Anna MAŁASIEWICZ

Teoria i praktyka statystyki małych obszarów

Działalność władz lokalnych, różnych instytucji i podmiotów gospodarczych uzależniona jest od posiadania kompleksowej informacji z różnych obszarów i dziedzin życia. Powoduje to, że z jednej strony niezbędna staje się informacja dostępna na niskim poziomie agregacji przestrzennej (np. gmin) czy też dla bardziej szczegółowych domen, z drugiej zaś strony uzyskanie tego typu informacji uniemożliwiają rosnące koszty badań, ograniczenia czasowe czy obciążenia respondentów.

Te pozornie sprzeczne cele, jakimi są niskie koszty badań przy jednoczesnym zapewnieniu wysokiej jakości informacji, mogą być osiągnięte dzięki zastosowaniu najnowszych osiągnięć metodologicznych z zakresu statystyki małych obszarów (SMO). Określenia „mały obszar" nie należy rozumieć dosłownie. Termin ten odnosić się może do domeny czy subpopulacji, dla których wielkość próby jest na tyle mała, że estymacja bezpośrednia może być nieefektywna. Metody SMO odgrywają współcześnie istotną rolę w kształtowaniu techniki uzyskiwania informacji. Dzięki swoim własnościom umożliwiają one uzyskanie wiarygodnych szacunków na niższych poziomach agregacji i bardziej szczegółowych domen, w sytuacji niewielkiej liczebności próby lub nawet braku obserwacji w próbie dla danego przekroju. Ich znaczenie nieustannie wzrasta, co przy postępującej informatyzacji, gromadzeniu coraz większych wolumenów danych z różnych źródeł (spisy, rejestry administracyjne czy badania reprezentacyjne) daje szerokie spektrum zastosowania tej statystyki.

ESTYMATORY STATYSTYKI MAŁYCH OBSZARÓW

W przypadku gdy liczebność próby w badaniu reprezentacyjnym w odpowiednio zdefiniowanych domenach jest duża, można wykorzystać estymatory bezpośrednie. Problem pojawia się w sytuacji, gdy domena jest reprezentowana w próbie przez niewielką liczbę jednostek, a nawet brak jest reprezentantów. W takich przypadkach stosowanie estymatora bezpośredniego obarczone jest zbyt dużym błędem, co związane jest zazwyczaj z jego większą wariancją. W konsekwencji, nie chcąc zwiększać kosztów badania przez zwiększanie próby, można rozważyć użycie estymacji pośredniej. Jej wykorzystanie może zapewnić zmniejszenie wariancji stosowanych estymatorów poprzez użycie informacji z innych domen i danych z innych źródeł, np. rejestrów administracyjnych czy spisów.

W konstrukcji estymatorów SMO można korzystać tylko z wyników przeprowadzonego badania lub też z informacji dodatkowych, np. z rejestrów administracyjnych. W zależności od przyjętego sposobu postępowania wnioskowanie może wykorzystywać jedynie schemat losowania (*design-based*), być wspomagane modelem (*model-assisted*) lub też być całkowicie oparte na modelu (*model-based*).

Wnioskowanie wykorzystujące schemat losowania całkowicie opiera się na prawdopodobieństwie dostania się jednostek do próby zgodnie z określonym schematem losowania. Do narzędzi wnioskowania tego typu można zaliczyć m.in. estymator bezpośredni, będący podstawą metodologii SMO i stanowiący punkt odniesienia dla innych estymatorów z tej statystyki. Rozszerzeniem tego podejścia jest **wnioskowanie wspomagane modelem**. Wyróżnia się w tym przypadku estymatory, które nie tylko wykorzystują informacje opierające się na strukturze prawdopodobieństwa inkluzji, ale również wynikają z modelu, korzystając z dodatkowych informacji. W grupie tej znajdują się m.in. estymator regresyjny i złożony.

Najbardziej rozwiniętym jest **wnioskowanie oparte na modelu**. Wyróżnia się takie estymatory, które na szeroką skalę wykorzystują odpowiednio skonstruowany model. Rozważany model statystyczny może opisywać powiązanie badanej zmiennej losowej z jedną lub większą liczbą zmiennych losowych. Można zaliczyć tutaj przede wszystkim estymatory wykorzystujące modele Faya-Herriota oraz Poissona (empiryczne oraz hierarchiczne estymatory bayesowskie).

Estymacja bezpośrednia

Jak wspomniano wcześniej, podstawowym estymatorem SMO jest estymator bezpośredni. Często jest on traktowany również jako punkt odniesienia w porównywaniu efektywności innych metod. Estymacja bezpośrednia opiera się jedynie na informacjach pochodzących z badanego obszaru, choć dopuszcza się w pewnych przypadkach użycie informacji dodatkowych (np. uogólniony estymator regresyjny GREG).

Klasycznym estymatorem bezpośrednim (zwanym inaczej ekspansyjnym) wartości globalnej w obszarze $i = 1, 2, \ldots, m$ jest suma ważona obserwacji pochodzących z próby s wielkości $n = \Sigma_i n_i$, a zatem ma on postać:

$$\hat{y}_i = \sum_{j \in s_i} w_{ij} y_{ij} \qquad i = 1, 2, \ldots, m \qquad j = 1, 2 \ldots, n_i \tag{1}$$

gdzie:

w_{ij} — wagi obserwacji, zależne od schematu losowania, części próby s_i należącej do i-tej domeny i elementu j danej części próby,

$\hat{y}_i$ — szacowana wartość globalna dla i-tej domeny,

y_{ij} — wartość dla j-tej jednostki w próbie.

Przyjmując za π_j prawdopodobieństwo wystąpienia j-tej jednostki w próbie, wagi mogą mieć postać $w_{ij} = \pi_{ij}^{-1}$, wtedy estymator bezpośredni zwany jest estymatorem Horvitza-Thompsona. Nieujemnym nieobciążonym estymatorem wariancji dla estymatora bezpośredniego wartości globalnej w przypadku stałej wielkości próby jest:

$$var(\hat{y}_i) = -\sum_{j<k} \sum_{j,k \in s_i} w_{ijk}\pi_{ij}\pi_{ik}\left(\frac{y_{ij}}{\pi_{ij}} - \frac{y_{ik}}{\pi_{ik}}\right)^2 \qquad (2)$$

gdzie:

$i = 1, 2, ..., m$; $\quad j < k = 1, 2, ..., n_i$; $\quad w_{ijk} = \dfrac{\pi_{ijk} - \pi_{ij}\pi_{ik}}{\pi_{ijk}\pi_{ij}\pi_{ik}}$; $\quad \pi_{ijk} = \sum_{s_i : j,k \in s_i} p(s_i)$

$p(s_i)$ — prawdopodobieństwo wyboru próby s_i.

Estymator bezpośredni jest nieobciążony i efektywny w przypadku odpowiedniej wielkości próby. Jednak w badaniach statystyki publicznej zdarzają się także przypadki braku jednostek w próbie dla obszaru lub domeny. Jednym ze sposobów wnioskowania o zbiorowości w takiej sytuacji jest wykorzystanie SMO.

Wykorzystanie zmiennych pomocniczych — estymacja regresyjna

Metody bezpośrednie korzystają jedynie z danych pochodzących z rozważanej domeny. W wyniku prac nad poprawą własności estymatorów z tej grupy, jak np. redukcji dużej wariancji w przypadku niewystarczającej wielkości próby, powstały bardziej złożone metody zwane metodami pośrednimi. Wykorzystują one, poza danymi dostępnymi z określonej domeny, informacje z innych obszarów i mogą łączyć dane ze spisów powszechnych, z badań próbkowych czy też rejestrów administracyjnych.

Przykładem wykorzystania zależności między badaną zmienną y a jedną lub większą liczbą zmiennych pomocniczych jest **estymacja regresyjna**. Załóżmy, że dla domeny i dostępne są zmienne, których znane są wartości globalne dla rozważanego małego obszaru $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, gdzie $x_{ik} = (x_{ik1}, \ldots, x_{ikn_i})^T$, $k = 1, 2, \ldots, p$ są wektorami zawierającymi informacje dodatkowe. Ze względu na dowolność p dla ułatwienia opisów w dalszej części artykułu przyjęto $X_i = (x_{i1}, \ldots, x_{in_i})^T$, tzn. $k = 1$.

Postać ogólnych estymatorów regresyjnych przedstawia się następująco:

$$\hat{y}_i^{reg} = \hat{y}_i + \left(X_i - \hat{X}_i\right)^T \hat{B}_i \tag{3}$$

gdzie:

$\hat{y}_i^{reg}$ — estymator regresyjny wartości globalnej w domenie $i = 1, 2, \ldots, m$,
$\hat{y}_i$ — estymator bezpośredni w domenie i,
$\hat{B}_i$ — rozwiązanie równania:

$$\left(\sum_{j \in s} w_{ij} x_{ij} x_{ij}^T\right) \hat{B}_i = \sum_{j \in s} w_{ij} x_{ij} y_{ij}$$

X_i — macierz znanych wartości globalnych zmiennych pomocniczych oraz $\hat{X}_i$ wyraża się wzorem:

$$\hat{X}_i = \sum_{j \in s} w_{ij} x_{ij} \tag{4}$$

gdzie:

w_{ij} — waga j-tej jednostki w i-tym obszarze, określonymi jak we wzorze (1),
x_{ij} — wektory zawierające informacje dodatkowe.

Estymator ten jest w przybliżeniu nieobciążony dla dużej próby. Wzór (3) opisuje w rzeczywistości uogólniony estymator regresyjny, będący estymatorem bezpośrednim używającym informacji dodatkowych.

Szczególnym przypadkiem uogólnionych estymatorów są estymatory regresyjne i ilorazowe. Wymienione dalej estymatory należą do grupy estymatorów syntetycznych, tzn. takich, które mogą czerpać informacje również spoza domeny, przy założeniu podobieństwa między małym obszarem i większym, zawierającym w sobie ten pierwszy.

Estymator regresyjny syntetyczny wartości globalnej ma postać:

$$\hat{y}_i^{reg,\ synt} = \beta X_i \tag{5}$$

gdzie β — wektor współczynników regresji obliczany analogicznie do $\hat{\beta}_i$ dla większego obszaru.

Estymator będzie efektywny, gdy mały obszar nie wykazuje silnego działania jednostek w odniesieniu do współczynników regresji. Po odpowiednim przekształceniu estymatora syntetycznego regresyjnego można otrzymać **estymator syntetyczny ilorazowy** wartości globalnej w postaci:

$$\hat{y}_i^{ilor,\,synt} = \frac{\hat{y}}{\hat{X}} X_i \tag{6}$$

gdzie:

$\hat{y}$ i $\hat{X}$ — estymatory wartości globalnych dla większego obszaru, obliczone zgodnie ze wzorami (1) i (4) bez subskryptu *i*, gdzie *j* = 1, 2, …, *n*,
n — liczebność większego obszaru.

Estymator syntetyczny regresyjny (5) będzie miał małe obciążenie, gdy wektor współczynników regresji β_i obliczony jedynie dla domeny będzie bliski wektorowi populacyjnemu. Analogicznie — estymator syntetyczny ilorazowy (6) — gdy iloraz $\hat{y}_i \hat{X}_i^{-1}$ obliczony jedynie dla domeny będzie bliski ilorazowi populacyjnemu. Ważną cechą tych estymatorów jest fakt „sumowania się" do szacunków na wyższym poziomie agregacji (to znaczy dla większego obszaru zawierającego rozważane małe obszary), gdzie oszacowania uważa się już za wiarygodne.

Wykorzystanie zmiennych pomocniczych — estymacja złożona

Kolejnym krokiem w polepszaniu precyzji oszacowań jest estymacja złożona. Jest to technika pozwalająca na redukcję wariancji całkowitej lub zbalansowanie potencjalnego obciążenia estymatora syntetycznego $\hat{y}_i^{synt}$ przeciw prawdopodobnej niestabilności estymatora bezpośredniego $\hat{y}_i^{bezp}$ (Pfeffermann, 2013).

Estymator złożony wartości globalnej stanowi kombinację liniową wspomnianych estymatorów — bezpośredniego i syntetycznego, a zatem:

$$\hat{y}_i^{\,złoż} = \gamma_i \hat{y}_i^{bezp} + (1-\gamma_i)\hat{y}_i^{synt} \tag{7}$$

gdzie γ_i — odpowiednio dobrana waga, która określa udział każdego ze składników w końcowym oszacowaniu.

Sposób ustalania wag γ_i może być różnorodny (Rao, 2003). Wartość wagi może być nadana z góry i być uzależniona od zmienności cechy dodatkowej, od liczebności populacji w domenach lub od wielkości próby. Może być również zależna od średniego błędu kwadratowego estymatorów składowych. Waga po-

winna być jednak dobrana w taki sposób, aby obciążenie utrzymane było na odpowiednim poziomie. Przykładem wagi nadanej *a priori* jest $\gamma_i = \frac{1}{2}$, a zatem wynikowy estymator złożony jest średnią arytmetyczną estymatorów bezpośredniego i syntetycznego.

Innym podejściem jest uzależnienie wagi od wielkości próby lub populacji czy też od wartości zmiennej dodatkowej X_i lub $\hat{X}_i$. Ustalono je tak, żeby w domenach, w których oczekiwana wielkość próby jest na tyle duża, estymator bezpośredni spełniał wymogi wiarygodności. W takim przypadku większa waga przypisywana jest estymatorowi bezpośredniemu. Waga spełniająca ten warunek może mieć postać:

$$\gamma_i = \begin{cases} 1 & \text{gdy } \hat{N}_i / N_i \geq \delta \\ \dfrac{\delta^{-1}\hat{N}_i}{N_i} & \text{w przeciwnym przypadku} \end{cases} \tag{8}$$

gdzie:

δ — subiektywnie wybrana wielkość dobrana tak, aby kontrolować wpływ estymatora syntetycznego oraz $\hat{N}_i = \sum_{s_i} w_{ij}$,

N_i — liczebność populacji *i*-tej domeny.

Rozważając wartości zmiennej dodatkowej zamiast wielkości populacji otrzymujemy wagi analogiczne do poprzednich, powstające przez zastąpienie ilorazów $\hat{N}_i / N_i$ przez $\hat{X}_i / X_i$. Wielkość δ można wybrać dowolnie, na ogół stosowane jest δ=1.

Inną stosowaną powszechnie postacią wag zależnych od wielkości populacji jest:

$$\gamma_i = \begin{cases} 1 & \text{gdy } \hat{N}_i / N_i \geq 1 \\ \left(\hat{N}_i / N_i\right)^{h-1} & \text{w przeciwnym przypadku} \end{cases} \tag{9}$$

dla wybranego *h* (na ogół stosowane jest *h* = 2).

Wadą stosowania wag zależnych od wielkości próby i populacji jest nieuwzględnianie zróżnicowania obszarowego w ramach zmienności obszarowej dla interesującej nas zmiennej, tzn. wszystkie cechy otrzymują te same wagi bez względu na ich niejednorodność między obszarami.

Nadawanie wag na podstawie danych opiera się na założeniu, że wagi zależne są od błędu średniokwadratowego estymatorów (MSE) i ich kowariancji. Możliwe jest znalezienie estymatora optymalnego, w którym wagi dobrane są w taki sposób, by minimalizować całkowitą wartość MSE, tzn. MSE ($\hat{y}_i^{złoż}$). Optymalna waga (Rao, 2003) ma postać:

$$\gamma_i^{opt*} = \frac{\mathrm{MSE}\left(\hat{y}_i^{synt}\right) - \mathrm{E}\left(\hat{y}_i^{bezp} - y_i\right)\left(\hat{y}_i^{synt} - y_i\right)}{\mathrm{MSE}\left(\hat{y}_i^{bezp}\right) + \mathrm{MSE}\left(\hat{y}_i^{synt}\right) + 2\mathrm{E}\left(\hat{y}_i^{bezp} - y_i\right)\left(\hat{y}_i^{synt} - y_i\right)} \tag{10}$$

co przy założeniu, że wartość oczekiwana iloczynu zmiennych losowych $\mathrm{E}\left(\hat{y}_i^{bezp} - y_i\right)\left(\hat{y}_i^{synt} - y_i\right)$ jest mała w porównaniu z $\mathrm{MSE}\left(\hat{y}_i^{synt}\right)$ można zapisać jako:

$$\gamma_i^{opt} = \frac{\mathrm{MSE}\left(\hat{y}_i^{synt}\right)}{\mathrm{MSE}\left(\hat{y}_i^{bezp}\right) + \mathrm{MSE}\left(\hat{y}_i^{synt}\right)} + \frac{1}{1 + F_i} \tag{11}$$

gdzie $F_i = \mathrm{MSE}\left(\hat{y}_i^{bezp}\right) / \mathrm{MSE}\left(\hat{y}_i^{synt}\right)$ jest ilorazem MSE składników bezpośredniego i syntetycznego. W praktyce wagę estymuje się na podstawie danych.

Składnik syntetyczny występujący we wzorze na estymator złożony może być dowolnie wybranym estymatorem syntetycznym. Korzystając z wymienionych wcześniej estymatorów syntetycznych można otrzymać m.in. następujące estymatory złożone oparte na estymatorach:

- Horvitza-Thompsona i syntetycznym regresyjnym:

$$\hat{y}_i^{złoż} = \gamma_i \hat{y}_i^{bezp} + (1 - \gamma_i)\hat{y}_i^{reg} = \gamma_i \sum_{j \in s_i} w_{ij} y_{ij} + (1 - \gamma_i)\beta X_i \tag{12}$$

- Horvitza-Thompsona i syntetycznym ilorazowym:

$$\hat{y}_i^{złoż} = \gamma_i \hat{y}_i^{bezp} + (1 - \gamma_i)\hat{y}_i^{ilor} = \gamma_i \sum_{j \in s_i} w_{ij} y_{ij} + (1 - \gamma_i)\frac{\hat{y}}{\hat{X}} X_i \tag{13}$$

Estymacja oparta na modelu

Metody estymacji pośredniej, takie jak regresyjna czy syntetyczna, wykorzystują domyślne modele wiążące małe obszary z danymi uzupełniającymi. W estymacji opartej na modelu tworzone są jawne modele angażujące dodatkowo w opisie badanego zjawiska zmienność między obszarami. Badane modele można podzielić na dwie główne kategorie:

1) modele poziomu zagregowanego (obszarowego), odnoszące się do małych obszarów i korzystające ze zmiennych dodatkowych poziomu obszaru, a zatem użyteczne w przypadku niedostępności danych indywidualnych,
2) modele poziomu jednostkowego, które oparte są na wartościach zmiennych objaśniających poziom jednostki.

Popularnym modelem — w szczególności w Stanach Zjednoczonych — jest **model Faya-Herriota**. Jest to prosty model poziomu obszaru mający dobre właściwości empiryczne. Jedną z jego zalet jest to, że model może opierać się jedynie na danych zagregowanych bez potrzeby użycia danych jednostkowych. Zastosowanie tej metody prowadzi do projektowo zgodnego estymatora (*design consistent estimator*), własności sprawiającej, że estymator oparty na modelu jest bliższy bezpośredniemu oszacowaniu dla dużej próbki, niezależnie od rzeczywistego modelu.

Modelem opierającym się na danych jednostkowych jest model poziomu jednostki zaproponowany po raz pierwszy przez Battesego, Hartera i Fullera. Ze względu na podobieństwo w konstrukcji do modelu Faya-Herriota często nazywa się go modelem Faya-Herriota poziomu jednostki.

Niech μ_i oznacza szukaną charakterystykę dla *i*-tej domeny, będącą funkcją wartości zmiennej *y*. Użycie modelu poziomu jednostki wymaga, aby znane były średnie obszarowe dla wszystkich domen $\overline{X}_i = \sum_{j=1}^{N_i} x_{ij} / N_i$. Model poziomu jednostki ma postać:

$$y_{ij} = x_{ij}^T \beta + v_i + \varepsilon_{ij} \tag{14}$$

gdzie:

y_{ij} — wartość badanej zmiennej dla *j*-tej jednostki w obszarze *i*,

v_i — (efekt obszaru *i*) oraz ε_{ij} (błędy losowe) — wzajemnie niezależne ze średnią zero i wariancjami σ_v^2 i $\sigma_{\varepsilon,\,i}^2$ odpowiednio,

x_{ij}^T — macierz zmiennych objaśniających,

β — wektor współczynników regresji.

Często szukaną charakterystyką danej domeny jest średnia małego obszaru. Przyjmując model (14) prawdziwą wartością średnią dla małego obszaru *i* jest $\overline{Y}_i = \overline{X}_i^T \beta + v_i + \overline{\varepsilon}_i$. Przy takich założeniach oraz dla dużych N_i mamy $\overline{\varepsilon}_i \cong 0$, stąd szukana wartość globalna dla domeny *i* często definiowana jest jako $\mu_i = \overline{X}_i^T \beta + v_i$.

Estymatorem wartości globalnej dla domeny *i* opartym na modelu będzie średnia ważona uogólnionego estymatora regresyjnego i syntetycznego regresyjnego, a zatem:

$$\hat{\mu}_i = \gamma_i \left[\bar{y}_i + \left(\bar{X}_i - \bar{x}_i \right)^T \beta \right] + (1 - \gamma_i) \bar{X}_i \hat{\beta} \tag{15}$$

gdzie:

$\hat{\mu}_i$ — estymator wartości globalnej szukanej charakterystyki dla *i*-tej domeny,

$\hat{\beta}$ — oszacowanie wektora współczynników regresji na podstawie przyjętego modelu obliczone ze wszystkich obserwacji,

$\bar{y}_i$ i $\bar{x}_i$ — wartości średnie obserwowanych y_{ij} i x_{ij} po obszarze *i*, tj. $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$, $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij} / n_i$, odpowiednio, gdzie n_i jest wielkością próby obszaru *i*. Waga γ_i wyrażona jest wzorem $\gamma_i = \sigma_v^2 \left(\sigma_v^2 + \sigma_{\varepsilon,i}^2 \right)^{-1}$, gdzie σ_v^2 jest wariancją efektu obszarowego, a $\sigma_{\varepsilon,i}^2 = \sigma_\varepsilon^2 / n_i$ jest wariancją błędu losowego w obszarze *i*.

Widać zatem, że estymator oparty na modelu jest estymatorem złożonym postaci (7), zbudowanym tak, by minimalizować całkowitą wariancję. Dla obszaru *k* bez jednostki w próbie, ze względu na znaną średnią $\bar{X}_k$, estymatorem jest $\hat{\mu}_k = \bar{X}_k \hat{\beta}$.

Model Faya-Herriota poziomu obszaru ma postać:

$$y_i = x_i^T \beta + v_i + e_i \tag{16}$$

a szukaną wartością globalną w domenie *i* jest $\mu_i = x_i^T \beta + v_i$.

Związany z modelem (16) estymator jest średnią ważoną estymatorów bezpośredniego i syntetycznego regresyjnego, a zatem klasycznym estymatorem złożonym o wagach dobranych tak, by minimalizować całkowitą wariancję:

$$\hat{\mu}_i = \gamma_i \hat{y}_i + (1 - \gamma_i) x_i^T \hat{\beta} \tag{17}$$

gdzie:

$\hat{\mu}_i$ — estymator wartości globalnej w domenie *i*,

$\hat{y}_i$ — estymator bezpośredni wartości globalnej w domenie *i*,

x_i — wektor znanych wartości oraz $\hat{\beta}$ oszacowaniem wektora współczynników regresji na podstawie przyjętego modelu.

Oszacowanie błędu średniokwadratowego estymatorów opartych na modelu Faya-Herriota składa się z części odpowiadającej za błąd losowy oraz za zmienność szacowanego wektora współczynników regresji w modelu. Zależy ono również od przyjętej metody szacowania wariancji będących częścią wag. Kompleksowe informacje na ten temat można znaleźć w opracowaniu J. N. K. Rao (2003).

Kolejnym przykładem estymacji opartej na modelu jest szacowanie na podstawie **modelu Poissona**. W modelu tym przyjęte jest założenie, że jeśli wielkość próby w domenie $n_i > 0$, to liczba osób z pewną cechą w obszarze spełnia rozkład Poissona z parametrem λ_i. Dodatkowo zakładana jest znajomość pewnej funkcji g(·), takiej że:

$$g(\lambda_i) = x_i^T \beta + v_i \tag{18}$$

gdzie:

λ_i — nieznany parametr rozkładu Poissona,
x_i — macierz zmiennych objaśniających,
β — wektor współczynników regresji oraz v_i efektami obszaru z wariancjami σ_v^2.

Załóżmy, że liczba ludności $\hat{N}_i$ z pewną cechą w domenie ze zmiennymi objaśniającymi jest związana relacją postaci:

$$\log\left(\hat{N}_i\right) = \sum_j \beta_j x_{ij} \tag{19}$$

naturalne zatem jest przyjęcie, że funkcja g(·) jest logarytmem, $g(\lambda_i)=\log(\lambda_i)$. Ostatecznie estymator liczby osób z badaną cechą oparty na modelu Poissona ma postać:

$$\hat{N}_i = \exp\left(x_i^T \hat{\beta} + \hat{v}_i\right) \tag{20}$$

dla jednostek biorących udział w dopasowaniu modelu, a zatem w przypadku, gdy $n_i > 0$. Gdy liczebność próby w domenie nie spełnia tego warunku, a zatem w domenie brak reprezentanta, estymatorem będzie:

$$\hat{N}_i = \exp\left(x_i^T \hat{\beta}\right) \tag{21}$$

SMO oraz wykorzystanie alternatywnych źródeł danych są przykładem dziedzin, w których procedury oparte na modelu stanowią cenny wkład

w tworzenie oficjalnej krajowej statystyki na świecie. Powodzenie każdej metody modelowej zależy jednak od dostępności dobrych danych pomocniczych. Należy zwrócić uwagę na dobór takich zmiennych dodatkowych, które są dobrymi wskaźnikami badanych zmiennych, zatem istotne jest angażowanie specjalistów w danej dziedzinie, którzy powinni mieć wpływ na wybór właściwych cech.

STOSOWANIE SMO

SMO zaczyna odgrywać coraz większą rolę w działalności krajowych urzędów statystycznych i wielu instytucji międzynarodowych — może być stosowana w wielu dziedzinach. Obecnie stosowana jest głównie poza Polską. SMO pozwala na otrzymywanie aktualnych i szczegółowych informacji dla niskich poziomów agregacji. Najczęściej wykorzystywane są: estymacja syntetyczna, regresyjna i złożona, a także oparta na modelu.

Do najważniejszych dziedzin, w których z powodzeniem stosuje się technikę SMO zalicza się:

- ochronę zdrowia — prowadzono badania dotyczące m.in. niezdolności psychicznej ludności w krótkich okresach, liczby urodzeń żywych i martwych, informacji o niepełnosprawnych, zachorowalności na raka;
- demografię — szacowanie m.in. liczby ludności i liczby mieszkań w różnych przekrojach;
- rynek pracy i bezrobocie — w tym m.in. wskaźniki bezrobocia, liczba bezrobotnych, siła robocza;
- badania rynkowe — m.in. dochody osobiste, zatrudnienie pełne i w niepełnym wymiarze, dochody rodzin, wskaźniki i zasięg ubóstwa;
- rolnictwo — m.in. szacowanie powierzchni, wielkości plonów, wartości hodowli i zbiorów.

Wymienione dziedziny są jedynie przykładowymi dziedzinami życia codziennego, w których SMO stosowana jest na szeroką skalę. Wspomniane oszacowania oparte są nie tylko na wynikach badań statystycznych, ale także wykorzystują rejestry administracyjne jako źródło dodatkowych zmiennych objaśniających.

Cytując przykłady za J. N. K. Rao (2003), Biuro Spisowe Stanów Zjednoczonych estymując dochody małych obszarów tworzyło szacunki dla stanowych i lokalnych władz w celu alokacji środków pieniężnych. Wykorzystany został model Faya-Herriota, gdzie dochód małego obszaru (tzn. zmienną objaśnianą y) otrzymano na podstawie zmiennych objaśniających, którymi były m.in. dochody terenu zawierającego mały obszar, dane z deklaracji podatkowych, a także dane ze spisu ludności i mieszkań — posługując się m.in. wartością zajmowanych mieszkań.

Innym przykładem jest szacowanie powierzchni uprawnych z daną rośliną (kukurydzą lub soją) i rozmiarów plonów. W tym przypadku wykorzystywano zdjęcia satelitarne, a zmiennymi objaśniającymi była liczba pikseli na obrazie sklasyfikowanych jako uprawa kukurydzy albo soi. Zdjęcia satelitarne mogą także posłużyć np. do szacowania wskaźnika urbanizacji.

W opracowaniu G. Dehnel (2003) omówiono estymację pośrednią jako narzędzie oceny rozwoju ekonomicznego. Porównane zostały najpopularniejsze sposoby SMO, a do oszacowania wskaźników rozwoju (sprzedaży detalicznej czy budowlanej, przeciętnej płacy, podatku od wynagrodzeń czy wskaźnika syntetycznego) posłużono się zmiennymi pochodzącymi ze sprawozdań DG-1 (źródło podstawowe) oraz z kartoteki przedsiębiorstw (źródło pomocnicze) i zmiennymi pochodzącymi z Banku Danych Lokalnych (źródło pomocnicze), w którym zamieszczono wyniki otrzymywane na podstawie m.in. badań statystycznych, rejestrów sądowych, rejestrów urzędów stanu cywilnego czy też PESEL.

Kolejnym przykładem użycia SMO jest działalność Banku Światowego, który wykorzystuje metodę estymacji pośredniej do estymacji ubóstwa na niskich poziomach agregacji przestrzennej w wielu krajach. Stosuje się w tym celu estymatory wykorzystujące model Faya-Herriota, w zależności od stopnia dostępnych danych — na poziomie jednostkowym lub obszaru.

Wyrazem coraz większego znaczenia SMO są liczne konferencje, publikacje oraz projekty o charakterze międzynarodowym, w których rozwija się teorię, ale również stosuje na szeroką skalę w praktyce różnego rodzaju estymatory. Oto niektóre międzynarodowe projekty:

- **ESSnet for Small Area Estimation** — ogólnym jego celem jest stworzenie ram umożliwiających szacunki dla małych obszarów w ramach Europejskich Badań Społecznych dotyczących m.in. demografii, bezrobocia, ubóstwa, warunków życia, zdrowia i edukacji. W projekcie powstały publikacje, w których znaleźć można szczegółowe opisy zastosowań estymacji małych obszarów i rodzaj wykorzystywanego oprogramowania;
- **The EURAREA project** — program badawczy finansowany przez Eurostat w ramach Programu Fifth Framework Programme FP5 Unii Europejskiej. Głównym celem było tu opracowanie techniki estymacji oraz odpowiedniego oprogramowania wspierającego ESS w zakresie szacowania podstawowej charakterystyki rynku pracy, jak też poziomu życia ludności w skali lokalnej;
- **SAMPLE — Small Area Methods for Poverty and Living Condition Estimates** — program badawczy finansowany przez Komisję Europejską w ramach Seventh Framework Programme (FP7). Celem jego było opracowanie nowych wskaźników zapewniających lepsze zrozumienie nierówności społecznej i ubóstwa w ujęciu lokalnym oraz metod ich estymacji z wykorzystaniem SMO. Dodatkowym celem projektu było wdrożenie procedur konstruk-

cji i interpretacji wskaźników ubóstwa i określenie stopnia ich przydatności dla samorządów.

Zgodnie z zaleceniami zawartymi w raportach sporządzonych z tych projektów należy rozpoczynać pracę od najprostszych estymatorów (np. klasycznego estymatora ekspansyjnego Horvitza-Thompsona), a dopiero później dokonywać rozszerzania estymatorów.

Podsumowanie

Zainteresowanie SMO w Polsce znacząco wzrosło w ostatnich latach, głównie ze względu na zwiększającą się rolę samorządów lokalnych. Określenia „mały obszar" nie należy rozumieć dosłownie. Termin ten odnosić się może do domeny czy subpopulacji, dla których wielkość próby jest na tyle mała, że estymacja bezpośrednia może być nieefektywna. Wykorzystanie metod SMO nie gwarantuje estymacji nieobarczonej błędem, ważne jest tu dokonanie wyboru właściwej techniki. Na poprawę jakości szacunków niewątpliwie wpłynąć może dostępność dodatkowych danych z istniejących już źródeł, co w dobie postępującej informatyzacji stanowi atut techniki pośredniej. Użycie metod estymacji pośredniej zazwyczaj pozwala uzyskać porównywalną lub znacznie lepszą precyzję szacunków w stosunku do klasycznych metod estymacji.

Ważną kwestią jest szacowanie precyzji estymatorów SMO. W tej dziedzinie prowadzone są przez statystyków na świecie badania mające na celu udoskonalenie sposobów szacowania błędów. W publikacji J. N. K. Rao (2003), uznawanej za jedną z podstawowych pozycji w tej tematyce, można znaleźć przykłady sposobów szacowania MSE i wariancji poszczególnych estymatorów. Przegląd metod w zakresie estymacji oraz oceny precyzji otrzymanych wyników w zależności od przyjętego schematu losowania znajduje się również m.in. w publikacji Cz. Domańskiego i K. Pruskiej (2001). Zarówno dla estymacji złożonej jak i syntetycznej, oceny precyzji otrzymanych estymatorów można dokonać szacując ich błąd średniokwadratowy. Dobrym sposobem jest także zastosowanie metody bootstrap.

SMO jest rozwijającą się dziedziną badań. Wyrazem coraz większego znaczenia metod estymacji pośredniej było utworzenie Ośrodka Statystyki Małych Obszarów w Urzędzie Statystycznym w Poznaniu. Ośrodek ten prowadzi prace badawcze w zakresie teorii i zastosowań metod tej statystyki w różnych dziedzinach. Można tu wymienić m.in. rynek pracy, statystykę przedsiębiorstw czy np. (we współpracy z Bankiem Światowym) estymację ubóstwa.

---

mgr Anna Małasiewicz — *Urząd Statystyczny w Poznaniu*

LITERATURA

Dehnel G. (2003), *Statystyka małych obszarów jako narzędzie oceny rozwoju ekonomicznego regionów*, Wydawnictwo Akademii Ekonomicznej w Poznaniu

Domański Cz., Pruska K. (2001), *Metody statystyki małych obszarów*, Wydawnictwo Uniwersytetu Łódzkiego

Pfeffermann D. (2013), *New Important Developments in Small Area Estimation*, „Statistical Science", Vol. 28, No. 1

Rao J. N. K. (2003), *Small Area Estimation*, Wiley, New York

SUMMARY

The article presents theoretical methodological considerations on estimation using small area statistics. Examples of applications of small area estimators in practice statistical surveys in Poland and in the world are presented.

РЕЗЮМЕ

В статье представляются теоретические рассуждения по методологии оценивания с использованием статистики малых домэн. В статье представляются примеры использования оценок статистики малых домэн в практике статистических обследований в Польше и в мире.