12. Kalton G., Kordos J., Platek R. (1993-red), *Small Area Statistics and Survey Designs*. GUS, Warszawa
13. Klimanek T. (2003), *Wielopoziomowa analiza struktury agrarnej gminy w systemie GEOINFO*. Maszynopis pracy doktorskiej w Katedrze Statystyki Akademii Ekonomicznej w Poznaniu
14. Kordos J. (1997), *Efektywne wykorzystanie statystyki małych obszarów*. „Wiadomości Statystyczne", nr 1, s. 11—19
15. Kordos J., Paradysz J. (2000), *Some experiments in small area estimation in Poland*. „Statistics in Transition", Vol. 4, Nr 4, s. 679—697
16. Kordos J. (2002), *Nowy projekt zastosowania estymacji dla małych obszarów w krajach europejskich*. (w:) Paradysz J. (2002-red), *Statystyka regionalna w służbie samorządu lokalnego i biznesu*. Internetowa Oficyna Wydawnicza Centrum Statystyki Regionalnej, Poznań, s. 38—50
17. Lehtonen R., Veijanen A. (1999), *Domain Estimation with Logistic Generalized Regression and Related Estimators*. (w:) *Small Area Estimation, International Association of Survey Statisticians Satellite Conference Proceedings*. Ryga 20 i 21 sierpnia 1999 r., Łotwa
18. Levy P. S. (1979), *Small Area Estimation — Synthetic and Other Procedures, 1968—1978*. (w:) *Synthetic Estimates for Small Areas*. Pod red. C. Steinberg (National Institute on Drug Abuse, Raport nr 24), U. S. Government Printing Office, Waszyngton, s. 4—19
19. Paradysz J. (2001), *Estymacja dla małych obszarów w statystyce regionalnej*. (w:) *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*. Pod red. A. Zeliasia. Wydawnictwo Akademii Ekonomicznej w Krakowie, 2001, s. 193—209
20. Pfeffermann D. (1999), *Small Area Estimation — Big Developments*. (w:) *Small Area Estimation, International Association of Survey Statisticians Satellite Conference Proceedings*. Ryga 20 i 21 sierpnia 1999 r., Łotwa
21. Rao J. N. K. (2003), *Small area estimation. Wiley series in survey methodology*. Wiley — Interscience, New Jersey
22. de Santis G. (1989), *Un'analisi della fecondità in Italia nel 1967—81 con il metodo dei figli propri*. Dep. Statystyki Uniwersytetu we Florencji, Serie „Ricerche empiriche", nr 15
23. Särndal C. E., Hidiroglou M. A. (1989), *Small, Domain Estimation: A Conditional Analysis*. „Journal of the American Statistical Association", vol. 84, s. 266—275
24. Särndal C. E., Swensson B., Wretman J. (1992), *Model Assisted Survey Sampling*. Springer Verlag
25. Singh M. P., Gambino J., Mantel H. J. (1994), *Issues and Strategies for Small Area Data*. Survey Methodology, nr 1, s. 3—14

Jacek WESOŁOWSKI

Problemy estymacji dla małych obszarów

Przedstawiane uwagi są uzupełnieniem artykułu „Zasilanie statystyki regionalnej za pomocą estymacji dla małych obszarów w perspektywie wykorzystania rejestrów administracyjnych" autorstwa J. Paradysza. Oto najistotniejsze punkty omawianego zagadnienia.

1. Zadanie estymacji dla małych obszarów to:
   a) tworzenie „dobrych" estymatorów parametrów dla takich podpopulacji, w których badana próbka jest nieliczna (nawet zerowa);
   b) ocena błędu estymacji.
2. Ogólna idea poszukiwania rozwiązań — „pożyczanie informacji" z innych zbiorów danych:
   a) z „podobnych obszarów" z tego samego badania;
   b) z tego samego „obszaru" lub „podobnych obszarów" z innych badań, np. prowadzonych wcześniej.
3. Rozwiązania problemu estymacji dla małych obszarów proponowane w ramach metody reprezentacyjnej (*design based approach*) są niesatysfakcjonujące. Wykorzystują one różne rodzaje estymatorów:

a) estymatory bezpośrednie (proste, regresyjne, ilorazowe) o małym (zerowym) obciążeniu i dużej, najczęściej za dużej, wariancji;
b) estymatory syntetyczne (proste, regresyjne, ilorazowe) o małej wariancji i dużym, niemożliwym do określenia, obciążeniu. Proponowane w tym kontekście założenie jednorodności małych obszarów sprowadza sytuację do klasycznego paradygmatu „catch 22", który najlepiej widać na przykładzie estymacji średniej i odpowiedniego estymatora syntetycznego — średniej z całej próbki. Albo obszary są jednorodne i wtedy nie ma problemu estymacji w małych obszarach, ponieważ średnia w całej populacji jest równa średniej w małym obszarze, czyli pozornie stosujemy estymator syntetyczny, albo obszary nie są jednorodne i estymatora syntetycznego nie należy stosować, bo może mieć duże obciążenie;
c) estymatory złożone (*composite*), które są kombinacją wypukłą estymatorów bezpośredniego i syntetycznego. Mają one mniejsze obciążenie niż estymatory syntetyczne i mniejszą wariancję niż estymatory bezpośrednie. Jednocześnie mają większe obciążenie niż estymatory bezpośrednie i większy błąd średniokwadratowy niż estymatory syntetyczne. Co więcej, bez założeń modelowych nie ma sposobu oszacowania tego błędu i obciążenia.

4. Nieuniknione jest podejście modelowe (*model based approach*), a co za tym idzie sięgnięcie po aparat statystyki matematycznej. Często jest to aparat bardzo zaawansowany teoretycznie (np. odtwarzanie skomplikowanych rozkładów warunkowych) i obliczeniowo (np. próbnik Gibbsa). Takie podejście umożliwia konstrukcję efektywnych estymatorów, obliczanie wariancji czy konstrukcję przedziałów ufności. Podstawowa formalna konsekwencja podejścia modelowego to fakt, że parametr, traktowany w podejściu reprezentacyjnym jako nieznana liczba (wektor), przy podejściu modelowym jest wielkością losową (zmienną losową, wektorem losowym). Co za tym idzie, problem estymacji formalnie zostaje zamieniony w problem predykcji.
5. Stosowane w podejściu modelowym metody konstrukcji estymatorów zakwalifikować można do trzech ogólnych schematów:
   a) empiryczny najlepszy liniowy predyktor nieobciążony, tzw. EBLUP (*empirical best linear unbiased predictor*), tworzony w następujący sposób: w założonym modelu znajdowany jest najlepszy nieobciążony predyktor liniowy, przy założeniu że parametry modelu, jak i wariancje zmiennych losowych występujących w modelu są znane. Następnie obliczany jest najlepszy nieobciążony estymator liniowy, tzw. BLUE dla parametrów modelu, przy znanych wariancjach zmiennych losowych występujących w modelu i wstawiany w miejsce parametru modelu w otrzymanym wcześniej wzorze na predyktor. Na koniec wariancje zmiennych losowych występujących w modelu obliczane są na podstawie całej próbki, przy zastosowaniu wybranej standardowej metody statystycznej i wstawiane w odpowiedni, otrzymany wcześniej wzór na predyktor;
   b) empiryczny estymator bayesowski, tzw. EB (*empirical Bayes*), w którym założenia modelowe dotyczące rozkładów warunkowych pozwalają na znalezienie estymatora bayesowskiego, natomiast parametry rozkładów warunkowych (*a priori*) są estymowane z całej próbki i następnie wstawiane do otrzymanego wzoru na estymator bayesowski. W przypadku normalnych rozkładów a priori podejście EB daje takie wyniki jak EBLUP. W obu tych metodach są trudności z prawidłową oceną wariancji ze względu na brak możliwości oceny zmienności estymatorów parametrów rozkładów a priori;
   c) hierarchiczny estymator bayesowski, tzw. HB (*hierarchical Bayes*), w modelu, w którym parametry modelu są także zmiennymi losowymi z zadanymi rozkładami a priori (tzw. hiperparametry). Założenia modelowe prowadzą do znalezienia

odpowiednich dalszych rozkładów warunkowych. Tutaj mogą się kryć znaczące trudności matematyczne. Te rozkłady warunkowe są podstawą do obliczenia wartości estymatora oraz oceny wariancji tego estymatora za pomocą symulacji Monte Carlo odpowiedniego łańcucha Markowa, tzw. MCMC (*Monte Carlo Markov chain*). Należy podkreślić, że są to zaawansowane metody obliczeniowe, za którymi stoją głębokie twierdzenia rachunku prawdopodobieństwa. Stosowane powinny być bardzo ostrożnie, ze względu na problemy związane ze zbieżnością.

P R Z Y K Ł A D. Model efektów losowych na poziomie małego obszaru (*area level random effects model*). Mamy m małych obszarów indeksowanych liczbami $i = 1, ..., m$. Interesują nas parametry θ_i, $i = 1, ..., m$, np. średnie w małych obszarach. Dany jest również na poziomie małego obszaru wektor $\boldsymbol{x}_1 = (x_{i,1}, ..., x_{i,p})^T$ zmiennych pomocniczych, $i = 1, ..., m$. Zakładamy, że estymator bezpośredni $\tilde{\theta}_i$ ma postać:

$$\tilde{\theta}_i = \theta_i + e_i \qquad \theta_i = \boldsymbol{x}_i^T \boldsymbol{b} + u_i$$

gdzie e_i i u_i są zmiennymi losowymi, $i = 1, ..., m$, a $\boldsymbol{b}$ — nieznanym stałym wektorem;

a) w podejściu EBLUP zakłada się, że e_i ma średnią 0 i wariancję V_i (traktowana jest jako błąd próbkowania), u_i ma średnią 0 i wariancję τ^2 (traktowana jest jako modelowy efekt losowy), przy czym zmienne te są nieskorelowane, $i = 1, ..., m$, natomiast $\boldsymbol{b}$ jest nieznanym wektorem. Pierwszy krok polega na znalezieniu $BLUP_b(\theta_i)$, $i = 1, ..., m$, przy znanym $\boldsymbol{b}$. Otrzymuje się:

$$BLUP_b(\theta_i) = \gamma_i \tilde{\theta}_i + (1 - \gamma_i) \boldsymbol{x}_i^T \boldsymbol{b} \qquad \gamma_i = \frac{\tau^2}{\tau^2 + V_i} \qquad i = 1, ..., m$$

Następnie obliczany jest z całej próbki $BLUE(\boldsymbol{b}) = \hat{\boldsymbol{b}}$ parametru $\boldsymbol{b}$ i wstawiany do tego równania w miejsce $\boldsymbol{b}$. W ten sposób otrzymywany jest $BLUP(\theta_i)$, tzn.:

$$BLUP(\theta_i) = \gamma_i \tilde{\theta}_i + (1 - \gamma_i) \boldsymbol{x}_i^T \hat{\boldsymbol{b}} \qquad i = 1, ..., m$$

Kolejny krok polega na estymacji wariancji V_i oraz τ^2. Uzyskane estymatory prowadzą do estymowanych wartości wag $\hat{\gamma}_i$. Ostatecznie:

$$EBLUP(\theta_i) = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) \boldsymbol{x}_i^T \hat{\boldsymbol{b}} \qquad i = 1, ..., m$$

b) zakłada się, np. że:

i. $\tilde{\theta}_1, ..., \tilde{\theta}_m$ są warunkowo niezależne pod warunkiem $\theta = (\theta_1, ..., \theta_m)$;

ii. rozkład warunkowy:

$$\tilde{\theta}_i \mid \theta = \tilde{\theta}_i \mid \theta_i \sim N(\theta_i, V_i) \qquad i = 1, ..., m$$

iii. zmienne losowe $\theta_1, ..., \theta_m$ są niezależne i mają rozkłady normalne:

$$\theta_i \sim N(\boldsymbol{x}_i^T \boldsymbol{b}, \tau^2) \qquad i = 1, ..., m$$

iv. parametr modelu $\boldsymbol{b}$ oraz τ^2 są nieznane, natomiast wariancje V_i, $i = 1, ..., m$ są znane.

Wtedy znajduje się rozkład a posteriori, który jest rozkładem normalnym:

$$\theta_i | \tilde{\theta}_i \sim N[(1 - B_i)\tilde{\theta}_i + B_i \boldsymbol{x}_i^T \boldsymbol{b}, V_i(1 - B_i)]$$

gdzie $B_i = V_i \tau^2 + 1/V_i$, $i = 1, ..., m$. Zatem estymator bayesowski ma postać:

$$BE(\theta_i) = (1 - B_i)\tilde{\theta}_i + B_i \boldsymbol{x}_i^T \boldsymbol{b}$$

i wariancję $V_i(1 - B_i)$. Nieznane parametry $\boldsymbol{b}$ i τ^2 trzeba estymować. W tym celu stosuje się np. iteracyjne rozwiązywanie następującego układu równań:

$$\boldsymbol{b} = (\boldsymbol{X}^T \boldsymbol{D}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{D}^{-1} \tilde{\theta}$$

$$\sum_{i=1}^{m} (\tilde{\theta}_i - \boldsymbol{x}_i^T \boldsymbol{b})^2 / (\tau^2 + V_i) = m - p$$

gdzie: $\tilde{\theta} = (\tilde{\theta}_1, ..., \tilde{\theta}_m)^T$,

$\boldsymbol{X} = (\boldsymbol{x}_1, ..., x_m)^T$,

$D = diag(V_1 + \tau^2, ..., V_m + \tau^2)$.

Własności tak otrzymanych estymatorów $\boldsymbol{b}$ i τ^2 są trudne do analizowania, w szczególności ich wariancje. Wartości tych estymatorów wstawiane są w miejsca $\boldsymbol{b}$ i τ^2 w estymatorze *BE*, dając empiryczny estymator bayesowski:

$$EBE(\theta_i) = (1 - \hat{B}_i)\tilde{\theta}_i + \hat{B}_i \boldsymbol{x}_i^T \hat{\boldsymbol{b}}$$

gdzie: $\hat{B}_i = V_i \hat{\tau}^2 + 1/V_i$,

$i = 1, ..., m$.

Podobnie aktualizowana jest ocena wariancji $V_i(1 - \hat{B}_i)$, czego wynikiem jest pominięcie kwestii zmienności estymatorów $\hat{\boldsymbol{b}}$ i $\hat{\tau}^2$ w szacunku wariancji estymatora $EBE(\theta_i)$, $i = 1, ..., m$;

c) w podejściu HB zakłada się, że:

i. zmienne losowe $\tilde{\theta}_1, ..., \tilde{\theta}_m$ są warunkowo niezależne pod warunkiem $(\theta, \boldsymbol{b}, \tau^2)$, przy czym rozkład warunkowy jest normalny:

ii. zmienne losowe $\theta_1, ..., \theta_m$ są warunkowo niezależne pod warunkiem $(\boldsymbol{b}, \tau^2)$, przy czym rozkład warunkowy jest normalny:

$$\theta_i|\boldsymbol{b},\tau^2 \sim N(\boldsymbol{x}_i^T\boldsymbol{b},\tau^2) \qquad i=1,...,m$$

iii. zmienne losowe $\boldsymbol{b}$ i τ^2 są niezależne i mają rozkłady niewłaściwe: jednostajne na $\boldsymbol{R}^p$ oraz $(0,\infty)$, odpowiednio.

Celem jest znalezienie średniej i wariancji a posteriori w rozkładzie warunkowym $\theta_i|\widetilde{\theta}_i$. Gęstość tego rozkładu daje się napisać i można próbować zastosować procedurę EB do wzorów przybliżonych na $E(\theta_i \mid \widetilde{\theta})$.

Efektywniejsze jest podejście HB, które polega na wyliczeniu rozkładów warunkowych potrzebnych do zastosowania próbnika Gibbsa, a następnie przy pomocy metody MCMC symulowanie rozkładu łącznego. W tym przypadku są to następujące rozkłady warunkowe:

i. p-wymiarowy rozkład normalny:

$$\boldsymbol{b}|\theta,\tau^2,\widetilde{\theta} \sim N_p[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\theta,\tau^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}]$$

ii. odwrotny rozkład Gaussowski:

$$\tau^2|\theta,\boldsymbol{b},\widetilde{\theta} \sim IG\left(\frac{1}{2}\sum_{i=1}^{m}(\theta_i - x_i^T\boldsymbol{b})^2, \frac{m-2}{2}\right)$$

iii. m rozkładów jednowymiarowych normalnych, ponieważ zmienne $\theta_1, ..., \theta_m$ są warunkowo niezależne pod warunkiem $(\boldsymbol{b},\tau^2,\widetilde{\theta})$:

$$\theta_i\Big|\boldsymbol{b},\ \tau^2,\ \widetilde{\theta} \sim N[(V_i^{-1}+\tau^{-2})^{-1}(V_i^{-1}\widetilde{\theta}_i + \tau^{-2}\boldsymbol{x}_i^T\boldsymbol{b}),(V_i^{-1}+\tau^{-2})^{-1}] \qquad i=1,...,m$$

Symulacje za pomocą próbnika Gibbsa pozwalają na empiryczne odtworzenie warunkowej wartości oczekiwanej i warunkowej wariancji θ_i pod warunkiem $\widetilde{\theta}$, co daje wartość estymatora uzyskanego metodą HB i jego wariancji.

Możliwe jest również rozwiązanie zadania metodą HB w sytuacji, gdy parametry V_i, $i=1,...,m$ nie są znane. Wtedy jednak model znacznie się komplikuje. Pojawia się więcej rozkładów warunkowych, bo parametry $V_1, ..., V_m$ traktowane są jako zmienne losowe.

6. Niektóre modele rozważane w literaturze:
   a) model efektów losowych na poziomie obszaru (*area level random effect model*);
   b) model regresyjny z zagnieżdżonym błędem na poziomie jednostkowym (*nested error unit level regression model*);

c) uogólnione modele liniowe z efektami losowymi (*generalized linear models (GLM) with random effects*);
d) modele szeregów czasowych, np. model losowych współczynników regresji, model autoregresyjny pierwszego rzędu.

7. Trudności w podejściu modelowym:
a) możliwość niezgodności między modelem populacyjnym a modelem, który odpowiada próbce. W takiej sytuacji potrzebna jest adiustacja modelu bądź otrzymywanych estymatorów. Problematyka ta mieści się w dziedzinie zwanej próbkowaniem informatywnym (*informative sampling*);
b) problem wyboru modelu (brak zadowalających propozycji w tej dziedzinie);
c) problem weryfikacji modelu. Czynione są pewne próby konstruowania testów zgodności (*goodness-of-fit*), głównie opierają się one na teście χ^2.

Podsumowanie

1. Estymacja dla małych obszarów jest zagadnieniem ważnym, któremu warto poświęcić wiele czasu i środków ze względu na rosnące potrzeby w tej dziedzinie.
2. Estymacja dla małych obszarów jest dziedziną ciekawą i trudną teoretycznie, wymagającą znacznej fachowej wiedzy w dziedzinie teorii rozkładów wielowymiarowych i statystyki matematycznej, w szczególności statystyki bayesowskiej.
3. Estymacja dla małych obszarów jest dziedziną zasadniczo różną od metody reprezentacyjnej ze względu na podejście modelowe. W przypadku próbkowania informatywnego obie te dziedziny nakładają się i oba podejścia są konieczne do wypracowania dobrych metod estymacji.
4. Estymacja dla małych obszarów wymaga korzystania z danych pomocniczych wysokiej jakości, czyli ma szansę przynieść dobre efekty w konkretnych badaniach, tylko gdy dostępne są odpowiednie dane z rejestrów administracyjnych.
5. W estymacji dla małych obszarów stosowanie naiwnych estymatorów: bezpośrednich, syntetycznych i złożonych jest niewskazane. Należy wypracowywać estymatory na podstawie założeń modelowych.
6. Estymacja dla małych obszarów wymaga intensywnych obliczeń komputerowych, szczególnie w ramach metod MCMC.
7. Słabą stroną estymacji w małych obszarach są trudności w wyborze i weryfikacji modelu.

dr hab. inż. Jacek Wesołowski — *Politechnika Warszawska i GUS*

LITERATURA

1. Ghosh, M. (2001), *Model-dependent small area estimation — theory and practice.* W: *Lecture Notes on Estimation for Population Domains and Small Areas* (R. Lehtonen, K. Djerf, eds), s. 51—108
2. Ghosh, M., Rao, J.N.K. (1994), *Small area estimation: an appraisal.* Statistical Science No 9, s. 55—93
3. Pfeffermann, D. (2002), *Small area estimation — new developments and directions.* International Statistical Review No 70, s. 125—143
4. Pfeffermann, D. (2002), *Modelling of complex survey data under informative sampling.* Materiały Baltic-Nordic Conference on Survey Sampling, s. 1—24
5. Rao, J.N.K. (2002), *Small area estimation.* Materiały Baltic-Nordic Conference on Survey Sampling, s. 1—18
6. Rao, J.N.K. (2003), *Small area estimation,* Wiley Europe