6. ESTADÍSTICA OFICIAL

LA ESTIMACIÓN EN ÁREAS PEQUEÑAS PARA LA ESTADÍSTICA OFICIAL

Montserrat Herrador, Jorge Saralegui
Instituto Nacional de Estadística

Resumen

La necesidad de disponer sistemáticamente de información estadística para dominios pequeños, se ha venido consolidando en los últimos años entre los objetivos de los sistemas de estadísticas oficiales. El artículo aborda los problemas ligados a la utilización de estimadores asistidos por modelos como forma de superar la limitación de la escasez de muestra en operaciones estadísticas cuya información tiene interés en pequeños dominios, con especial atención a los aspectos de calidad estadística, como la relevancia, transparencia y fiabilidad de los datos oficiales a difundir.

1. 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 asistidos o basados en modelos, como forma de superar la limitación de la escasez de muestra en las operaciones estadísticas cuya información tiene interés en pequeños dominios.

Los problemas de la estimación en pequeños dominios surgen por el aumento en los costes y en la complejidad de los diseños muestrales que aspiren a alcanzar cotas aceptables de calidad de las estimaciones en todas las áreas o dominios de interés para los usuarios, lo cual, a su vez, puede tener consecuencias negativas en la calidad de las estimaciones para los dominios superiores previstos inicialmente en el proyecto de encuesta. Los límites, por razones de coste, para la ampliación sin restricciones de los tamaños muestrales en todos los dominios de interés, deben ser interpretados en un sentido amplio: bajo el punto de vista de la recogida y producción de datos, claro está, pero también bajo el punto de vista de la carga de respuesta a las unidades a contactar en la encuesta. El aumento de tamaños muestrales para mejorar la eficiencia en dominios pequeños debe tener en cuenta ambos costes, sin olvidar otras pérdidas de calidad debidas a los plazos de obtención de resultados y al impacto de determinados errores ajenos al muestreo (falta de respuesta, errores de medida, efectos entrevistador, etc.) de consecuencias más negativas cuanto mayores son los tamaños muestrales.

Tras hacer una exposición en las dos primeras secciones del problema general de la estimación en áreas o dominios pequeños, la ponencia continúa en la sección 3 con la situación en particular en nuestro país. La sección siguiente analiza la casuística ligada a la utilización de modelos en la estadística oficial, para pasar en las secciones 5 y 6 a una presentación de las principales clases de estimadores de áreas pequeñas actualmente en estudio para su aplicación a las encuestas del Instituto Nacional de Estadística (INE), y a los problemas relacionados con la calidad esperada de estas nuevas técnicas de producción de información. El artículo finaliza en la sección 7 con un análisis del papel de una oficina central de estadística en este campo.

2. Los dominios de estimación

Los dominios de estimación en una encuesta por muestreo son aquellos subconjuntos del ámbito poblacional para los que es necesario producir, bajo determinados requisitos de calidad, estimaciones de todas o algunas de las variables objetivo. Son por lo tanto unidades de análisis de los resultados de la encuesta, pero pueden o no ser también unidades de observación.

El estimador directo de diseño para el total de una característica en un dominio *d* suele tomar la forma general del estimador de Horwitz –Thomson; esdecir,

$$\hat{Y}_{d\pi} = \sum\nolimits_{S_d} \frac{y_k}{\pi_k} \quad \text{si } N_d \text{ es desconocido;} \quad \hat{Y}_{d\pi} = \sum\nolimits_{S_d} \frac{N_d}{\hat{N}_d} \frac{y_k}{\pi_k} \quad \text{si } N_d \text{ es conocido;}$$

donde S_d es la muestra en el dominio d, $\hat{N}_d = \sum_{S_d} \frac{1}{\pi_k}$ el estimador del total poblacional y π_k es la probabilidad de que la unida k pertenezca a la muestra.

En este tipo de estimadores, los valores muestrales de la característica y_k aparecen multiplicados por un 'peso inicial' o 'peso de diseño' que en fases posteriores es corregido, intentando mantener la forma lineal del estimador, por los efectos de cambio de estrato, incidencias de marco, falta de respuesta, equilibrado a poblaciones externas, u otros, de acuerdo a modelos implícitos más o menos complejos. El estimador directo de la media en el dominio queda así de la forma más general

$$\hat{\bar{Y}}_d^{direct} = \sum_{s_d} \frac{y_k w_k}{\sum_{s_d} w_k}, \text{ donde } w_k \text{ incorpora los diversos ajustes.}$$

No obstante, este tipo de 'modelos', por razones históricas difíciles de asimilar, no suelen ser considerados como tales en la literatura, y paradójicamente todavía es frecuente hoy en día encontrar polémicas académicas sobre la ventaja o no de utilizar modelos en la estimación del muestreo de poblaciones finitas, cuando, como queda dicho, en la mayoría de los proyectos, los modelos están presentes en diversas fases de la elaboración de resultados.

El pequeño dominio puede tener o no una dimensión territorial, pero en cualquier caso el problema se presenta cuando S_d es muy pequeño o nulo. Independientemente del procedimiento de estimación directa previsto, el pequeño dominio puede aparecer tanto en encuestas económicas como en encuestas de hogares, aunque en este caso los avances de la teoría hayan sido muy superiores. Baste citar como ejemplos del problema la gran presión de las peticiones dirigidas al INE para suministrar datos estructurales (o incluso coyunturales) de una encuesta económica a nivel de 4 ó 5 dígitos de la Clasificación de Actividades Económicas (CNAE), desagregados según la tipología del tamaño de las empresas, con lo que el estimador directo debe ser calculado con tamaños muestrales muy pequeños (pudiendo ser nulos si se introduce, además, como es frecuente, la dimensión territorial). En encuestas de hogares es más común el caso del pequeño dominio definido por la dimensión territorial exclusivamente, siendo cada vez más intensa la demanda de información (tasa de paro, índice de pobreza relativa, gasto per cápita) por pequeños territorios, como la comarca, la isla, el distrito, etc.

En España, prácticamente todos los diseños de encuesta contemplan la Comunidad Autónoma -NUT II, unidad estadística territorial de nivel II en terminología del sistema estadístico europeo- como dominio de observación y análisis, por lo que el problema de estimación en áreas pequeñas aparece a partir del nivel NUT III (provincia), aunque no en todos los casos. Por ejemplo, la Encuesta de Población Activa (EPA) publica datos por estimación directa por provincias. Sin embargo, la Encuesta de Presupuestos Familiares (EPF), sólo publica algunos indicadores de gasto y equipamiento por comunidades autónomas. En este último caso, para la publicación de datos provinciales se recomienda la acumulación de muestras de varios años/trimestres, procedimiento válido en ausencia de otros modelos que demuestren mayor robustez. Llama la atención el escaso recurso, por parte de los analistas y usuarios en España, a la utilización de muestras acumuladas de varios periodos para obtener estimaciones (en este caso provinciales), tanto en encuestas de hogares como en las estructurales de empresas.

Pero no hay que olvidar que las necesidades de información en áreas pequeñas territoriales en España, van más allá de la provincia. El INE está recogiendo en la actualidad, a través de las diversas autoridades con competencia en la materia, la situación administrativa del nivel NUT IV ("comarca" en España) con fines estadísticos Y, en este contexto, adquiere mayor urgencia la disponibilidad de métodos de estimación "poco exigentes" en tamaños muestrales para el nivel NUT IV.

El problema de las áreas pequeñas se puede presentar también ante demandas de información sobre subdominios ad hoc, como por ejemplo una cuenca (o parte de ella) hidrográfica, un cinturón industrial, una zona fronteriza …, por lo que los modelos disponibles deben contemplar también el hecho de la demanda estadística para dominios que no han sido tenidos en cuenta a priori en el diseño.

3. Usuarios y aplicaciones

Por principios de calidad, todo esfuerzo de innovación metodológica debe tener en cuenta la tipología y expectativas de los usuarios efectivos y potenciales. Destacan, en primer lugar, como demandantes de las estadísticas en dominios pequeños (con al menos una dimensión territorial), los centros de decisión de la propia administración local competente en el "pequeño territorio", así como los responsables de políticas más generales (CCAA, Estado) pero de aplicación en el ámbito local. Estas políticas cubren un amplio espectro, como las relativas a educación y formación, vivienda, salud y dependencia, lucha contra la exclusión social, medio

ambiente, desarrollo sostenible, detección y catalogación de "ecozonas", cubriendo diversos campos de acción política y de preocupación social.

Dependiendo de cada país y circunstancia, son unas u otras necesidades de información las que mayor peso aportan a la demanda inicial de estadísticas en áreas pequeñas. En el caso español, probablemente sea la iniciativa de las autoridades regionales (comunidades autónomas) la que mayor impulso está aportando, en paralelo al desarrollo de las competencias autonómicas generales.

Una parte de la actividad estadística orientada a atender esta demanda, sin embargo, no entra dentro de la especificidad de los modelos y estimadores en áreas pequeñas, sino en el campo más amplio de las "estadísticas para áreas pequeñas", de gran desarrollo también en nuestros días. Se trata de bancos de datos, asociados o no a sistemas de información geográfica (SIG), en los que se acumula transversal y longitudinalmente una gran masa de información y metainformación de fuentes diversas, incluidos indicadores simples y sintéticos, relativos a cada pequeña área, con un esquema de mantenimiento que puede responder a un modelo teórico más o menos desarrollado, como se puede observar en la gran diversidad de sistemas de indicadores disponibles. Por ejemplo, el sistema de información de este tipo que mantiene el INE, del que se extraen periódicamente informes publicados (Indicadores Sociales de España), integra también una dimensión territorial, que llega hasta el nivel municipal en algunos indicadores básicos.

Precisamente, los métodos de estimación en áreas pequeñas pueden servir para abastecer una pequeña parte de las necesidades de información de estos sistemas o, a su vez, ser receptores de una parte de información de estos bancos de datos de ámbito local, para ser utilizada como abastecedora de variables explicativas en los modelos subyacentes a los estimadores. Afortunadamente, la componente académica y de investigación universitaria, al menos a nivel internacional, aunque de manera muy incipiente en España, ha tenido –o está teniendo - un papel muy importante en el desarrollo y uso de estos modelos, lo que se percibe en la bibliografía que existe sobre el tema y que no deja de crecer día a día. Y no sólo en el campo de la estadística, sino también en otras disciplinas como las ciencias de la salud, de la educación, la sociología o la geografía del desarrollo.

4. Algunas limitaciones a la aplicación de modelos en la Estadística Oficial

Antes de aplicar sistemáticamente, aunque sea de forma parcial, los métodos de estimación asistida o basada en modelos es necesario ser muy conscientes de los riesgos y limitaciones que afectan a estas técnicas, lo cual explica en parte su relativamente lenta (y escasa) implantación en la estadística oficial, y no solo en la española.

En primer lugar, nos encontramos con el problema de la calidad y de la disponibilidad de fuentes externas que proporcionen las variables auxiliares que explican los modelos. Una fuente externa, de carácter exhaustivo, como un censo o un registro administrativo, puede muy bien alcanzar cotas de calidad aceptables para sus propios fines, pero ser inservible en la aplicación de un modelo, aunque en éste intervengan variables aparentemente análogas en la encuesta y en la fuente externa. En el caso de los censos, normalmente el principal problema es su larga periodicidad, sin olvidar algunos problemas de conceptualización en variables de uso posible en los modelos (ej.: situación de actividad económica de la persona). En los registros administrativos, como el registro de demandantes del INEM, por seguir con el ejemplo, de gran potencialidad a priori para explicar la variable estadística relativa a la tasa de desempleo, a la hora de su implantación pueden presentar inconvenientes - de cobertura e inestabilidad de la definición administrativa, principalmente- de cierto impacto en el rendimiento del modelo.

En el caso de los registros de población, entre los que podemos considerar al Padrón Continuo español, algunos fenómenos de importancia creciente, como la doble residencia o la formación de hogares complejos, disminuyen la propiedad explicativa de variables tan esenciales como la distribución de población por sexo y edad en un área pequeña. En el ámbito económico, una fuente exhaustiva como el Directorio Central de Empresas del INE, puede presentar limitaciones metodológicas graves a la hora de producir agregados para pequeños territorios, entre otras cosas por lo problemático de distribuir territorialmente la actividad de las empresas multilocalizadas o con vínculos complejos de unidades legales de actividad económica.

La utilización de modelos en estadística oficial debe hacer frente también a los problemas de falta de transparencia para determinados grupos de usuarios, como los medios de comunicación o algunos centros de decisión política, más acostumbrados a la estadística de origen administrativo u obtenida por estimación directa. Y ello en el contexto paradójico de la presencia habitual de modelos subyacentes en los estimadores finales de casi todas las encuestas, como se ha comentado anteriormente. Algunos modelos presentan cierta complejidad para la estimación de parámetros y sus varianzas, con software avanzado, apto solamente para personas expertas,

por lo que su uso e interpretación tiene asociados ciertos costes no presentes en la difusión de otras estadísticas.

Las grandes series de datos oficiales, con abundancia de interrelaciones multivariantes, que deben ser consistentes con los valores marginales, no se adaptan bien a la aplicación de estimadores basados en modelos para pequeños dominios, cuya consistencia a niveles más agregados, o con dimensión temporal, no siempre está garantizada, por lo que estos estimadores deben ser producidos conjuntamente con una metainformación y detalle metodológico difícil de asimilar por algunos usuarios.

En otro orden de cosas, especialmente en las estadísticas de empresas, el escaso tamaño muestral de algunas subpoblaciones impone un límite inferior al tamaño del dominio de estimación, por razones de confidencialidad. Algunos usuarios, por otra parte, buenos conocedores de un área pequeña, pueden no ser muy comprensivos con una estimación anómala en un dominio por ellos tan bien conocido, por muy "correcto" que sea el método de estimación empleado, lo cual contribuye a la preocupación por el riesgo de aparición de sesgos incontrolados, que afecta potencialmente a los estimadores basados en modelos.

5. Variables, modelos, estimadores

En los modelos de áreas pequeñas, las variables explicativas responden a dos tipologías muy distintas que condicionan el modelo: las que llamaremos aquí variables *X*, disponibles tanto en la observación muestral directa como en la fuente externa, a nivel de área pequeña, o las variables que llamaremos Z, o variables de área, disponibles solamente en los agregados procedentes de la fuente externa, pero que no son recogidas en la encuesta sobre la que se estimarán los parámetros del modelo. Un ejemplo clásico de variable X es la inscripción o no en la Oficina de Empleo, que se pregunta en la Encuesta de Población Activa a todas las personas mayores de 15 años, y que está siendo considerada en los modelos que están siendo investigados por el INE para la estimación en áreas pequeñas de la tasa de desempleo. Su correspondiente agregado por áreas, procedente de las estadísticas administrativas del INEM es el complemento necesario para la aplicación del estimador. Un ejemplo de variable Z en estudio para ser aplicada en modelos por el INE, es la base imponible del IRPF per cápita en el área, en nivel y en estructura según el origen de la renta (salarios, pensiones, desempleo, rentas del capital…). Para modelos de encuestas económicas, el volumen de negocio total de una actividad en el área, disponible en algunas fuentes externas, es un candidato firme a variable auxiliar (de área o de unidad elemental, según el tipo de encuesta).

En ocasiones, una variable tipo *X* puede dar mayor rendimiento en los modelos si se considera como *Z*, debido a que los errores de observación u otras diferencias conceptuales pueden hacer que la característica recogida en la encuesta difiera sustancialmente de la que constituye el agregado de la fuente externa. La decisión de considerar una variable explicativa presente en la encuesta como de tipo *X* o de tipo *Z*, es de gran trascendencia para la potencia explicativa de la componente sintética del modelo. Según sea el tipo de variables de las que disponemos en el mundo real de la encuesta y de las fuentes externas que proporcionan información para cada pequeña área, tendremos los diferentes tipos de modelos en los que se basan los estimadores frecuentistas.

Los *modelos de unidad elemental de dos niveles*, con efecto de áreas (índice *d*) y efecto de individuo (índice i) para explicar una variable *y* observada en la unidad elemental *i* del área *d,* pueden incluir tanto variables del tipo *X* como del tipo *Z* (por eso también se denominan 'modelos mixtos') según la casuística expuesta anteriormente:

$$y_{id} = Z_d^T\eta + x_{id}^T\beta + u_d + e_{id} \quad \text{con} \quad u_d \approx iid \ \ \mathrm{N}(0,\sigma_u^2); \quad e_{id} \approx \ \mathrm{iid}\ \mathrm{N}(0,\sigma_e^2)$$

Sin embargo los modelos de área contemplan solamente el efecto del área:

$$\overline{Y}_d = \overline{Z}_d^T\eta + u_d \ \text{ que se complementa con el modelo para el estimador } \ \hat{\overline{Y}}_d = \overline{Y}_d + e_d\,;$$

$$\text{con} \ \ u_d \approx iid \ \ \mathrm{N}(0,\sigma_u^2); \quad e_{id} \approx \ \mathrm{iid}\ \mathrm{N}(0,\psi_d);$$

En el ejemplo de la fórmula, el modelo explica el valor medio de la variable objetivo en el área, a partir de las variables auxiliares tipo *Z*, agregados de área.

Un concepto importante en la elección del modelo es el de grupo de regresión, o conjunto poblacional que engloba a las áreas pequeñas objetivo, y con cuya muestra se estiman los parámetros de los modelos (por ejemplo, una NUT II puede constituir el grupo de regresión para estimar parámetros a aplicar en estimadores al nivel NUT IV).

A la vista de los indicadores de evaluación, si su rendimiento se considera adecuado para estimar los parámetros, de estos modelos se derivan los estimadores sintéticos, como el *sintético de regresión*

$$\hat{\bar{Y}}_d^{rsyn} = \bar{X}_d^T \hat{\beta} + \bar{Z}_d^T \hat{\eta}$$

Un caso particular de los sintéticos de regresión es el sintético básico, en el que la estructura de una variable discreta en el dominio superior sirve para explicar la estructura en el pequeño dominio incluida en aquella

$$\hat{\bar{Y}}_d^{syn} = \frac{1}{N_d} \sum_g N_{gd} \hat{\bar{Y}}_g^{direct}$$

con lo que el estimador en el área se obtiene a partir de las estimaciones de los valores medios en un dominio superior dentro de los grupos g de una variable explicativa discreta (ej. clase de actividad, grupos de edad) cuyos subtotales poblacionales están disponibles también en el área pequeña, a partir de un registro actualizado, un Padrón o un Censo reciente.

Una clase de estimadores muy utilizados para áreas pequeñas es la de los *EBLUP (empirical best linear unbiased predictor)*

$$\hat{\bar{Y}}_d^{eblup} = \lambda_d \hat{\bar{Y}}_d^{greg} + (1 - \lambda_d) \bar{Y}_d^{rsyn} \quad (\text{ con modelo de unidad. elemental});$$

En el que interviene el estimador GREG, que se expone más adelante, entre los estimadores asistidos por el modelo. O bien el

$$\hat{\bar{Y}}_d^{eblup} = \lambda_d \hat{\bar{Y}}_d^{direct} + (1 - \lambda_d) \hat{\bar{Y}}_d^{rsyn}, \quad \lambda_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\varphi}_d} \quad (\text{con modelo de área}).$$

Estos estimadores combinan una componente directa o asistida por modelos, con el fin de asegurar que no se produzcan aumentos sustanciales de sesgo, con la componente sintética, para incorporar la ganancia en precisión debida a la aportación del modelo, cuya potencia explicativa se utiliza para disminuir la variabilidad del estimador combinado. El factor λ_d tiene en cuenta las dos fuentes de variabilidad contempladas en el modelo, e interviene en este tipo de estimadores combinados ponderando cada componente del estimador inversamente a cada aportación a las varianzas σ_u^2 y φ_d.

Los modelos no frecuentistas (bayesianos) para estimación en áreas pequeñas, disponibles ya en la literatura, no están siendo objeto de consideración para su desarrollo a corto plazo, aunque es previsible que se vaya a producir un avance en su utilización en los próximos años, también en el campo de la estimación en dominios pequeños.

Los modelos EBLUP parecen en principio los mejores candidatos a consolidarse para la obtención de estimadores de variables de encuestas en las que existe riesgo de que abunden las áreas con escasa muestra para el estimador directo. Un ejemplo serían los indicadores de ingresos o del gasto per capita procedentes del Panel de Hogares o de la EPF (muestras nacionales entre 10 ó 15.000 hogares) a nivel de comarcas (NUT IV) o incluso provincias (NUT III). Sin embargo, para parámetros de encuestas con muestra generosa en un área pequeña (como la EPA), incluso a nivel NUT IV se puede pensar en la eficiencia de estimadores basados en el diseño del tipo *GREG; es decir,*

$$\hat{\bar{Y}}_d^{greg} = \hat{\bar{Y}}_d^{direct} + (\bar{X}_d - \hat{\bar{X}}_d^{direct})^T \hat{\beta}$$

Un caso particular, muy utilizado, de estos estimadores es el postestratificado dentro de área, cuando las variables explicativas son dicotómicas de pertenencia o no a un grupo g de clasificación de una variable discreta, del que se dispone de subtotales en el área (ej. nivel de estudios, edad, etc.)

$$\hat{\bar{Y}}_d^{pps} = \frac{1}{N_d} \sum_g N_{gd} \hat{\bar{Y}}_{gd}^{direct} .$$

Éstos y otros estimadores (incluidas algunas versiones de la clase SPREE), para diversas opciones y encuestas, han sido investigados en el proyecto internacional de investigación EURAREA:

http://www.statistics.gov.uk/methods_quality/eurarea/

6. Efectos de calidad estadística de la componente sintética del estimador

Los estimadores sintéticos, o la componente sintética de un estimador combinado, tratan de "ganar fuerza" respecto al estimador directo, a partir de fuentes externas, de referencia coetánea o no, y de las estimaciones directas de conjuntos poblacionales 'próximos' (en el tiempo, en el espacio o en la analogía frente al fenómeno en estudio). Cuanto mayor sea la variabilidad que explique el modelo subyacente al estimador sintético, mayor reducción de varianza del estimador entre áreas. Es decir, el modelo sintético aporta una parte sustancial de "muestra efectiva" al estimador directo, aumentando 'virtualmente' el tamaño muestral en el área. Por ello, este tipo de estimadores es muy sensible a la omisión o a la mala calidad de las variables auxiliares que intervienen en el modelo. En cuanto al sesgo, estos estimadores, condicionados al modelo, son aproximadamente insesgados, pero sin esa condición son sesgados, produciendo sobreestimación en algunas áreas y subestimación en otras, con un error medio entre áreas cercano a cero.

La componente sintética de estos estimadores plantea así la disyuntiva clásica en muestreo relativa a la ganancia en eficiencia aportada en principio por el modelo, pero a cambio de la aparición de sesgos en el área pequeña. Este comportamiento da origen a dos de los aspectos más difíciles de transmitir con transparencia al usuario de estadísticas oficiales que utilicen modelos de este tipo: el suavizado del estimador sintético y el riesgo (mayor que en la estimación directa) de estimaciones outlier o anómalas. En efecto, el estimador sintético, por su propia definición, introduce cierto suavizado sobre la distribución en el muestreo del estimador directo en el área. Basta pensar en un estimador sintético básico que imputa a una comarca la media de la variable objetivo en la comunidad autónoma que la contiene; el riesgo de "suavizado" excesivo viene dado por una excesiva simplificación –omisión- en el número de variables que "explican" la especificidad del área dentro de la comunidad autónoma.

El riesgo de outliers presente en cualquier estimador, puede adquirir especial dimensión en los modelos sintéticos de pequeñas áreas, por la aparición de residuos grandes en la regresión. Determinados grupos de usuarios de estas estadísticas son especialmente vulnerables ante estos errores, que deben ser cuidadosamente controlados. Es el caso de aquellas áreas pequeñas a las que se dan valores inusuales de la variable objetivo de Y, pero que sin embargo presentan valor X o Z dentro de lo esperado en el dominio superior o en los dominios próximos. Por ejemplo, la estimación en un área de gran densidad, con residencias aparentemente secundarias (aunque de hecho se trate de principales), cuando se utilicen como variables explicativas del modelo las procedentes de un registro de población que no recoja bien estos fenómenos.

Un procedimiento, ya clásico, para evaluar un estimador basado en modelos es su comparación con el estimador directo en dominios con muestra suficiente, utilizando métodos de simulación bootstrap o con poblaciones artificiales. Entre los requisitos mínimos para evaluación que se imponen de manera natural en la elección de variables auxiliares y modelos para estimación en áreas pequeñas se encuentra el de que las estimaciones basadas en modelos deben ser consistentes respecto a estimadores insesgados directos, cuando haya muestra suficiente en el área. Además, es lógico aspirar a que un estimador combinado elimine una parte importante de la variabilidad entre áreas explicada por el modelo, por lo que es exigible que su error cuadrático medio sea sustancialmente menor que la variabilidad total del estimador directo en el área:

$$ECM\left(\hat{\bar{Y}}_d^{com}\right) << ECM\left(\hat{\bar{Y}}_d^{direct}\right)$$

A su vez, se deberá evaluar hasta qué punto la posible aportación al sesgo de la componente sintética del estimador combinado será compensada por la componente 'directa'. También se suele considerar como criterio de evaluación de la mejora que introduce el estimador con componente sintética, por razones obvias, la aportación o no de una mayor estabilidad temporal del estimador combinado del cambio, respecto a la que proporcionan las estimaciones directas.

7. El papel de una Oficina Central de Estadística en la estimación de áreas pequeñas

La demanda de estadísticas oficiales con un elevado detalle en la estimación en pequeños dominios plantea importantes disyuntivas en el proceso de decisión relativo a la producción/difusión sistemática de estimaciones en esos ámbitos, basadas o asistidas por modelos. Son varias las líneas de actuación necesarias para abordar esta demanda del sistema con criterios de calidad, algunas de ellas ya vigentes en los planes corrientes de los principales sistemas estadísticos internacionales y, en particular, en la Unión Europea y en España, por lo que es previsible que en el medio plazo, si se consiguen superar algunas incertidumbres –o si se encuentra satisfacción y estímulo en el otro lado de la balanza- veamos aparecer paulatinamente los estimadores para áreas pequeñas

basados en modelos, en el ámbito de la UE y, en concreto, en nuestro país.

En primer lugar, la aportación de la estadística oficial a la componente I+D de estimadores en áreas pequeñas debe ser suficientemente decidida e intensa, especialmente por la posibilidad de abastecer a la investigación con los datos primarios de encuestas, poblaciones reales o artificiales, esenciales para la evaluación de modelos y del rendimiento de los estimadores. Los modelos que contienen tiempo-espacio, la estimación de varianzas, el problema de la afijación óptima y la evaluación de modelos con diseños complejos o la consistencia de marginales en la estimación en áreas pequeñas, son campos en los que la teoría o la experimentación tienen todavía mucho que aportar, por lo que es de esperar que el camino ya iniciado tenga continuidad en los próximos años.

Es necesario también que la generalización de la estimación en áreas pequeñas en un país como España, en las que las administraciones territoriales están interesadas ó tienen competencias para producir estadísticas oficiales, se produzca sin despilfarro de recursos o pérdidas de calidad potencial. Cuando se trata de aplicar distintos modelos en distintos territorios no se debe olvidar que algunos modelos ganan fuerza cuanto mayor es el "grupo de regresión" donde se estiman los parámetros del modelo, por lo sería absurdo una auto limitación a priori de los objetivos de calidad. La incorporación de las mejores prácticas internacionales y la armonización de metodologías a nivel nacional son también parte importante del papel a jugar por una oficina central de estadística en este campo.

Un aspecto muy importante para analistas, productores eventuales de estimadores para áreas pequeñas, es la incorporación en las primeras fases del proyecto estadístico de una encuesta por muestreo del objetivo de estimación en pequeños dominios. Un ejemplo sería la afijación por comarcas, NUT IV, en el rediseño de la EPA aplicado en el primer trimestre de este año, a fin de disponer de una suficiente componente "directa" para estimadores combinados, acotando en lo posible el riesgo de sesgos incontrolados. Este tipo de afijaciones ya se han aplicado, por ejemplo, en la EPF por provincias, con el fin de que las estimaciones plurianuales a ese nivel (una forma de "estimador sintético" al fin y al cabo, ganando fuerza en el tiempo) permitan estimar la estructura del consumo para la base del IPC.

Precisamente, tiene mucho interés evaluar el impacto en la eficiencia general del diseño cuando se consideran las áreas pequeñas de una determinada tipología (zonas, distritos, subgrupos de empresas de determinada actividad, hogares de ciertas categorías socioeconómicas, etc.) como unidades de observación o "dominios a priori" en el proceso de afijación/selección de la muestra. Esta evaluación, incorporada en la fase inicial de un proyecto del órgano estadístico, constituye una aportación más de calidad al producto final.

Por supuesto, un papel importantísimo de los órganos centrales del sistema de estadísticas oficiales es asegurar la calidad de las fuentes administrativas potenciales suministradoras de las variables auxiliares de los modelos, alma mater de la eficiencia de los estimadores sintéticos. Todo esfuerzo en este sentido frente a las administraciones responsables del mantenimiento de las bases de datos o registros de educación, salud, prestaciones sociales, mercado de trabajo, renta, actividad económica, etc., es fundamental. La asimilación por parte de las distintas administraciones de la conveniencia de alcanzar una buena calidad del dato primario administrativo en áreas pequeñas, es básica en el proceso.

Es necesario potenciar la calidad de la información geo-referenciada a muy bajo nivel de detalle a fin de posibilitar la utilización de las bases de datos físicos en combinación con otros cualitativos y cuantitativos: sin ir más lejos, la presencia de una coordenada UTM en las bases de datos de la Seguridad Social y de la Agencia Tributaria serían de gran utilidad. La colaboración entre administraciones (central, autonómica) es esencial en la consecución de este objetivo y en general, en la mejora de las diversas fuentes administrativas con potencialidad para su utilización con fines estadísticos. Las propias operaciones exhaustivas del sistema estadístico (censos de población, padrón continuo, Movimiento natural de la Población, etc.) deben conceder la importancia que se merece al detalle geográfico y a la disponibilidad de la información para dominios pequeños.

A nivel más "doméstico" se plantea en las oficinas centrales de estadística el problema de la ubicación de la unidad administrativa encargada de la estimación en áreas pequeñas. ¿Debe ser una actividad centralizada, o cada servicio promotor –encuestas sociales, laborales, de empresas,…- debe disponer de su propia capacidad de producción de estimaciones en áreas pequeñas? Otros problemas, como la 'normalización ' de la difusión de estimaciones en áreas pequeñas, con presentación de información 'consistente' con las estadísticas corrientes, o el suministro de microdatos y metainformación con relevancia y transparencia constituyen también retos a resolver en el contexto de esta nueva tipología de datos.