ESTADÍSTICA ESPAÑOLA
Vol. 47, Núm. 159, 2005, págs. 299 a 320

Diseño y evaluación empírica de una estrategia de predicción por muestreo en cooperativas agrarias

por
SANTIAGO MURGUI IZQUIERDO
Mª CONSUELO COLOM ANDRÉS
Mª CRUZ MOLÉS MACHÍ
Departamento de Economía Aplicada
Universidad de Valencia

RESUMEN

Sobre la base de un modelo de superpoblación, se propone una estrategia para estimar agregados poblacionales en las cooperativas agrarias valencianas. La disponibilidad de datos censales para distintos ejercicios económicos, permite efectuar un análisis de sensibilidad de la estrategia con especial referencia a la existencia de datos anómalos. La eficiencia de estimadores alternativos robustos se evalúa utilizando técnicas de replicación.

Palabras clave: modelo de superpoblación, dato anómalo, estimador robusto

Clasificación AMS: 62D05

1. INTRODUCCIÓN

El diseño de estrategias de investigación por muestreo en poblaciones finitas bajo la perspectiva que ofrecen los modelos de superpoblación, es conocido en la literatura estadística desde hace algunos años, cabe citar como una recopilación actualizada el texto de Valliant et al (2000). La problemática asociada con este planteamiento inferencial comprende varios aspectos fundamentales: la formulación del modelo, la obtención de estimadores óptimos, el cálculo de la precisión de los resultados, la especificación de un procedimiento de selección muestral y finalmente, el análisis crítico de toda la estrategia en su conjunto.

Una de las conclusiones más llamativas de este esquema metodológico es la conveniencia de utilizar en ocasiones muestras intencionadas, en las que las unidades son seleccionadas por procedimientos en los que no interviene el azar. Este tipo de propuestas, junto a la propia naturaleza del enfoque, inducen a que en la aplicación práctica de una estrategia de tales características se susciten cuestiones que implican tanto al modelo, como a los datos y al procedimiento de selección muestral.

En este trabajo se abordan estas cuestiones en el contexto de una aplicación al universo de las cooperativas agrarias valencianas. Desde 1994 la Consellería de Agricultura de la Comunidad Valenciana elabora y publica las Cuentas Económicas Agregadas del Cooperativismo Agrario (2002). Los datos utilizados son los que a través de sus balances y cuentas de pérdidas y ganancias presentan las citadas entidades con carácter anual y obligatorio.

El retraso con que suele presentase esta información censal obliga a elaborar las cuentas definitivas mucho tiempo después de haber finalizado el correspondiente ejercicio. Es por ello por lo que se propone la determinación de un grupo reducido de cooperativas, en las que pueda accederse a los datos inmediatamente después de cerrar sus cuentas anuales y que constituirán la base para elaborar un avance de los agregados globales.

El objetivo del trabajo es pues la selección de la muestra de entidades y la elaboración de un proceso inferencial que cada año permita la estimación de las cifras agregadas. Todo ello, teniendo en cuenta que las especiales circunstancias que concurren hacen poco apropiado recurrir al muestreo aleatorio.

La formulación del modelo, el cálculo de los estimadores, su precisión y el diseño muestral se abordan en el segundo apartado. En el tercero, se efectúa una valoración empírica de la estrategia y de los resultados que proporciona. En el cuarto, se abordan cuestiones de sensibilidad y robustez frente a la existencia de

datos anómalos. Finalmente, en el quinto se recogen las conclusiones más relevantes.

2. MODELO Y ESTIMACIÓN

Sea $U=(u_1,u_2,\ldots,u_N)$ el colectivo formado por las N entidades que han presentado actividad económica durante un ejercicio $(t-1)$ y cuyos resultados económicos se suponen ya conocidos. Denotemos por $(y_{1(t-1)},y_{2(t-1)},\ldots,y_{N(t-1)})$ el conjunto de valores declarados por las unidades de U acerca de una variable Y. Para efectuar la predicción del agregado de Y para el ejercicio siguiente t, supondremos que a lo largo de los dos años, la composición del citado universo no sufre alteraciones, de lo contrario se procedería a restringir lo que sigue al colectivo de entidades con actividad en ambos ejercicios. Denotemos por $(Y_{1t},Y_{2t},\ldots,Y_{Nt})$ el vector aleatorio asociado con los valores que adopta la variable Y en el ejercicio t.

La experiencia acerca de la forma en que las entidades suelen elaborar sus presupuestos anuales, así como la propia dinámica de la actividad económica sugieren que las variables Y_{it} verifican las siguientes hipótesis:

$$H_1: E[Y_{it}] = \beta_t y_{i(t-1)} \qquad \text{para } i=1,\ldots,N$$

$$H_2: V[Y_{it}] = \nu_{it} \qquad \text{para } i=1,\ldots,N \text{ siendo } \nu_{it}=\sigma_t^2 y_{i(t-1)}$$

$$H_3: C[Y_{it},Y_{jt}] = 0 \qquad \text{para } i\neq j$$

De manera simplificada, el modelo propuesto, ampliamente estudiado en sus aspectos teóricos en la literatura, puede expresarse mediante la ecuación $Y_{it}=\beta_t y_{i(t-1)}+\varepsilon_i$ para $i=1,\ldots,N$, siendo ε_i perturbaciones aleatorias independientes de media 0 y varianzas respectivas ν_{it}.

Supongamos que se pretende estimar el agregado $Y_t=\sum_{i=1}^{N} Y_{it}$ de los valores que adopta la variable Y sobre todas las entidades en el ejercicio t. Para ello, admitiremos que además de la información censal recogida en el ejercicio $(t-1)$, se dispone de los datos del ejercicio t proporcionados por una muestra s formada por n entidades, a los que denotamos por $(y_{1t},y_{2t},\ldots,y_{nt})$.

La búsqueda de un estimador $\hat{Y}_t$ lineal, insesgado con respecto al modelo –esto es, verificando la condición $E\left[\hat{Y}_t - Y_t\right] = 0$– y con error cuadrático medio $E\left[\left(\hat{Y}_t - Y_t\right)^2\right]$ mínimo, conduce a la expresión:

$$\hat{Y}_t = \sum_s y_{it} + \sum_{s'} y_{i(t-1)} \left(\frac{\sum_s \frac{y_{i(t-1)} y_{it}}{v_{it}}}{\sum_s \frac{y_{i(t-1)}^2}{v_{it}}} \right)$$

donde el primero de los sumatorios se extiende sobre las unidades seleccionadas en la muestra s y el segundo sobre las no seleccionadas s´.

Algunos autores proponen buscar un estimador lineal con mínimo error cuadrático medio, asumiendo para el mismo la expresión $\sum_s y_{it} + \hat{\beta}_t \sum_{s'} y_{i(t-1)}$, donde $\hat{\beta}_t$ es un estimador insesgado para β_t. Se comprueba que el estimador óptimo así obtenido, coincide con el anteriormente propuesto $\hat{Y}_t$. Además, se comprueba que el estimador $\hat{\beta}_t$ que minimiza el error cuadrático medio del citado estimador, coincide con el que se obtiene al aplicar el método de mínimos cuadrados ponderados en base al modelo, esto es, el que minimiza $\sum_s \frac{\left(y_{it} - \beta_t y_{i(t-1)}\right)^2}{v_{it}}$.

Considerando que $v_{it} = \sigma_t^2 y_{i(t-1)}$ el estimador óptimo de β_t se reduce al cociente $\hat{\beta}_t = \frac{\sum_s y_{it}}{\sum_s y_{i(t-1)}}$ y el estimador del total poblacional es del tipo razón $\hat{Y}_t = \frac{\sum_s y_{it}}{\sum_s y_{i(t-1)}} \sum_U y_{i(t-1)}$.

El error cuadrático medio de este estimador bajo el modelo propuesto, adopta la expresión $E\left[\left(\hat{Y}_t - Y_t\right)^2\right] = \sigma_t^2 \frac{\sum_U y_{i(t-1)} \sum_{s'} y_{i(t-1)}}{\sum_s y_{i(t-1)}}$ en la que σ_t^2 debe ser estimado a

partir de los datos, siendo $\hat{\sigma}_t^2 = \frac{1}{n-1}\sum_s \frac{\left(y_{it} - \hat{\beta}_t y_{i(t-1)}\right)^2}{y_{i(t-1)}}$ un estimador insesgado del mismo.

Observar que el cálculo del estimador óptimo y su error se ha efectuado exclusivamente desde el soporte que ofrece el modelo de superpoblación, sin necesidad de hacer ninguna referencia al procedimiento de selección de la muestra. Bajo esta perspectiva, es pues posible plantearse la determinación de un diseño muestral para el que se minimice el error cuadrático medio del estimador.

Asumiendo que la variable económica Y toma valores positivos, la solución a este problema consiste en un diseño intencionado, que con probabilidad uno selecciona aquellas unidades del universo sobre las que la variable Y en el ejercicio $(t-1)$ alcanzó los valores más altos. En esencia, se trata de que el agregado $\sum_s y_{i(t-1)}$ sea lo más grande posible frente al agregado $\sum_{s'} y_{i(t-1)}$. Es conveniente hacer notar que cualquier desviación sobre esta condición óptima, se traducirá en un incremento sobre el error cuadrático del estimador, aunque la magnitud del mismo no tiene porque ser excesivamente alta si se sustituyen algunas unidades por otras sobre las que la variable Y también adopta valores elevados.

3. EVALUACIÓN EMPÍRICA DE LA ESTRATEGIA EN EL UNIVERSO COOPERATIVO VALENCIANO

El punto de partida para el análisis que sigue es la información censal de los datos económicos de las entidades cooperativas agrarias valencianas referidos a los ejercicios comprendidos entre 1996 y 2000. Las variables sobre las que se aplicará la metodología anteriormente descrita son: inmovilizado, activo, cifra de negocio y gasto de personal.

El objetivo es el diseño de una estrategia para estimar los valores agregados de las variables citadas, inicialmente en el ejercicio 2001 y posteriormente en ejercicios sucesivos. Sin embargo, con el fin de evaluar la fiabilidad de la propuesta y los resultados que con ella se generen, se ha procedido a efectuar su aplicación de manera experimental con carácter retroactivo a los ejercicios 1997, 1998, 1999 y 2000. Para cada uno de estos cuatro ejercicios se calcularán las estimaciones de los correspondientes agregados, partiendo para ello de los datos restringidos a una muestra y la información censal completa del ejercicio precedente. Posteriormente, los valores estimados para cada ejercicio serán comparados con los verdaderos valores asociados al censo completo del mismo, obteniendo de esta forma una imagen de la fiabilidad real de la estrategia.

Existen dos razones que desaconsejan, en este caso, proceder a la selección de una muestra aleatoria cada año. En primer lugar, el reducido tamaño de las muestras –se sugirió una cifra entorno a 20 entidades– lo que se traduciría en niveles excesivamente bajos para la precisión de los resultados. En segundo lugar, la necesidad de controlar la identidad de las cooperativas seleccionadas, asegurando la colaboración de su gerencia aportando los datos económicos en plazos razonablemente cortos.

Todo ello, sugiere el interés por plantear la estrategia en el contexto de un modelo de superpoblación. Para cada par de ejercicios y cada variable se ha asumido un modelo como el propuesto en el apartado 2, por lo que el estimador a utilizar en todos los casos será del tipo razón. No obstante, antes de seleccionar la muestra con el criterio óptimo ya descrito, conviene efectuar alguna reflexión.

De acuerdo con la argumentación anterior, el error cuadrático medio de cada estimador sería mínimo si para cada año y variable se eligieran las entidades que en el ejercicio anterior hubieran presentado los valores más altos. Sin embargo, en la práctica sería más interesante mantener la misma muestra para todas las variables durante varios años consecutivos, puesto que de esta forma estaría asegurada la colaboración de las entidades seleccionadas. Es obvio que ambas posturas han de compatibilizarse.

La solución finalmente adoptada ha consistido en seleccionar 20 cooperativas que en el ejercicio 2000 presentaron un valor elevado de la cifra de negocio, ajustando la selección para asegurar que los valores correspondientes a las variable indicadas en los ejercicios anteriores seguían estando entre los más elevados.

Para facilitar la aplicación retrospectiva de la estrategia a los años para los que se dispone de información, se admite que las entidades que presentan actividad cada año son las mismas que ya presentaban actividad en el precedente. Lógicamente, esto equivale a afirmar que las predicciones se efectuarán siempre sobre el supuesto de que no existen variaciones en el directorio de cooperativas de dos años consecutivos. En la práctica, en la medida en que se disponga de información específica sobre este particular, podrán introducirse las correcciones oportunas en los valores estimados.

Asumiendo la aproximación a la Normal, el error de estimación máximo relativo del agregado de cada variable y ejercicio Y_t, para una confianza del 95% se ha calculado a través de la expresión $\pm \dfrac{1{,}96\sqrt{\hat{E}\left[\left(\hat{Y}_t - Y_t\right)^2\right]}}{\hat{Y}_t}$. Este es el error que podemos considerar teórico, asociado con el modelo asumido. Además, en cada caso, partiendo de la información censal completa se ha calculado el verdadero

valor del agregado Y_t, por lo que el error real relativo de cada estimación puede medirse por $\frac{\hat{Y}_t - Y_t}{Y_t}$.

En el Cuadro 1 se recogen ambos errores (en porcentaje) para los cuatro ejercicios y las cuatro variables. En cada caso, se ha añadido una medida de la bondad del ajuste de los datos al respectivo modelo utilizando para ello la expresión $1 - \frac{\sum_s \left(y_{it} - \hat{\beta}_t y_{i(t-1)}\right)^2}{\sum_s (y_{it} - \bar{y}_t)^2}$, cuyos valores están siempre comprendidos en el intervalo $]-\infty, 1]$.

Los resultados recogidos en el Cuadro 1 permiten observar que en la mayor parte de los casos el error teórico se mantiene por debajo del 5%, mientras que el error real alcanza su máximo en 4,22%. Además, el error real casi siempre respeta la cota máxima que establece el error teórico, lo que puede interpretarse como una ratificación de la adecuación del modelo a los datos.

Cuadro 1

ESTIMACIÓN CON 20 COOPERATIVAS GRANDES

	Inmovilizado	*Activo*	*Cifra*	*Personal*
		Año 2000		
Predicción	38.844.412	155.450.171	133.904.579	23.906.715
Error teórico	3,83	2,81	3,39	3,41
Error real	-0,32	-1,58	-4,22	-0,91
Bondad ajuste	0,95	0,98	0,98	0,98
		Año 1999		
Predicción	54.145.201	186.532.578	173.958.468	29.383.003
Error teórico	5,73	8,19	4,32	4,32
Error real	-2,12	3,22	3,38	1,77
Bondad ajuste	0,91	0,84	0,98	0,97
		Año 1998		
Predicción	48.917.153	194.767.654	187.789.950	34.322.468
Error teórico	7,99	12,19	4,31	4,71
Error real	-2,91	-1,49	-2,55	-3,63
Bondad ajuste	0,83	0,50	0,96	0,96
		Año 1997		
Predicción	41.550.925	199.992.890	170.400.419	24.798.865
Error teórico	8,21	9,61	3,76	2,78
Error real	-0,11	0,51	3,34	3,61
Bondad ajuste	0,87	0,58	0,99	0,99

Es interesante destacar que los casos en que se detectan los niveles más bajos para la precisión de las estimaciones, coinciden con los que ofrecen un menor ajuste de los datos al modelo. No obstante, en estos mismos casos el error real observado continua adoptando valores aceptables.

Para comprobar la posible influencia de las unidades seleccionadas en la precisión de las estimaciones, en el Cuadro 2 se han calculado los resultados correspondientes a una muestra formada por 20 entidades con valores intermedios de la cifra de negocio, observándose que la selección de entidades con características que se alejan de las que debe verificar la muestra óptima, conlleva un elevado incremento tanto del error teórico como del real.

Cuadro 2

ESTIMACIÓN CON 20 COOPERATIVAS MEDIANAS

	Inmovilizado	*Activo*	*Cifra*	*Personal*
		Año 2000		
Predicción	38.063.921	171.224.665	134.305.913	19.648.679
Error teórico	31,93	13,03	18,71	10,51
Error real	1,69	-11,89	-4,54	17,07
Bondad ajuste	0,95	0,80	-11,01	0,87
		Año 1999		
Predicción	58.845.685	129.913.099	185.650.547	38.450.382
Error teórico	23,74	36,25	14,97	11,49
Error real	-10,99	32,60	-3,12	-28,55
Bondad ajuste	0,77	-4,40	0,03	0,91
		Año 1998		
Predicción	40.991.630	177.761.412	182.377.841	27.350.597
Error teórico	3,30	12,55	15,86	11,24
Error real	13,76	7,37	0,40	17,42
Bondad ajuste	1,00	0,91	-0,06	0,92
		Año 1997		
Predicción	44.312.420	211.989.362	173.581.007	29.211.653
Error teórico	11,45	10,26	21,03	18,34
Error real	-6,76	-5,45	1,54	-13,54
Bondad ajuste	0,86	0,99	-0,05	0,83

Puesto que la precisión de las estimaciones está ligada tanto al tamaño como a las características de la muestra seleccionada, no es posible prever las consecuencias que sobre la fiabilidad de los resultados induciría un incremento del tamaño

muestral. Observar que para aumentar el número de entidades en la muestra es necesario recurrir a valores de la variable auxiliar cada vez más bajos, lo que dificultaría cualquier intento por conocer de antemano la disminución del error que esto conllevará. Para ilustrar la variación de la precisión con el tamaño de la muestra en el universo de las cooperativas, en el Cuadro 3 se han calculado las estimaciones asociadas con una muestra formada por 40 entidades con valor elevado de la cifra de negocio.

Cuadro 3

ESTIMACIÓN CON 40 COOPERATIVAS GRANDES

	Inmovilizado	*Activo*	*Cifra*	*Personal*
		Año 2000		
Predicción	39.112.718	154.500.893	132.272.625	23.813.254
Error teórico	3,44	2,00	2,66	2,04
Error real	-1,02	-0,96	-2,95	-0,51
Bondad ajuste	0,95	0,98	0,97	0,99
		Año 1999		
Predicción	52.205.951	188.879.068	173.268.243	28.937.967
Error teórico	4,55	4,86	4,17	4,63
Error real	1,53	2,00	3,76	3,26
Bondad ajuste	0,94	0,90	0,97	0,95
		Año 1998		
Predicción	47.864.320	183.918.047	187.449.535	33.637.581
Error teórico	10,59	8,92	3,37	3,12
Error real	-0,69	4,16	-2,37	-1,56
Bondad ajuste	0,87	0,60	0,96	0,97
		Año 1997		
Predicción	41.204.465	203.721.193	169.803.625	24.801.452
Error teórico	5,04	5,35	2,80	2,70
Error real	0,73	-1,34	3,68	3,60
Bondad ajuste	0,92	0,77	0,99	0,99

Los errores contenidos en el Cuadro 3 en general, son menores que los que recoge el Cuadro 1. La disminución media del error teórico de las estimaciones se sitúa entorno al 20%, valor sensiblemente inferior al 30% que cabría esperar de haber tomado muestras aleatorias. Curiosamente, en algunos casos, el error llega a incrementarse con el tamaño de la muestra, circunstancia atribuible a variaciones en la estimación del parámetro σ_t^2.

4. ANÁLISIS DE SENSIBILIDAD Y ROBUSTEZ

4.1 Sensibilidad y valores anómalos

La estrategia inferencial desarrollada anteriormente se apoya exclusivamente en las hipótesis que especifican el modelo de superpoblación, en consecuencia el grado de adecuación del modelo a la realidad de los datos tendrá una incidencia fundamental en la fiabilidad de las predicciones. La propia naturaleza de los datos y el análisis de los resultados que se recoge en los cuadros anteriores, nos inducen a concentrar las críticas en dos direcciones distintas: la validez de la segunda hipótesis acerca de la heteroscedasticidad del modelo y la posible existencia de datos anómalos o "outliers".

Para contrastar la heteroscedasticidad se ha recurrido al test de White, aplicándolo a todas las muestras que se consideran en el estudio. Los resultados obtenidos en general no inducen a rechazar la hipótesis propuesta, por lo que procede aceptar la validez del modelo.

Aceptada pues la idoneidad de las hipótesis, las posibles causas que justificarían la falta de precisión de algunas de las estimaciones obtenidas en el Cuadro 1 pueden ser: un valor elevado de los parámetros σ_t^2, o bien, la existencia de datos anómalos o 'butliers". A continuación centraremos el análisis en la detección y tratamiento de estos últimos.

Utilizando la terminología de Chambers (1986) pueden distinguirse dos tipos de datos anómalos, los representativos y los no representativos. Los primeros, están asociados con estructuras poblacionales complejas cuyo modelo responde a la integración de al menos dos formulaciones diferentes, una que genera la mayor parte de los datos y otra que genera valores extremos sistemáticamente separados de la mayoría. Los segundos, son datos excepcionales que pueden aparecer de manera muy puntual en la población, en ocasiones atribuidos a errores de transcripción o de registro, y que por su naturaleza merecen ser analizados separadamente del resto.

No es fácil reconocer en la práctica cuándo estamos frente a uno u otro tipo de dato anómalo, ni siquiera cuándo debemos juzgar un dato como anómalo. Observar que bajo la estrategia de estimación propuesta, el carácter anómalo de un dato debe referirse al valor del ratio entre las observaciones de la variable de interés y la auxiliar. Si en la muestra se detecta algún valor extremo aislado del resto, cabe admitir con reservas que se trata de un dato anómalo no representativo, posiblemente único en toda la población. Si por el contrario, se detecta un número significativo de datos anómalos con similares características, puede interpretarse que

éstos son reflejo de un colectivo más amplio existente en la población y la perspectiva de análisis debe ser diferente.

Aplicando propuestas generalistas al contexto de las poblaciones finitas, el análisis inferencial en presencia de datos anómalos, pasa por la construcción de estimadores robustos que proporcionen estimaciones poco sensibles a la aparición de valores extremos. Aunque la construcción de estos estimadores no siempre exige una especificación previa de los datos que se consideran anómalos, en este trabajo se ha optado por proceder a su identificación en algunas de las poblaciones analizadas.

Siguiendo las directrices de Prescott (1975), el test que se propone para la detección de datos anómalos es de tipo secuencial y se apoya en el estadístico $\max_i \left| \frac{y_{it} - \hat{\beta}_t y_{i(t-1)}}{s_{it}} \right|$ donde s_{it}^2 es una estimación insesgada de $V\left[y_{it} - \hat{\beta}_t y_{i(t-1)} \right]$. Después de calcular esta última varianza sobre nuestro modelo, se propone la siguiente expresión para el denominador:

$$s_{it}^2 = \frac{y_{i(t-1)}}{n-1} \left(1 - \frac{y_{i(t-1)}}{\sum_s y_{i(t-1)}} \right) \sum_s \frac{\left(y_{it} - \hat{\beta}_t y_{i(t-1)} \right)^2}{y_{i(t-1)}}$$

Fijado un nivel de significación, el test determina la existencia de un dato anómalo cuando el valor del estadístico indicado supera una cota convenientemente calculada a partir de su distribución, localizándolo en la misma unidad sobre la que el citado estadístico alcanza su valor. La distribución del estadístico ha sido estudiada de forma aproximada por Barnett y Lewis (1980), apoyándose en las desigualdades de Bonferroni y estudios de simulación.

Para su aplicación en este trabajo, se utilizan las cotas de admisibilidad elaboradas por Lund (1975) para distintos tamaños muestrales y un nivel de significación del 5%. No obstante, puesto que las tablas de Lund no proporcionan valores críticos para tamaños muestrales superiores a 100, la determinación de datos anómalos sobre las poblaciones censales se ha realizado recurriendo a la aproximación $\sqrt{(N-1)F/(N-2+F)}$ para las citadas cotas, donde F es el valor de una distribución F de Snedecor con 1 y $N-2$ grados de libertad que acumula a su izquierda una probabilidad igual a $1 - \frac{0,05}{N}$.

La aplicación del test se efectúa en pasos sucesivos, cuando se detecta un primer dato anómalo, de manera secuencial debe proseguirse con la búsqueda de otros nuevos posibles datos anómalos. Para ello, se elimina sobre el colectivo

correspondiente el dato ya detectado y de nuevo se estima el parámetro β_t. A continuación, se calcula el valor del estadístico que define el test sobre la muestra reducida –sin los datos anómalos ya detectados– y de nuevo vuelve a aplicarse el test, repitiéndose el proceso mientras siga detectándose nuevos valores extremos anómalos.

En el Cuadro 4 se indica el número de datos anómalos obtenidos con un nivel del 5% para los seis casos en que se dispone de menor precisión en las estimaciones. En primer lugar se indican los datos anómalos detectados al aplicar el test sobre cada muestra seleccionada y a continuación, los detectados al aplicar el test sobre la población censal correspondiente. Completando la información anterior, en el Cuadro 4 también se incluye para cada caso el número de datos anómalos detectados en la población que forman parte de la muestra correspondiente.

Cuadro 4

NÚMERO DE VALORES ANÓMALOS DETECTADOS

Año		*Inmovilizado*	*Activo*
1999	Detectados muestra	0	1 (5%)
	Detectados población	51 (10,2%)	25 (5%)
	Incluidos muestra	6 (30%)	2 (10%)
1998	Detectados muestra	0	2 (10%)
	Detectados población	40 (8,1%)	26 (5,3%)
	Incluidos muestra	6 (30%)	2 (10%)
1997	Detectados muestra	1 (5%)	1 (5%)
	Detectados población	42 (8,6%)	21 (4,3%)
	Incluidos muestra	4 (20%)	2 (10%)

Los resultados que recoge el Cuadro 4 evidencian cierta ambigüedad en la determinación de datos anómalos. En todas las muestras se observa que existen datos de los considerados anómalos en la población, que sin embargo el test no los clasifica como tales en la muestra. Esto es especialmente llamativo en las poblaciones asociadas con la variable inmovilizado.

La causa de esta ambigüedad hay que buscarla en la naturaleza relativa del término, que cuestiona la aplicabilidad del test en el contexto finito. Un dato se considera anómalo si comparado con el resto del colectivo, presenta un valor extremo para el ratio. En consecuencia, si en la muestra coincide un número significativo de datos extremos, es posible que frente a este colectivo más pequeño y

heterogéneo algunos datos considerados como anómalos en la población, el test no los identifique como tales en la muestra.

La conclusión que se deduce de estas apreciaciones es que en poblaciones **f**-nitas, la detección de datos anómalos debe interpretarse únicamente en sentido orientativo y por lo tanto, para la inferencia con datos anómalos preferentemente debe optarse por soluciones robustas que no exijan la identificación de los mismos.

Otro hecho destacable en los resultados recogidos en el Cuadro 4 es la sobre-rrepresentación de datos anómalos en la muestra con respecto a la población. En principio, los resultados sugieren que los datos anómalos tienden a concentrarse en las entidades grandes, algo que iría en contra de la recomendación efectuada de seleccionar este tipo de entidades para formar la muestra. Sin embargo, debido al reducido tamaño de la muestra y el escaso número de datos anómalos que se detectan, este resultado no es estadísticamente significativo.

4.2 Estimadores robustos

Una primera opción para dotar de robustez a las estimaciones construidas a partir de muestras que contienen datos anómalos consiste en modificar el peso originalmente asignado a las unidades seleccionadas. Su aplicación es especialmente interesante cuando los citados datos son no representativos.

Una vez determinados los datos anómalos existentes en la muestra, se procede a eliminarlos, calculando con las unidades restantes una estimación del agregado poblacional. Posteriormente, al resultado obtenido se incorporan de manera separada los valores asociados con los datos anómalos que previamente se habían suprimido. Es fácil comprobar que el estimador final resultante, o estimador corregido, es equivalente al que determina la expresión $\dot{Y}_t = \sum_s y_{it} + \dot{\beta}_t \sum_{s'} y_{i(t-1)}$ donde $\dot{\beta}_t$ es el estimador de razón restringido a la muestra sin los datos anómalos. Los errores relativos reales asociados a las estimaciones obtenidas por este procedimiento en las distintas poblaciones se recogen en el Cuadro 5.

La segunda opción para la construcción de estimadores robustos se basa en la corrección o suavizado de los datos muestrales considerados extremos. Su aplicación está más indicada en el caso de valores anómalos representativos y sobre la primera opción tiene la ventaja de no requerir la determinación de los mismos.

La alternativa robusta más común al estimador óptimo $\hat{Y}_t$ se obtiene utilizando los M-estimadores $\ddot{\beta}_t$, introducidos por Huber (1981), para los parámetros β_t. Estos estimadores se obtienen como solución a la ecuación

$\sum_s \frac{y_{i(t-1)}}{v_{i(t-1)}} \phi\left(\frac{y_{it} - \beta_t y_{i(t-1)}}{\sqrt{v_{it}}}\right) = 0$, en la que como antes v_{it} denota la varianza de Y_{it} y ϕ es una de las funciones de influencia referenciadas por Hampel et al (1986). En este trabajo se han ensayado tres funciones de influencia diferentes:

– función bicuadrada: $\phi(r) = \begin{cases} r\left(1 - \frac{r^2}{c_1^2}\right)^2 & \text{si } |r| \le c_1 \\ 0 & \text{si } |r| > c_1 \end{cases}$

– función de Huber: $\phi(r) = \begin{cases} r & \text{si } |r| \le c_2 \\ c_2 \times \text{sgn}(r) & \text{si } |r| > c_2 \end{cases}$

– función de Hampel: $\phi(r) = \text{sgn}(r) \begin{cases} |r| & \text{si } 0 \le |r| < a \\ a & \text{si } a \le |r| < b \\ a(c - |r|)/(c - b) & \text{si } b \le |r| < c \\ 0 & \text{si } c \le |r| \end{cases}$

donde los valores de las constantes c_1, c_2, a, b y c se han elegido de acuerdo con propuestas fundamentadas en argumentos de tipo empírico, que delimitan la proporción de datos a suavizar en la muestra.

El cálculo de los estimadores $\ddot{\beta}_t$ se efectúa a través de algoritmos numéricos. En particular, en este trabajo se ha recurrido al paquete informático S-plus, y como c_1, c_2, a, b y c se han tomado los valores estándar que utiliza el propio programa. La predicción de los agregados poblacionales se calcula a través de la expresión $\ddot{Y}_t = \sum_s y_{it} + \ddot{\beta}_t \sum_{s'} y_{i(t-1)}$.

En el Cuadro 5 se recogen los errores relativos reales asociados a las distintas estimaciones obtenidas para las tres propuestas de funciones de influencia y las distintas poblaciones de cooperativas.

La observación de los errores reales asociados con las distintas estimaciones robustas ofrece una imagen de la fiabilidad del resultado. Sin embargo, para poder establecer conclusiones acerca de su precisión, es necesario evaluar la eficiencia de los distintos estimadores.

Cuadro 5

ERRORES REALES DE LAS ESTIMACIONES ROBUSTAS

	Inmovilizado	*Activo*	*Cifra*	*Personal*
		Año 2000		
Razón	-0,32	-1,58	-4,22	-0,91
Corregida	-0,32	0,09	-4,22	-0,91
HUBER	-0,15	-0,95	-6,00	-1,35
HAMPER	-0,39	-0,63	-5,74	-1,92
BICUADRADA	0,25	-1,05	-5,95	0,06
		Año 1999		
Razón	-2,12	3,22	3,38	1,77
Corregida	-2,12	-2,24	4,55	1,77
HUBER	0,59	1,15	3,45	2,05
HAMPER	-0,10	-0,33	3,38	2,09
BICUADRADA	0,60	1,13	3,48	1,51
		Año 1998		
Razón	-2,91	-1,50	-2,56	-3,63
Corregida	-2,91	-15,52	-2,56	-3,63
HUBER	-2,41	-7,99	-2,18	-3,18
HAMPER	-4,17	-9,75	-2,42	-3,67
BICUADRADA	-2,84	-9,67	-2,00	-3,37
		Año 1997		
Razón	-0,11	0,51	3,35	3,61
Corregida	2,10	-8,44	4,28	3,61
HUBER	2,29	-4,19	4,27	2,93
HAMPER	1,27	-5,23	4,13	3,12
BICUADRADA	2,49	-4,82	4,20	2,49

4.3 Análisis de la eficiencia de los estimadores robustos

Ante la dificultad que supone calcular el error cuadrático medio de los estimadores propuestos sobre bases teóricas, se ha optado por analizar la eficiencia recurriendo a procedimientos empíricos de replicación. Puesto que la estrategia inferencial propuesta se apoya en muestras intencionadas no aleatorias, no es posible en este caso proceder a la replicación de las muestras. En su lugar, se han efectuado 500 replicaciones de cada una de las seis poblaciones con menor precisión del estimador óptimo, utilizando un procedimiento que consta de los siguientes pasos:

1º. Después de eliminar en cada una de las poblaciones censales de partida todos los datos anómalos detectados, se procede a estimar los parámetros que especifican el correspondiente modelo.

2º. Con base en el modelo estimado y manteniendo siempre los mismos valores originales para la variable auxiliar, se generan 500 poblaciones, todas ellas con el mismo tamaño que el de la población utilizada para estimar el correspondiente modelo.

3º. A cada una de las poblaciones así generadas, se le incorporan los datos anómalos previamente eliminados en la población origen. Con ello se asegura que todas las poblaciones replicadas con una misma referencia, poseerán el mismo tamaño que ésta e idéntico nivel de contaminación por datos anómalos.

4º. En todos los casos se mantendrá la misma muestra inicialmente seleccionada, formada por 20 cooperativas grandes. Esto significa que los valores muestrales de la variable auxiliar serán siempre los mismos en las 500 replicaciones de una misma población origen y que además, todas las muestras poseerán los mismos datos anómalos.

En esencia, las poblaciones replicadas y sus respectivas muestras se han construido respetando el modelo generador de los datos considerados normales, asegurando en cada caso el mismo nivel de contaminación detectado en la población original.

El sesgo de cada estimador $\ddot{Y}_{jt}$ puede aproximarse por $\frac{1}{500}\sum_{j=1}^{500}\left(\ddot{Y}_{jt} - Y_{jt}\right)$,

donde Y_{jt} es el agregado de la replicación j de la población asociada con el ejercicio t para cada variable Y. El error cuadrático medio puede aproximarse por

$\frac{1}{500}\sum_{j=1}^{500}\left(\ddot{Y}_{jt} - Y_{jt}\right)^2$ y la eficiencia se ha calculado a través de la raíz del cociente

entre el error cuadrático medio de cada estimador y el del estimador óptimo sin correcciones. La comparación entre las distintas alternativas puede efectuarse a través de los resultados recogidos en el Cuadro 6.

Del análisis de la eficiencia relativa se desprende que en términos generales, salvo en la predicción del activo para 1999 que a continuación se comenta, el estimador de razón es preferible a cualquiera de los estimadores robustos que se proponen. Únicamente el estimador de Huber para el inmovilizado de 1998 tiene menor error cuadrático, pero la disminución es realmente poco significativa.

Cuadro 6

EFICIENCIA DE LOS ESTIMADORES ROBUSTOS

	Razón	*Corrección*	*Huber*	*Hampel*	*Bisquare*
			Inmovilizado 1999		
Sesgo	407.271	-2.585.492	-1.431.550	-864.106	-1.983.136
E.C.M.	3,43E+11	6,95E+12	2,46E+12	2,11E+12	4,61E+12
Eficiencia relativa	100	450	268	248	367
			Activo 1999		
Sesgo	-7.851.652	1.092.275	-47.009	763.949	1.013.524
E.C.M.	6,53E+13	4,84E+12	4,31E+12	4,99E+12	5,80E+12
Eficiencia relativa	100	27	26	28	30
			Inmovilizado 1998		
Sesgo	832.425	-2.074.701	-218.374	1.224.595	-1.553.547
E.C.M.	9,147E+11	4,72E+12	7,11E+11	2,27E+12	3,59E+12
Eficiencia relativa	100	227	88	158	198
			Activo 1998		
Sesgo	-3595646	14144218	11858967	14122345	14095466
E.C.M.	1,66E+13	2,05E+14	1,47E+14	2,05E+14	2,05E+14
Eficiencia relativa	100	351	297	351	351
			Inmovilizado 1997		
Sesgo	626.511	-1.313.307	-803.281	-696.451	900.886
E.C.M.	5,34E+11	1,91E+12	8,60E+11	7,62E+11	1,67E+12
Eficiencia relativa	100	189	127	119	177
			Activo 1997		
Sesgo	-4.561.150	5.619.190	5.243.961	6.288.602	5.660.966
E.C.M.	2,30E+13	3,41E+13	3,07E+13	4,32E+13	3,58E+13
Eficiencia relativa	100	122	116	137	125

Estudiando con detalle la población de la variable activo en el bienio 98/99, se observa la presencia de un dato anómalo con un ratio interanual muy alejado del resto. Este dato singular con un alto poder de contaminación, también forma parte de la muestra y su localización puede apreciarse en las Figuras 1a y 1b. El estimador de razón pondera este dato anómalo con el mismo factor de elevación que los demás datos, produciendo así una distorsión en la estimación correspondiente. Ésta podría ser la causa de que en este caso específico, cualquiera de los estimadores robustos se comporte mejor que el de razón.

De este comentario podría pues concluirse que, únicamente son preferibles los estimadores robustos en un caso, aquél en el que sobre la muestra se detecta algún dato anómalo no representativo, totalmente alejado no sólo de los datos normales sino incluso del resto de los que pueden considerarse anómalos.

Figura 1 a
POBLACIÓN ACTIVO 98/99

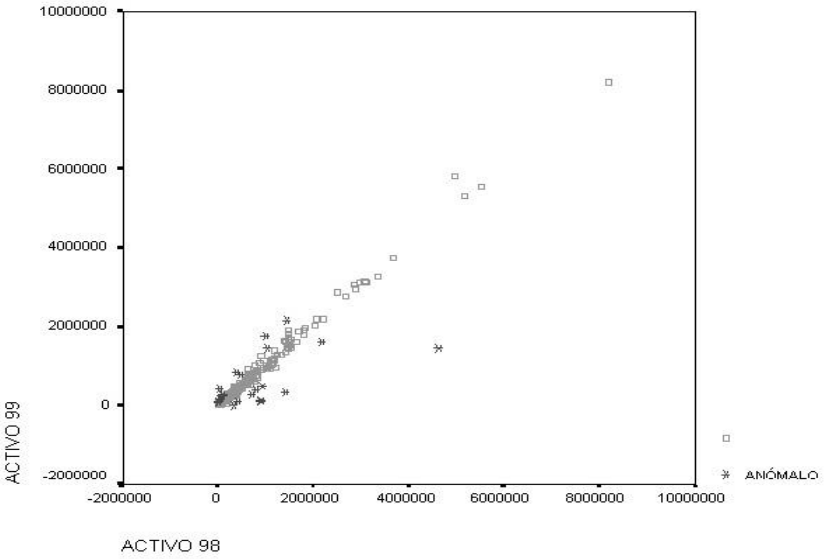
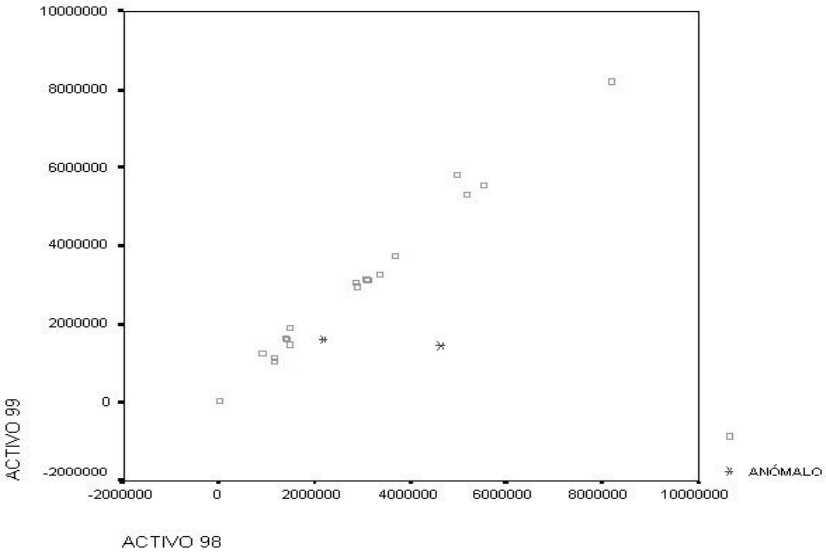


Figura 1 b
MUESTRA ACTIVO 98/99



4.4 Un modelo de contaminación

Todos los resultados anteriores se han obtenido asumiendo que las poblaciones están formadas fundamentalmente por valores que se ajustan a un modelo, más un conjunto de datos contaminantes sobre el que no se plantea hipótesis alguna. Una forma de ilustrar teóricamente la eficiencia de los estimadores robustos propuestos puede conseguirse si se admite que los datos anómalos son generados por uno o varios modelos alternativos al que genera la población principal. Puesto que en general, los valores atípicos del ratio interanual pueden aparecer tanto en el extremo superior como en el inferior, una propuesta razonable sería admitir que una parte de los datos anómalos se ajustan a un modelo lineal alternativo al de la población principal y el resto a otro con las mismas características pero diferentes parámetros. Recurriendo a una formulación global, el modelo para la población asociada a la variable Y en el ejercicio t podría plantearse en la siguiente forma:

$$Y_{it} = \Delta_{1it}\left(\mu_{1it} + \sqrt{\nu_{1it}}\,\varepsilon_{1it}\right) + \Delta_{2it}\left(\mu_{2it} + \sqrt{\nu_{2it}}\,\varepsilon_{2it}\right) + \Delta_{3it}\left(\mu_{3it} + \sqrt{\nu_{3it}}\,\varepsilon_{3it}\right)$$

donde cada Δ_{jit} tomará exclusivamente los valores 1 ó 0, verificándose $\Delta_{1it} + \Delta_{2it} + \Delta_{3it} = 1$. Además se admite que $P(\Delta_{jit} = 1) = \theta_{jt}$ para cada i y $j = 1, 2, 3$, siendo $\mu_{jit} = \beta_{jt}\, y_{i(t-1)}$ y $\nu_{jit} = \sigma^2_{jt}\, y_{i(t-1)}$ para cada i y $j = 1, 2, 3$.

Un estimador robusto para el total poblacional adopta la expresión $\tilde{Y}_t = \sum_s y_{it} + \tilde{\beta}_t \sum_{s'} y_{i(t-1)}$, donde el estimador $\tilde{\beta}_t$ se calcula por cualquiera de los procedimientos descritos, siempre buscando la condición $E\left[\tilde{\beta}_t\right] = \beta_{1t}$.

Desde esta perspectiva, se comprueba que el sesgo del estimador robusto está determinado por $E\left[Y_t - \tilde{Y}_t\right] = \sum_{s'} y_{i(t-1)}\left[\theta_2(\beta_{2t} - \beta_{1t}) + \theta_3(\beta_{3t} - \beta_{1t})\right]$ y se verifica que una estimación insesgada del mismo es $\dfrac{\sum_{s'} y_{i(t-1)}}{\sum_s y_{i(t-1)}} \sum_s \left(y_{it} - \tilde{\beta}_t\, y_{i(t-1)}\right)$.

Si se procede a incorporar esta última expresión a la originalmente propuesta para el estimador robusto, se conseguiría corregir el carácter sesgado de este último. Sin embargo, es fácil comprobar que el estimador que finalmente se obtiene de esta forma coincide con el estimador óptimo o de razón asociado con todos los datos Y_t, por lo que perderá las garantías de robustez que se pretendía conseguir. En definitiva, bajo el modelo indicado puede proponerse un estimador que sea robusto frente a la aparición de datos anómalos, pero éste será sesgado y en

consecuencia no puede asegurarse que su eficiencia será inferior a la del estimador original. Algo perfectamente compatible con los resultados empíricos obtenidos por simulación en las poblaciones replicadas.

5. CONCLUSIONES

Los resultados obtenidos a lo largo del trabajo se apoyan en distintos escenarios, unos son de tipo teórico y se establecen bajo la propuesta de modelos específicos, otros son de naturaleza empírica y se basan en datos censales disponibles para ejercicios ya cerrados. Además, se han obtenido resultados por simulación de poblaciones y muestras. Tras su valoración, las principales conclusiones que se derivan son las siguientes:

– La estrategia propuesta para estimar los agregados poblacionales, a pesar del reducido tamaño muestral, presenta unos niveles de precisión aceptables en la mayor parte de los casos ensayados.

– En las aplicaciones retrospectivas se constata que en general, el error real de las predicciones es coherente con la precisión teórica prevista. Lo que de alguna forma evidencia la validez del modelo propuesto.

– Los niveles de precisión más bajos van acompañados de un menor ajuste de los datos muestrales al modelo, atribuyéndose esta circunstancia a la existencia de datos anómalos.

– Teniendo en cuenta la naturaleza relativa del concepto de dato anómalo, el test generalmente aplicado para detectar la existencia de tales datos, no es apropiado en el ámbito de las poblaciones finitas. Los mismos datos que se clasifican como anómalos de acuerdo con el test en el contexto global de una población, pueden ser considerados como ordinarios al aplicar el test en el entorno de un conjunto de datos muestrales.

– Como norma general y consecuentemente con las limitaciones inherentes al proceso de detección de datos anómalos, la construcción de estimadores robustos debe realizarse a través de funciones de influencia, suavizando los valores extremos sin necesidad de especificar si los mismos deben considerarse como anómalos o no.

– Los estimadores robustos habitualmente propuestos en la literatura, definidos para protegerse frente a la influencia negativa de los valores extremos, en general no se comportan mejor que el estimador óptimo del tipo razón construido con los datos muestrales originales, por lo que en base a su eficiencia su utilización no está justificada.

– Únicamente en un caso, en el que se detecta la existencia de un dato anómalo no representativo, con un valor del ratio interanual muy alejado del resto de valores, se ha comprobado un mejor comportamiento de las alternativas robustas frente a la utilización del estimador óptimo sobre los datos originales. Por lo tanto, en la práctica, antes de decidir si se efectúan correcciones en las estimaciones motivadas por la existencia de datos anómalos, es conveniente intentar averiguar la naturaleza de tales valores extremos, puesto que sólo en el caso en que fueran no representativos estaría indicada la corrección.

REFERENCIAS

BARNETT, V.; LEWIS, T. (1980): «Outliers in Statistical Data», New York: Wiley.

CHAMBERS, R.L. (1986): «Outlier Robust Finite Population Estimation», *Journal of the American Statistical Assosiation*, **81**, pp. 1063-1069.

CONSELLERIA D'AGRICULTURA, PEIXCA I ALIMENTACIÓ, (2002): «Cuentas Económicas Agregadas del Cooperativismo Agrario», Ed. Generalitat Valenciana.

HAMPEL F.R.; RONCHETTI, E.M.; ROUSSEEUW, P.J. and STAHEL, W.A. (1986): «Robust Statistics: The Approach Based on Influence Functions», New York: Wiley.

HUBER, P.J. (1981): «Robust Statistics», New York: Wiley.

LUND, R.E. (1975): «Tables for an Approximate Test for Outliers in Linear Models», *Technometrics*, **17**, nº 4, pp. 473-476.

PRESCOTT, P. (1975): «An Approximate Test for Outliers in Linear Models», *Technometrics*, **17**, nº 1, pp. 129-132.

VALLIANT, R; DORFMAN, A.H. and ROYAL, R.M. (2000): «Finite Population Sampling and Inference. A Prediction Approach», New York: Wiley.

DESIGN AND EVALUATION OF A SAMPLING PREDICTION STRATEGY IN AGRARIAN COOPERATIVE SOCIETIES

ABSTRACT

This paper proposes a strategy based on a superpopulation model in order to estimate the population aggregate totals of Valencian agrarian cooperative societies. We have census data for several economic exercises that allow us to carry out a sensitivity analysis of this strategy that focuses on the presence of outliers. The efficiency of robust alternative estimators is evaluated by using replication techniques.

Key words: superpopulation model, outlier, robust estimator

AMS classification: 62D05