A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 271, 2012

Tomasz Żądło[*]

O PREDYKCJI WARTOŚCI GLOBALNEJ W DOMENIE Z WYKORZYSTANIEM INFORMACJI O ZMIENNYCH DODATKOWYCH PRZY ZAŁOŻENIU MODELU FAYA-HERRIOTA

Streszczenie. W pracy zostaną zaprezentowane najlepsze liniowe nieobciążone predyktory (*ang.* Best Linear Unbiased Predictors – BLUP) i empiryczne najlepsze liniowe nieobciążone predyktory (*ang.* Empirical Best Linear Unbiased Predictors – EBLUP), ich błędy średniokwadratowe (*ang.* Mean Squared Errors – MSE) oraz estymatory MSE dla modelu Faya-Herriota (Fay, Herriot (1979)). Model ten należy do klasy ogólnych mieszanych modeli liniowych typu A, co oznacza, że jest on zakładany dla wartości estymatorów bezpośrednich charakterystyk w domenach. Ponadto przyjmuje się, że wartości wariancji estymatorów bezpośrednich są znane. W artykule będzie analizowany symulacyjnie z wykorzystaniem rzeczywistych danych wpływ zastąpienia nieznanych wariancji estymatorów bezpośrednich ich nieobciążonymi estymatorami i estymatorami otrzymanymi przy wykorzystaniu ogólnych funkcji wariancji na obciążenia predyktorów, wartość MSE oraz obciążenia estymatorów MSE. Ponadto będzie uwzględniony problem niespełnienia założeń o normalności rozkładu składników losowych specyficznych dla domen. Analiza symulacyjna zostanie przeprowadzona w oparciu o dane dotyczące 8624 gospodarstw rolnych z powiatu Dąbrowa Tarnowska, które zostały uzyskane w spisie rolnym w 1996 roku.

Słowa kluczowe: BLUP, EBLUP, model Faya-Herriota, estymatory MSE.

I. PODSTAWOWE OZNACZENIA

Populacja N-elementowa oznaczana przez Ω dzieli się na D rozłącznych podpopulacji Ω_d (d=1,...,D) nazywanych dziedzinami badania lub domenami o liczebnościach N_d (d=1,...D) każda. Z populacji wybierana jest (losowo lub celowo) próba s o liczebności n. Część wspólna d-tej domeny i próby będzie oznaczana przez $s_d = s \cap \Omega_d$ a liczebność tego zbioru przez n_d. Zbiór elementów d-tej domeny, które nie znalazły się w próbie, będzie oznaczany przez $\Omega_{rd} = \Omega_d - s_d$ a liczebność tego zbioru przez $N_{rd} = N_d - n_d$. Średnia wartość badanej zmiennej w d-tej domenie oznacza będzie przez μ_d a wartość globalna badanej zmiennej w d-tej domenie przez $\theta_d = N_d \mu_d$.

[*] Dr, Katedra Statystyki, Uniwersytet Ekonomiczny w Katowicach.

II. MODEL NADPOPULACJI

W pracy będzie analizowany model Faya-Herriota, który skrótowo będziemy oznaczać przez F-H (Fay, Herriot (1979)). Należy on do klasy modeli typu A (zob. Rao (2003)), co oznacza, że jest zakładany dla wartości estymatorów bezpośrednich charakterystyk w domenach. Dodajmy, że modele typu B (Rao (2003)) zakładane są dla zmiennych losowych, których realizacjami są wartości badanej zmiennej. Odnośnie wartości bezpośredniego estymatora średniej w domenie, oznaczanego przez $\hat{\mu}_d$, zakładamy, że

$$\hat{\mu}_d = \mu_d + e_d \tag{1}$$

gdzie $\mu_d = \mathbf{x}_d^T\boldsymbol{\beta} + v_d$ jest średnią w domenie, $\mathbf{x}_d$ wektorem wartości p-zmiennych dodatkowych w d-tej domenie, $\boldsymbol{\beta}$ wektorem p nieznanych parametrów, e_d jest błędem wynikającym z planu losowania oraz e_d i v_d (d=1,...,D) są niezależne, przy czym $e_d \overset{iid}{\sim} N(0, W_d)$ i $v_d \overset{iid}{\sim} N(0, A)$ i przyjmuje się, że wariancje W_d są znane. Podkreślmy, że zaprezentowany model jest szczególnym przypadkiem następujących modeli: ogólnego modelu liniowego, ogólnego mieszanego modelu liniowego oraz ogólnego mieszanego modelu nadpopulacji z blokowo-diagonalną macierzą wariancji i kowariancji, przy czym zamiast zmiennych losowych Y_i (i=1,...,N) mamy $\hat{\mu}_d$ (d=1,...,D) oraz $\forall_{i,d} Z_{id} = 1$.

III. NAJLEPSZY LINIOWY NIEOBCIĄŻONY PREDYKTOR

Dla znanego A i bez konieczności założeń normalności rozkładów składników losowych, predyktor typu BLU średniej w d-tej domenie oraz jego MSE dane są wzorami (Datta, Rao, Smith (2005), Datta, Lahiri (2000), Lahiri, Rao (1995), Prasad, Rao (1990)):

$$\hat{\mu}_d^{BLUP} = \hat{\mu}_d - B_d(A)\left(\hat{\mu}_d - \mathbf{x}_d^T\hat{\boldsymbol{\beta}}\right) \tag{2}$$

gdzie

$$B_d(A) = W_d(A + W_d)^{-1} \tag{3}$$

$$\hat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \frac{B_d(A)}{W_d} \mathbf{x}_d \mathbf{x}_d^T \right)^{-1} \left(\sum_{d=1}^{D} \frac{B_d(A)}{W_d} \mathbf{x}_d \hat{\mu}_d \right) \tag{4}$$

$$MSE_{\xi}(\hat{\mu}_d^{BLUP}) = g_{1d}(A) + g_{2d}(A) \tag{5}$$

gdzie

$$g_{1d}(A) = A W_d (A + W_d)^{-1} \tag{6}$$

$$g_{2d}(A) = W_d^2 (A + W_d)^{-2} \mathbf{x}_d^T \left(\sum_{u=1}^{D} (A + W_d)^{-1} \mathbf{x}_u \mathbf{x}_u^T \right)^{-1} \mathbf{x}_d \tag{7}$$

Stąd, przy znanych liczebnościach domen N_d, predyktor typu BLU wartości globalnej i jego MSE dla modelu F-H dane są wzorami: $\hat{\theta}_d^{BLUP} = N_d \hat{\mu}_d^{BLUP}$ i $MSE_{\xi}(\hat{\theta}_d^{BLUP}) = N_d^2 MSE_{\xi}(\hat{\mu}_d^{BLUP})$, gdzie $\hat{\mu}_d^{BLUP}$ dane jest wzorem (2), a $MSE_{\xi}(\hat{\mu}_d^{BLUP})$ dane jest wzorem (5). Dodajmy, że prezentowane wyniki są przypadkami szczególnymi twierdzenia Hendersona (1950).

IV. ESTYMACJA PARAMETRÓW MODELU NADPOPULACJI

Przy wyprowadzeniu postaci predyktora typu BLU zakłada się, że parametr A modelu nadpopulacji jest znany. W praktyce jest on szacowany na podstawie danych z próby. Poniżej przedstawione zostaną wykorzystywane w praktyce metody estymacji tego parametru.

Pierwszymi dwoma metodami szacowania parametru A są metody największej wiarygodności (ang. *Maximum Likelihood*) i metody największej wiarygodności z ograniczeniami (ang. *Restricted Maximum Likelihood*). Estymatory parametru A otrzymane tymi metodami przy założeniu normalności rozkładu zmiennych losowych oznaczać będziemy odpowiednio przez $\hat{A}_{ML}$ oraz $\hat{A}_{RE}$. Ponieważ rozważany model nadpopulacji jest szczególnym przypadkiem ogólnego modelu liniowego, można zastosować znane w literaturze procedury np. Rao (2003) s. 100–102. Dodajmy, że do iteracyjnego rozwiązania równań nieliniowych, które pojawiają się w obu metodach, zostanie wykorzystany algorytm wyrównujący (ang. *scoring algorithm*), który jest również prezentowany w pracy Rao (2003) s. 100. Datta, Rao i Smith (2005) zwracają uwagę, że w ich rozważaniach symulacyjnych algorytm ten charakteryzuje się lepszymi własnościami niż algorytmy EM i Newtona-Raphsona. Algorytm ten różni się od meto-

dy Newtona-Raphsona wyłącznie uwzględnieniem zamiast hesjanu logarytmu funkcji wiarygodności wartości oczekiwanej tej macierzy. Taka modyfikacja zmniejsza czas wykonywania jednej iteracji, ze względu na prostszą formę wartości oczekiwanej hesjanu w porównaniu z hesjanem, choć liczba iteracji może wzrosnąć.

Oprócz estymacji parametru A wspomnianymi dwoma metodami będzie również rozważana metoda Faya-Herriota (1979). Estymator $\hat{A}_{FH}$ parametru A otrzymuje się jako otrzymane w sposób iteracyjny rozwiązanie równania:

$$\frac{1}{D-p}\mathbf{Y}^T Q(\hat{A}_{FH})\mathbf{Y} - 1 = 0 \tag{8}$$

gdzie $\mathbf{Y}^T Q(A)\mathbf{Y} = \sum_{d=1}^{D}(W_d + A)^{-1}(\hat{\mu}_d - x_d^T\hat{\boldsymbol{\beta}})^2$, $\hat{\boldsymbol{\beta}}$ jest dane wzorem (4), p jest liczbą parametrów wektora $\boldsymbol{\beta}$. Ponadto rozważany będzie estymator parametru A uzyskany metodą zaproponowaną przez Prasada-Rao (1990), który obliczany jest ze wzoru $\max(0, \hat{A}_{PR})$ gdzie:

$$\hat{A}_{PR} = (D-p)^{-1}\left[\sum_{d=1}^{D}(\hat{\mu}_d - x_d^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\mu}})^2 - \sum_{d=1}^{D}W_d(1 - x_d^T(\mathbf{X}^T\mathbf{X})^{-1}x_d)\right] \tag{9}$$

a $\hat{\boldsymbol{\mu}} = col_{1\le d\le D}(\hat{\mu}_d)$ i $\mathbf{X} = col_{1\le d\le D}(x_d^T)$. Podkreślmy, że w przypadku stosowania estymatorów $\hat{A}_{FH}$ i $\hat{A}_{PR}$ nie jest wymagana normalność rozkładu składników losowych.

Należy przypomnieć, że w modelu F-H przyjmuje się, że wariancje $\hat{\mu}_d$ oznaczane przez W_d są znane nawet w przypadkach empirycznych (tj. gdy inne nieznane parametry zastępowane są wartościami estymatorów). W praktyce, zastępuje się je wartościami estymatorów lub wartościami estymatorów po wygładzeniu (choć przy wyprowadzeniach przyjmuje się, że są one znane), co może jednak mieć wpływ na obciążenia predyktorów i estymatorów MSE oraz na wartość MSE. Problem ten będzie studiowany w badaniu symulacyjnym. Ze względu na założenie niezależności składników losowych e_d w badaniu symulacyjnym domeny będą warstwami i wówczas estymatory W_d dane będą wzorami:

$$\hat{W}_d = \frac{N_d - n_d}{N_d n_d} \frac{1}{n_d - 1} \sum_{i=1}^{n_d} (Y_i - \bar{Y}_{sd})^2 \tag{10}$$

Ponadto, jak podają np. Lahiri i Rao (1995), często wygładza się wartości estymatorów wariancji z wykorzystaniem uogólnionych funkcji wariancji. Opis tej metody można znaleźć w pracy Woltera (1985). Poniżej przedstawiamy jedną z możliwych technik, która zostanie wykorzystana w tym artykule. Należy zauważyć, co podkreśla również Wolter (1985), że brak jest teoretycznych uzasadnień postaci różnych funkcji wykorzystywanych do modelowania wariancji a ich dobór ma charakter empiryczny. W opracowaniu oceny wariancji W_d będziemy wygładzać wykorzystując funkcję (Wolter (1985) s. 203):

$$\log\left(W_d \mu_d^{-2}\right) = \alpha - \beta \log(\mu_d) \tag{11}$$

gdzie α i β są szacowane metodą najmniejszych kwadratów w oparciu o równanie (11), gdzie W_d i μ_d zastępujemy odpowiednio $\hat{W}_d$ i $\hat{\mu}_d$. Następnie w równaniu (11) α, β i μ_d zastępujemy ich ocenami i z (11) obliczamy wygładzone wartości wariancji W_d, które będziemy oznaczać przez $\hat{W}_{GVF\,d}$.

V. EMPIRYCZNE NAJLEPSZE LINIOWE NIEOBCIĄŻONE PREDYKTORY I ICH BŁĘDY ŚREDNIOKWADRATOWE

Predyktory typu EBLU średniej oraz wartości globalnej w domenie dla modelu F-H, które oznaczać będziemy przez $\hat{\mu}_d^{EBLUP}$ oraz $\hat{\theta}_d^{EBLUP}$, dane są odpowiednio wzorami (2) i $N_d \hat{\mu}_d^{BLUP}$, gdzie parametr A zastępowany jest jednym z omówionych w poprzednim rozdziale estymatorów. Warto nadmienić, że predyktory typu EBLU pozostają nieobciążone, m.in. dlatego że omawiane estymatory parametru A są parzystymi, niezmienniczymi ze względu na przesunięcie funkcjami $\hat{\mu}_d$. Wówczas zachodzi poniższe twierdzenie.

Twierdzenie 1. (Kackar i Harville (1981)). Rozważmy predyktor typu EBLU i załóżmy, że spełnione są założenia ogólnego mieszanego modelu liniowego. Jeśli spełnione są trzy następujące warunki:

(i) wartość oczekiwana predyktora typu EBLU jest skończona,

(ii) $\hat{A}$ jest dowolnym estymatorem mającym własność parzystości i niezmienniczości względem przesunięcia tj. $\hat{A}(-\hat{\mu}_d) = \hat{A}(\hat{\mu}_d)$ i $\hat{A}(\hat{\mu}_d + \mathbf{Xb}) = \hat{A}(\hat{\mu}_d)$ dla dowolnego $\mathbf{b} \in R^p$,

(iii) rozkłady e_d i v_d są symetryczne względem 0 (niekoniecznie normalne),

wówczas predyktor typu EBLU jest ξ-nieobciążony.

Zakładając, że (Datta, Rao, Smith (2005) s.186) wartości **X** są jednostajnie ograniczone, wartości $\mathbf{X}^T\mathbf{V}_{ss}^{-k}(A)\mathbf{X}$ (k=1,2,3) są rzędu $O(D)$ oraz że W_d są ograniczone od góry i od zera, MSE predyktora typu EBLU wartości średniej w domenie dla modelu F-H można wyrazić wzorem (Prasad, Rao (1990), Datta, Lahiri (2000)):

$$MSE_{\xi}(\hat{\mu}_d^{EBLUP}(\hat{A})) = g_{1d}(A) + g_{2d}(A) + g_{3d}(A) + o(D^{-1}) \tag{12}$$

gdzie $g_{1d}(A)$ dane jest wzorem (6), $g_{2d}(A)$ dane jest wzorem (7). Postać $g_{3d}(A)$ we wzorze (12) zależy od użytego estymatora parametru *A*. Dla estymatora zaproponowanego przez Prasada i Rao (1990) przyjmuje postać:

$$g_{3d}(A) = g_{3dPR}(A) = 2W_d^2(A+W_d)^{-3}D^{-2}\sum_{u=1}^{D}(A+W_u)^2, \tag{13}$$

dla estymatorów parametru *A* uzyskanych metodą największej wiarygodności i metodą największej wiarygodności z ograniczeniami (Datta, Lahiri (2000)):

$$g_{3d}(A) = g_{3dML}(A) = g_{3dRE}(A) = 2W_d^2(A+W_d)^{-3}\left(\sum_{u=1}^{D}(A+W_u)^{-2}\right)^{-1}, \tag{14}$$

a dla estymatorów F-H parametru *A* (Datta, Rao, Smith (2005)):

$$g_{3d}(A) = g_{3dFH}(A) = 2DW_d^2(A+W_d)^{-3}\left(\sum_{u=1}^{D}(A+W_d)^{-1}\right)^{-2}. \tag{15}$$

Błąd średniokwadratowy predyktora typu EBLU wartości globalnej w domenie dla modelu F-H ma postać: $MSE_\xi(\hat{\theta}_d^{EBLUP}(\hat{A})) = N_d^2 MSE_\xi(\hat{\mu}_d^{EBLUP}(\hat{A}))$, gdzie $MSE_\xi(\hat{\mu}_d^{EBLUP}(\hat{A}))$ ma postać (12).

VI. ESTYMATORY BŁĘDÓW ŚREDNIOKWADRATOWYCH

W niniejszym rozdziale przedstawione zostaną estymatory MSE oznaczane przez $\hat{MSE}_\xi(\hat{\mu}_d^{EBLUP}(\hat{A}))$, które są w przybliżeniu nieobciążone w następującym sensie: $E_\xi\left(\hat{MSE}_\xi(\hat{\mu}_d^{EBLUP}(\hat{A}))\right) - MSE_\xi(\hat{\mu}_d^{EBLUP}(\hat{A})) = o(D^{-1})$. Datta i Lahiri (2000) podają następującą postać estymatora MSE predyktora typu EBLU wartości średniej w domenie:

$$\hat{MSE}_\xi(\hat{\mu}_d^{EBLUP}(\hat{A})) = g_{1d}(\hat{A}) + g_{2d}(\hat{A}) + 2g_{3d}(\hat{A}) - (B_d(\hat{A}))^2 b_{\hat{A}}(\hat{A}) \quad (16)$$

gdzie $b_{\hat{A}}(A)$ to asymptotyczne obciążenie estymatora $\hat{A}$ (z dokładnością do składnika $o(D^{-1})$), $B_d(A)$ dane jest wzorem (3).

Dla estymatorów $\hat{A}_{PR}$ oraz $\hat{A}_{RE}$, które są asymptotycznie nieobciążone (tj. z dokładnością do składnika $o(D^{-1})$) wzór (16) upraszcza się do postaci

$$\hat{MSE}_\xi(\hat{\mu}_d^{EBLUP}(\hat{A}_{PR})) = g_{1d}(\hat{A}_{PR}) + g_{2d}(\hat{A}_{PR}) + 2g_{3dPR}(\hat{A}_{PR}) \quad (17)$$

$$\hat{MSE}_\xi(\hat{\mu}_d^{EBLUP}(\hat{A}_{RE})) = g_{1d}(\hat{A}_{RE}) + g_{2d}(\hat{A}_{RE}) + 2g_{3dRE}(\hat{A}_{RE}) \quad (18)$$

Ponieważ asymptotyczne obciążenie estymatora $\hat{A}_{ML}$ wynosi (Datta, Lahiri (2000)):

$$b_{\hat{A}_{ML}}(A) = -\left(\sum_{u=1}^{D}(A+W_u)^{-2}\right)^{-1} tr\left[\left(\sum_{u=1}^{D}(A+W_u)^{-1}\mathbf{x}_u\mathbf{x}_u^T\right)^{-1}\left(\sum_{u=1}^{D}(A+W_u)^{-2}\mathbf{x}_u\mathbf{x}_u^T\right)\right] \quad (19)$$

w przypadku, gdy wykorzystujemy estymator $\hat{A}_{ML}$, estymator MSE dany wzorem (16) przyjmuje postać:

$$M\hat{S}E_{\xi}(\hat{\mu}_d^{EBLUP}(\hat{A}_{ML})) = g_{1d}(\hat{A}_{ML}) + g_{2d}(\hat{A}_{ML}) +$$
$$+2g_{3dML}(\hat{A}_{ML}) + (B_d(\hat{A}_{ML}))^2\left(\sum_{u=1}^{D}(A_{ML}+W_u)^{-2}\right)^{-1} \times \tag{20}$$
$$\times tr\left[\left(\sum_{u=1}^{D}(A_{ML}+W_u)^{-1}\mathbf{x}_u\mathbf{x}_u^T\right)^{-1}\left(\sum_{u=1}^{D}(A_{ML}+W_u)^{-2}\mathbf{x}_u\mathbf{x}_u^T\right)\right].$$

Asymptotyczne obciążenie estymatora $\hat{A}_{FH}$ dane jest wzorem (Datta, Rao, Smith (2005)):

$$b_{\hat{A}_{FH}}(A) = 2\left[\left(D\sum_{u=1}^{D}(A+W_u)^{-2}\right)-\left(\sum_{u=1}^{D}(A+W_u)^{-1}\right)^2\right]\left(\sum_{u=1}^{D}(A+W_u)^{-1}\right)^{-3} \tag{21}$$

Stąd, gdy wykorzystujemy estymator $\hat{A}_{FH}$, estymator MSE dany wzorem (16) ma postać:

$$M\hat{S}E_{\xi}(\hat{\mu}_d^{EBLUP}(\hat{A}_{FH})) = g_{1d}(\hat{A}_{FH}) + g_{2d}(\hat{A}_{FH}) + 2g_{3dML}(\hat{A}_{FH}) - 2(B_d(\hat{A}_{FH}))^2 \times$$
$$\times\left[\left(D\sum_{u=1}^{D}(A_{FH}+W_u)^{-2}\right)-\left(\sum_{u=1}^{D}(A_{FH}+W_u)^{-1}\right)^2\right]\left(\sum_{u=1}^{D}(A_{FH}+W_u)^{-1}\right)^{-3} \tag{22}$$

Estymatory MSE predyktorów typu EBLU wartości globalnej w domenie dane są następującym wzorem: $M\hat{S}E_{\xi}(\hat{\theta}_d^{EBLUP}(\hat{A})) = N_d^2 M\hat{S}E_{\xi}(\hat{\mu}_d^{EBLUP}(\hat{A}))$.

Warto pokreślić, że (Lahiri, Rao (1995), Datta, Rao, Smith (2005)) estymatory MSE uzyskane w przypadku, gdy estymacji parametru A metodą Prasada-Rao oraz F-H charakteryzują się pewną odpornością na brak normalności rozkładu składników losowych.

VII. ANALIZA MONTE CARLO

Zostaną przedstawione wyniki analizy Monte Carlo przygotowanej z wykorzystaniem pakietu R (R Development Core Team, 2007). Analizujemy

dane dotyczące 8624 gospodarstw rolnych z powiatu Dąbrowa Tarnowska, które zostały uzyskane w spisie rolnym w 1996 roku. W rozważanym powiecie znajduje się D=79 miejscowości, które będą traktowane jako domeny. Liczebności domen wahają się od 20 do 610 gospodarstw rolnych. Ze względu na założenie niezależności składników losowych e_d zdecydowano się na rozważanie losowania warstwowego (w warstwach losowanie proste bez zwracania), gdzie warstwami są domeny. Alokacja próby w warstwach jest w przybliżeniu proporcjonalna – z każdej warstwy losowanych jest ok. 10% elementów. Celem jest predykcja wartości globalnej powierzchni zasiewów (w arach) w domenach. W modelu F-H za $\hat{\mu}_d$ przyjmujemy oceny średniej powierzchnia zasiewów (w arach) w domenach (średnie arytmetyczne dla danych z próby w domenach). Zmienną dodatkową będzie rzeczywista średnia powierzchnia gospodarstwa rolnego (w arach) w domenie. Model ze zmienną dodatkową uwzględnia stałą. Wartość współczynnika korelacji liniowej Pearsona pomiędzy tymi charakterystykami w rozważanym zbiorze danych wynosi 0,5498.

Liczba iteracji w symulacji to M=5000. W każdym kroku symulacji generowane są wartości $\hat{\mu}_d$ zgodnie z modelem F-H. Za wartości W_d przyjmujemy:

$$W_d = \frac{N_d - n_d}{N_d n_d} \frac{1}{N_d - 1} \sum_{i=1}^{N_d} \left(y_i - N_d^{-1} \sum_{i=1}^{N_d} y_i \right)^2 , \tag{23}$$

gdzie wartości y_i są wartościami badanej zmiennej w rozważanym zbiorze danych. Wartości składników losowych e_d generowane są niezależnie zgodnie z rozkładem $N(0, \sqrt{W_d})$. Za wartość parametru A w symulacji przyjmujemy wartość obliczaną z wykorzystaniem metody największej wiarygodności z ograniczeniami (przy założeniu rozkładu normalnego zmiennych losowych), przy czym zamiast $\hat{\mu}_d$ wykorzystujemy rzeczywiste wartości średnie w domenach dostępne w rozważanym zbiorze danych przyjmując jednocześnie zera za wariancje W_d estymatorów $\hat{\mu}_d$. Wówczas składniki losowe v_d generowane są niezależnie zgodnie z rokładami normalnym, jednostajnym i przesuniętym wykładniczym (tak aby wartość oczekiwana wynosiła 0) z wariancją A. Za $\boldsymbol{\beta}$ przyjmujemy wartości obliczone na podstawie danych z populacji zgodnie ze wzorem (4), gdzie zamiast $\hat{\mu}_d$ wykorzystujemy rzeczywiste wartości średnie w domenach dostępne w rozważanym zbiorze danych przyjmując jednocześnie zera za wariancje W_d estymatorów $\hat{\mu}_d$.

W poniższych tablicach zostaną przedstawione wartości różnych statystyk uzyskane dla wszystkich 79 domen. Aby ograniczyć wielkość tablic wynikowych zostaną zaprezentowane wyłącznie wartości minimalne (min), pierwszego kwartyla (Q_1), mediany (Me), średniej arytmetycznej (średnia), trzeciego kwartyla (Q_3) oraz maksymalne dla wyników uzyskanych dla wszystkich domen. Skróty używane w kolumnach dotyczą rodzaju użytego estymatora parametru A: PR – estymator Prasada-Rao, FH – Faya-Herriota, ML – estymator uzyskany metodą najwięjszej wiarygodności, RE - estymator uzyskany metodą najwięjszej wiarygodności z ograniczeniami. W kolumnach wprowadzono również informcje o rozkładzie składników losowych v_d: N- rozkład normalny, J – rozkład jednostajny i W – przesunięty rozkład wykładniczy.

Omawiając wartości prezentowane w tablicy 1 warto zwrócić uwagę, że różnice między wartościami funkcji $g_3(.)$ są znaczące i co ważne $g_{3dPR}(A) < g_{3dFH}(A) < g_{3dML}(A) = g_{3dRE}(A)$. Porównując jednak wartości funkcji $g_3(.)$ z przybliżonymi wartościami MSE okazuje się, że różnice między wartościami funkcji $g_3(.)$ (a $g_3(.)$ są niższego rzędu od sumy pozostałych dwóch komponentów przybliżonego MSE) mają mniejsze znaczenie.

Tablica 1. Wartości funkcji $g_3(A)$ i przybliżonego MSE

	PR		FH		ML, RE	
	g_3	MSE	g_3	MSE	g_3	MSE
min	8,955	484,368	3,020	478,433	2,560	477,973
Q_1	89,445	1795,923	30,163	1737,376	25,567	1732,836
Me	158,651	2818,701	53,502	2708,221	45,349	2699,655
średnia	145,320	3032,170	49,006	2935,856	41,538	2928,388
Q_3	211,743	4225,032	71,405	4070,569	60,524	4058,593
max	233,638	6365,049	78,789	6285,721	66,783	6279,570

Analizując wartości prezentowane w tablicy 2 należy podkreślić, że wybór estymatora parametru A, estymacja parametrów W_d a nawet rozkład składników losowych v_d ma niewielki wpływ na względne obciążenia predyktorów, które we wszystkich omawianych przypadkach nie przekraczają co do modułu 1,3%. Przedstawione w tablicy 3 względne wartości pierwiastków MSE o wartościach od ponad 5% do ponad 39% (m.in. w zależności od domeny) sugerują konieczność poszukiwania modeli charakteryzujących się lepszą jakością dopasowania. Jednocześnie, w rozważanym badaniu największy wpływ na zmiany wartości względnej pierwiastka MSE miała ocena parametrów W_d, a sposób estymacji parametru A jak i rozkład składników losowych v_d miały wpływ znacznie mniejszy.

Tablica 2. Wartości względnych obciążeń (w %) predyktorów typu EBLU

Est. A	PR			ML			RE			FH		
	znane Wd											
	N	J	W	N	J	W	N	J	W	N	J	W
min	–0,54	–0,66	–0,47	–0,54	–0,63	–0,68	–0,53	–0,63	–0,68	–0,52	–0,64	–0,60
Q_1	–0,09	–0,14	–0,08	–0,10	–0,14	–0,11	–0,10	–0,14	–0,11	–0,10	–0,14	–0,08
Me	0,02	–0,02	0,05	–0,01	0,00	0,01	–0,01	0,00	0,02	–0,02	–0,00	0,04
śred.	0,05	0,00	0,06	0,04	0,00	0,01	0,04	0,01	0,01	0,04	0,00	0,03
Q_3	0,21	0,16	0,19	0,17	0,15	0,13	0,16	0,15	0,13	0,17	0,15	0,15
max	0,63	1,03	0,82	0,66	1,01	0,75	0,66	1,01	0,78	0,63	1,00	0,84
	szacowane Wd											
min	–0,66	–0,60	–0,67	–0,83	–0,60	–0,74	–0,84	–0,61	–0,74	–0,84	–0,61	–0,73
Q_1	–0,06	–0,11	–0,09	–0,08	–0,12	–0,13	–0,09	–0,12	–0,12	–0,08	–0,12	–0,12
Me	0,01	0,02	0,06	–0,01	0,01	0,03	–0,01	0,02	0,03	0,00	0,02	0,04
śred.	0,08	0,03	0,06	0,06	0,03	0,04	0,06	0,03	0,04	0,06	0,03	0,05
Q_3	0,26	0,16	0,18	0,18	0,15	0,18	0,18	0,15	0,18	0,18	0,15	0,18
max	0,91	1,04	0,75	0,97	1,00	0,91	0,98	1,00	0,91	0,98	1,00	0,90
	wygładzone, szacowane Wd											
min	–1,23	–0,73	–0,70	–1,21	–0,69	–0,72	–1,22	–0,70	–0,72	–1,23	–0,72	–0,70
Q_1	–0,11	–0,10	–0,10	–0,10	–0,11	–0,10	–0,10	–0,10	–0,10	–0,11	–0,10	–0,10
Me	–0,02	0,02	0,07	–0,02	0,02	0,04	–0,02	0,02	0,04	–0,02	0,02	0,06
śred.	0,05	0,03	0,07	0,05	0,03	0,06	0,05	0,03	0,06	0,05	0,03	0,07
Q_3	0,21	0,17	0,20	0,20	0,16	0,19	0,19	0,16	0,19	0,20	0,17	0,19
max	1,13	1,00	0,83	1,08	1,00	0,76	1,10	1,00	0,77	1,11	1,00	0,81

Tablica 3. Wartości względnych RMSE (w %) predyktorów typu EBLU

Est. A	PR			ML			RE			FH		
	znane Wd											
	N	J	W	N	J	W	N	J	W	N	J	W
min	6,34	6,28	6,44	6,06	5,98	5,92	6,06	5,98	5,92	6,06	5,98	5,94
Q_1	12,51	12,53	12,51	12,11	12,09	12,09	12,09	12,08	12,08	12,11	12,10	12,08
Me	15,12	15,36	15,12	14,85	14,97	14,84	14,85	14,97	14,83	14,85	14,98	14,83
śred.	15,57	15,60	15,54	15,16	15,17	15,10	15,15	15,16	15,09	15,17	15,19	15,08
Q_3	18,37	18,17	17,84	17,88	17,72	17,51	17,87	17,71	17,50	17,90	17,74	17,48
max	28,83	29,78	29,45	28,44	29,15	29,13	28,40	29,13	29,10	28,40	29,17	28,98
	szacowane Wd											
min	6,46	6,40	6,65	6,08	6,02	6,00	6,09	6,03	6,00	6,08	6,02	6,00
Q_1	13,20	13,25	12,99	12,66	12,63	12,67	12,64	12,62	12,59	12,65	12,63	12,60
Me	15,62	15,95	15,75	15,79	15,74	15,88	15,89	15,78	15,98	15,82	15,77	15,91
śred.	16,88	16,86	16,84	16,95	16,93	16,87	17,01	16,98	16,93	16,97	16,95	16,88
Q_3	19,41	19,24	19,26	20,35	20,20	20,18	20,48	20,30	20,23	20,38	20,23	20,16
max	35,68	34,65	34,74	37,43	36,05	36,67	37,52	36,14	36,76	37,39	36,02	36,64
	wygładzone, szacowane Wd											
min	7,69	7,56	7,47	7,94	7,77	7,71	7,81	7,65	7,58	7,73	7,58	7,48
Q_1	12,73	12,77	12,64	12,82	12,79	12,75	12,74	12,75	12,65	12,70	12,74	12,62
Me	15,68	15,81	15,52	15,59	15,80	15,61	15,59	15,77	15,60	15,61	15,76	15,56
śred.	16,92	16,93	16,84	16,74	16,74	16,66	16,78	16,78	16,70	16,83	16,84	16,74
Q_3	20,58	20,48	20,43	20,39	20,29	20,25	20,46	20,31	20,36	20,49	20,37	20,35
max	38,88	39,14	37,84	38,09	38,29	36,98	38,54	38,73	37,43	38,73	38,96	37,67

Podsumowując wyniki prezentowne w tablicy 4 warto zwrócić uwagę, że w przypadku gdy parametry W_d są szacowane obciążenia estymatorów MSE mogą być bardzo wysokie. Dalej omawiać będziemy wyniki wyłącznie w sytuacji, gdy W_d są znane. Warto zauważyć, że w przypadku modelu Faya-Herriota (ze zmiennymi dodatkowymi) wykorzystanie estymatora Prasada-Rao może w niektórych domenach prowadzić do wysokich obciążeń (wysoka wartość maksimum). Pomijając wyniki uzyskiwane w przypadku estymacji parametru A metodą Prasada-Rao, w przypadku znanych W_d w żadnym z rozpatrywanych przypadków obciążenie estymatora MSE nie przekroczyło co do modułu 10%. Ponadto, w przypadku gdy parametr A jest szacowany metodą Faya-Herriota estymatory MSE zazwyczaj charakteryzują się niższymi obciążeniami niż w przypadku estymacji parametru A innymi metodami. Warto również podkreślić, że obciążenia MSE uzyskane w przypadku jednostajnego rozkładu składników losowych v_d są zbliżone do przypadku, gdy v_d mają rozkład normalny. Natomiast, gdy v_d mają przesunięty rozkład wykładniczy, obciążenia MSE są większe.

Tablica 4. Wartości względnych obciążeń estymatorów MSE (w %) predyktorów typu EBLU

Est. A	PR			ML			RE			FH		
	N	J	W	N	J	W	N	J	W	N	J	W
min	–3,9	–2,6	–4,7	–3,3	–3,2	–9,5	–3,2	–3,2	–9,0	–2,9	–3,3	–7,3
Q_1	1,4	0,5	1,4	–1,5	–0,8	–3,6	–1,2	–0,6	–3,2	–1,2	–1,1	–2,2
Me	3,8	3,2	5,2	–0,3	0,2	–1,1	–0,1	0,3	–0,8	0,0	0,3	–0,2
śred.	5,1	4,3	9,1	0,0	0,3	–1,6	0,2	0,4	–1,3	0,2	0,2	–0,5
Q_3	6,1	5,7	9,6	1,2	1,2	0,7	1,5	1,3	0,8	1,6	1,2	1,4
max	47,9	31,9	109,3	5,6	4,9	5,1	5,8	5,0	5,0	5,5	4,7	5,1
min	–89,9	–90,2	–82,7	–96,2	–96,0	–96,1	–96,2	–96,0	–96,1	–96,2	–96,0	–96,1
Q_1	–26,9	–27,9	–14,1	–51,8	–52,7	–50,7	–52,2	–53,1	–51,5	–52,1	–53,1	–51,4
Me	–1,3	–1,1	7,3	–20,8	–22,9	–20,5	–21,5	–23,2	–20,9	–21,5	–23,2	–20,7
śred.	5,7	1,6	32,4	–14,6	–14,4	–14,9	–14,8	–14,5	–15,0	–15,8	–15,6	–15,7
Q_3	23,8	25,4	46,1	23,7	25,1	22,7	23,4	24,7	23,0	21,9	23,1	22,1
max	311,3	224,0	785,8	106,8	108,0	105,3	109,5	110,7	108,4	104,7	105,3	105,6
min	–89,5	–89,3	–89,7	–89,5	–89,3	–89,7	–89,7	–89,5	–89,9	–89,6	–89,4	–89,9
Q_1	–23,2	–24,0	–23,2	–22,1	–22,8	–23,0	–23,3	–24,0	–24,2	–23,2	–24,2	–23,8
Me	23,0	21,7	21,2	21,5	19,5	20,6	21,1	19,2	20,3	21,8	20,0	20,9
śred.	23,3	23,5	23,8	19,5	20,0	19,0	20,0	20,5	19,6	21,5	21,9	21,8
Q_3	58,6	59,1	61,1	50,4	53,6	51,9	52,0	54,9	53,4	55,0	57,4	58,0
max	183,4	181,4	187,0	170,9	170,9	177,9	176,3	176,0	182,8	180,3	179,5	186,2

Przejdźmy do wyników dotyczących stosunku MSE predyktorów typu EBLU do MSE predyktorów typu BLU dla przypadku modelu Faya-Herriota, które nie są prezentowane w tablicach wynikowych. W rozważanym badaniu symulacyjnym spadek dokładności predyktora ze względu na estymację parame-

tru A jest niewielki, ale w przypadku gdy dodatkowo szacowane są parametry W_d spadek ten jest znacznie większy. W przypadku, gdy v_d mają rozkład normalny i pomijając sytuacje, gdy parametr A jest szacowany metodą Prasada-Rao (co zazwyczaj prowadzi do gorszej efektywności predyktorów), średnie wartości rozważanych współczynników efektywności w przypadku modelu Faya-Herriota nie przekraczają: gdy W_d są znane 1,02 a gdy W_d są szacowane (lub szacowane i wygładzane) 1,27. Podobne wyniki dotyczące efektywności uzyskano dla modelu Faya-Herriota w przypadku, gdy składniki losowe v_d mają rozkład jednostajny. W tym przypadku średnia efektywność , gdy parametry W_d są znane, jest poniżej 1,02 (za wyjątkiem gdy parametr A jest szacowany metodą Prasada-Rao – wówczas wynosi prawie 1,08). Gdy parametry W_d są szacowane, średnia efektywność jest poniżej 1,27, a gdy szacowane W_d są dodatkowo wygładzane – poniżej 1,29. Dla modelu Faya-Herriota w przypadku, gdy składniki losowe v_d mają przesunięty rozkład wykładniczy a parametry W_d są znane, średnia efektywność jest poniżej 1,01 (za wyjątkiem, gdy parametr A jest szacowany metodą Prasada-Rao – wówczas nie przekracza 1,08). Gdy parametry W_d są szacowane średnia efektywność jest poniżej 1,27, a gdy szacowane W_d są dodatkowo wygładzane poniżej 1,28.

Podsumowując wyniki symulacji, należy podkreślić, że szacowanie parametrów W_d ma znaczący wpływ zwłaszcza na efektywność estymatora i na obciążenia prezentowanych estymatorów MSE (wyprowadzonych przy założeniu, że te parametry są znane). Niezbędne więc jest zaproponowanie alternatywnych estymatorów MSE (np. jackknife, bootstrap) lub wyprowadzenie postaci (przybliżonych) MSE w sytuacji, gdy parametry W_d są szacowane i zaproponowanie estymatorów MSE.

BIBLIOGRAFIA

Datta, G. S., Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistica Sinica*, 10, 613–627.

Datta G.S., Rao J.N.K, Smith D.D. (2005), On measuring the variability of small area estimators under basic area level model, *Biometrika*, 92, 1, 183–196.

Fay R.E., Herriot R.A. (1979), Estimates of income for small places: An application of James-Stein procedures to census data, *Journal of the American Statistical Association*, 74, 269–277.

Henderson, C.R. (1950), Estimation of genetic parameters (Abstract), *Annals of MathematicalStatistics*, 21, 309–310.

Kackar, R.N., Harville, D.A. (1981), Unbiasedness of two-stage estimation and prediction procedures for mixed linear models, *Communications in Statistics, Series A,* 10, 1249–1261.

Lahiri, Rao (1995), Robust estimation of mean squared error of small area estimators, *Journal of the American Statistical Association*, 90, 430, 758–766

Prasad, N.G.N, Rao, J.N.K. (1990), The estimation of mean the mean squared error of small area estimators, *Journal of the American Statistical Association*, 85, 163–171.

Rao, J.N.K. (2003), *Small area estimation*. John Wiley & Sons, New York.

Wolter K.M. (1985), *Introduction to variance estimation*, Springer-Verlag, New York

Tomasz Żądło

ON SOME PROBLEM OF PREDICTION OF DOMAIN TOTAL UNDER FAY-HERRIOT MODEL

Abstract

In the paper BLUPs and EBLUPs, their MSEs and estimators of MSEs under Fay-Herrior model (Fay, Herrior (1979)) are presented. This model belongs to the class of general linear mixed model type A, what means that is assumed for direct estimates of domain characteristics. What is more, it is assumed that variances of direct estimates are known. In the paper the influence of replacing the variances by their unbiased estimates and by genereal variance function's estimates on biases of predictors, MSEs and biases of estimators of MSEs is studied in the simulation based on the real data. The problem of nonormality of area specific random components is also included.