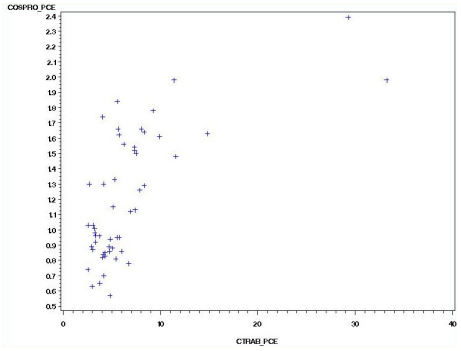
7.2. Análisis y evolución del modelo de estimación de censo y padrón

En una primera aproximación a la estimación de DEE, hasta principios de 2010 la predicción para los DEE necesarios en cada delegación era el resultado

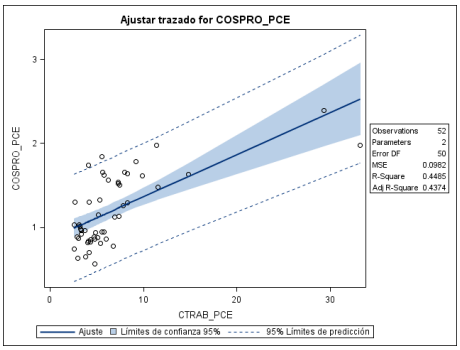
[5] Si consideraramos un modelo con término independiente (ctrab y numoper como independientes) tiene un R^2 de 0.6335, pero ni este es significativo ni lo es el número de operaciones. El modelo sin término indep (vale 0.255) tiene 0.9280 de R^2 (0.63 si lo calculamos como $1 - \frac{SR}{ST}$), pero este R^2 no es comparable con el anterior, por lo que hay que usar el criterio de akaike para decidir entre los dos modelos y aquel cuyo valor sea menor será el elegido. En el modelo sin término independiente el número de operaciones sí que es significativa.
Calculando el criterio de akaike en ambos da un valor menor en el caso sin término independiente, por lo que elegiremos este último modelo (-95,98 sin indep y -94,1065 con indep). Para realizar estos cálculos se ha utilizado la siguiente opción de SAS:

```
proc reg data=agrupacion_p1_filtrado_demo_D outest=demograficas_est_agrup1_D;
    model cospro_demo_D = ctrab_demo_D num_oper_demo_D / noint selection=adjrsq aic bic spec collinoint dw ;
    output out = demograficas_agrup1_conpred_D p=demograficas_D_ndees r=demograficas_D_resid_ndees;
run;
quit;
```

de sumar las predicciones individuales de cada modelo, y en concreto el modelo relativo a censo electoral y padrón era una regresión simple de cospro frente a las cargas de trabajo. Las nubes de puntos y la recta pueden verse en la Figura 15.



Nube de puntos inicial



Recta de Regresión

Figura 15: Modelo inicial de censo electoral y padrón

Como una primera mejora, y a la vista de esta nube se ha incluído en el modelo una variable ficticia o dummie, que toma el valor 1 para Madrid y Barcelona y 0 para el resto de provincias. Mediante una transformación box-cox se ha asegurado que se cumplen las hipótesis del modelo lineal y se ha conseguido aumentar el R^2 de 0.448 a 0.487, es decir, casi un 10 %.

Por tanto el modelo en una primera fase queda de la siguiente forma:

$$cospro_{pce}(\lambda) = \beta_0 + \beta_1 * ctrab_{pce} + \beta_2 * fict_{MB} \tag{16}$$

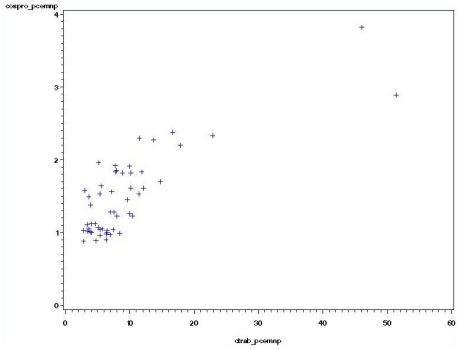
El parámetro de la transformación box-cox es $\lambda = 0{,}3$, y los coeficientes que acompañan al término independiente, a las cargas de trabajo de censo y padrón y a la variable ficticia respectivamente en el modelo transformado son -0.324612732, 0.077253884, -1.213802255. Una matización que conviene recordar es que estas operaciones siempre tienen en cuenta que tanto en cospro como en cargas haya datos significativos, es decir, distintos de cero.

En una aproximación posterior se decidió agrupar los datos de censo y padrón con los datos de movimiento natural de población por establecer una coherencia entre áreas de trabajo relacionadas y similares. De esta forma se compactaría más la nube de puntos y sería posible obtener mejores resultados.

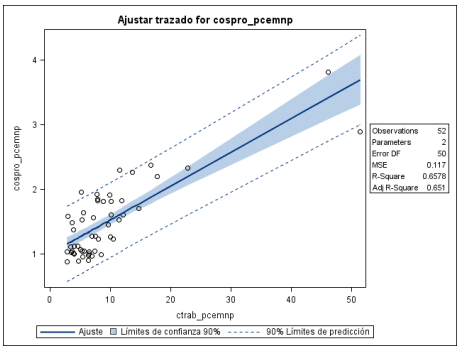
En el modelo utilizado se crea una variable, que es la interacción entre la variable ficticia para Madrid y Barcelona y las cargas de trabajo para censo, padrón y mnp. Tenemos pues:

$$cospro_{pcemnp}(\lambda) = \beta_0 + \beta_1 * ctrab_{pcemnp} + \beta_2 * fict_{MB} + \beta_3 * ctrab_{pcemnp} * fict_{MB} \tag{17}$$

Al igual que en el caso anterior, se filtran los datos para eliminar los ceros en cargas o en cospro y se busca una transformación box-cox. Con estas modificaciones, hemos conseguido un R^2 de 0.7297050301. (Sin realizar la transformación se tenía un R^2 de 0.7327, y con una regresión lineal simple, con término independiente, de cospro frente a cargas, agrupando con mnp los datos de censo y padrón, el R^2 era 0.657). Representando gráficamente esta agrupación se puede apreciar la mejora que se consigue en los resultados, según la Figura 16:



Nueva nube de puntos



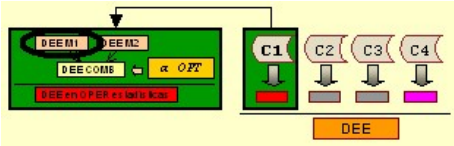
Recta de Regresión

Figura 16: Modelo simplificado de censo electoral, padrón y movimiento natural de población

Por tanto, en el modelo de estimación el parámetro de la transformación box-cox es $\lambda = 0{,}5$, y los coeficientes que acompañan al término independiente, a las cargas de trabajo de censo, padrón y mnp, a la variable ficticia y a la interacción entre ambas (en el modelo transformado) son respectivamente -0.114281755, 0.0605963838, 6.4366090389 y -0.156341005. Dicha modelización es la que se ha utilizado en la estimación de DEE en el ejercicio 2011, puesto que la propuesta que se describe en el siguiente punto está pendiente de aprobación.

Como puede apreciarse, la ganancia en la explicación de la variable dependiente es significativa con estas modificaciones. Bien es verdad que no es la misma variable que al principio del estudio, puesto que se ha agrupado con movimiento natural de población, pero puesto que el objetivo del trabajo es una estimación global esto no afecta demasiado al objetivo y le da más robustez a la estimación final.

7.3. Propuesta de estimación de DEE en censo electoral, padrón y movimiento natural de población mediante técnicas de análisis multivariante



7.3.1. Planteamiento de la situación inicial

Dada la diversidad de operaciones y tareas distintas llevadas a cabo en las áreas de censo electoral, padrón y movimiento natural de población, ha sido necesario identificar la mayor parte de las mismas y solicitar información a las subdirecciones encargadas de estos tres bloques.

Para centrar el análisis en un año se ha considerado el 2009 como punto de partida ya que se disponía de información completa y revisada para el mismo, así como datos de cospro anuales de los DEE en dicho periodo. Con este escenario, el objetivo final es construir un modelo econométrico capaz de estimar los DEE necesarios en estas 3 parcelas.

El objetivo de plantear y estudiar un modelo para estas actividades se justifica por el hecho de que alguna de las variables utilizadas en el libro de cargas para calcular la carga de trabajo en estas áreas están basadas en opiniones "subjetivas"de los delegados de estadística del instituto, lo cual no parece muy conveniente, puesto que la idea es trabajar con datos objetivos de cargas. Estas estimaciones se sustituirían por las del modelo actual, que junto con las estimaciones en operaciones económicas, demográficas y bienales, así como la estimación del modelo con todas las variables y la combinación de ambas predicciones daría el número de DEE necesarios en cada delegación provincial.

7.3.2. Estudio y elección de las variables de censo electoral y padrón

Por parte de la Oficina del Censo Electoral se preparó un fichero con información de variables referidas a 2009, con el cuál el grupo ha trabajado, identificando posibles anomalías en los mismos que desvirtuaran un análisis posterior.

El resultado de esta fase se concreta en la siguiente imagen, que contiene el conjunto de variables de censo electoral que participarán en la metodología:

VARIABLE	OBSERVACIONES	Unidad de medida	Registros por persona y día
modif_territoriales	Número de cambios de municipio, seccionado y resto de variables territoriales en	numero de cambios, que se pueden hacer 12 por persona y hora	72
pere	Número de registros del Pere analizados en 2009	15	90
reclamaciones	Número de reclamaciones al censo electoral en un año por cada delegación	10 por persona y hora	60
tcd	Número de Tarjetas censales devueltas	90 por persona y hora, con lápiz óptico	540
rechazados_2009	Número de registros rechazados en la base de datos de censo electoral	30 por persona y hora	180
Variables excluídas			
certif_inscripcion	Número de certificados de inscripción en el censo electoral	6 por persona y hora	36
excluidos_jurado	Excluídos en el año 2010 en el Tribunal del Jurado	170 por persona y día	170

Figura 17: Variables de Censo Electoral

Las variable relativa a los certificados de inscripción en el censo se ha excluído porque los datos presentaban una disparidad enorme entre delegaciones, y aunque se ha intentado clarificar la situación y entender la cumplimentación en cada caso, al final se ha considerado que hace falta más homogeneización en las definiciones relativas a este apartado, porque su inclusión produciría gran distorsión de los resultados finales.

Respecto al trabajo relacionado con el Tribunal del Jurado, además de los excluídos se ha considerado el trabajo de envío de listas a municipios, al bop y a la audiencia, pero al ser una operación bienal, en el cálculo de personas por año este correspondía entre un 1 y un 2 por ciento de una persona, con lo que se decidió no incluirla en el resumen final.

Como comentario de este apartado cabe decir que hay algunas tareas que no están medidas, como la carga en censo electoral de determinadas cintas que se envían puntualmente a las delegaciones, relativas por ejemplo a DNI, etc...

Por otra parte, la subdirección de Padrón proporcionó toda la información de 2009 disponible relativa a las delegaciones provinciales, en la cuál algunas de las variables no repercutían o no se traducían directamente en trabajo para dichas delegaciones. Con algunas recomendaciones de los responsables de esta parcela pudimos identificar y resumir las variables relavantes, que se muestran en la imagen.

VARIABLE	OBSERVACIONES	Registros por persona y día
err100	Suma por provincia de todos los errores 100 del año 2009 que han tenido que gestionar en las DDPP. Viene del fichero Incidencias mensuales tratamiento manual_DDPP_2009_2010.xls de Ana Jurado	240
err72	Suma por provincia de todos los errores 72 del año 2009 que han tenido que gestionar en las DDPP. Viene del fichero Incidencias mensuales tratamiento manual_DDPP_2009_2010.xls de Ana Jurado	240
err77	Suma por provincia de todos los errores 77 del año 2009 que han tenido que gestionar en las DDPP. Viene del fichero Incidencias mensuales tratamiento manual_DDPP_2009_2010.xls de Ana Jurado	240
fich_variac_reci_deleg	Número de ficheros que se reciben en las DDPP de variaciones para su tratamiento, validación, etc... que no llegan por IDA_WEB. Viene del fichero anterior también	16 ficheros al día
err82	Suma por provincia de todos los errores 82 del año 2009 que han tenido que gestionar en las DDPP. Viene del fichero intercambio provincia_09.xls de Ana Jurado	240
err81	Suma por provincia de todos los errores 81 del año 2009 que han tenido que gestionar en las DDPP. Viene del fichero intercambio provincia_09.xls de Ana Jurado	240
numalegtratmanual09	Número de alegaciones para tratamiento manual en las DDPP para las cifras. Los datos están en el fichero: Recuento provincial ficheros Cifras, Reparos y Alegaciones.xls (columna AA)	60 registros por persona y día
fichCpordelegacion09	Número de ficheros C recibidos en las DDPP para las cifras. Los datos están en el fichero: Recuento provincial ficheros Cifras, Reparos y Alegaciones.xls (columna Q)	28 ficheros al día
num_muni_con_REO_devporIne	Número de ficheros R, E, O, que devuelve el Ine a través de las DDPP para las cifras. Los datos están en el fichero: Recuento provincial ficheros Cifras, Reparos y Alegaciones.xls (columna S,U,W)	28 ficheros al día
numfichaleg_reci_09	Número de ficheros de alegaciones recibidos en las DDPP para las cifras. Los datos están en el fichero: Recuento provincial ficheros Cifras, Reparos y Alegaciones.xls (columna Y)	21 ficheros al día
discre_intrapromun09	Número de discrepantes intraprovinciales e intramunicipales, que están en el fichero ResDiscre2009Prov.xls	240
num_fich_EO_reciporAytos	Cuenta el número de ficheros E y O recibidos de los aytos para el proceso de cifras. Viene del fichero Recuento provincial Cifras ficheros E y O.xls	42 ficheros al día
diccio	Tratamiento de diccionarios	80
consultas	En esta variable estarían las consultas tanto de censo electoral como las de padrón	15 minutos por consulta con cada ayto
bajas_SPCE	Aquí estarían los movimientos tratados en la SPCE, tanto altas, bajas, casos extraños	14 expedientes por persona y día

Figura 18: Variables de Padrón

Como puede verse en los dos últimos gráficos, se han establecido ratios por persona y día para ser capaces de traducir la carga de trabajo a personas con las que poder operar después. Dichos ratios pensamos que son conservadores, tras consensuar muchos de ellos en varias delegaciones.

Como análisis futuro se podrían estudiar las variables relacionadas con las asociaciones, DNI, extranjeros, que por el momento no están consideradas.

7.3.3. Reducción de las variables mediante Análisis de Componentes Principales

El análisis de componentes principales tiene como objetivo reducir la dimensionalidad, es decir, describir los valores de p variables por un pequeño subconjunto $r \leq p$ de ellas, con una pequeña pérdida de información. Este nuevo conjunto de variables tiene la peculiaridad de que serán combinaciones lineales de las originales y además son incorreladas, facilitando así la interpretación de los datos. Esta técnica es debida fundamentalmente a Hotelling (1933), aunque originariamente K. Pearson (1901) ya la dibujó con trabajos sobre ajustes ortogonales por mínimos cuadrados.

Así pues, disponemos de un conjunto de variables, las cuales queremos reducir para poder construir después un modelo estadístico de estimación de DEE.

Tras algunas pruebas se ha comprobado que el determinante de la matriz de correlaciones es muy bajo, lo cual indica que las variables están correlacionadas y es un indicio de la idoneidad de este análisis.

El test de Bartlett y la medida de adecuación muestral de Kaiser-Meyer-Olkin son los siguientes, que como puede verse son bastante aceptables para poder realizar esta metodología.

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,791
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	892,443
	gl	91
	Sig.	,000

Figura 19: Pruebas sobre adecuación de este análisis

Se han realizado varios análisis para determinar el número de componentes adecuados en la extracción, decidiendo finalmente considerar 2 componentes, los cuales explican el porcentaje de varianza descrito en la figura 20

Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	7,713	55,095	55,095	7,713	55,095	55,095
2	2,604	18,599	73,694	2,604	18,599	73,694
3	,976	6,973	80,668			
4	,676	4,830	85,497			
5	,525	3,750	89,247			
6	,415	2,962	92,209			
7	,340	2,428	94,637			
8	,305	2,180	96,817			
9	,174	1,242	98,059			
10	,142	1,018	99,077			
11	,065	,463	99,540			
12	,037	,263	99,802			
13	,018	,131	99,933			
14	,009	,067	100,000			

Método de extracción: Análisis de Componentes principales.

Figura 20: Extracción de las componentes

El gráfico de sedimentación es la representación gráfica de los autovalores, y se suele usar para decidir el número de factores o componentes y es el siguiente:

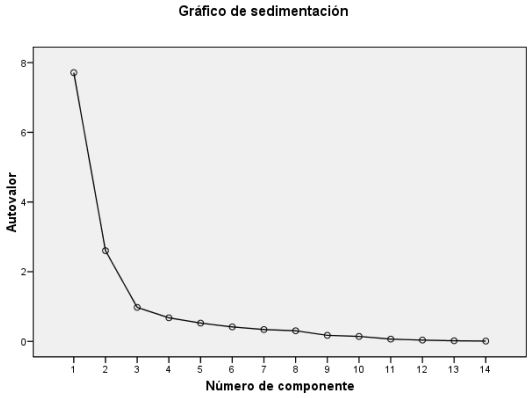


Figura 21: Gráfico de sedimentación

Los coeficientes de los componentes son los siguientes:

Matriz de componentes[a]

	Componente	
	1	2
Puntua(modif_ territoriales)	,749	-,093
Puntua(pere)	,880	-,273
Puntua(reclamaciones)	,918	-,275
Puntua(tcd)	,669	-,468
Puntua(rechazados_ 2009)	,849	-,167
Puntua(errores_y_ duplicados)	,962	-,097
Puntua(numalegtratmanu al09)	,766	-,219
Puntua(fich Cpordelegacion09)	,110	,866
Puntua(num_muni_con_ REO_devporIne)	,625	,706
Puntua(numfichaleg_ reci_09)	,671	,582
Puntua(num_fich_EO_ reciporAytos)	,321	,662
Puntua(diccio)	,957	-,162
Puntua(consultas)	,665	,201
Puntua(bajas_SPCE)	,736	,219

Método de extracción: Análisis de componentes principales.

a. 2 componentes extraídos

Figura 22: Componentes principales

Finalmente, consideramos las dos componentes que resumen toda la información de censo y padrón, que junto con la variable de movimiento natural de

población (esta variable proviene de traducir a personas la parte de matrimonios, defunciones y nacimientos, en base a unos ratios estipulados de trabajo por persona y día, y sumar las tres partes) nos permitirá establecer un modelo de estimación.

7.3.4. Modelo econométrico de estimación de censo, padrón y mnp

Para la construcción de este modelo se han dispuesto de los datos de cospro de 2009 relativos al tiempo empleado por los DEE en todas las tareas relacionadas con las tres áreas y en cada delegación provincial. Ésta será la variable dependiente en dicho estudio, la cuál deberá ser explicada por las tres componentes y la variable relativa al movimiento natural de población.

Se ha analizado cómo debería entrar cada variable en el modelo y se han probado varias combinaciones de modelos, viendo los inconvenientes y bondades de cada uno, resultando como modelo propuesto por el grupo el siguiente:

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	1,208	,154		7,824	,000	,898	1,519
	mnp	,117	,054	,741	2,154	,036	,008	,226
	c1_cubo	-,015	,006	-,425	-2,585	,013	-,027	-,003
	c2_cubo	-,031	,016	-,168	-1,867	,068	-,064	,002
	REGR factor score 1 for analysis 2	,238	,175	,444	1,360	,180	-,114	,589

a. Variable dependiente: pcemnp_A2

Figura 23: Salida de SPSS sobre modelo de estimación

Es decir, el modelo tendría la siguiente expresión:

$$pcemnp_{A2} = 1{,}208 + 0{,}117mnp + 0{,}238(comp1) - 0{,}015(comp1)^3 - 0{,}031(comp2)^3 \quad (18)$$

Las características del mismo son:

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,855[a]	,731	,708	,28894701	,731	31,882	4	47	,000

a. Variables predictoras: (Constante), REGR factor score 1 for analysis 2, c2_cubo, c1_cubo, mnp

b. Variable dependiente: pcemnp_A2

Figura 24: Coeficiente de determinación

Como puede comprobárse, el R^2 es 0.731, mejorando por tanto el del modelo inicial, que era 0.448, y el del modelo utilizado en la estimación de 2011, y aumentando así la explicación de la variabilidad del número de DEE, con la ventaja de no estar sujeto a variables subjetivas y sí a un conjunto de variables anuales fácilmente medibles.

Sobre este modelo se han hecho contrastes de heteroscedasticidad (White y Breusch-Pagan) arrojando p-valores altos, que no invalidad la hipótesis de homoscedasticidad. También se han hecho contrastes de normalidad de residuos, test reset de Ramsey y algunas comprobaciones adicionales arrojando resultados positivos y reforzando la validez de este modelo.

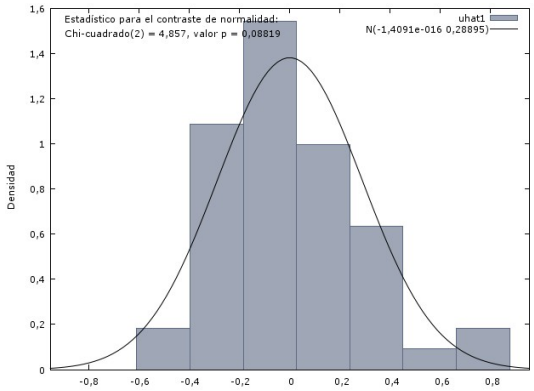
```
Contraste de heterocedasticidad de Breusch-Pagan
MCO, usando las observaciones 1-52
Variable dependiente: uhat^2 escalado

            Coeficiente   Desv. Tipica   Estadístico t   Valor p
  ------------------------------------------------------------
  const       1,45874       0,827490          1,763       0,0844   *
  mnp        -0,146361      0,290082         -0,5046      0,6162
  c1          0,775754      0,935719          0,8290      0,4113
  c1_cubo    -0,0257823     0,0312804        -0,8242      0,4140
  c2_cubo     0,0543866     0,0882320         0,6164      0,5406

  Suma de cuadrados explicada = 7,56371

Estadístico de contraste: LM = 3,781855,
con valor p  = P(Chi-cuadrado(4) > 3,781855) = 0,436333
```

Contraste de Breusch-Pagan



Contraste de normalidad de residuos

Figura 25: Comprobaciones de las hipótesis del modelo lineal general

Es conveniente aclarar que este modelo sólo estima los DEE necesarios para censo electoral, padrón y movimiento natural de población y para que cumpla su función es necesario disponer de las variables necesarias en el año para el cuál se requiera el cálculo. Estas son las relativas al movimiento natural de población y las variables que aparecen en las dos componentes del modelo. En cierto modo esto no es posible porque siempre se van a disponer, como mucho, de las variables descritas en el último año, por lo que cualquier modelo que se

construya con estas variables estimaría cuántas personas hubieran sido necesarias en ese período, y, en el caso de que las cargas se mantuvieran, los DEE que se necesitarían en el año en curso.

Por otra parte la matriz de coeficientes para el cálculo de las puntuaciones en las componentes es la siguiente:

	Componente	
	1	2
Puntua(modif_ territoriales)	,097	-,036
Puntua(pere)	,114	-,105
Puntua(reclamaciones)	,119	-,106
Puntua(tcd)	,087	-,180
Puntua(rechazados_ 2009)	,110	-,064
Puntua(errores_y_ duplicados)	,125	-,037
Puntua(numalegtratmanual09)	,099	-,084
Puntua(fich Cpordelegacion09)	,014	,333
Puntua(num_muni_con_ REO_devporlne)	,081	,271
Puntua(numfichaleg_ reci_09)	,087	,223
Puntua(num_fich_EO_ reciporAytos)	,042	,254
Puntua(diccio)	,124	-,062
Puntua(consultas)	,086	,077
Puntua(bajas_SPCE)	,095	,084

Figura 26: Coeficientes para el cálculo de las puntuaciones en las componentes

También hay que tener en cuenta que con la metodología actual, esta estimación sería sólo una parte del proceso, puesto que junto a ella, habría que calcular las estimaciones de DEE en operaciones económicas, bienales y demográficas, para obtener por suma los DEE estimados totales por delegación. Esta predicción se combinariá después con la estimación obtenida con el modelo agregado para obtener la estimación final.

Por tanto con estos dos modelos descritos en la figura 14 estamos en condiciones de proporcionar, para cada delegación, el número de DEE necesarios en el periodo deseado. Lo que veremos ahora es cómo combinar ambas estimaciones para obtener el valor final asignado a cada una de las delegaciones en tareas estadísticas.

7.4. Combinación de predicciones

El análisis de la calidad de las predicciones de un modelo no siempre resulta una labor sencilla, ya que, en ocasiones, se puede prescindir de un modelo de predicción por no ser el mejor de acuerdo con el criterio de selección que se haya elegido. Según lo anterior, se puede perder información valiosa contenida en el modelo descartado debido, entre otros motivos, a que algunas variables exógenas contenidas en el modelo descartado pueden no estar incluidas en el supuesto mejor modelo.

La combinación de predicciones, según diversos autores como Clemen, R. (1989) y Collopy, F. y Armstrong, J.S. (1992), incrementa la exactitud de los resultados obtenidos de las previsiones individuales de los modelos considerados.

Un factor importante en la combinación de predicciones es la adecuada elección del coeficiente de ponderación que va a afectar al modelo individual de predicción. En este trabajo se seguirá el procedimiento de asignación de pesos propuesto por Granger (1980) que se expone a continuación.

Supongamos que z_i y t_i son dos predicciones para el valor y_i, usando dos procedimientos diferentes. Una nueva predicción para dicho valor, combinando las predicciones anteriores, sería:

$$y_i = \alpha * z_i + (1 - \alpha) * t_i \tag{19}$$

Lo que nos propondremos es calcular ese valor de α óptimo de forma que minimice la varianza de los errores de ambas estimaciones separadas. En adelante omitiremos los subíndices para facilitar la notación. Demostremos pues la siguiente proposición.

Proposición 7.1. *Supongamos que z y t son dos predicciones para el valor y, usando dos procedimientos diferentes. Ambas predicciones se supondrán calculadas utilizando los mismos datos. Una nueva predicción para dicho valor, combinando las predicciones anteriores, sería:*

$$y = \alpha * z + (1 - \alpha) * t \tag{20}$$

El valor del parámetro α que minimiza la varianza de los errores de la predicción combinada es:

$$\alpha = \frac{SR_2 - Pi}{SR_2 + SR_1 - 2 * Pi} \tag{21}$$

, siendo $Pi = \sum_{i=1}^{n} (y_i - \hat{z_i}) * (y_i - \hat{t_i})$

Demostración. Antes de comenzar los cálculos denotaremos por SR_1 la suma residual de la primera de las predicciones, es decir, la de z y por SR_2 la suma residual de la segunda de las predicciones, o sea, la de t. Recordemos que la suma residual se define por la suma de los cuadrados de las diferencias entre los valores de la variable dependiente y los valores estimados:

$$SR_1 = \sum_{i=1}^{n} (z_i - \hat{z_i})^2 \tag{22}$$

Sabemos que en modelos con término independiente se verifica que $ST = SR + SE$, es decir, que la suma total es la suma residual más la suma explicada, y lo que pretendemos es minimizar SR_y en función de α. Veamos:

$$\begin{aligned} SR_y = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} (y_i - (\alpha * \hat{z_i} + (1-\alpha) * \hat{t_i}))^2 = \\ = \sum_{i=1}^{n} (\alpha * \hat{y_i} + (1-\alpha) * \hat{y_i} - (\alpha * \hat{z_i} + (1-\alpha) * \hat{t_i}))^2 = \\ = \sum_{i=1}^{n} (\alpha * (y_i - \hat{z_i}) + (1-\alpha) * (y_i - \hat{t_i}))^2 = \\ = \alpha^2 \sum_{i=1}^{n} (y_i - \hat{z_i})^2 + (1-\alpha)^2 \sum_{i=1}^{n} (y_i - \hat{t_i})^2 + 2\alpha(1-\alpha) * Pi = \\ = \alpha^2 * SR_1 + (1-\alpha)^2 * SR_2 + 2\alpha(1-\alpha) * Pi \end{aligned}$$

Para calcular el mínimo tenemos que derivar e igualar a cero:

$$\begin{aligned} 0 = \frac{\partial SR_y}{\partial \alpha} = 2\alpha * SR_1 + 2\alpha * SR_2 - 2 * SR_2 + 2Pi - 4\alpha * Pi = \\ = \alpha * (2 * SR_1 + 2 * SR_2 - 4Pi) + 2Pi - 2 * SR_2 \end{aligned}$$

Si despejamos α de la igualdad anterior queda:

$$\alpha = \frac{SR_2 - Pi}{SR_1 + SR_2 - 2 * Pi} \tag{23}$$

Para comprobar que efectivamente este es el valor de α buscado, hay que sustituir este valor en la suma residual de y (SR_y) para obtener un Valor Mínimo (VM) y demostrar que es menor que SR_1 y que SR_2, concluyendo así la demostración.
Tenemos pues que

$$\alpha = \frac{SR_2 - Pi}{SR_1 + SR_2 - 2 * Pi} \qquad 1 - \alpha = \frac{SR_1 - Pi}{SR_1 + SR_2 - 2 * Pi} \tag{24}$$

$$
\begin{aligned}
SR_y &= \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \\
&= \sum_{i=1}^{n}(\frac{SR_2 - Pi}{SR_1 + SR_2 - 2 * Pi} * \hat{z} + \frac{SR_1 - Pi}{SR_1 + SR_2 - 2 * Pi} * \hat{t} - \frac{SR_1 + SR_2 - 2 * Pi}{SR_1 + SR_2 - 2 * Pi} * y)^2 \\
&= \sum_{i=1}^{n}(\frac{(SR_2 - Pi) * \hat{z} - (SR_2 - Pi) * y + (SR_1 - Pi) * \hat{t} - (SR_1 - Pi) * y}{SR_1 + SR_2 - 2 * Pi})^2 \\
&= \sum_{i=1}^{n}(\frac{(SR_2 - Pi) * (\hat{z} - y) + (SR_1 - Pi) * (\hat{t} - y)}{SR_1 + SR_2 - 2 * Pi})^2 \\
&= \frac{(SR_2 - Pi)^2}{(SR_1 + SR_2 - 2 * Pi)^2} * \sum_{i=1}^{n}(\hat{z} - y)^2 + \frac{(SR_1 - Pi)^2}{(SR_1 + SR_2 - 2 * Pi)^2} * \sum_{i=1}^{n}(\hat{t} - y)^2 + \\
&+ \frac{2(SR_1 - Pi)(SR_2 - Pi)}{(SR_1 + SR_2 - 2 * Pi)^2} * \sum_{i=1}^{n}(\hat{z} - y)(\hat{t} - y) \\
&= \frac{(SR_2 - Pi)^2}{(SR_1 + SR_2 - 2 * Pi)^2} * SR_1 + \frac{(SR_1 - Pi)^2}{(SR_1 + SR_2 - 2 * Pi)^2} * SR_2 + \\
&+ \frac{2(SR_1 - Pi)(SR_2 - Pi)}{(SR_1 + SR_2 - 2 * Pi)^2} * Pi \\
&= \frac{1}{(SR_1 + SR_2 - 2 * Pi)^2} * ((SR_2^2 + Pi^2 - 2Pi * SR_2) * SR_1 \\
&+ (SR_1^2 + Pi^2 - 2Pi * SR_1) * SR_2 + 2 * (SR_1 - Pi)(SR_2 - Pi) * Pi)) \\
&= \frac{1}{(SR_1 + SR_2 - 2 * Pi)^2} * (SR_2^2 \dot{S}R_1 + Pi^2 \dot{S}R_1 - 2Pi * SR_2 \dot{S}R_1 \\
&+ SR_1^2 \dot{S}R_2 + Pi^2 \dot{S}R_2 - 2Pi * SR_1 \dot{S}R_2 + 2 * SR_1 SR_2 Pi - 2 * Pi^2 \dot{S}R_2 \\
&- 2 * Pi^2 \dot{S}R_1 + 2Pi^3) \\
&= \frac{1}{(SR_1 + SR_2 - 2 * Pi)^2} * (SR_1 SR_2 (SR_1 + SR_2) - Pi^2 (SR_1 + SR_2) \\
&- 2PiSR_1 SR_2 + 2Pi^3) \\
&= \frac{1}{(SR_1 + SR_2 - 2 * Pi)^2} * ((SR_1 + SR_2)(SR_1 SR_2 - Pi^2) \\
&- 2 * Pi(SR_1 SR_2 - Pi^2)) \\
&= \frac{SR_1 SR_2 - Pi^2}{SR_1 + SR_2 - 2 * Pi} = Minimo = V.M.
\end{aligned}
$$

Una vez que hemos visto el valor mínimo de la suma residual de y hay que comprobar que efectivamente se cumple la siguiente desigualdad:

$$V.M. < SR_1 \tag{25}$$

$\frac{SR_1 SR_2 - Pi^2}{SR_1 + SR_2 - 2 * Pi} < SR_1 \Leftrightarrow SR_1 SR_2 - Pi^2 < SR_1^2 + SR_1 SR_2 - 2 * PiSR_2$

$\Leftrightarrow 0 < SR_1^2 + Pi^2 - 2 * PiSR_1 \Leftrightarrow 0 < (SR_1 - Pi)^2$
, lo cual se cumple siempre, por lo que $V.M. < SR_1$.

De manera similar puede demostrarse que $V.M. < SR_2$, con lo que queda demostrada la proposición[6].

□

Vamos a desarrollar dos expresiones que utilizaremos posteriormente:

$$R_M^2 = 1 - \frac{V.M}{ST} = 1 - \frac{SR_1SR_2 - Pi}{ST(SR_1 + SR_2 - 2Pi)} \tag{26}$$

$$\begin{aligned}
\alpha(1-\alpha) &= \frac{SR_2 - Pi}{SR_1 + SR_2 - 2 * Pi} * \frac{SR_1 - Pi}{SR_1 + SR_2 - 2 * Pi} \\
&= \frac{SR_1SR_2 - Pi^2 - Pi * SR_1 - Pi * SR_2 + 2Pi^2}{(SR_1 + SR_2 - 2 * Pi)^2} \\
&= \frac{SR_1SR_2 - Pi^2 - Pi * (SR_1 + SR_2 - 2Pi)}{(SR_1 + SR_2 - 2 * Pi)^2} \\
&= (V.M - Pi) * \frac{1}{(SR_1 + SR_2 - 2 * Pi)}
\end{aligned}$$

Corolario 7.2. $R_M^2 = \alpha R_1^2 + (1-\alpha)R_2^2 + \frac{\alpha(1-\alpha)(SR_1+SR_2-2*Pi)}{ST}$

Demostración.

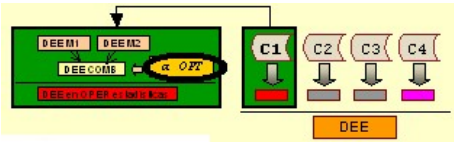
$$\begin{aligned}
\alpha R_1^2 + (1-\alpha)R_2^2 &= \frac{SR_2 - Pi}{SR_1 + SR_2 - 2Pi}R_1^2 + \frac{SR_1 - Pi}{SR_1 + SR_2 - 2Pi}R_2^2 \\
&= \frac{(SR_2 - Pi)(\frac{ST - SR_1}{ST}) + (SR_1 - Pi)(\frac{ST - SR_2}{ST})}{SR_1 + SR_2 - 2Pi} \\
&= \frac{SR_2ST - Pi * ST - SR_1SR_2 + Pi * SR_1 + SR_1ST - SR_2SR_1 - Pi * ST + Pi * SR_2}{SR_1 + SR_2 - 2Pi} \\
&= \frac{ST(SR_1 + SR_2 - 2 * Pi) + Pi * SR_1 + Pi * SR_2 - 2 * SR_1SR_2}{ST(SR_1 + SR_2 - 2Pi)} \\
&= 1 - \frac{-Pi(SR_1 + SR_2) + 2SR_1SR_2}{ST(SR_1 + SR_2 - 2Pi)} = 1 - \frac{-Pi(SR_1 + SR_2) + 2Pi^2 - 2Pi^2 + 2SR_1SR_2}{ST(SR_1 + SR_2 - 2Pi)} \\
&= 1 - \frac{-Pi(SR_1 + SR_2 - 2Pi) + 2(SR_1SR_2 - Pi^2)}{ST(SR_1 + SR_2 - 2Pi)} = 1 + \frac{Pi}{ST} - \frac{2}{ST}V.M \\
&= 1 + \frac{Pi - 2 * V.M}{ST} = 1 - \frac{2 * V.M - Pi}{ST} = 1 - \frac{V.M}{ST} - \frac{V.M - Pi}{ST} = R_M^2 - \frac{V.M - Pi}{ST} \\
&= R_M^2 - \frac{\alpha(1-\alpha)(SR_1 + SR_2 - 2 * Pi)}{ST}
\end{aligned}$$

[6] No se ha utilizado en ningún momento que tenga que ser $0 < \alpha < 1$

Por tanto queda demostrado que se cumple la siguiente propiedad:

$$R_M^2 = \alpha R_1^2 + (1-\alpha)R_2^2 + \frac{\alpha(1-\alpha)(SR_1 + SR_2 - 2 * Pi)}{ST} \tag{27}$$

□



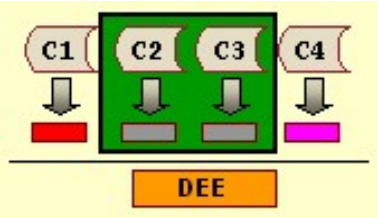
En el caso que nos ocupa y siguiendo el esquema descrito en en la figura anterior se ha calculado el valor óptimo de α resultando ser -0.155945892, lo cual indica que en la ponderación, como es lógico, puesto que el R^2 es mayor, va a pesar más el modelo del total que el desglosado.

Cabe comentar que la estimación eficiente de las ponderaciones dependerá del cumplimiento de los supuestos del modelo lineal de regresión. Además de comprobar estas hipótesis y para descartar la presencia de problemas muestrales se han realizado dos comprobaciones fundamentalmente:

1. Se ha comprobado que en la regresión que combina los pronósticos no aparecen errores autocorrelacionados, puesto que en este caso habría que recurrir a una combinación dinámica de ellos según propone Diebold (1985), con la que se obtendrían pronósticos mejorados. Para ello se han calculado el estadístico de Durbin-Watson y al tomar un valor con el que no es posible determinar si existe o no autocorrelación (puesto que es 2,41 y los valores críticos son $4 - d_U = 2,37$ y $4 - d_L = 2,53$) se ha ejecutado el test de Breusch-Godfrey, el cual ha arrojado p-valores hasta el orden 4 no significativos, lo cual hace pensar que no hay tal presencia.

2. Se han estudiado las relaciones de colinealidad entre los pronósticos, resultando el Índice de condición para cada uno de los valores propios menor que 30 en todos los casos, por lo que la presencia de multicolinealidad se considera baja según proponen Belshey, Kuh y Welsch. Este precepto tambiés es importante verificarlo puesto que de no cumplirse habría que recurrir al uso de la regresión de Raíces Latentes (Webster et al, 1974, Gunst et al, 1976), de componentes principales o la regresión Ridge (Vinod y Ullah, 1981, y Hoerl, Kennard y Baldwin, 1975), para obtener una estimación más eficiente de las ponderaciones.

7.5. Consideraciones sobre el resto de componentes de la estimación final



En esta nueva fase de estimación se ha hecho un estudio más detallado de las componentes Resto de operaciones no estadísticas e Incidencias, puesto que aunque está claro que es necesario utilizarlas, había disparidad de opiniones en cuanto a determinar un criterio lógico de su tratamiento.

La consideración de mayor peso en la decisión de actuar en la forma que se describe a continuación es el hecho de dar la mayor importancia posible a la parte de estimación que proporcionan los modelos estadísticos. Es decir, la ordenación del dato final de DEE estimados no debe verse afectada por las incidencias ni por el resto de operaciones no estadísticas, cuestión que en la primera versión metodológica no funcionaba así.

Figura 27: Componente 2 y 3 de la estimación de DEE

Como resultado de estos estudios se ha decidido analizar el Resto de no estadísticas separadamente para cada grupo de delegaciones con igual número de DEE. Esto se justifica por el hecho de existir mucha variabilidad en la cumplimentación de cospro respecto a estas tareas, según puede apreciarse en la Figura 28, de forma que delegaciones pequeñas tenían datos mayores que delegaciones bastante más grandes que ellas, cuestión aparentemente poco lógica.

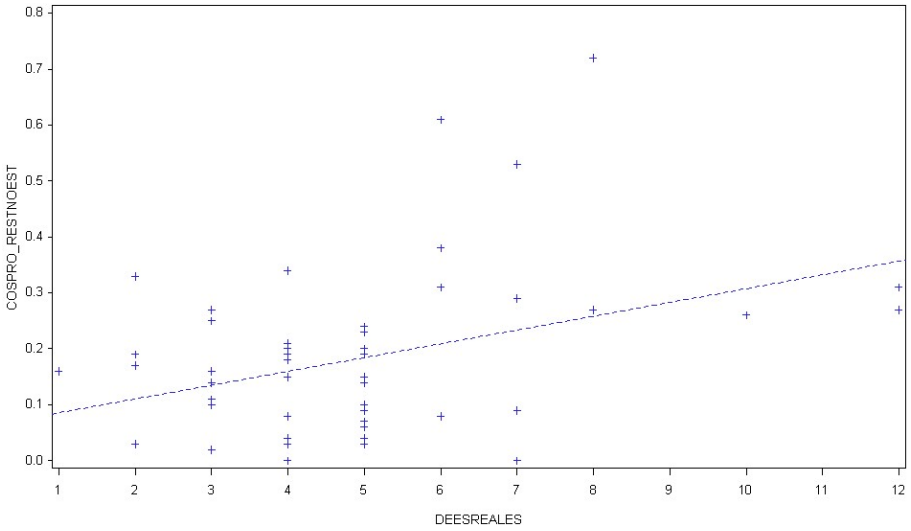


Figura 28: Variabilidad de la componente Resto de tareas no estadísticas intra y entre delegaciones con un número fijo de DEE

Por tanto, lo que ha hecho el grupo de trabajo es considerar la media por grupo de delegaciones con igual número de DEE, de forma que en la estimación final, según este número se le asignará la parte de DEE que le corresponda en estas tareas no estadísticas.

Respecto a las incidencias, se han analizado los datos de cospro, tanto del año 2008 como 2009, arrojando resultados totales muy similares (ver Figura 29), lo cual ha conducido a replantear la parte que se suma para cada delegación. Con el objeto de que la suma total de incidencias sea constante e igual a 14.41 DEE en 2008, que es el valor total ajustado con las consultas realizadas[7], se ha calculado la probabilidad de estar enfermo como el cociente entre ese valor y los DEE reales, que eran 252 en 2008, es decir, 0,05718254. De esta forma, la parte de incidencias que se sumará a cada delegación será el producto entre el número de DEE reales en 2008 multiplicado por este valor.

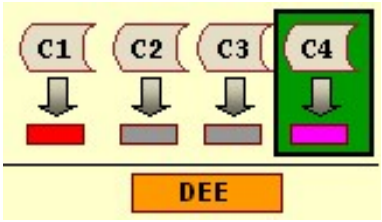
ID	PROVINCIA	RESTEST_08	RESTEST_09	DIF_RESTEST	RESTNOEST_08	RESTNOEST_09	DIF_RESTNOEST	INCIDEN_08	INCIDEN_09	DIF_INCID
	TOTAL	0,75	0,54	0,21	9,38	6,40	2,98	17,99	16,85	1,14
ID	PROVINCIA	RESTEST		2008-2009	RESTNOEST		2008-2009	INCIDEN		2008-2009

Figura 29: Valores totales de incidencias y resto de operaciones estadísticas y no estadísticas

[7] Puesto que los datos de partida de cospro corresponden a A2 corregimos estas incidencias por la relación existente en cada delegación de DEE respecto de A2. Dado que hay valores considerables en algunas delegaciones se han analizado estos casos individualmente en cada centro, viendo si esas incidencias habían sido de diplomados en estadística o no.

7.6. Estimación de los DEE que son necesarios para el año 2011

Siguiendo el esquema metodológico descrito se han calculado las estimaciones con los 2 modelos y se han combinado utilizando el α óptimo de -0.155945892 descrito anteriormente de manera que se minimiza la suma residual del modelo combinado.



Como comentario de la columna referente a las tareas propias del 2011, que se englobarían en el cuarto componente de nuestro esquema según la ilustración anterior, de las cuales no se disponen de datos de cospro, éstas engloban el censo de población, los trabajos previos del mismo relativos a la prueba piloto y la actualización de la cartografía y a algunos trabajos extras en las delegaciones que tienen CATI. Las estimaciones relativas al censo de población se han elaborado con la información suministrada a las delegaciones desde la subdirección de muestreo y recogida de datos en el mes de marzo. En estos cálculos se ha estipulado un ratio de 25 agentes por DEE y sólo durante un mes y medio en este año.

De esta forma se puede obtener un cuadro final con el número de DEE que son necesarios para el año 2011.

Se puede utilizar todo lo expuesto en este documento con las cargas de trabajo de cualquier año para estimar el componente C1, que unido a los componentes C2, C3 y a la particularidad de la estimación del componente C4 daría lugar al número de DEE estimados que son necesarios para ese año de estudio.

8. Bibliografía

Referencias

[1] Pengfei Li. Box-Cox Transformations: An Overview. Department of Statistics, University of Connecticut, Apr 11, 2005.

[2] Jeffrey M. Wooldridge: Introductory Econometrics. A modern approach. Ed: Prentice Hall, 2010

[3] The Box-Cox transformation technique: a review. R.M.Sakia. The Statistician (1992) 41, pp. 169-178.

[4] Time-series analysis supported by Power Transformations. Victor M.Guerrero. Journal of forecasting, Vol. 12, 37-48 (1993)

[5] Análisis estadístico de series de tiempo económicas. Victor Manuel Guerrero Guzmán. Ed: Thomson, 2003

[6] The Retransformed Mean After a Fitted Power Transformation. Jeremy M.G.Taylor. Journal of the American Statistical Association, Vol. 81, No 393. (Mar.,1986),pp.114-118.

[7] Retransformation Bias: A look at the Box-Cox Transformation to Linear Balanced Mixed ANOVA Models. Sakia, R.M. Metrika (1990) 37:345-351

[8] Reducing transformation bias in curve fitting. Miller DM (1984). The American Statistician 38: 124-126.

[9] A Monte Carlo investigation of the Box-Cox transformation in small samples. Spitzer JJ (1978).

[10] Correction for bias introduced by a transformation of variables. Jerzy Neyman. Elizabeth L.Scott. The Annals of Mathematical Statistics Vol 31. No 3 (Sep 1960), pp 643-655

[11] Estimación de la media en Distribuciones asimétricas. Elkin Castaño V.Revista Colombiana de estadística No 33 y No 34, 1996

[12] Nuevos métodos de análisis multivariante. Carles. M. Cuadras. 2010

[13] Un contraste de normalidad basado en la transformación Box-Cox. Daniel P. Sánchez de Rivera, Juan Ignacio. P. Sánchez de Rivera. Estadística española, num 110, 1986.

[14] El modelo lineal sin término independiente y el coeficiente de determinación. Un estudio de Monte Carlo. Rafaela Dios Palomares. Universidad de Córdoba.1996.

[15] Stability of some model selection criteria. Carmen García Olaverri. Universidad pública de Navarra. 1996.

[16] SAS code to select the best multiple linear regression model for multivariate data using information criteria. Dennis J.Beal, Science applications international corporation, Oak Ridge, TN.

[17] Errores frecuentes en la interpretación del coeficiente de determinación lineal. Elena Martínez Rodríguez. Centro universitario San Lorenzo del Escorial.

[18] Procedimiento de verificación y aplicaciones para la especificación de modelos econométricos. Antonio García Ferrer, Departamento de econometría. Universidad Autónoma de Madrid. Estadística Española num 102, 1984 pags 13 a 34.

[19] Greene W. (1998). Análisis econométrico, Prentice Hall.

[20] Detection of model specification, outlier, and multicollinearity in multiple linear regression model using partial regression/residual plots. George C.J. Fernandez, Department of Applied Economics and Statistics. University of Nevada.

[21] Robust regression and outlier detection with the Robustreg procedure. Colin Chen, SAS Institute Inc.

[22] Using heterokedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. Andrew F.Hayes (Ohio State University), and Li Cai (University of North Carolina). 2007

[23] Bates, J.M. and Granger, C.W.J., (1969), "The combination of Forecasts". Operational Research Quarterly, Vol.20, 451-468.

[24] Métodos de combinación de pronósticos: una aplicación a la inflación colombiana. Lecturas de economía No. 52. Medellín, enero 2000.

[25] Combinación de pronósticos y variables predictoras con error. Lecturas de economía, Vol. 41. Castaño. E. (1994).

[26] Metodología del Índice de precios de vivienda. INE (2009).

[27] El Sistema Estadístico SAS, Prentice Hall 2007. César Pérez.

[28] SAS graph 9.1 Reference 2004. SAS Institute Inc,Cary, NC, USA.

[29] Ucla university: http://www.ats.ucla.edu/stat/sas/webbooks/reg/default.htm

[30] R graph gallery. Eric Lecoutre, december 2003.

[31] Gorsuch, R. 1983. Factor Analysis. Second Edition. LEA.

[32] Garcia Jiménez, E.; Gil Flores, J. y Rodriguez Gomez, G. (2000). Análisis Factorial. Cuadernos de Estadística. Editorial La Muralla.

[33] Creating statistical graphics in SAS 9.2: what every statistical user should know. Robert N.Rodriguez and Tonya E. Balan. SAS Institute Inc. Cary, North Carolina.

[34] Técnicas estadísticas con SPSS. Prentice-Hall.2001

[35] OECD Microdata Project. Technological and non-technological innovation. Proposal for indicators of modes of innovation. Fructuoso van der Veen. Statistics Netherlands. September 2007.

[36] Regionalización pluviométrica de la España peninsular a partir de métodos multivariantes. Tesina de José Miguel Sánchez Llorente, Universidad de Salamanca, 1996

[37] Revista de Estudios, Revsa, Salamanca. Estadística Descriptiva y modelos de predicción de diversos contaminantes en la atmósfera urbana de Salamanca. Panero Santos, J.M. Sanchez. No 38, 1996, pag 383-428.

[38] Edición de textos científicos Latex. Walter Mora. F , Alex Borbón. A. Abril 2011.

[39] The not so short. Introduction to Latex2ϵ, by Tobias Oetiker, Hubert Partl, Irene Hyna and Elisabeth Schalegl. June 2010.