Islas Canarias
Del 15 al 19 de noviembre de 2021

istac | INSTITUTO CANARIO DE ESTADÍSTICA

evm

Estimación en áreas pequeñas de indicadores clave con R

Andrés Rodríguez González
Instituto Canario de Estadística (ISTAC)
arodgonx@gobiernodecanarias.org

Elisa Jorge González
Ecosistemas Virtuales y Modulares (EVM)
elisa.jorge@evm.net

PALABRAS CLAVE
Áreas pequeñas, R, EPA

INTRODUCCIÓN

La demanda de estadísticas oficiales con un gran detalle en la desagregación, tanto en el campo de la estadística económica como en el de la estadística social y laboral, no deja de crecer. En consecuencia, la necesidad de disponer sistemáticamente de datos publicados para dominios pequeños, se ha venido consolidando en los últimos años entre los objetivos de los sistemas de estadística oficiales, al mismo tiempo que se han venido desarrollando diversas líneas de investigación sobre la utilización de estimadores en áreas pequeñas como forma de superar la limitación de la escasez de muestra en las operaciones estadísticas por muestreo y cuya información tiene interés en pequeños dominios. A grandes rasgos, la estimación en áreas pequeñas emplea datos obtenidos en grandes dominios para construir estimaciones para los pequeños.

Con esto en mente, el diseño de las encuestas por muestreo se realiza con el objetivo de obtener estimaciones fiables de las principales variables de interés con un mínimo de precisión prefijada en agregados poblacionales definidos a priori. No obstante, se suele demandar información de magnitud inferior a los considerados en el diseño de las encuestas. Por ejemplo, puede resultar de interés el disponer de estimaciones municipales en una encuesta cuyo agregado poblacional mínimo considerado en el diseño haya sido la provincia.

Por tanto, nos encontramos ante la problemática de la insuficiencia del número de puntos de muestreo o incluso ausencia total en algunos casos para poder dar cualquier tipo de estimación fiable referida a estos ámbitos no previstos en el diseño.

En este sentido, el Instituto Canario de Estadística (ISTAC) ha venido ejecutando desde el 2004 la estimación a nivel insular y comarcal de las principales variables de la Encuesta de Población Activa (EPA) del Instituto Nacional de Estadística (INE). Desde entonces, el trabajo realizado por parte del ISTAC conjuntamente con el Departamento de Estadística, Investigación Operativa y Computación de la Universidad de La Laguna dio paso a la publicación periódica de estimaciones trimestrales de las 27 comarcas estadísticas de la Comunidad Autónoma de Canarias sobre la población a partir de 16 años y su relación con la actividad económica por sexos y grupos de edad.

El presente documento tiene como finalidad exponer de forma sintética la metodología de las estimaciones insulares y comarcales de las principales variables de la EPA, haciendo uso del software libre R.

OBJETIVOS

Con el propósito de proporcionar estadísticas con el mayor nivel de detalle posible que describa de forma pormenorizada la realidad de la Comunidad Autónoma de Canarias, el poder suministrar información trimestral sobre características ligadas a variables sociodemográficas, como la actividad, la ocupación y el paro, tanto a nivel insular como comarcal, ha demostrado en los últimos años la utilidad de realizar estimaciones a un nivel de desagregación mayor en algunas de sus encuestas.

Por otro lado, debido al trabajo de modernización que se está desarrollando por parte del ISTAC en materia de producción estadística, durante los últimos años se ha promovido el uso de software libre, el cual evoluciona a una velocidad sorprendente. Dentro del software libre, el lenguaje de programación R es el más utilizado por la cantidad de rutinas que tiene implementadas y la capacidad de realizar cálculos estadísticos muy complejos, estando

formado por un conjunto de herramientas flexibles que pueden ampliarse fácilmente mediante paquetes, librerías o la definición de funciones propias.

Por ello en el último año, se ha realizado un trabajo de revisión de la metodología que se venía aplicando hasta el momento, con intención de versionar el programa que se ha estado empleando hasta día de hoy para la estimación en áreas pequeñas de las principales variables de la EPA, programado en los comienzos de la implementación de dicha metodología en el lenguaje de programación FORTRAN, con vistas a usar un lenguaje ampliamente utilizado en la comunidad estadística como es el lenguaje de software libre R, donde se disponen de diversas librerías desarrolladas con el objetivo de aplicar estimación en áreas pequeñas.

El trabajo realizado para la migración de un lenguaje a otro ha tenido como objetivo no solo la migración a un lenguaje ampliamente utilizado por la comunidad estadística, sino también con idea de analizar las tareas que se hacían hasta ahora para, posteriormente, generar una metodología y programación que permita su extrapolación para la estimación de otras posibles variables de interés.

METODOLOGÍA

El trabajo realizado se basa en la metodología estadística de estimación en áreas pequeñas de variables laborales a nivel local. Cabe destacar que dentro de la metodología de áreas pequeñas existen diferentes enfoques, que se pueden englobar en dos grandes grupos: *estimadores basados en diseño* y *basados en modelos.*

Estimador

El estimador empleado para la obtención de estimaciones insulares y comarcales es el estimador *sintético-región dinámico con precisión variable*. Para ello, se construyen regiones como uniones de comarcas que presentan un comportamiento homogéneo en cuanto a las variables objetivo a estudiar, con la finalidad de reducir el sesgo de la estimación y muy probablemente su error cuadrático. Se construyen distintas regiones homogéneas para cada una de las siguientes combinaciones de variables objetivo: hombre ocupado, mujer ocupada, hombre parado y mujer parada. Para la construcción de dichas regiones, se construyen regiones borrosas (fuzzy), formadas por la unión de aquellas comarcas que más próximas están a la que se desea estimar. De esta forma la aportación de información indirecta para la estimación de una comarca podrá variar en cada realización de la encuesta, y por lo tanto esto hace que dichas regiones dinámicas dependan de la comarca a estimar y de la realización particular de la EPA.

Para la construcción de las regiones dinámicas se indica cuál es la distancia entre comarcas, y más concretamente el orden de cercanía. Una vez fijado los órdenes de cercanía entre comarcas para cada una de las variables objetivo que se pretenden estimar, se fijan cuántas de ellas prestarán su información a la estimación, para lo que a su vez se fija un método de parada en la utilización de la información de las comarcas vecinas, de forma que una vez definidas las matrices de distancias, una comarca utilizará la información de las comarcas más próximas hasta que la muestra alcance el nivel muestral obtenido con la precisión dada. En este caso, dicha precisión es variable, variando entre comarcas, variable objetivo y sexo.

Calibrado a posteriori

Una vez obtenidas las estimaciones $\hat{Y}_d$ para cada uno de los dominios pequeños, éstas deben ser corregidas con la finalidad de que sean consistentes con los datos publicados por el INE para cada provincia. Para ello se emplea el factor de calibrado

$$F_{yp} = \hat{y}_p{}^{epa} / \sum_{d \epsilon P} \quad \hat{Y}_d \quad ,$$

siendo $\hat{y}_p{}^{epa}$ la estimación según la EPA para la provincia P. Además, para la estimación de la población inactiva se procede de la siguiente forma:

1. Se calibran las estimaciones de población ocupada y parada para cada provincia y sexo.

2. Se obtiene la población inactiva para cada dominio d como la diferencia entre el total poblacional de dicho dominio, N_d, menos la suma de población ocupada y parada estimada.

Adicionalmente se utiliza un calibrado mediante un procedimiento tipo CalMar para obtener consistencia de las estimaciones por sexo, grupos de edad y variable objetivo a nivel de isla con las estimaciones de las provincias

Estimación en pequeñas áreas de la población actual

El Reglamento Nº 105/2007 establece que el INE debe ofrecer estimaciones anuales de la población actual a nivel insular, datos que se proporcionan a través de la operación estadística *"Estimación de la Población Actual"* (ePOBa). La información que ofrece dicha operación es la estimación anual del total poblacional de las islas por sexo y grupo de edad (<25, 25-54, >54 años). No obstante, las Estimaciones en pequeñas áreas de la EPA requieren adicionalmente estimaciones trimestrales de totales poblacionales a nivel comarcal.

Las estimaciones de datos poblacionales de los diferentes trimestres se obtienen a partir de los datos insulares anuales ofrecidos por la ePOBa siendo modificados siguiendo un método similar al CalMar como ya se mencionó anteriormente, de tal manera que se minimicen las perturbaciones al cuadrado bajo las restricciones siguientes:

- Los datos ajustados deben concordar con las estimaciones poblacionales por sexo a nivel de provincia suministradas por la correspondiente realización trimestral de la EPA.

- Los datos ajustados deben concordar con los totales poblacionales por grupo de edad y sexo a nivel de la Comunidad Autónoma de Canarias.

Una vez obtenidas las estimaciones insulares trimestrales, se usan los datos proporcionados por el Padrón Municipal de Habitantes a 1 de enero del respectivo año, con información de comarcas por grupo de edad y sexo, para estimar los datos comarcales mediante reparto proporcional.

Estimación del error cuadrático medio

El método de remuestreo empleado para la estimación del Error Cuadrático Medio (MSE) es el método de Jack-knife. Este método consiste en tomar tantas submuestras como clusters (secciones censales en el caso de la EPA) existen en la muestra. Cada submuestra se forma así con la muestra original al dejar fuera una sección censal cada vez.

RESULTADOS

La metodología anteriormente descrita ha sido implementada en R para la obtención de las estimaciones insulares y comarcales de la Encuesta de Población Activa (EPA), que aportan trimestralmente datos básicos de la población y su relación con la actividad económica, así como las correspondientes tasas de actividad, empleo y paro; permitiendo de este modo, su comparación con los datos nacionales y de otras Comunidades Autónomas.

El programa desarrollado está diseñado para calcular las estimaciones de los totales de la población ocupada, parada, inactiva y activa, y de las tasas de actividad, paro y empleo así como sus respectivos coeficientes de variación (aplicando el método de Jack-knife) a nivel de Comunidad Autónoma, provincias e islas (total, por sexo, por grupo de edad y, por grupo de edad y sexo) para una realización de Encuesta de Población Activa (EPA) de un trimestre particular.

Para poder realizar las estimaciones de áreas pequeñas de la EPA se parte del fichero de microdatos de la realización de la EPA a analizar, el cual es un documento de texto siguiendo el formato descrito en el *Diseño de registro y valores válidos* suministrados por el Instituto Nacional de Estadística (INE).

Además del uso de los microdatos como la fuente principal de información, para el cálculo de las estimaciones es necesario disponer de la siguiente información auxiliar:

- Los ***tamaños poblacionales por islas*** (en general, se obtienen a partir de la ePOBa, y coinciden para los cuatro trimestres de un mismo año), sexos y los grupos de edad comprendidos entre 16-24, 25-54 y más de 55. Dicha información se lee a partir de un fichero de texto con siete columnas separadas por espacios, donde la primera columna es un texto no superior a cinco caracteres que hace referencia a la isla, y las siguientes tres columnas representan población de hombre por grupos de edad y las últimas tres la población de mujeres por grupos de edad.

- Los ***tamaños poblacionales por comarcas*** (en general, se obtienen a través del Padrón Continuo de población y coincide para los cuatro trimestres de un mismo año), sexos y los grupos de edad comprendidos entre 16-24, 25-54 y más de 55. La estructura del fichero es equivalente al de la población de islas por grupo de edad.

- Información de ***proximidades entre comarcas*** que contiene el orden de proximidad entre las distintas comarcas que serán empleados para tomar muestra “prestada” de una comarca parecida, para la construcción de regiones homogéneas de comportamiento para hombres ocupados, mujeres ocupadas, hombres parados y mujeres paradas.

- Valores iniciales de ***precisión (en porcentaje) para ocupados y parados***. Estos valores proporcionarán una cuota inferior de las precisiones a utilizar para obtener el tamaño de muestra necesario para estimar una determinada comarca. El programa a partir de estos valores introducidos adicionalmente prueba con valores mayores a estas cantidades, de medio punto en medio punto, quedándose con aquellas que proporcionan los coeficientes de variación mínimos. Se pueden utilizar tres formas de introducir las precisiones:

    1. Valores por defecto, precisión para ocupados de 1% y para parados del 2,5%, independientemente del sexo y comarcas.

    2. Introducción de valores por el usuario, uno para ocupados y otro para parados, siendo asignados independientemente del sexo y comarcas.

    3. Lectura desde un fichero. En este caso, el usuario puede elegir precisiones distintas por comarcas, sexo y variable objetivo.

De modo simplificado y esquemático, el programa realiza los siguientes pasos para la obtención de las estimaciones:

1. Se leen los ficheros de entrada necesarios (microdatos e información auxiliar).

2. Se calcula el número de secciones para utilizar el método Jack-knife.

3. Se ajustan los valores de las islas a la provincia correspondiente y a la Comunidad Autónoma.

4. Se ajustan los datos comarcales a los datos por islas.

5. Se inicializan las precisiones dependiendo del método de introducción seleccionado.

6. Se calcula el tamaño muestral necesario en función de los tamaños ajustados de las comarcas.

7. Se asignan las comarcas que se necesitan para estimar una comarca.

8. Se calcula la estimación según las regiones formadas por las comarcas.

9. Se calibran los estimadores.

10. Se calculan los inactivos por sexo por diferencia con respecto al total por sexo en la comarca e isla.

11. Se generan los ficheros de salida.

Una vez finalizado el procedimiento para la obtención de las estimaciones, como resultado se generan tres ficheros que contienen la siguiente información:

- Un fichero que contiene las estimaciones.
- Un fichero que contiene los coeficientes de variación.
- Un fichero con los tamaños muestrales.

CONCLUSIÓN

La aplicación de técnicas de áreas pequeñas permite resolver el problema de obtener los principales indicadores de la relación con la actividad en las islas para áreas locales. Para ello, 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, haciendo uso del estimador sintético-región dinámico con precisión variable.

Además, el trabajo realizado a lo largo de este proyecto ha traído en consecuencia la identificación de futuras líneas de trabajo, así como mejoras a implementar:

- A lo largo del proceso de migración de un lenguaje a otro se ha hecho notoria la mejora computacional, al pasar de un lenguaje de programación especialmente adaptado al cálculo numérico como es el FORTRAN, a un lenguaje de programación orientado a objetos. Además, R aporta otras ventajas como la integración con distintas bases de datos y el amplio abanico de herramientas estadísticas y gráficas.

- La nueva implementación de la metodología abre la puerta a su uso en otras encuestas. A pesar de ser necesario un análisis específico para el uso de áreas pequeñas en otras variables de interés, el trabajo realizado en generar funciones en R que apliquen paso por paso la metodología necesaria para la estimación en áreas pequeñas, permite hacer pruebas y aplicarlo en otras encuestas sin necesidad de crear un nuevo programa ad-hoc para cada una de ellas.

- En esta primera etapa de migración entre lenguajes de programación, a pesar de haber realizado mejoras de cómputo y simplificaciones en el código, no se han implementado todas las mejoras que el uso de R permiten. Queda claro que a futuro se debe seguir trabajando en esta línea, con idea de automatizar lo máximo posible todo el procedimiento, desde la lectura de la información necesaria en base de datos, hasta la generación de los ficheros de publicación.

BIBLIOGRAFÍA

[1] Enhancing Small Area Estimation Techniques to meet European Needs (EUROAREA Project) (2005).

[2] INE. (2021). Encuesta de Población Activa. Metodología 2021. Madrid.

[3] ISTAC. (2008). Encuesta de Población Activa. Metodología para la estimación en pequeñas áreas de Canarias. Gran Canaria.

[4] Platek, R.; Rao, J.N.K.; Särndal, C. E. and Singh, M. P. (1987). Small Area Statistics. New York: John Wiley.

[5] Rao, J.N.K. (2003). Small Area Estimation. John Wiley & Sons, New Jersey.