ESTIMACIÓN EN ÁREAS PEQUEÑAS. APLICACIÓN A LA ESTIMACIÓN DE LA TASA DE PARO EN LAS PROVINCIAS ESPAÑOLAS

Antonio Vaamonde[1] y Ricardo Luaces
Departamento de Estadística y Desarrollo de Operaciones, Universidad de Vigo, España

RESUMEN
En este trabajo se desarrollan algunos métodos para estimación en áreas pequeñas, utilizando el enfoque predictivo del muestreo de poblaciones finitas. Se construye un modelo, empleando para ello el teorema de Royall sobre estimadores óptimos, con el objetivo principal de estimar la tasa de paro -un indicador socioeconómico que muestra la situación del mercado laboral- para las provincias españolas, a partir de las estimaciones obtenidas en las Comunidades Autónomas (las cuales agrupan varias provincias) mediante la Encuesta de Población Activa, utilizando además diferentes variables auxiliares obtenidas para las áreas pequeñas mediante registro o censo. También se discute la validación del modelo, un aspecto de especial importancia en las estimaciones de área pequeña. Los resultados obtenidos son mejores que los que se consiguen con el enfoque tradicional del muestreo basado en el diseño.

ABSTRACT
Some methods for small area estimation are developed, using predictive approach for finite populations sampling. A model is built, using Royall theorem for optimal estimators, with the aim of estimating unemployment rate (an economic indicator for the state of the labour market) for provinces using greater Autonomous Communities estimations, from Spanish Active Population Survey, and some information about different au xiliary variables obtained by means of public registration or census. Validation of the model is discussed. Results obtained are better than those with usual design based sampling approach.

Key words: Small area estimation, predictive approach, finite populations sampling, unemployment.

MSC: 62D05

1. INTRODUCCIÓN

Las autoridades nacionales, regionales y locales planifican y actúan con el objetivo de promover el pleno empleo. El paro es un problema que afecta a todos los niveles de la Administración Pública y en general las medidas políticas de carácter global no suelen ser satisfactorias para las entidades locales, que también pueden desarrollar sus propias estrategias de empleo. Para ello necesitan algunas herramientas que les permitan determinar -con precisión, fiabilidad y puntualidad aceptables- las principales variables e indicadores del mercado laboral: empleo, paro, población activa, tasa de empleo y paro, y la ocupación en función de sexo, edad y sector de actividad entre otros.

Las estimaciones de los indicadores de empleo no se obtienen a menudo fácilmente; aunque algunas variables demográficas pueden ser determinadas con precisión mediante registro, otras como la tasa de paro solo se pueden estimar mediante técnicas de muestreo, en general para grandes áreas como una nación o una región autónoma, y generalmente no alcanzan el nivel local (municipio o provincia), debido a las limitaciones en la metodología y el tamaño muestral.

En España, como en otros países europeos, la estimación se realiza mediante la Encuesta de Población Activa (EPA), que utiliza un diseño estratificado con el criterio principal del tamaño del municipio. La mayoría de los municipios no están representados en la muestra, y muchos de los que sí lo están tienen un tamaño muestral muy reducido, lo que hace que las estimaciones a nivel municipal tengan una precisión inaceptable.

El interés por desarrollar técnicas de estimación para áreas pequeñas que permitan resolver razonablemente estos problemas es creciente entre los investigadores de la Estadística. El término “área pequeña”, es utilizado frecuentemente para referirse a zonas geográficas, pero también puede ser aplicado a otros dominios de interés con límites no geográficos, como grupos de edad, sectores de actividad, etc. Es el tamaño muy reducido de la muestra en el dominio, y consecuentemente la gran varianza de los estimadores directos lo que define el área pequeña, y no el tamaño del área en sí mismo.

E-mail: [1]vaamonde@uvigo.es

El diseño de las encuestas por muestreo en general está sometido a limitaciones de tiempo y de presupuesto, lo que reduce las posibilidades de obtener estimaciones fiables en áreas pequeñas, utilizando las técnicas estándar de muestreo. Sin embargo, la creciente demanda de tales estimaciones ha conducido a notables esfuerzos para desarrollar nuevos métodos que permitan obtener los mejores resultados posibles de las muestras existentes. Tales métodos utilizan datos obtenidos en grandes dominios para construir estimaciones para los pequeños e incluyen técnicas como la estimación bayesiana o el enfoque predictivo.

Los primeros intentos de estimación en áreas pequeñas se encuentran ya en el texto clásico de Hansen, Hurwitz y Madow publicado por primera vez en 1953, que proponían para ello la utilización de estimadores de regresión; estos métodos fueron difundidos principalmente por Ericksen (1973). Platek **et al**. (1987) y Schaible (1996) resumen algunas técnicas y aplicaciones; Ghosh y Rao (1994) hacen una revisión de los diferentes métodos existentes, en un trabajo que es referencia obligada en este campo. Ericksen y Kadane (1985) y Ericksen, Kadane y Tukey (1989) aplican algunas técnicas de estimación de área pequeña para resolver la falta de cobertura de algunos grupos demográficos minoritarios en el censo USA.

El problema consiste en estimar el total de una variable Y, es decir la suma de los valores de la variable para todos los elementos, en un dominio o subpoblación especificado dentro de una población finita más grande. La muestra, en la población grande, puede haber sido seleccionada para cualquier propósito alejado de nuestro interés local, y puede contener pocas unidades (o ninguna) en el área pequeña. Utilizaremos el enfoque predictivo del muestreo de poblaciones finitas, basado en modelos, debido a que el enfoque tradicional basado en el diseño no ofrece soluciones satisfactorias para este problema.

El teorema general de predicción (Royall, 1976), que se muestra a continuación, proporciona el Estimador Lineal Insesgado Óptimo (ELIO):

Sea $Y = (Y_1 ,.. \ Y_N)$ un vector aleatorio con el siguiente modelo lineal general:

$$E_M [Y] = X\beta$$

$$VAR_M [Y] = V$$

donde

- X es una matriz N x p de variables auxiliares
- β es un vector px1 de parámetros desconocidos
- V es una matriz de covarianzas definida positiva

Sea $T = \gamma'Y$ una combinación lineal que queremos estimar.

Teorema: Entre los estimadores lineales de T, insesgados en la predicción, la varianza del error es mínima para:

$$T_e = \gamma_s' Y_s + \gamma_r' [X_r\beta_e + V_{rs} V_{ss}^{-1} (Y_s - X_s\beta_e)]$$

donde

$$\beta_e = (X_s' V_{ss}^{-1} X_s)^{-1} X_s' V_{ss}^{-1} Y_s$$

Y la varianza del error de T_e es:

$$VAR_M (T_e - T) = \gamma_r' (V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) \gamma_r + \gamma_s' (X_r - V_{rs} V_{ss}^{-1} X_s)(X_s' V_{ss}^{-1} X_s)^{-1} (X_r - V_{rs} V_{ss}^{-1} X_s)' \gamma_r$$

donde X_r, X_s ,Y_s, V_{rr}, V_{rs},V_{sr} ,V_{ss} ,γ_r ,γ_s se obtienen mediante partición de X, V, y γ en unidades que están en la muestra (s) y que no están en la muestra (r).

Consideraremos una población en la que podemos establecer una clasificación cruzada. Cada unidad pertenece a una clase c = 1..C y a un dominio d = 1..D. Una clase puede ser un grupo de edad/sexo, y un dominio es nuestra área pequeña (en general una subpoblación cualquiera de interés).

Sea s_{cd} la muestra de la casilla (c,d) y r_{cd} el conjunto de unidades no muestreadas en la misma. El total en el dominio es la suma para las unidades muestreadas y no muestreadas:

$$T_d = \sum_c \sum_{i \in S_{cd}} + \sum_c \sum_{i \in r_{cd}} Y_i$$

2. CONSTRUCCIÓN DEL MODELO

Nuestro objetivo es estimar la tasa de paro en las provincias españolas. La tasa de paro solo puede ser determinada mediante la Encuesta de Población Activa realizada por el Instituto Nacional de Estadística (INE). Es una encuesta por muestreo aleatorio estratificado, trimestral, con un tamaño de muestra aproximado de 65 000 hogares, lo que permite estimar la tasa de paro en cada Comunidad Autónoma y en cada provincia. La precisión es muy elevada para el conjunto de la nación, pero es baja en las provincias, especialmente en aquellas con menores tamaños de muestra.

En paralelo a las estimaciones de la EPA, existe un registro de parados, realizado por el Instituto Nacional de Empleo (INEM), en el cual se inscriben las personas que buscan empleo. Aunque todos los contratos de trabajo deben realizarse a través del INEM por norma legal, no todos los demandantes de empleo se registran, especialmente aquellos que buscan su primer empleo (por ejemplo estudiantes que han finalizado sus estudios). Por esta razón, y también por algunas diferencias en la definición de parado, el paro registrado es en general menor que el paro estimado por la EPA.

Para los ocupados también existe un registro, el de la Seguridad Social, que incluye a todos los trabajadores excepto los empleados de la Administración Pública. Pero este registro considera el domicilio de la empresa en lugar del domicilio del trabajador, lo que produce grandes diferencias –especialmente a nivel local- entre el número registrado de trabajadores ocupados y el número estimado por la EPA.

Estas dos variables, paro registrado y altas en la Seguridad Social, pueden ser obtenidas mensualmente para cada municipio, pero la tasa de paro calculada con ellas no es útil, debido a que existen diferencias mayores de un 20% con la tasa real de paro, en algunos casos.

La tasa estimada de paro es un indicador relativo, lo que permite establecer comparaciones entre distintas áreas, y evaluar los resultados de las políticas locales de empleo. Es el mejor indicador que podemos utilizar con este objeto.

La muestra de la EPA se distribuye dentro de cada provincia en diferentes estratos según el tamaño de los municipios. Nuestro objetivo a largo plazo es desarrollar una metodología estadística que permita la estimación de la tasa de paro en cada municipio, incluso aquellos que tienen muy pocas unidades en la muestra o ninguna, utilizando los resultados muestrales para la provincia y la comunidad autónoma (que incluye varias provincias) junto con variables auxiliares a nivel local, utilizando la reciente metodología de estimación en áreas pequeñas. El presente trabajo tiene sin embargo un objetivo más simple: obtener estimaciones para las provincias utilizando el mismo esquema. Dado que para estas áreas si existen resultados muestrales de la EPA, es posible comparar ambas estimaciones, directa y de área pequeña, lo que nos permitirá averiguar si estas técnicas pueden ser utilizadas razonablemente para nuestro objetivo principal, la estimación a nivel municipal.

Para la estimación del paro, utilizaremos como variables auxiliares las siguientes: sector de actividad, sexo y edad

SECTOR: Agricultura/Pesca, Construcción, Industria, Servicios, Sin sector.

EDAD y SEXO: 16-19 años, 20-24, 25-54, y 55 o más años, en total 8 segmentos.

Tenemos por lo tanto una clasificación cruzada, con 8 grupos de edad/sexo, y cinco sectores (clases c) y las provincias (dominios d). Sea el siguiente modelo:

$$E[Y_i] = \mu_c \qquad \text{si } i \text{ está en } (c,d)$$

$$\text{cov}[Y_i, Y_i] = \sigma^2_{cd} \qquad \text{si } i \text{ está en } (c,d)$$

$$\text{cov}[Y_i, Y_j] = \sigma^2_{cd} \cdot \rho_{cd} \qquad \text{si } i,j \text{ están en } (c,d)$$

$$\text{cov}[Y_i, Y_j] = 0 \quad \text{en otro caso}$$

La varianza de la media muestral de la casilla (c,d), cuando $n_{cd} > 0$ es:

$$v_{cd} = \sigma^2_{cd} \; [\, 1 + (n_{cd} - 1)\, \rho_{cd}] \,/\, n_{cd}$$

El predictor lineal insesgado óptimo (ELIO) para el total de la casilla (c,d) es, de acuerdo con el Teorema de Royall:

$$T_{cd} = \sum_{S_{cd}} Y_i + (N_{cd} - n_{cd})[w_{cd} Y_{mscd} + (1 - w_{cd})\,\mu'_c]$$

donde

$$w_{cd} = n_{cd}\rho_{cd} / [1 + (n_{cd} - 1)\,\rho_{cd}] ,$$

$$\mu'_c = \sum_{j}^{D} u_{cj} Y_{mscj}$$

con

$$u_{cj} = v_{cj}^{-1} / \sum_{D} v_{cj}^{-1}$$

y

$$v_{cj}^{-1} = 0;\ Y_{mscj} = 0;\ \sum Y_i = 0 \quad \text{si} \quad n_{cd} = 0$$

para las casillas sin muestra:

$$T_{cd} = N_{cd}\,\mu'_c$$

siendo μ'_c una estimación de la media de la clase c utilizando todas las unidades de la muestra en c, de los diferentes dominios en los que hay unidades muestrales, considerando la diferente variabilidad clase/dominio.

El estimador para todo el dominio se obtiene sumando todas las clases dentro del dominio:

$$T_d = \sum_{c} T_{cd}$$

Este estimador utiliza información de todos los dominios, dado que la media μ_c depende solo de la clase.

La varianza del estimador T_d es:

$$VAR_M\,(T_d) = \sum_{c} N_{cd}^2 \sigma_{cd}^2\,(1 - f_{cd})(1 - \rho_{cd})\,[1 - (1 - f_{cd})\,(1 - w_{cd})(1 - u_{cd})] / n_{cd}$$

siendo

$$[1-(1 - f_{cd})\,(1 - w_{cd})(1 - u_{cd})] < 1$$

Cuando el modelo es adecuado, el estimador ELIO para el total en los dominios que no tienen muestra coincide con el estimador sintético, una técnica que “en cierto modo reproduce a escala las estimaciones insesgadas de subpoblaciones grandes en proporción a la incidencia de cada sub-clase dentro del área pequeña de interés” (González, 1973):

$$T_d = \sum_{c} N_{cd} \cdot \mu_c$$

donde el sumatorio se realiza para todas las clases c, y además:

$\mu_c = \sum_{j=1}^{D} l_{cj} Y_{mscj}$ es una media ponderada de las medias de casilla.

El estimador sintético puede escribirse:

$$T_d = \sum_c n_{cd} \cdot \mu_c + \sum_c (N_{cd} - n_{cd})\mu_c$$

con lo que queda claro que la suma muestral $\sum n_{cd}\, Y_{mscd}$ (conocida si hay muestra en el dominio) está siendo estimada por $\sum n_{cd} \cdot \mu_c$. Desde luego siempre podemos evitar estimar lo que ya es conocido y utilizar:

$$T_d = \sum_c n_{cd} \cdot Y_{mscd} + \sum_c (N_{cd} - n_{cd})\mu_c$$

La media ponderada del estimador sintético y del estimador post-estratificado para cada casilla, calculando el estimador agregado como la suma para todas las casillas se conoce como estimador compuesto.

El número de parados se obtiene como el total de clase para una variable Y_i con valores 1 y 0 (sin/con empleo) con probabilidad μ_c y $(1 - \mu_c)$ respectivamente de acuerdo con el modelo. En esta situación la varianza es $\sigma^2_{cd} = \mu_c(1-\mu_c)$, para la clase c independientemente del dominio. Además podemos suponer que las variables Y_i son también independientes dentro de la casilla.

Con este estimador ELIO (según el modelo) hemos obtenido el número total de parados para cada provincia de España.

Para la estimación de la ocupación aplicamos el mismo modelo, con sector de actividad como criterio de clasificación. Con ambas estimaciones, número de parados y número de ocupados, se calcula la tasa de desempleo, que es el cociente:

$$\text{TASA DE DESEMPLEO} = \frac{\text{PARADOS}}{\text{PARADOS} + \text{OCUPADOS}}$$

3. VERIFICACIÓN DEL MODELO

La verificación del modelo es muy importante en la estimación en áreas pequeñas, debido a que los estimadores solamente son insesgados y de varianza mínima si el modelo es válido. Es necesario investigar la validez a priori de las hipótesis del modelo, y también podemos utilizar información disponible en diferentes fuentes para contrastar los resultados.

a) Consistencia interna del modelo

Utilizamos, para la estimación de la ocupación, una clasificación basada en el sector de actividad. Este es un criterio realista: muchos sectores (por ejemplo pesca, construcción naval, conservas de pescado, etc. actualmente en crisis) son afectados por la evolución económica de forma similar, independientemente de la zona en la que se ubican las empresas, de modo que las tendencias de empleo son similares dentro de cada sector. Las diferencias son mayores entre sectores que entre provincias, y la tasa de paro provincial es distinta básicamente porque cada provincia tiene una estructura sectorial diferente. Por otra parte la estructura sectorial es estable a corto plazo.

b) Comparación con la estimación directa donde se dispone de datos.

Para cada provincia hemos estimado la tasa de paro utilizando la estimación de la correspondiente Comunidad Autónoma, y las variables auxiliares, como si no hubiese muestra directa en la misma. Pero ya que existe una muestra, es posible comparar ambos resultados: La estimación de área pequeña, con el método propuesto en este trabajo, y la estimación directa elaborada por el Instituto Nacional de Estadística. La Tabla 1 muestra los resultados obtenidos con ambas estimaciones y la diferencia para cada provincia.

Como se puede observar, las diferencias son razonablemente pequeñas: en el 70% de las provincias el error o diferencia entre ambas estimaciones es menor que 0,02 y en el 85% menor que el 0,03. En la gran mayoría de las provincias la diferencia está por lo tanto dentro del margen de error de estimación atribuible al muestreo directo.

Para conseguir errores de estimación tan reducidos como estos se requieren en general muestras directas superiores a 1000 unidades en cada provincia, mientras que los resultados mostrados en la tabla (estimación modelo) se han logrado con una muestra común para toda la Comunidad Autónoma, que puede ser así mucho más pequeña, ya que no se requiere una muestra suficientemente grande para cada provincia. Esto permite reducir coste y tiempo. El modelo que se propone parece ser, por lo tanto, razonablemente adecuado.

4. CONCLUSIONES

La aplicación de técnicas de área pequeña basadas en el enfoque predictivo del muestreo de poblaciones finitas –con el apoyo de modelos- permite resolver el problema de obtener tasas de paro estimadas para áreas locales. Bajo las condiciones del modelo los estimadores son insesgados y de varianza mínima, mientras que los habituales estimadores basados en el diseño son únicamente insesgados. La estimación para zonas con una participación muy pequeña o nula en la muestra se consigue "pidiendo prestada" la fuerza o capacidad para estimar, basada en la información muestral, a otras zonas, aunque deben hacerse algunas hipótesis simplificadoras para construir un modelo operativo. La verificación del modelo –no siempre fácil o incluso posible- se hace en este trabajo, comparando las estimaciones de área pequeña con las estimaciones directas, lo que hace posible validar el modelo que permite obtener estimaciones de similar precisión con un coste mucho más reducido, y que podrá ser aplicado –este es nuestro objetivo a largo plazo- a zonas más pequeñas donde no exista muestra directa.

REFERENCIAS

ERICKSEN, EUGENE P. (1973): "A method for combining Sample Survey Data and Symptomatic indicators to obtain Population estimates for local areas", **Demography** 10, 137-159.

ERICKSEN, E.P. and J.B. KADANE (1985): "Estimating the population in census year (with discussion)", **J. Amer. Statist. Assoc**. 80, 98-131.

ERICKSEN, E.P.; J.B. KADANE and J.W. TUKEY (1989): "Adjusting the 1980 Census of Population and Housing", **Journal of the American Statistical Assoc**., 84:927-944.

GHOSH, M. and J.N.K. RAO (1994): "Small area estimation: an appraisal", **Statistical Science** 9(1), 55-93.

GONZÁLEZ, M.E. (1973): **Use and Evaluation of Synthetic Estimates**. Proceedings of Social Statistics Section, American Statistical Association.

HANSEN, MORRIS H.; WILLIAM N. HURWITZ and WILLIAM G. MADOW (1993): **Sample survey methods and theory**; New York : Wiley, primera edición en 1953}

PLATEK, R., J.N.K. RAO; C. E. SÄRNDAL and M. P. SINGH (1987): **Small Area Statistics**. New York: John Wiley.

ROYALL, R.M. (1979): Prediction Models in Small Area Estimation; in Steinberg, J.(Ed) Synthetic Estimates for Small Areas. National Institute on Drug Abuse, Research Monograph 24;; Washington D.C.

SCHAIBLE, WESLEY L. (1996): **Indirect Estimators in U.S. Federal Programs**; Springer Verlag, Berlín.

VALLIANT, RICHARD; ALAN H. DORFMAN and RICHARD M. ROYALL (2000): **Finite population sampling and inference; a prediction approach**; John Wiley, New York.

APÉNDICE

Tabla 1. Estimación de área pequeña y estimación directa.

TASA DE PARO	ESTIMACIÓN modelo	E.P.A.	dif.		TASA DE PARO	ESTIMACIÓN modelo	E.P.A.	dif.
ANDALUCÍA					**CATALUÑA**			
Almería	0,118	0,085	0,033		Barcelona	0,102	0,103	0,001
Cádiz	0,239	0,224	0,015		Girona	0,060	0,074	0,014
Córdoba	0,216	0,224	0,008		Lleida	0,069	0,053	0,016
Granada	0,165	0,166	0,001		Tarragona	0,083	0,067	0,016
Huelva	0,202	0,178	0,024		**COMUNIDAD VALENCIANA**			
Jaén	0,153	0,213	0,060		Alicante	0,107	0,122	0,014
Málaga	0,143	0,161	0,018		Castellón de la Plana	0,072	0,059	0,013
Sevilla	0,193	0,179	0,015		Valencia	0,120	0,112	0,008
ARAGÓN					**EXTREMADURA**			
Huesca	0,047	0,022	0,025		Badajoz	0,177	0,153	0,025
Teruel	0,064	0,044	0,020		Cáceres	0,147	0,186	0,039
Zaragoza	0,070	0,078	0,008		**GALICIA**			
CANARIAS					Coruña (A)	0,122	0,127	0,005
Palmas (Las)	0,125	0,109	0,016		Lugo	0,093	0,077	0,016
Santa Cruz de Tenerife	0,109	0,122	0,013		Orense	0,122	0,071	0,051
CASTILLA - LA MANCHA					Pontevedra	0,125	0,139	0,014
Albacete	0,113	0,063	0,049		**PAÍS VASCO**			
Ciudad Real	0,117	0,110	0,007		Álava	0,078	0,081	0,004
Cuenca	0,080	0,099	0,019		Guipúzcoa	0,074	0,054	0,020
Guadalajara	0,066	0,093	0,027		Vizcaya	0,106	0,115	0,009
Toledo	0,088	0,110	0,022		Toledo	0,088	0,110	0,022
CASTILLA Y LEÓN					**CASTILLA Y LEÓN**			
Ávila	0,107	0,115	0,008		Ávila	0,107	0,115	0,008
Burgos	0,078	0,071	0,008		Burgos	0,078	0,071	0,008
León	0,128	0,112	0,016		León	0,128	0,112	0,016
Palencia	0,119	0,087	0,032		Palencia	0,119	0,087	0,032
Salamanca	0,129	0,156	0,027		Salamanca	0,129	0,156	0,027
Segovia	0,064	0,110	0,045		Segovia	0,064	0,110	0,045
Soria	0,050	0,039	0,012		Soria	0,050	0,039	0,012
Valladolid	0,119	0,123	0,003		Valladolid	0,119	0,123	0,003
Zamora	0,152	0,141	0,011		Zamora	0,152	0,141	0,011