SaberEs, Núm. 5 (2013), 37-47. Sección Artículos

ESTIMACION DE LA FUNCION GENERALIZADA DE VARIANCIAS PARA LA ENCUESTA ANUAL DE HOGARES DE LA CIUDAD AUTONOMA DE BUENOS AIRES, AÑO 2010

Cristina Cuesta*

Gonzalo Marí*

Matias Rivero**

Resumen. La función generalizada de variancia (FGV) es un método muy difundido entre organismos gubernamentales encargados de dirigir y planificar encuestas de gran magnitud, que permite obtener de manera sencilla el error muestral de una estimación dada. Consiste en un modelo que pretende reproducir la relación entre un estimador y alguna medida de su variación, por ejemplo la Variancia Relativa (VR). Si bien pueden plantearse varios modelos para la FGV, en la práctica usualmente se consideran dos modelos muy simples (modelo lineal y modelo logarítmico). En este trabajo se explora el uso de otro modelo, basado en la regresión *spline* penalizada (*P-spline*). Se comparan dichos modelos utilizando datos provistos por la Dirección General de Estadística y Censos de la Ciudad Autónoma de Buenos Aires. Para ello se utilizarán las estimaciones de totales de características de interés provenientes de la "Encuesta Anual de Hogares" realizada por dicho organismo, correspondiente al año 2010.

Palabras clave: Errores de estimación; Diseños muestrales complejos; Modelos *P-spline.*

* Docentes Investigadoras del Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario.

** Estadístico de la Dirección General de Estadística y Censos de la Ciudad Autónoma de Buenos Aires.

Contacto: cbcuesta@gmail.com.

Abstract. Generalized Variance Functions (GVF) is a highly disseminated methodology among government agencies entrusted with planning and implementing complex surveys. This method allows obtaining the sample error of estimation in a simple way. It is a model which main objective is to reproduce the relationship between an estimator and a measure of its variance, for example, the Relative Variance (RV). It is possible to adjust several models for de GVF but in practice two simple models are usually considered (linear and logarithmic models). In this article we explore the use of other model based in *P-spline* regression. Models are compared using data from the General Direction of Statistics and Census of Buenos Aires City. The estimations of totals of characteristics from "Annual Home Survey" of 2010 will be used.

Keywords: Estimation errors; Complex survey designs*; P-spline* models.

Original recibido el 08-05-2013

Aceptado para su publicación el 31-07-2013

1. Introducción

Los organismos gubernamentales que planifican y llevan adelante una encuesta de gran magnitud tienen por objetivo brindar estimaciones (por ejemplo totales o proporciones) para un gran número de características que surgen tanto de las variables primarias recolectadas en el operativo, como del cruce entre dos o más de ellas. Estas estimaciones sólo tienen una validez real si están acompañadas de una medida del error de las mismas. El cálculo de estos errores se puede complejizar mucho según sea el diseño muestral utilizado. Es preciso remarcar también que a menudo el número de características a estimar es muy grande y por lo tanto el proceso de estimar los errores muestrales considerando el diseño muestral puede ser una tarea ardua de realizar. En definitiva, la obtención de un medidas del error de una estimación puede dificultarse debido a que: 1) debe respetarse la estructura del estimador y el diseño muestral llevado a cabo en la encuesta, que por lo general son complejos. 2) la cantidad de características de interés que provienen de las encuestas es muy numeroso y resulta casi imposible calcular y publicar todas las estimaciones con sus errores muestrales respectivos.

La función generalizada de variancia (FGV) es un método muy difundido en estos casos ya que permite obtener de manera sencilla el error muestral de una estimación. La misma consiste en ajustar un modelo con el objetivo de encontrar una relación entre el estimador y alguna medida de su variación, por ejemplo la Variancia Relativa (VR). Si bien pueden plantearse varios modelos para la FGV, en la práctica usualmente se consideran dos modelos muy simples (modelo lineal y modelo logarítmico) (Jang *et al,* 1997; Jang *et al,* 2000; Valliant, 1987). Estos modelos están ampliamente difundidos por su simplicidad de interpretación e implementación. Sin embargo, no siempre logran un ajuste adecuado, especialmente para valores extremos de la característica de interés (por ejemplo cuando se quieren estimar totales muy pequeños). Para subsanar este inconveniente se propone postular un modelo de regresión *P-spline* a los datos el cual se espera que se ajuste mejor a los datos en todo su campo de variación y que por otro lado también sea simple de llevar a cabo. La ventaja de este último es que la forma del mismo no está pre-establecida ya que se construye a partir de los propios datos.

La comparación entre los modelos se lleva a cabo con datos provistos por la Dirección General de Estadística y Censos (DGEyC) de la Ciudad Autónoma de Buenos Aires (CABA). Para ello se utilizan las estimaciones de totales de características de interés provenientes de la "Encuesta Anual de Hogares" realizada por dicho organismo, correspondiente al año 2010.

2. Aspectos metodológicos

2.1 FGV para el estimador de un Total Poblacional

Usualmente se emplea el Coeficiente de Variación (CV) como una medida del error muestral de una estimación. Otra opción válida es calcular el cuadrado del CV, que se conoce como Variancia Relativa (VR) y se define como el cociente entre la variancia del estimador y el estimador al cuadrado. La mayoría de las FGV consideradas están basadas en la premisa que la

variancia relativa es una función decreciente de la esperanza del estimador del total (Wolter, 2007)

En el caso de considerar un estimador del total, se tiene que

$$VR(\hat{t}_x) = \frac{Var(\hat{t}_x)}{\hat{t}_x^2} \qquad (1)$$

donde:

t_x es el estimador de un total poblacional t_x y
Var (t_x) es la variancia del estimador t_x correspondiente al diseño muestral empleado.

A fin de estimar la VR en función del total poblacional estimado (evitando el cálculo de la variancia de dicho total) se plantean modelos del tipo:

$$\hat{VR}(\hat{t}_x) = \alpha + \frac{\beta}{\hat{t}_x} + \varepsilon \qquad (2)$$

$$\log\left(\hat{VR}(\hat{t}_x)\right) = \alpha + \beta \log(\hat{t}_x) + \varepsilon \qquad (3)$$

dondeα y β son parámetros de la regresión que deben ser estimados. Los métodos de estimación más usados para estimar estos parámetros son Mínimos Cuadrados Ordinarios (MCO) y Mínimos Cuadrados Ponderados (MCP). Nótese además que ambas funciones son decrecientes en t_x. Usualmente se denomina al modelo (2) como el "Modelo Lineal" y (3) como el "Modelo Logarítmico".

En este trabajo se ajusta el modelo (2) mediante MCO y MCP ponderando por la inversa de VR t_x^2 existen otras alternativas de ponderación (Wolter, 2007) pero en este trabajo no se tendrán en cuenta). El modelo (3) se ajusta únicamente mediante MCO, ya que al aplicar la transformación logaritmo se reduce la inestabilidad que puede presentarse para valores pequeños de t_x de esta forma es redundante el ajuste por MCP.

2.2 Regresión *Spline* Penalizada

Un modelo de regresión *Spline* (o regresión por partes) es aquel donde se divide el campo de variación de la característica estudiada en subregiones tales que en cada una de ellas se ajusta un modelo de regresión polinómica (en general de bajo orden) y que están unidos en los extremos ("nodos") para dar continuidad a la curva. La expresión de este modelo para la FGV es:

$$\hat{VR}(\hat{t}_x) = \beta_0 + \beta_1 \hat{t}_x + \ldots + \beta_p \hat{t}_x^{\,p} + \sum_{k=1}^{K} \beta_{pk} (\hat{t}_x - N_k)_+^p + \varepsilon \qquad (4)$$

donde las expresiones de la forma $t_x - N_K \overset{P}{+}$ se conocen como funciones de "bases truncadas"; las mismas toman el valor $t_x\text{-}N_K{}^P$ cuando es mayor que N_K y cero en otro caso.

El problema de este tipo de modelos es que dependen de la cantidad de regiones que se consideren y de su amplitud. Por ello se postulan modelos de regresión *Spline* penalizados (*P-Spline*) que permiten solucionar este conflicto definiendo un número grande de regiones y luego ponderando (penalizando) la importancia de que las regiones sean consideradas distintas. El grado de suavizado está controlado por un parámetro de suavizado (λ).

El modelo de regresión *Spline* Penalizada puede re-expresarse como un modelo mixto considerando β_{PK} como un coeficiente aleatorio. Esto fue profundizado y difundido en los últimos años por diversos autores como Ruppert *et al.* (2003) y Ngo *et al.* (2004), y es muy útil tanto por el estudio de sus propiedades teóricas como por su implementación práctica ya que la mayoría de los programas de computación estadísticos tienen rutinas para realizar estimaciones y pruebas de hipótesis con modelos mixtos. Una gran ventaja de este enfoque es que el problema de la cantidad de nodos a utilizar deja de ser importante y además el analista no debe preocuparse por decidir el parámetro de penalidad (λ) que mejor represente sus datos, ya que el mismo queda determinado por la estimación de las componentes de variancia respectivas $\lambda_{opt} = {}^{2p}\overline{\sigma_\varepsilon^2\ \sigma_\mu^2}$ Si bien existen diversos procedimientos e estimación de las componentes de variancia, los más razonables están basados en la verosimilitud: Máxima Verosimilitud (ML) o Máxima Verosimilitud Restringida (REML).

2.3 Medidas empleadas para la evaluación de los modelos

La comparación de los modelos para decidir cual es aquél que suministra predicciones más adecuadas de las VR, se realiza a través de las medidas de diagnóstico ER y ERA.

$$ER = \frac{\hat{CV} - \hat{CV}_{pred}}{\hat{CV}} * 100 \qquad (5)$$

El error relativo (ER) indica el porcentaje que representa la diferencia entre el CV estimado de manera directa y el CV predicho por el modelo CV_{pred} sobre su CV estimado directamente.

Además permite detectar patrones de sub y sobre estimación:

Sobreestimación: ER<0 se tiene cuando $CV_{pred} > CV$

Subestimación: ER>0 se tiene cuando $CV > CV_{pred}$.

Se pretende que el ER sea en promedio cercano a cero, con la menor dispersión posible, que la mediana también sea cercana a cero y los porcentajes de sub y sobre estimación sean similares.

El error relativo absoluto (ERA) se define como el valor absoluto del ER y al igual que el anterior, mide la pérdida de precisión relativa en valor absoluto debido al uso de los CV_{pred}.

$$ERA = |ER| \qquad (6)$$

El ERA permite evaluar los CV_{pred} analizando el número de ERA's mayores a determinados porcentajes, tomando como valor crítico el 20%. El promedio de estos mide la distancia promedio entre las estimaciones directas y las predichas por el modelo y se expresa como un porcentaje del estimado de forma directa. Se pretenden valores pequeños del promedio de ERA's lo que indica que la FGV ajusta adecuadamente.

2.4 Encuesta Anual de Hogares

La Encuesta Anual de Hogares (EAH) es llevada a cabo por la Dirección General de Estadísticas y Censos de la Ciudad Autónoma de Buenos Aires desde el año 2002. Tiene como objetivo recabar datos para conocer y analizar la situación socioeconómica y demográfica de la población y de los hogares de la Ciudad. Los temas incluidos en la encuesta corresponden a educación, salud, aspectos demográficos y mercado de trabajo.

El diseño llevado a cabo en la EAH es un diseño muestral estratificado en dos etapas, que considera como ultima unidad de muestreo a la vivienda.

Las estimaciones de interés empleadas para el ajuste de los modelos de FGV corresponden a totales de características de personas y de viviendas-hogares de la Ciudad Autónoma de Buenos Aires (CABA) de la EAH del año 2010. En ambos casos se estiman para cada característica, los totales y sus variancias para obtener las VR y CV con los cuales se realizan los ajustes y evaluación de los modelos.

Si bien valores de CV<10% se consideran como el límite de una precisión buena o aceptable y valores superiores representan un precisión mediocre o mala, se toma como criterio de inclusión aquellas estimaciones de CV<20% para garantizar un tamaño considerable de datos para realizar el ajuste. De esto último, se tiene que el número de estimaciones referidas a características de las personas es de 311 estimaciones (tales como "Total de hombres", "Total de mujeres nacidas en CABA", "Total de hombres afiliados a algún sistema de salud", etc.) y en el caso de características de viviendas-hogares es de 125 estimaciones (tales como "Total de viviendas tipo inquilinato", "Total de viviendas con más de 3 habitaciones", etc.).

3. Resultados

Los modelos se ajustan, para dos perfiles de interés, personas por un lado, y por otra parte, viviendas-hogares. Sin embargo, se presentan los resultados del ER y ERA para cuatro grupos formados por los cuartiles de la distribución de las estimaciones del total, ya que interesa evaluar en particular como es el desempeño de los modelos en los tramos iniciales, parciales y finales del ajuste.

Se ajustan dos modelos lineales (MCO y MCP), un modelo logarítmico (MCO) y un modelo de regresión spline penalizado cuadrático considerando 7 nodos en el caso de estimaciones de totales correspondientes a personas, y 5 nodos en las correspondientes a viviendas y hogares. Los totales son expresados en miles de unidades.

3.1 Resultados correspondientes a Personas

En la Tabla 1, se puede observar que de acuerdo a las medidas de diagnóstico que los modelos que presentan un mejor ajuste a los datos son los modelos *P-Spline* y Logaritmo.

Con respecto al ER, el modelo Logaritmo posee media y mediana de –0,69 y –1,20 respectivamente, mientras que el modelo *Spline* tiene valores mayores, con una media –1,89 y mediana –1,87, y además presenta menor dispersión que el modelo Logaritmo. Ambos presentan similares porcentajes de sobreestimación de las VR 54,02 % para el modelo Logaritmo y 59,81 % para el modelo *Spline.*

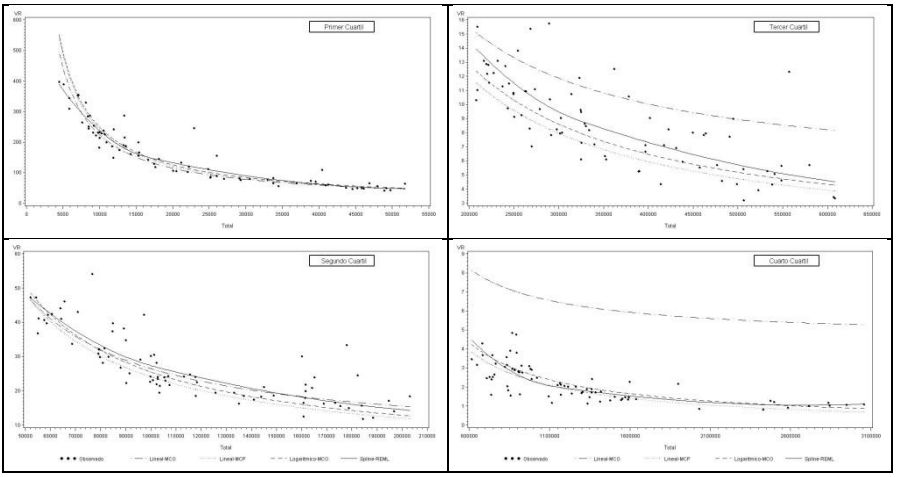
Tabla 1. Estadísticas descriptivas del ER y ERA para los modelos ajustados. FGV para CABA. Totales de Personas.

Modelo	Método de Ajuste	ER				ERA		
		Media	Desvío	Mediana	% de VR Sobreestimadas	Media	Mediana	Percentil 90
Lineal	MCO	-24,53	39,10	-7,66	72,99	28,67	10,47	89,15
Lineal	MCP	3,18	11,22	2,97	37,30	8,73	6,60	19,88
Logaritmo	MCO	-0,69	11,69	-1,20	54,02	8,58	6,16	19,42
P-*Spline*	REML	-1,89	11,13	-1,87	59,81	8,37	6,13	19,00

Fuente: Elaboración propia con datos de Dirección General de Estadística y Censos de la Ciudad Autónoma de Buenos Aires

En relación al ERA, el modelo Logaritmo tiene una media y mediana de 8,58 y 6,16 respectivamente, mientras que el modelo *Spline* presenta valores menores, con una media de 8,37 y mediana 6,13. Además se observa en ambos, que el 90% de los ERA son menores al 20%.

Figura 1. FGV correspondientes a características de Personas (para cada uno de los cuartiles de la distribución de los totales).



Fuente: Elaboración propia con datos de Dirección General de Estadística y Censos de la Ciudad Autónoma de Buenos Aires

Como estas medidas involucran a todo el campo de variación resulta útil observar cómo se desempeñan los modelos en las distintas regiones del ajuste, sobre todo en el tramo inicial y final. Para ello se muestran en la Figura 1 los ajustes por tramos (en cada uno de los cuartiles de la distribución del estimador) donde se observa que el ajuste del modelo *Spline* es similar a los restantes modelos, pero ajusta mejor los datos correspondientes a estimaciones de totales pequeños. Esto último es importante destacar, ya que la tabla de la FGV que se emplea en la DGE y C correspondiente a características de personas, considera estimaciones de totales a partir de los valores estudiados. Se observa también que el modelo *Spline* ajusta mejor los datos en el rango de valores entre 2,5 y 3 millones de personas, a pesar de que las diferencias respecto al Logaritmo son mínimas. Si bien ambos modelos tienen comportamientos similares, el *P-Spline* parece más adecuado para ajustar los datos en los tramos iniciales y finales. A pesar de que no resulta ser el más parsimonioso, la importancia práctica de un buen ajuste en los tramos mencionados determina la elección.

Por lo señalado en último lugar, se prefiere el modelo *P-spline* cuya ecuación estimada es:

$$\hat{VR}(\hat{t}_x) = 563.37 - 44.29(\hat{t}_x) + 1.17(\hat{t}_x{}^2) - 1.10(\hat{t}_x - 16.62)_+^2 - 0.07(\hat{t}_x - 49.76)_+^2$$
$$- 3.62\text{x}10^{-3}(\hat{t}_x - 102.83)_+^2 - 4.30\text{x}10^{-4}(\hat{t}_x - 196.82)_+^2 - 1.70\text{x}10^{-4}(\hat{t}_x - 324.95)_+^2$$
$$- 2.00\text{x}10^{-5}(\hat{t}_x - 607.01)_+^2 - 6.74\text{x}10^{-6}(\hat{t}_x - 1115.80)_+^2$$

3.2 Resultados correspondientes a Viviendas-Hogares

En la Tabla 2 se observa que, al igual que en el caso de personas, los modelos que presentan un mejor ajuste a los datos son los modelos *Spline* y Logaritmo, ya que éstos presentan los valores que más se aproximan a las características pretendidas por las medidas de diagnóstico ER y ERA.

Tabla 2. Estadísticas descriptivas del ER y ERA para los modelos ajustados. FGV para CABA. Totales de Viviendas-Hogares.

Modelo	Método de Ajuste	ER				ERA		
		Media	Desvío	Mediana	% de VR Sobreestimadas	Media	Mediana	Percentil 90
Lineal	MCO	-63,24	106,98	-17,45	76	68,35	24,01	205,88
Lineal	MCP	3,53	11,46	2,47	37,60	8,87	7,23	20,35
Logaritmo	MCO	-1,92	21,44	2,25	41,60	15,15	11,90	25,83
P-*Spline*	REML	-1,50	12,81	-2,00	55,20	10,29	8,65	20,93

Fuente: Elaboración propia con datos de Dirección General de Estadística y Censos de la Ciudad Autónoma de Buenos Aires

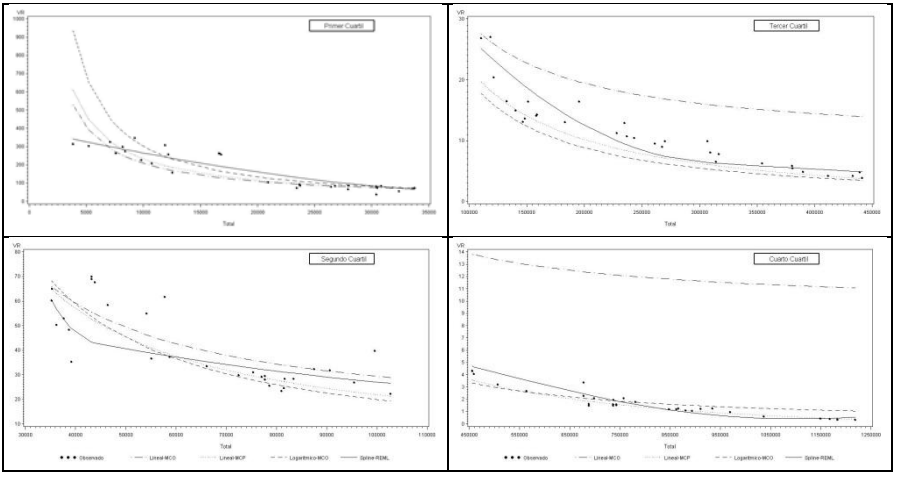
Con respecto al ER, el modelo Logaritmo posee media y mediana de -1,92 y 2,25 respectivamente, mientras que el modelo *P-Spline* tiene valores menores, con una media –1,50, mediana –2,00 y menor dispersión. El porcentaje de sobreestimación de las VR, es de 41,60% para el modelo Logaritmo y 55,20% para el modelo *P-Spline*.

En relación al ERA, el modelo Logaritmo tiene una media y mediana de 15,15 y 11,90 respectivamente, mientras que el modelo *P-Spline* presenta valores más pequeños, con una media de 10,29 y mediana 8,65. Además se observa que el modelo *P-Spline* tiene el 90% de los ERA con valores menores a 21%, siendo menor a 25,83% observado en el caso del Logaritmo.

A pesar de que el modelo *P-Spline* es levemente superior al modelo Logaritmo, los resultados son muy similares. En la Figura 2 puede observarse que el modelo *P-Spline* tiene un mejor desempeño general que el modelo Logaritmo. Se observa que ambos modelos difieren considerablemente y el *P-Spline* ajusta mejor los datos alrededor de estimaciones en el rango de 4000 a 8000 viviendas. Esto último es importante destacar, ya que la tabla de la FGV que se emplea en la DGEyC correspondiente a características de viviendas, considera estimaciones de totales similares a los valores analizados, y si bien en la práctica se acepta que el modelo elegido pueda sobreestimar los datos, el modelo Logaritmo arrojaría estimaciones muy superiores a las que se obtendrían con cualquier otro modelo.

Se puede remarcar también que el modelo *P-Spline* ajusta mejor que el Logaritmo en el rango de valores cercanos a 1,2 millones de viviendas.

Figura 2. FGV correspondientes a características de Viviendas-Hogares (para cada uno de los cuartiles de la distribución de los totales).



Fuente: Elaboración propia con datos de Dirección General de Estadística y Censos de la Ciudad Autónoma de Buenos Aires

En efecto el modelo *P-Spline,* en general, tiene un mejor ajuste de los datos en todas las regiones, hecho que no se observa con el modelo Logaritmo. Nuevamente el modelo *P-Spline* parece más adecuado para ajustar los datos. A pesar de que no resulta ser el más parsimonioso, la importancia práctica de un buen ajuste en los tramos iniciales y finales determina la elección. Por esto último, se prefiere el modelo *P-Spline* cuyo ajuste resultante es:

$$\hat{VR}(\hat{t}_x) = 395.76 - 14.45(\hat{t}_x) + 0.13(\hat{t}_x{}^2) + 0.06(\hat{t}_x - 23.73)_+^2 - 0.19(\hat{t}_x - 43.81)_+^2$$

$$- 1.27\text{x}10^{-3}(\hat{t}_x - 99.51)_+^2 - 4.20\text{x}10^{-4}(\hat{t}_x - 306.67)_+^2 + 7.76\text{x}10^{-6}(\hat{t}_x - 698.70)_+^2$$

4. Consideraciones finales

Una de las ventajas de utilizar las FGV es que permite obtener una medida del error a través de un modelo en aquellas encuestas donde el total de estimaciones resulta ser grande y por cuestiones económicas y operativas no resulta viable la publicación de cada una de las estimaciones con sus correspondientes medidas de precisión.

En este trabajo se contrastaron los resultados obtenidos de modelos lineales paramétricos tradicionalmente utilizados para modelar la FGV con un modelo semiparamétrico *(P-Spline)* hasta el momento no utilizado para el ajuste de FGV y que mostro resultados muy satisfactorios.

El ajuste se llevó a cabo con totales de características derivadas de la Encuesta Anual de Hogares realizada por la DGEyC de la Ciudad Autónoma de Buenos Aires. De los ajustes se observa que para ambos perfiles de interés (personas y viviendas-hogares), el modelo *P-Spline* en general compite con el Logarítmico (el de mejor resultados entre los modelos tradicionales). Se observó que las discrepancias entre ambos modelos son mínimas en cuanto al ER y ERA, y que las diferencias de ajuste en los tramos iniciales y finales pueden definir la elección del modelo. En este sentido, el modelo *P-Spline*, para los datos del año 2010, parece más adecuado para ajustar los datos en los tramos iniciales y finales, pero tiene como desventaja que no resulta ser el más parsimonioso. Para este trabajo se tuvo presente la importancia de un buen ajuste en los tramos mencionados y en consecuencia se prefiere el modelo *Spline* Penalizado en ambos perfiles de interés, a pesar de que el modelo Logaritmo también resulta una opción válida.

En general, la relación entre las estimaciones de totales y sus VR presenta leves variaciones en cada año, que no podrían captarse adecuadamente por los modelos clásicos. En este sentido, el modelo *Spline* Penalizado permite más flexibilidad en el ajuste de los datos, siendo una ventaja respecto de los restantes modelos.

Referencias bibliográficas

Jang D. S., Cox B. G. and Edson D. J. (1997). Generalized Variance Function for data from Multi-frame Surveys: the SESTAT experience. *Proceedings of the Survey Research Methods Section, American Statistical Association* 2, 158-163. Recuperado de http://www.amstat.org/sections/srms/proceedings/papers/1997_024.pdf

Jang D., Garrett J. K., Piotrowski F. W. , Owens W. B. (2000). Generalized Variance Function Methodology for ACNielsen's Homescan Household Panel Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 811-815. Disponible en http://www.amstat.org/sections/srms/proceedings/papers/2000_138

Ngo, L. and Wand, M. P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, 9 (1), 1-54.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression.Cambridge.* New York: University Press.

Valliant, R. (1987). Generalized Variance Functions in Stratified Two-Stage Sampling. *Journal of the American Statistical Association,* 82 (398) 499-508.

Wolter, K. M. (2007). *Introduction to Variance Estimation.* New York: Springer.