Jan KUBACKI

Zastosowanie hierarchicznej estymacji bayesowskiej w szacowaniu wartości dochodów ludności w powiatach

Metody estymacji parametrów na poziomie małych obszarów są zwykle stosowane w przypadku, gdy konieczne jest wyznaczenie szacunków na podstawie badania reprezentacyjnego, jednak nie dysponuje się próbą o dostatecznej wielkości. Powoduje to wzrost wartości błędów szacunku. Wtedy zachodzi potrzeba tzw. „pożyczenia mocy" (*borrow strength*).

Pod pojęciem „mały obszar" można rozumieć małe jednostki administracyjne (np. powiaty) lub specyficzne grupy wyodrębnione z populacji (np. grupy socjoekonomiczne), dla których liczebność próby nie jest wystarczająca. Problem ten może dotyczyć też tzw. cech rzadkich, które są obserwowane dużo rzadziej i przez to ich szacunki, które są odpowiednie dla cech powszechnych (np. dla województw), nawet w skali mogą nastręczać trudności. Przykładem takiej cechy może być szacunek wartości dochodu na osobę z zasiłków dla bezrobotnych wyznaczany dla powiatów. Gdy jest on szacowany dla województw z *Badania budżetów gospodarstw domowych* (BBGD), względny błąd szacunku bywa niekiedy dość duży i może nawet przekraczać 20%. Zastosowanie w takim przypadku metod szacunków dla małych obszarów może być zasadne.

Metodologia szacowania parametrów dla małych obszarów jest systematycznie rozwijana już od lat 80. ub. stulecia. Powstało kilka szerszych opracowań, również w postaci monografii, które ujmują te zagadnienia w dość wyczerpujący sposób. Wspomnieć tutaj można książki J. N. K. Rao (2003) oraz N. T. Longforda (2005). Również w polskiej literaturze dostępne są opracowania zawierające omówienie zagadnień estymacji dla małych obszarów, w tym opracowania: Cz. Brachy i in. (2004); Cz. Domańskiego, K. Pruskiej (2001); E. Gołaty (2004); T. Żądło (2008).

Zagadnienia estymacji dla małych obszarów były też przedmiotem wielu konferencji naukowych. Wspomnieć tutaj można jedną z konferencji na ten temat, która odbyła się w 1992 r. w Warszawie (Kalton i in., 1993) oraz cykl konferencji *Small Area Estimation*, które odbywają się cyklicznie co 2 lata. Począwszy od 2005 r. odbyły się kolejno: w 2005 r. w Jyväskylä w Finlandii (http://www.stat.jyu.fi/sae2005/ /index.html), w 2007 r. w Pizie (http://sae2007.dsm.unipi.it/), w 2009 r. w Elche (http://icio.umh.es/congresos/sae2009/), w 2011 r. w Trewirze (http://www.uni--trier.de/index.php?id=30789) oraz w Bangkoku w 2013 r. (http://www.math.sc.chula.ac.th/math/www.math.sc.chula.ac.th/sae2013/index.html).

Konferencje, na których omawiane są zagadnienia estymacji dla małych obszarów odbywały się też w Polsce. Mam tu na myśli konferencje pt. *Metoda*

reprezentacyjna w badaniach ekonomiczno-społecznych, które organizuje Uniwersytet Ekonomiczny w Katowicach (zob. http://web2.ue.katowice.pl/ /metoda/) Kongres Statystyki Polskiej zorganizowany w 2012 r. w Poznaniu (http://www.stat.gov.pl/pts/kongres2012/) oraz cykl konferencji pt. *Wielowymiarowa analiza statystyczna*, które organizuje Uniwersytet Łódzki (http://www.msa.uni.lodz.pl/). Literatura dotycząca estymacji dla małych obszarów zawiera obszerny materiał teoretyczny wraz z przykładami zastosowań, co pozwala na wdrożenie opisywanych metod do praktyki statystycznej.

Metoda hierarchicznej estymacji bayesowskiej jest jedną z częściej stosowanych metod estymacji dla małych obszarów. O wzroście zainteresowania tą techniką świadczą ukazujące się ostatnio liczne publikacje, w tym prace doktorskie autorstwa M. Vogta (2010) i B. Liu (2009) oraz M. Karaganis (2009). Również w Polsce dostępna jest publikacja dotycząca zastosowania metod hierarchicznych w estymacji dla małych obszarów autorstwa W. Niemiro oraz J. Wesołowskiego (2010). Metoda ta zakłada znajomość rozkładów *a priori* $f(\lambda)$ zarówno dla parametrów rozważanego modelu, jak i rozkładów warunkowych $f(\mu, y|\lambda)$ parametrów małych obszarów μ (przy zadanych wartościach parametrów modelu) z uwzględnieniem danych y pochodzących z badania.

Wykorzystując twierdzenie Bayesa można uzyskać rozkład *a posteriori* $f(\mu|y)$. W prostych przypadkach rozkład taki można wyznaczyć analitycznie, jednak bardziej złożone przypadki wymagają zastosowania specjalnych metod MCMC (*Markov Chain Monte Carlo* — metoda Monte Carlo dla łańcuchów Markowa), które najczęściej wykonywane są numerycznie za pomocą tzw. próbnika Gibbsa.

HIERARCHICZNA ESTYMACJA BAYESOWSKA (HB) — ZASTOSOWANIE DLA MAŁYCH OBSZARÓW

Przedstawimy teraz założenia dotyczące metody HB nieco dokładniej. Załóżmy, że musimy otrzymać następujący rozkład *a posteriori*:

$$f(\mu|y) = \int f(\mu, \lambda|y)\, d\lambda \tag{1}$$

Używając wnioskowania bayesowskiego otrzymujemy następującą zależność:

$$f(\mu, \lambda|y) = \frac{f(y, \mu|\lambda) f(\lambda)}{f_1(y)} \tag{2}$$

gdzie $f_1(y)$ — rozkład brzegowy i ma postać:

$$f_1(y) = \int f(y, \mu|\lambda) f(\lambda)\, d\mu\, d\lambda \tag{3}$$

Jak już wspomniano, w konkretnych przypadkach do przeprowadzenia takich obliczeń potrzebna jest znajomość rozkładów *a priori*, które można uwzględnić przy konstruowaniu modeli dla małych obszarów. W rozważanym przypadku bierzemy pod uwagę model dla małych obszarów typu A, czyli tzw. podstawowy model poziomu obszaru, który ma postać:

$$\hat{\theta}_i = z_i^T \beta + b_i v_i + e_i \tag{4}$$

gdzie:

$\hat{\theta}_i$ — estymator badanej cechy θ_i dla małego obszaru i, dla której zachodzi zależność $\theta_i = z_i^T \beta + b_i v_i$,
z_i — wektor zmiennych objaśniających,
β — wektor współczynników regresji,
b_i — znane dodatnie stałe,
v_i — określa błąd modelu,
e_i — określa błąd wynikający ze schematu losowania.

Zakłada się ponadto często, że v_i ($i = 1, 2, \ldots$) są niezależnymi zmiennymi o jednakowych rozkładach (*independent, identically distributed* — iid) o następujących własnościach:

$$E_m(v_i) = 0 \qquad Var(v_i) = \sigma_v^2 \tag{5}$$

gdzie:

E_m — wartość oczekiwana składnika v dla modelu,
Var — wariancja dla modelu.

Z kolei przy błędach wynikających ze schematu losowania (dla ocen bezpośrednich) uwzględnia się, że:

$$E(e_i|\theta_i) = 0 \qquad Var(e_i|\theta_i) = \psi_i \tag{6}$$

Przyjmuje się, że błędy szacunku dla ocen bezpośrednich ψ_i są również znane. Biorąc pod uwagę założenia (4)—(6) oraz przyjmując, że znany jest oprócz parametrów rozkładów ocen bezpośrednich, w tym ich wariancji, także rozkład błędu modelu σ_v^2, który ma postać odwrotnego rozkładu Gamma $G^{-1}(a, b)$ o parametrach a i b (gdzie a jest współczynnikiem kształtu, zaś b współczynnikiem skali), możemy zapisać model hierarchiczny w następującej postaci:

$$
\begin{aligned}
&(\mathrm{i})\,\hat{\theta}_i|\theta_i,\beta,\sigma_v^2 \overset{ind}{\sim} N(\theta_i,\psi_i) \qquad i=1,\ldots,m\\
&(\mathrm{ii})\,\theta_i|\beta,\sigma_v^2 \overset{ind}{\sim} N\left(z_i^T\beta, b_i^2\sigma_v^2\right) \quad i=1,\ldots,m\\
&(\mathrm{iii})\,f(\beta)<1\\
&(\mathrm{iv})\,\sigma_v^2 \sim G^{-1}(a,b)
\end{aligned}
\tag{7}
$$

Rozważamy tutaj przypadek znanego rozkładu zmiennej σ_v^2 oraz „płaskiego" rozkładu *a priori* dla β, dany przez $f(\beta)<1$. Ponadto zakładamy, w odróżnieniu od modelu (10.3.1) z podręcznika Rao, że znana jest wartość parametrów *a* i *b* dla rozkładu σ_v^2, co naszym zdaniem jest dobrym przybliżeniem dla modelu (10.3.3)[1] z podręcznika Rao. Wartości tych parametrów mogą być wyznaczone na podstawie empirycznego rozkładu ocen błędów modelu pochodzących z modeli regresji liniowej. W tym przypadku bierzemy pod uwagę identyczne co do zmiennych objaśniających modele dla powiatów oraz zbliżone co do zmienności wartości ocen zarówno ocen bezpośrednich, jak i parametrów regresji liniowej β dla modelu. Można zatem sądzić, że przybliżenie takie może prowadzić do poprawnych ocen *a posteriori* dla rozważanego modelu.

Dodać należy, że zgodnie z sugestią Rao[2]: *gdy zakłada się o* σ_v^2*, że jest znane oraz* $f(\beta)<1$*, podejścia HB oraz BLUP w warunkach normalności prowadzą do identycznych estymacji punktowych oraz miar zmienności*. Należy jednak zwrócić uwagę, że model (10.3.1) u Rao nie uwzględnia w pełni zmienności σ_v^2, co skutkuje zgodnością, ale dla bardziej uproszczonych miar wariancji — zob. wyrażenie (7.1.6) u Rao[3]. Zatem uwzględnienie tej zmienności prowadzi do ocen zgodnych z szacunkami EBLUP (a więc uwzględniających pełną zmienność modelu).

METODY MONTE CARLO DLA ŁAŃCUCHÓW MARKOWA

Przyjmując, że $\eta=\left(\mu^T,\lambda^T\right)^T$ jest wektorem parametrów dla małych obszarów μ oraz parametrów modelu λ, należy zwrócić uwagę, że dla bardziej złożonych modeli, do których zaliczyć można z pewnością model (7), wylosowanie próby z łącznego rozkładu postaci (2) może być trudne ze względu na złożoną postać mianownika $f_1(y)$[4]. Zastosowanie w takim przypadku metody MCMC pozwala na uniknięcie takich trudności. Konstruuje się tu łańcuch Markowa $\left\{\eta^{(k)},\ k=0,1,2,\ldots\right\}$ taki, że rozkład $\eta^{(k)}$ jest zbieżny do rozkładu stacjonarnego

[1] Rao J. N. K. (2003), s. 240 i 241.
[2] Tamże, s. 237.
[3] Tamże, s. 117.
[4] Tamże, s. 224 i 225.

łańcucha Markowa, dla którego zachodzi zależność $f(\eta|y)=\pi(\eta)$. Stąd, pomijając początkowo wylosowane wartości (w ramach tzw. symulacji wstępnej — *burn in*, której długość wynosi *d*), otrzymujemy *D* zależnych prób $\eta^{(d)},...,\eta^{(d+D)}$ wylosowanych na podstawie rozkładu celu $f(\eta|y)$. Próba taka jest niezależna od punktu startowego $\eta^{(0)}$.

Każda konstrukcja łańcucha Markowa wymaga, aby prawdopodobieństwo przejścia łańcucha ze stanu $\eta^{(k)}$ w chwili *k* do stanu $\eta^{(k+1)}$ w chwili *k*+1 miało postać $P\left(\eta^{k+1}=\eta^{k+1}\middle|\eta^0=\eta^0,...,\eta^k=\eta^k\right)=P\left(\eta^{k+1}=\eta^{k+1}\middle|\eta^k=\eta^k\right)$, czyli nie zależało od ewolucji łańcucha do chwili *k*. W takim przypadku spełniony musi być warunek stacjonarności dla prawdopodobieństwa przejścia:

$$\int \pi\left(\eta^{(k)}\right)P\left(\eta^{(k+1)}\middle|\eta^{(k)}\right)d\eta^{(k)}=\pi\left(\eta^{(k+1)}\right) \tag{8}$$

Równanie (8) pokazuje, że jeśli $\eta^{(k)}$ można uzyskać z $\pi(\cdot)$, wtedy również $\eta^{(k+1)}$ można uzyskać z $\pi(\cdot)$. Konieczne jest również, aby zapewnić, że prawdopodobieństwo przejścia w *k* krokach $P\left(\eta^k=\eta^k\middle|\eta^0=\eta^0\right)$ jest dla $k\to\infty$ zbieżne do rozkładu stacjonarnego $\pi(\eta)$, niezależnie od tego, jakiego dokonamy wyboru dla $\eta^{(0)}$. Rozważany tu łańcuch powinien być nieprzywiedlny (*irreducible*) oraz nieperiodyczny (*aperiodic*). Nieprzywiedlność oznacza, że dla wszystkich punktów początkowych $\eta^{(0)}$ łańcuch osiąga dowolny niepusty zbiór przestrzeni stanów z dodatnim prawdopodobieństwem. Nieperiodyczność oznacza, że łańcuch nie powinien oscylować między różnymi zbiorami stanów w sposób periodyczny.

PRÓBNIK GIBBSA

Realizację obliczeniową metod MCMC można przeprowadzić przy pomocy tzw. próbnika Gibbsa[5]. Zakłada on, że ciąg próbek $\eta^{(k)}$ uzyskujemy dzieląc wektor η na bloki $\eta_1,...,\eta_r$. Bloki te mogą zawierać jeden lub więcej elementów, przykładowo, w podstawowym modelu obszaru, w którym $\mu=(\theta_1,...,\theta_m)^T=\theta$ oraz $\lambda=\left(\beta^T,\sigma_v^2\right)^T$. W takim przypadku η może składać się z następujących bloków: $\eta_1=\beta$, $\eta_2=\theta_1$, ..., $\eta_{m+1}=\theta_m$, $\eta_{m+2}=\sigma_v^2$, przy czym *r*=*m*+2. Wymagane jest, aby znany był następujący zbiór rozkładów warunkowych Gibbsa $f(\eta_1|\eta_2,...,\eta_r,y)$, $f(\eta_2|\eta_1,\eta_3,...,\eta_r,y)$, ..., $f(\eta_r|\eta_1,...,\eta_{r-1},y)$.

Próbnik Gibbsa wykorzystuje prawdopodobieństwo warunkowe do skonstruowania jądra przejścia $\mathrm{P}(\cdot|\cdot)$ takiego, że rozkład stacjonarny otrzymanego łań-

[5] Rao J. N. K. (2003), s. 225 i 226.

cucha Markowa jest równy $\pi(\eta) = f(\eta|y)$. Wynik ten jest konsekwencją faktu, że $f(\eta|y)$ jest jednoznacznie określone przez zbiór warunków Gibbsa.

Algorytm Gibbsa można przedstawić w następujących krokach:

0 — wybór punktu startowego $\eta^{(0)}$ dla komponentów $\eta_1^{(0)}, ..., \eta_r^{(0)}$, przyjmując k jako równe 0. Można np. wybrać jako punkty początkowe szacunki metodą REML (*restricted maximum likelihood*) dla parametrów modelu λ oraz szacunki EB (*empirical Bayes*) dla parametru μ. Mogą to być jednak dowolnie dobrane punkty;

1 — wygenerowanie $\eta^{(k+1)} = \left(\eta_1^{(k+1)}, ..., \eta_r^{(k+1)}\right)$ poprzez wylosowanie $\eta_1^{(k+1)}$, korzystając z rozkładu $f\left(\eta_1 \middle| \eta_2^{(k)}, ..., \eta_r^{(k)}, y\right)$, następnie trzeba wylosować $\eta_2^{(k+1)}$ z rozkładu $f\left(\eta_2 \middle| \eta_1^{(k+1)}, \eta_3^{(k)}, ..., \eta_r^{(k)}, y\right)$ aż do $\eta_r^{(k+1)}$ z rozkładu $f\left(\eta_r \middle| \eta_1^{(k+1)}, ..., \eta_{r-1}^{(k+1)}, y\right)$;

2 — przyjęcie $k = k+1$ oraz powrót do kroku 1.

Kroki 1 i 2 określają jeden cykl dla każdego k. Sekwencja $\{\eta^{(k)}\}$ wygenerowana przez próbnik Gibbsa jest łańcuchem Markowa ze stacjonarnym rozkładem $\pi(\eta) = f(\eta|y)$.

PRZYJĘTE ZAŁOŻENIA DO MODELU HIERARCHICZNEGO, W TYM PRZYJĘTE PARAMETRY ROZKŁADÓW

Tak jak już pokazano w (7), model hierarchiczny powinien zawierać kilka założeń dotyczących rozkładów *a priori* uwzględniających zarówno schemat losowania, przyjęty model objaśniający zjawisko oraz jego zmienność. W opracowaniu tym rozważane są szacunki dla małych obszarów — powiatów. Oznacza to konieczność pokonania kilku trudności, które wynikają m.in. z niedostatecznej wielkości próby. Szacunki bezpośrednie (w tym ich błąd estymacji) wyznaczono stosując specyficzną technikę, która zakładała określenie błędów szacunku z użyciem metody półprób zrównoważonych (BRR — *balanced repeated replication*) tam, gdzie zastosowanie tej metody jest możliwe oraz techniki bootstrap, w przypadku gdy nie da się zastosować metody BRR.

Technikę BRR stosowano, gdy dostępne były dwie podpróby dla konkretnego powiatu. Sytuacja taka zachodzi zwykle w przypadku większych powiatów, dla których dostępna jest próba o większej liczebności. W sytuacji kiedy nie można było wyodrębnić takich podprób, stosowano prostą metodę bootstrap w analogiczny sposób do metody opisanej w podręczniku Woltera[6]. W przeprowadzeniu takich obli-

[6] Wolter K. M. (2007), s. 214—217.

czeń dobierano każdą z prób bootstrapowych niezależnie (jest to jeden z warunków poprawnego przeprowadzenia takich obliczeń), zaś liczba wylosowanych prób bootstrapowych była równa 500, co jest wartością często stosowaną w praktyce.

Należy jednak zaznaczyć, że w przypadku kiedy stosuje się zasadę bootstrap dla małych prób, ocena zgodności szacunków z rzeczywistym błędem szacunku może być problematyczna[7], gdyż brak jest odpowiedniej teorii, która wystarczająco objaśniałaby takie szacunki. Dlatego obliczenia, które tutaj przedstawiono mogą mieć przybliżony charakter. Przeprowadzona wcześniej analiza tej metody (Kubacki i in., 2011[8]; Kubacki, Jędrzejczak, 2012[9]) wykazała skuteczność takiego podejścia. Porównanie szacunków bootstrapowych oraz otrzymanych z użyciem innych metod, w tym metody linearyzacji Taylora, wskazało na zgodność obydwu tych metod. Metoda BRR jest stosowana obecnie w BBGD do określenia precyzji szacunków[10].

Rozważono następujące zmienne określające kilka kategorii dochodu, wyznaczono na podstawie BBGD:

- rozporządzalny,
- z pracy najemnej,
- z pracy na własny rachunek,
- ze świadczeń z ubezpieczeń społecznych,
- z emerytur,
- z rent z tytułu niezdolności do pracy,
- z rent rodzinnych,
- ze świadczeń z pomocy społecznej,
- z zasiłków dla bezrobotnych.

Zmiennymi objaśniającymi w modelu regresji były zmienne pochodzące z rejestru POLTAX, określające następujące kategorie dochodu:

1) przychód z wynagrodzenia ze stosunku: pracy, służbowego, spółdzielczego itp.,
2) przychód z emerytury — renty krajowej, zagranicznej, renty strukturalnej itp.,
3) przychód z działalności wykonywanej osobiście,
4) przychód z praw majątkowych,
5) przychód z najmu, podnajmu lub dzierżawy,
6) przychód z pozostałych źródeł,
7) przychód z działów specjalnych produkcji rolnej,
8) odliczenia od dochodu (przychodu) składek na ubezpieczenia społeczne,
9) odliczenia od podatku (ryczałtu) składki na powszechne ubezpieczenie zdrowotne.

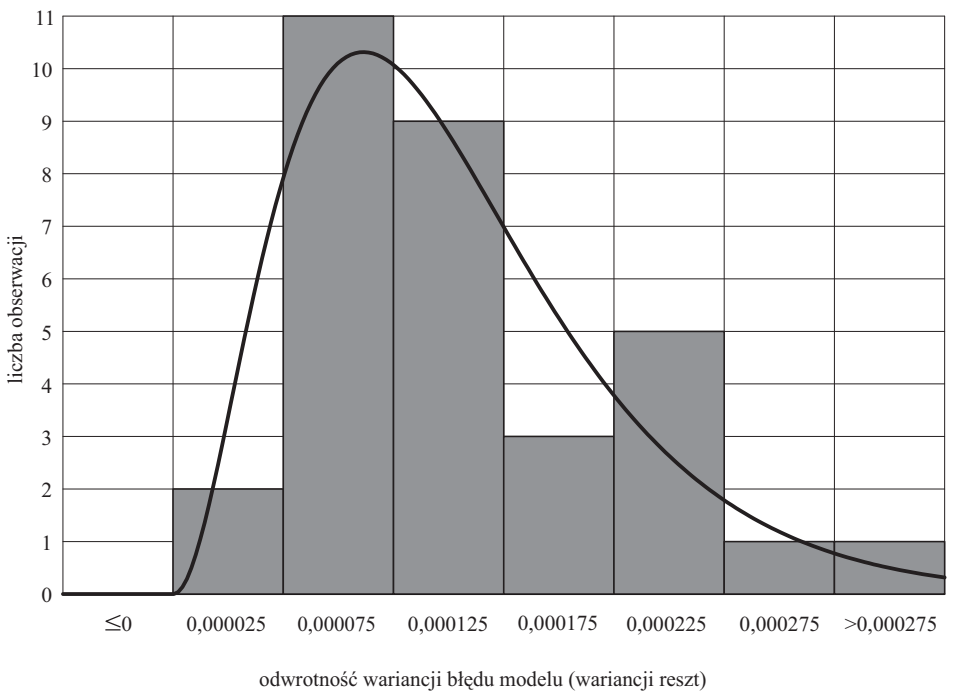
Zmienne 5), 6) i 7) były przedstawione w postaci sumy. Dane te zostały zagregowane na poziomie powiatów, a następnie wyznaczono wskaźniki średnich dochodów z wymienionych źródeł dzieląc sumy dla powiatów przez liczbę osób faktycznie zamieszkałych w danym powiecie. Wszystkie tak określone zmienne

[7] Tamże, s. 217.
[8] Strona 69 i 70.
[9] Strona 58—69.
[10] *Metodologia*… (2011), s. 15 —17.

objaśniające wykorzystywano w modelach dla zmiennych z BBGD. Nie oznacza to jednak, że liczba rozważanych zmiennych objaśniających nie mogłaby zostać uzupełniona np. o dane z ZUS lub z urzędów pracy. Można to jednak potraktować jako kolejną propozycję badawczą, tym bardziej że tylko część zmiennych z rejestru POLTAX odpowiada, co do określenia, zmiennym z BBGD. Może to osłabiać zgodność modeli dla małych obszarów z danymi empirycznymi.

Parametry rozkładu (iv) w modelu (7) wyznaczono określając wartość współczynników kształtu i współczynnika skali z wykorzystaniem rozkładu empirycznego, uzyskanego na podstawie błędu modelu dla regresji liniowej. Przykład takiego rozkładu pokazano na wykr. 1, przedstawiono rozkład błędu modelu otrzymanego z regresji liniowej za pomocą wzoru $\sigma^2 = 1/(m-p)\sum_{i=1}^{m}\left(\hat{\theta}_i - z_i^T\beta\right)^2$, przy czym m oznacza liczbę powiatów w modelu, p oznacza liczbę parametrów w równaniu regresji, zaś $\left(\hat{\theta}_i - z_i^T\beta\right)^2$ oznacza kwadrat reszty dla jednostki o indeksie i z równania regresji. Zaprezentowany histogram dotyczy odwrotności wielkości σ^2.

Wykr. 1. ROZKŁAD EMPIRYCZNY BŁĘDU MODELU UZYSKANY NA PODSTAWIE MODELI REGRESJI LINIOWEJ DLA DOCHODU ROZPORZĄDZALNEGO W POWIATACH Z UŻYCIEM DANYCH Z BBGD ORAZ ZMIENNYCH Z REJESTRU POLTAX W LATACH 2003 I 2004 (dopasowanie — rozkład gamma)



Ź r ó d ł o: obliczenia własne.

W obliczeniach wykorzystywano programy WinBUGS[11] oraz R-project[12] (w tym moduły R2WinBUGS[13], coda[14] i MASS). Specjalnie przygotowane makro w systemie R-project służyło jako łącznik pozwalający na wczytywanie danych, wykonywanie na ich podstawie niezbędnych obliczeń (w tym symulacji z użyciem programu WinBUGS) oraz ich automatyczną wizualizację.

Makro w systemie R-project ma (w uproszczeniu) postać następującą:

```
# określenie parametrów modelu:
model_HB <- paste("C:/Documents and Settings/PTS/Moje dokumenty/model_kongres_demo.txt",
sep = " ")
infile1 <- "coda1.txt"
infile2 <- "coda2.txt"
indfile <- "codaindex.txt"
burn_in <- 3000
zmienna <- " dochg"
a0 <- dochg_shape
b0 <- dochg_rate
data <- list(N=N, Y=Y, tau=tau, A=A, B=B, C=C, D=D, E=E, F=F, G=G, a0=a0, b0=b0)
model <- lm( Y ~ 1 + A + B + C + D + E + F + G)
mod_smry <- summary(model)
alpha <- as.vector(mod_smry$coefficients[,1])
sigma_2 <- (mod_smry$sigma)*(mod_smry$sigma)
precu <- 1/sigma_2
u <- vector(mode = "numeric", length = N)
inits <- list(list(alpha=alpha, precu=precu, u=u),list(alpha=alpha, precu=precu, u=u))
parameters <- c("mu", " alpha", "precu", "u")
# symulacja - wywołanie programu WinBUGS i wczytywanie wyników wg formatu coda
sim_HB <- bugs(data, inits, parameters, model_HB,n.chains=2, n.burnin = 1, n.iter=10000, n.thin = 1,
codaPkg=TRUE)
results1 <- read.coda(infile1, indfile, 2, 10000, 1)
results2 <- read.coda(infile2, indfile, 2, 10000, 1)
```

W szacunkach użyto schematu obliczeniowego stosowanego również we wcześniejszych pracach na temat wykorzystania metod hierarchicznych w szacunkach dla małych obszarów. Wspomnieć tu można dwie prace, które opracowali V. Gomez-Rubio (2008)[15] oraz M. Vogt (2010)[16]. W tym przypadku oceny $Y[p]$ dotyczą ocen na podstawie estymacji bezpośredniej wraz z oceną błędu oceny bezpośredniej $tau[p]$, wartości od $A[p]$ do $G[p]$ określone są przez warto-

[11] Spiegelhalter D. J. i in. (2003).
[12] Venables W. N., Ripley B. D. (2002).
[13] Sturtz S., Ligges U., Gelman A. (2005).
[14] Plummer M. i in. (2006).
[15] Strona 10.
[16] Strona 134.

ści zmiennych objaśniających z modelu, zaś parametry $a0$ i $b0$ pochodzą z rozkładu empirycznego *a priori* uzyskanego na podstawie rozkładu błędu modelu dla regresji liniowej. Oto schemat obliczeniowy:

```
model
{
 for(p in 1 : N) {
  Y[p] ~ dnorm(mu[p], tau[p])
  mu[p] <- alpha[1] + alpha[2] * A[p] + alpha[3] * B[p] + alpha[4] * C[p] + alpha[5] * D[p] + al-
pha[6] * E[p] + alpha[7] * F[p] + alpha[8] * G[p] + u[p]
  u[p] ~ dnorm(0, precu)
 }
 precu ~ dgamma (a0,b0)
 alpha[1] ~ dflat()
 alpha[2] ~ dflat()
 alpha[3] ~ dflat()
 alpha[4] ~ dflat()
 alpha[5] ~ dflat()
 alpha[6] ~ dflat()
 alpha[7] ~ dflat()
 alpha[8] ~ dflat()
 sigmau<-1/precu
}
```

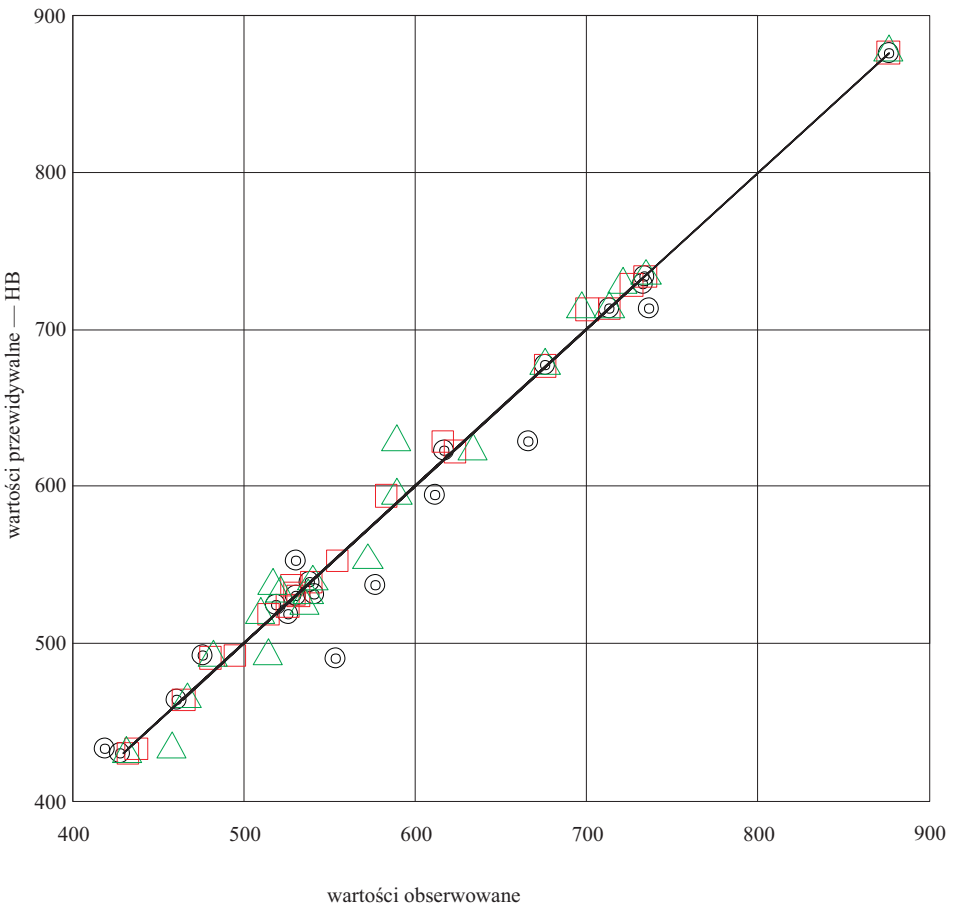
Diagnostyka zbieżności oparta na wyjściu MCMC dla $\{\eta^{(k)}\}$ jest często stosowana w praktyce w celu określenia długości symulacji wstępnej d. Popularna jest obecnie miara diagnostyczna oparta na wielokrotnych przebiegach oraz klasycznej analizie wariancji[17].

Załóżmy, że $h(\eta)$ określa skalarną miarę η, np. $h(\eta)=\mu_i$, czyli średnia dla i-tego małego obszaru. Oznaczając wartość $h(\eta)$ dla l-tego przebiegu (zakładamy, że skonstruowane są przynajmniej 2 przebiegi obliczeniowe, w ogólności jest ich L) jako $h_{l,d+1}, \ldots, h_{l,2d}$; $l=1, \ldots, L$, obliczamy wariancję między przebiegami (*between-run variance*) $B = d\sum_{l=1}^{L}\left(\bar{h}_{l\cdot} - \bar{h}_{\cdot\cdot}\right)^2 / (L-1)$ oraz wariancję wewnątrz przebiegu (*within-run variance*) $W = \sum_{l=1}^{L} \sum_{k=d+1}^{2d} \left(h_{l,k} - \bar{h}_{l\cdot}\right)^2 / [L(d-1)]$, gdzie $\bar{h}_{l\cdot} = \sum_k h_{lk} / d$ oraz $\bar{h}_{\cdot\cdot} = \sum_l \bar{h}_{l\cdot} / L$. Używając dwu komponentów wariancyjnych B oraz W można określić szacunek wariancji $h(\eta)$ w docelowym rozkładzie jako:

$$\hat{V} = \frac{d-1}{d} W + \frac{1}{d} B \tag{9}$$

[17] Gelman A., Rubin D. B. (1992), s. 457—472.

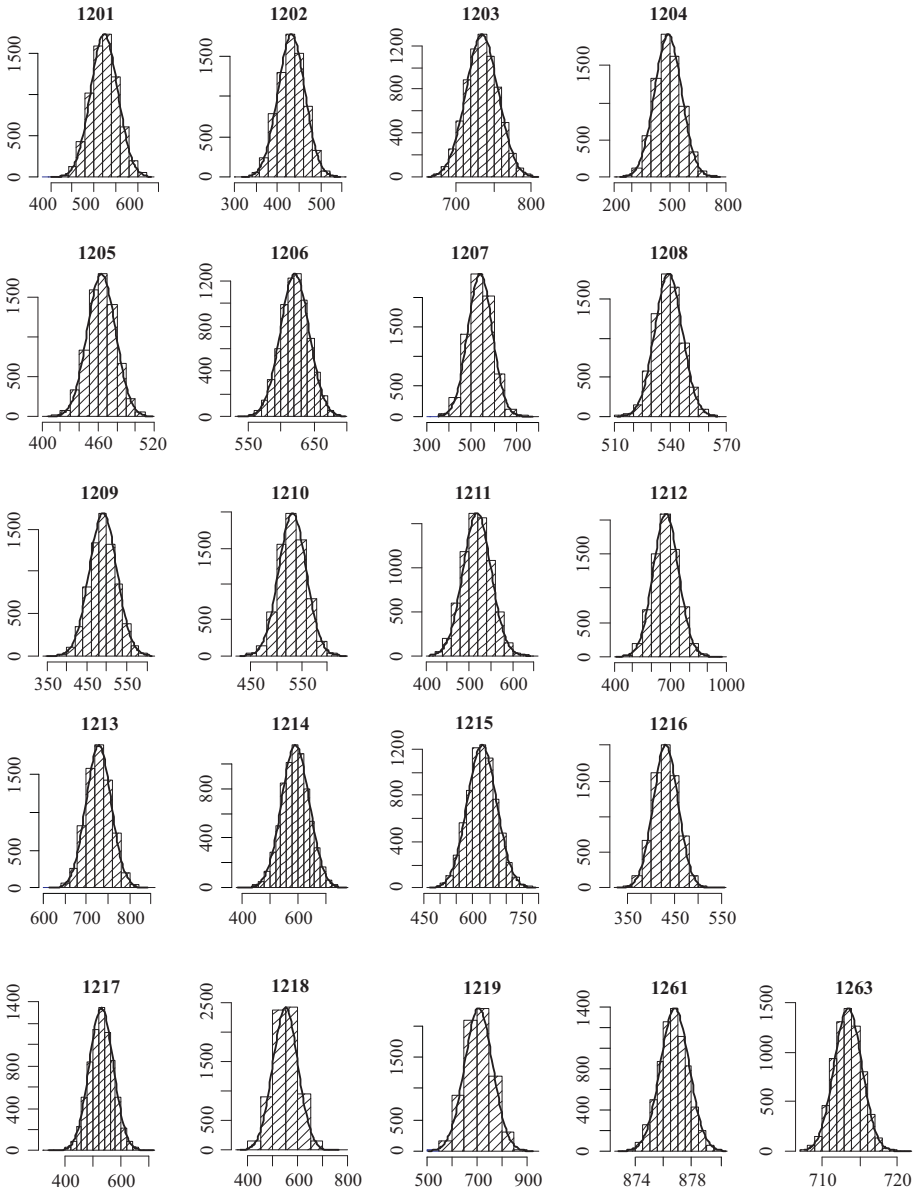
Wykr. 2. ROZRZUT DLA SZACUNKÓW DOCHODU ROZPORZĄDZALNEGO NA OSOBĘ NA PODSTAWIE BBGD ORAZ WYBRANYCH ZMIENNYCH Z REJESTRU POLTAX W 2003 R. DLA POWIATÓW W WOJ. MAŁOPOLSKIM OSZACOWANE Z UŻYCIEM ESTYMATORA BEZPOŚREDNIEGO (punkty czarne), EBLUP-REML (punkty czerwone), UPROSZCZONEJ ESTYMACJI EB (punkty zielone) ORAZ HIERARCHICZNEGO ESTYMATORA BAYESOWSKIEGO



kółka — oceny bezpośrednie,
kwadraty — EBLUP-REML,
trójkąty — naive EB

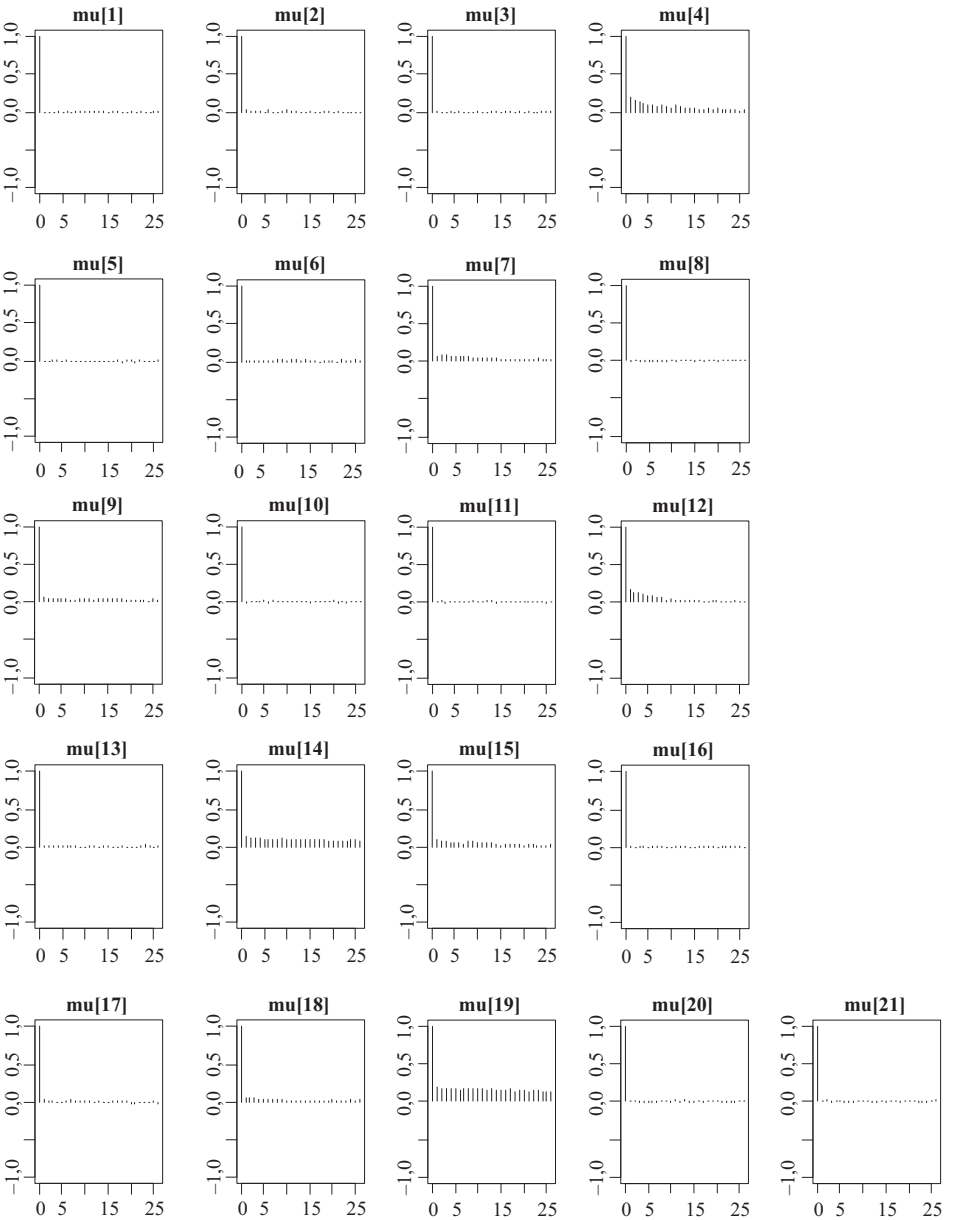
Ź r ó d ł o: jak przy wykr. 1.

Wykr. 3. ROZKŁAD SZACUNKÓW Z MODELU DLA DOCHODU ROZPORZĄDZALNEGO NA OSOBĘ NA PODSTAWIE BBGD ORAZ WYBRANYCH ZMIENNYCH Z REJESTRU POLTAX W 2003 R. DLA POWIATÓW W WOJ. MAŁOPOLSKIM OTRZYMANE W DRODZE SYMULACJI METODĄ MCMC Z UŻYCIEM PRÓBNIKA GIBBSA



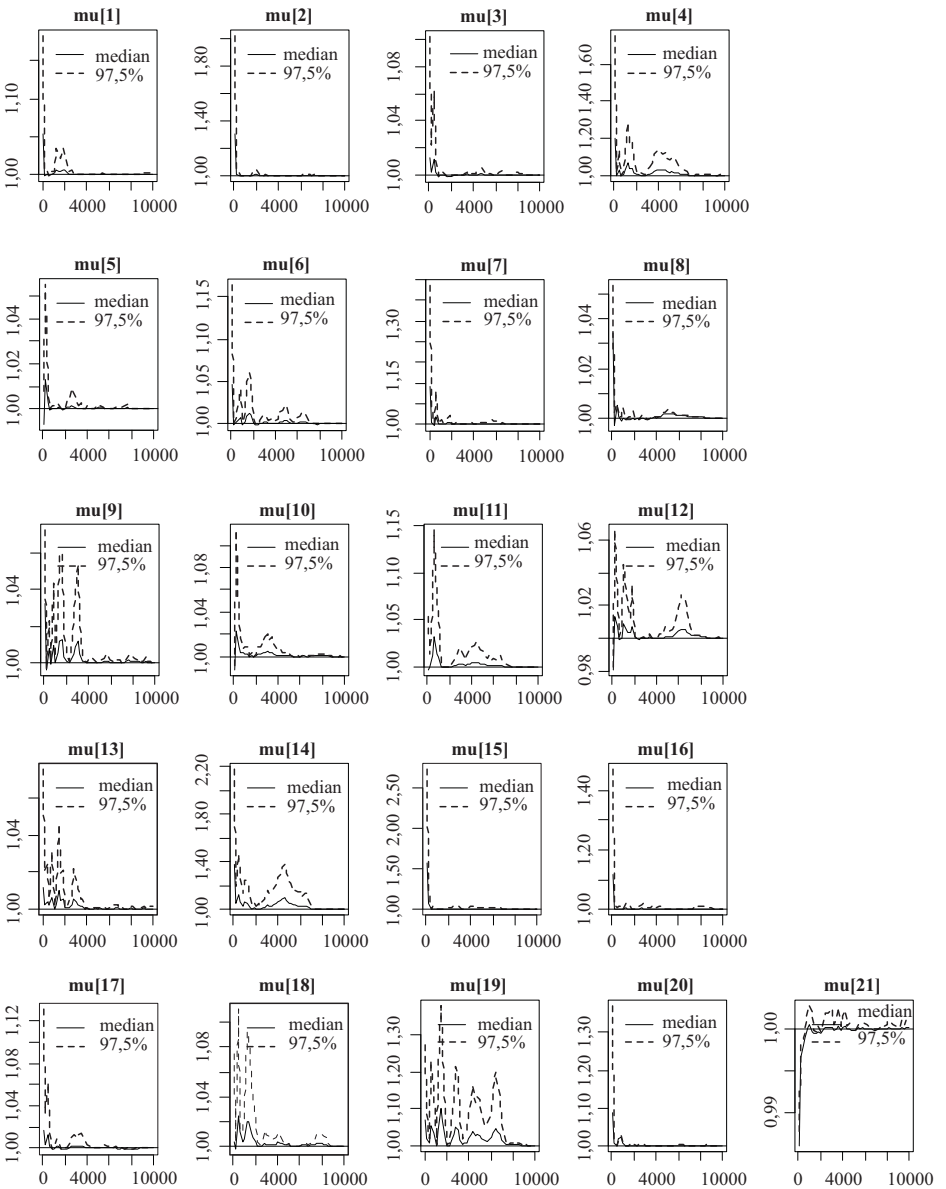
Ź r ó d ł o: jak przy wykr. 1.

Wykr. 4. DIAGNOSTYCZNA AUTOKORELACJA SZACUNKÓW Z MODELU DLA DOCHODU ROZPORZĄDZALNEGO NA OSOBĘ NA PODSTAWIE BBGD ORAZ WYBRANYCH ZMIENNYCH Z REJESTRU POLTAX W 2003 R. WEDŁUG POWIATÓW W WOJ. MAŁOPOLSKIM OTRZYMANE W DRODZE SYMULACJI METODĄ MCMC Z UŻYCIEM PRÓBNIKA GIBBSA



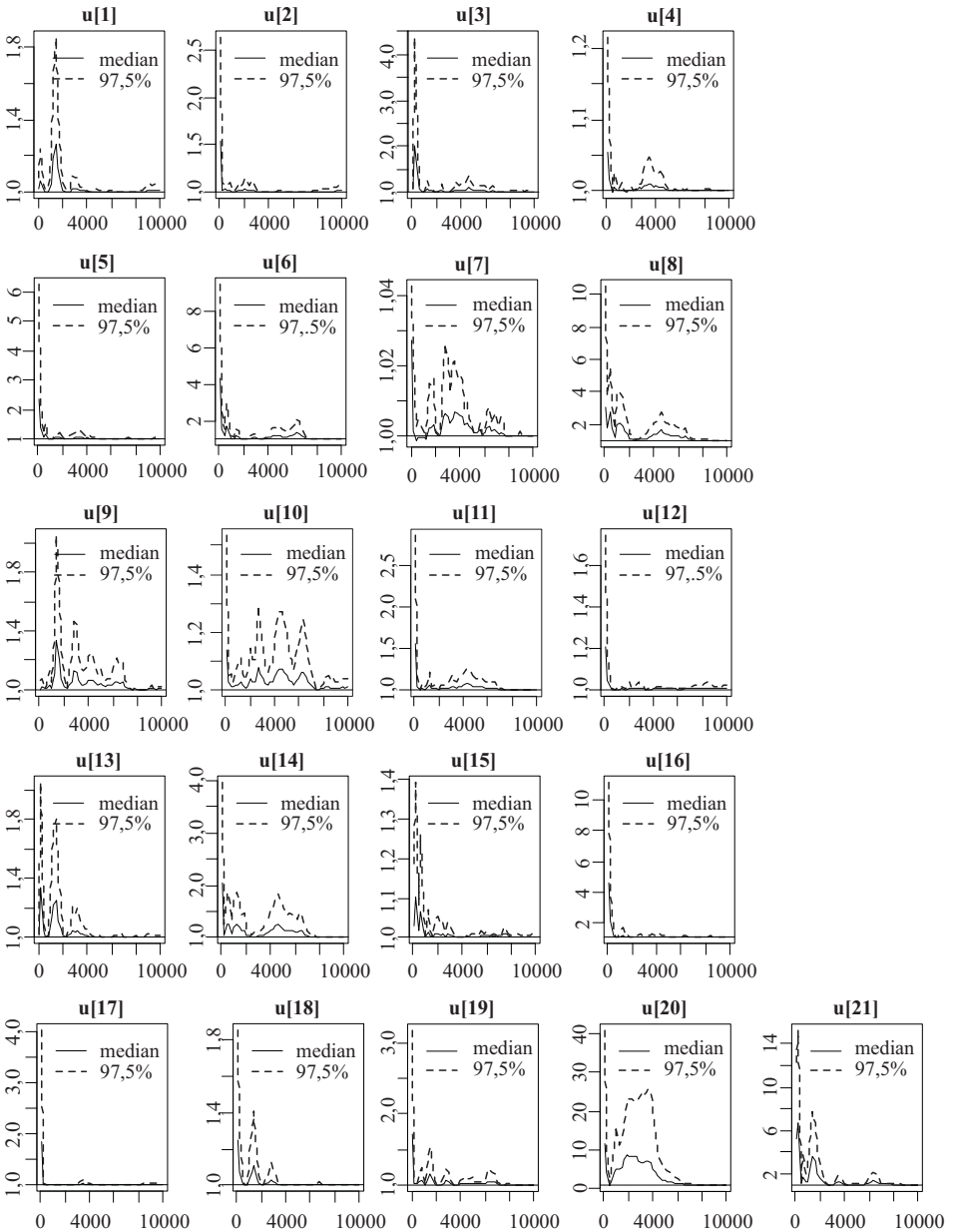
Ź r ó d ł o: jak przy wykr. 1.

Wykr. 5. DIAGNOSTYKA GRB SZACUNKÓW Z MODELU DLA DOCHODU ROZPORZĄDZALNEGO NA OSOBĘ NA PODSTAWIE BBGD ORAZ WYBRANYCH ZMIENNYCH Z REJESTRU POLTAX W 2003 R. WEDŁUG POWIATÓW W WOJ. MAŁOPOLSKIM OTRZYMANE Z UŻYCIEM PRÓBNIKA GIBBSA



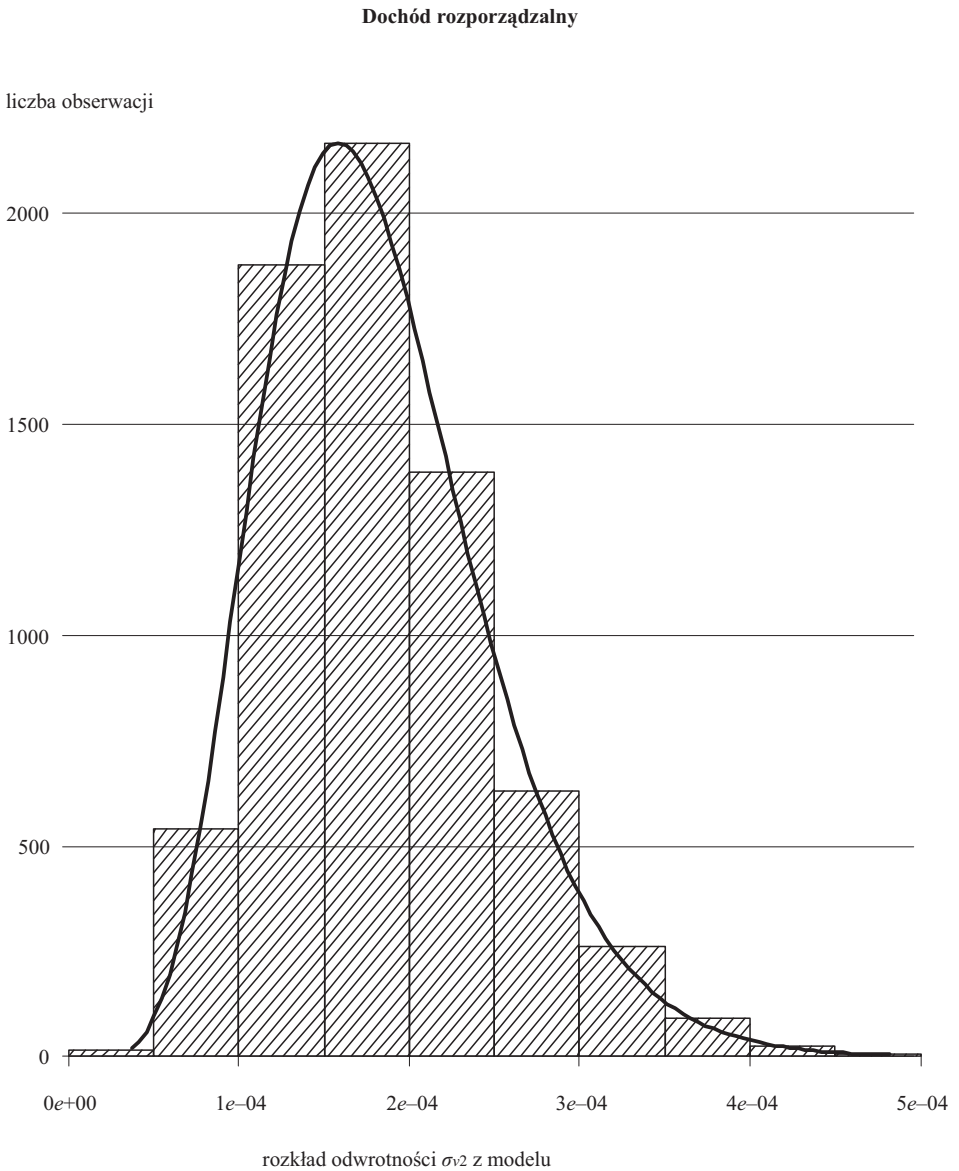
Ź r ó d ł o: jak przy wykr. 1.

Wykr. 6. DIAGNOSTYKA GRB SZACUNKÓW BŁĘDÓW MODELU (składniki losowe v_i) DLA DOCHODU ROZPORZĄDZALNEGO NA OSOBĘ NA PODSTAWIE BBGD ORAZ WYBRANYCH ZMIENNYCH Z REJESTRU POLTAX W 2003 R. DLA POWIATÓW W WOJ. MAŁOPOLSKIM OTRZYMANE W DRODZE SYMULACJI METODĄ MCMC Z UŻYCIEM PRÓBNIKA GIBBSA



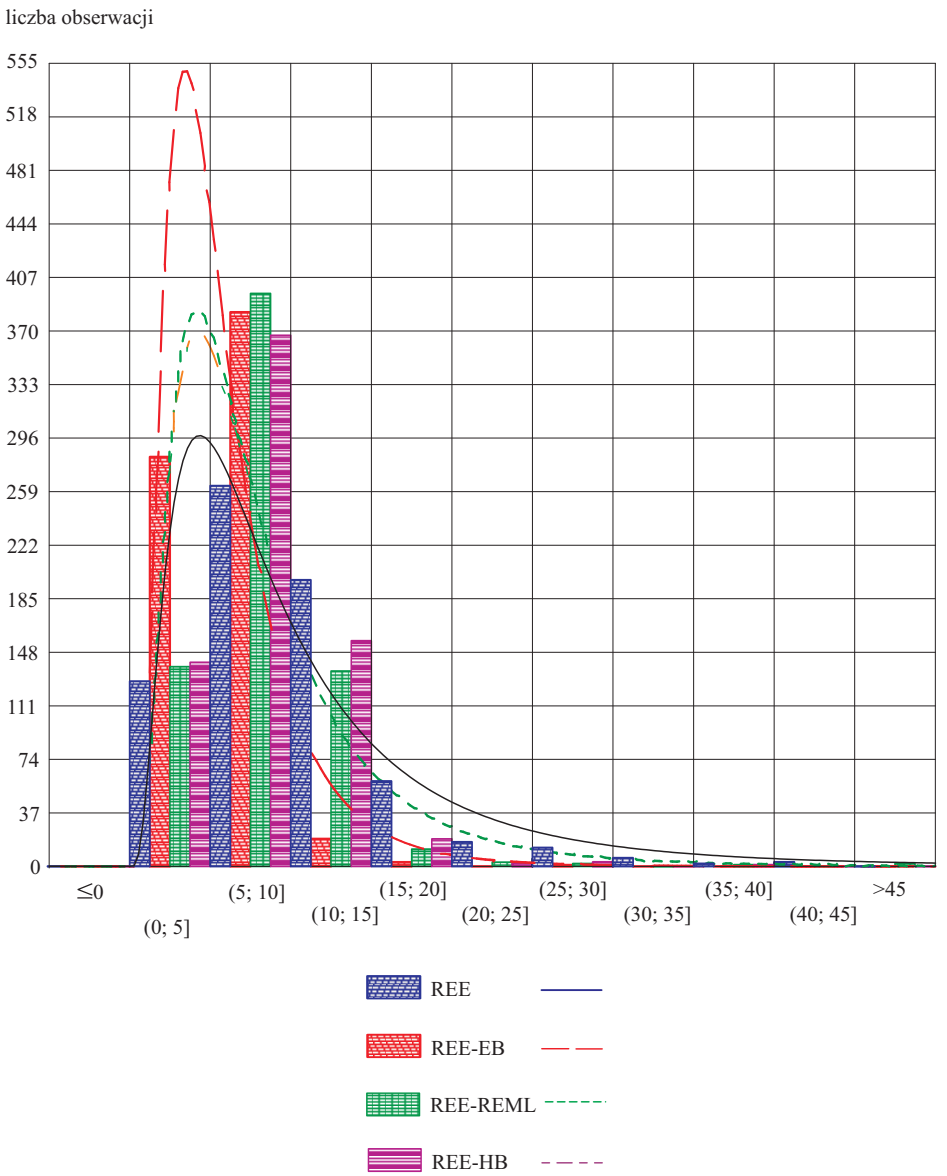
Ź r ó d ł o: jak przy wykr. 1.

Wykr. 7. ROZKŁAD BŁĘDU MODELU (odwrotność błędu modelu σ_v^2) UZYSKANY NA PODSTAWIE SYMULACJI METODĄ MCMC Z UŻYCIEM PRÓBNIKA GIBBSA ORAZ DANYCH Z BBGD I ZMIENNYCH Z REJESTRU POLTAX DLA WOJ. MAŁOPOLSKIEGO W 2003 R. (dopasowanie — rozkład gamma)



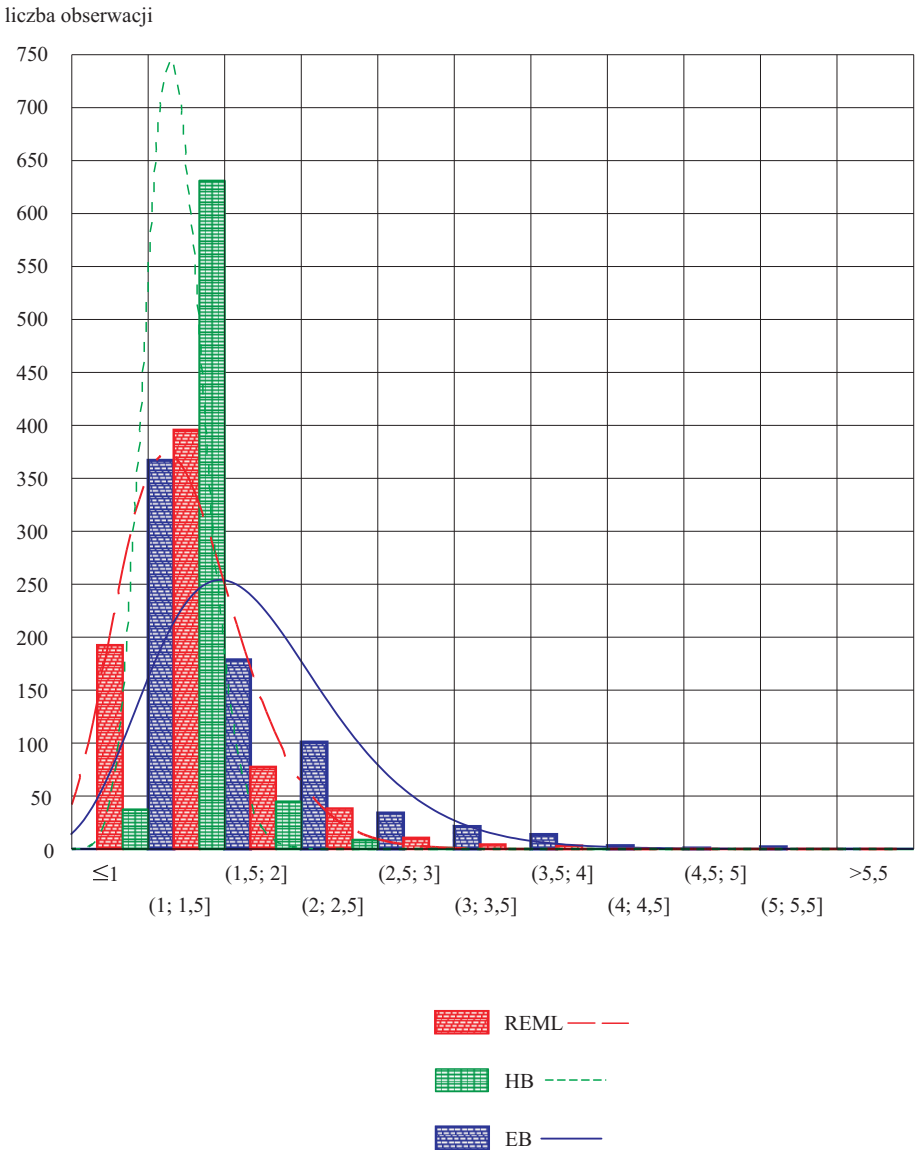
Ź r ó d ł o: jak przy wykr. 1.

Wykr. 8. ROZKŁAD WZGLĘDNEGO BŁĘDU SZACUNKU DLA ESTYMATORA BEZPOŚREDNIEGO ORAZ DLA ESTYMATORÓW EB, EBLUP (wariant REML) i HB DLA DOCHODU ROZPORZĄDZALNEGO W POWIATACH NA PODSTAWIE BBGD ORAZ DANYCH Z SYSTEMU POLTAX DLA LAT 2003 I 2004



Ź r ó d ł o: jak przy wykr. 1.

Wykr. 9. ROZKŁAD REDUKCJI WZGLĘDNEGO BŁĘDU SZACUNKU DLA ESTYMATORÓW EB, EBLUP (wariant REML) I HB DLA DOCHODU ROZPORZĄDZALNEGO W POWIATACH NA PODSTAWIE BBGD ORAZ DANYCH Z SYSTEMU POLTAX W LATACH 2003 I 2004



Ź r ó d ł o: jak przy wykr. 1.

Estymator docelowej wariancji $\widehat{V}$ parametrów używanych w modelu (np. μ lub v) przeszacowuje wartość $\widehat{V}$, przyjmując, że początkowy rozrzut dla przybiegów od 1 do L jest zbyt duży, ale w warunkach stacjonarności[18] jest nieobciążony (jeśli początkowy rozkład odpowiada rozkładowi docelowemu). Stosując tę drugą własność można obliczyć czynnik „potencjalnej redukcji skali" (*potential scale reduction factor*):

$$\widehat{R} = \widehat{V} / W$$

Jeśli osiągnięta jest zbieżność, $\widehat{R}$ będzie bliskie 1. W przeciwnym przypadku $\widehat{R}$ będzie znacząco większe niż 1, co może sugerować, że należy zwiększyć długość okresu *burn-in d*. Konieczne zatem jest wyznaczenie $\widehat{R}$ dla skalarnych parametrów, np. dla średnich i dla małych obszarów μ_i.

Wykresy dla autokorelacji pokazują korelację między zmienną $\eta^{(k)}$ oraz każdą ze zmiennych $\eta^{(k+t)}$ (gdzie t wskazuje na opóźnienie względem elementu k) w analizowanym łańcuchu. Korelacja dla opóźnienia t wyrażona jest zależnością $Corr(\eta^{(k)}, \eta^{(k+t)})$. Brak autokorelacji oznacza na wykresie, że dla $t = 0$ mamy korelację równą 1, zaś dla kolejnych wartości opóźnień wartości te są bliskie zeru. W rozważanych przykładach zazwyczaj kolejne wartości opóźnień wiążą się z coraz mniejszą wartością zależności $Corr(\eta^{(k)}, \eta^{(k+t)})$.

WYNIKI OBLICZEŃ I ICH DYSKUSJA

Jak wspomniano wcześniej, oceny z modelu dla estymatora HB mają (w przypadku uwzględnienia założeń dla modelu (7) zbliżoną postać do estymatora EBLUP i to zarówno jeśli chodzi o oceny punktowe, jak i oceny błędów szacunku. Zastosowana metoda pozwala ponadto na uzyskanie względnie stabilnego przebiegu symulacji. Rozkłady współczynników modelu liniowego, jak również oceny samych parametrów μ (wykr. 3) mają w niemal wszystkich przypadkach charakter normalny, zaś rozkład parametru σ_v^2 zachowuje odwrotny rozkład Gamma (wykr. 7). Normalny jest również dla przeważających przypadków rozkład błędów modelu. Poza tym przebieg symulacji zazwyczaj nie wykazuje autokorelacji (wykr. 4) i zachowuje stabilność praktycznie już od samego początku przebiegu (wykr. 5). Przedstawimy zatem wyniki dla woj. małopolskiego.

Pewną cechą szczególną w pokazanym przypadku jest istnienie widocznej autokorelacji szacunków błędów z modelu dla części powiatów. Jest to szczególnie widoczne w ocenach miast na prawie powiatów (Krakowa i Tarnowa). Wiąże się to z dość niską wartością błędu dla szacunku bezpośredniego w porównaniu z podobnymi przebiegami w przypadku innych powiatów. Takie za-

[18] Gelman A., Rubin D. B. (1992), s. 461.

chowanie się symulacji MCMC jest też obserwowane czasami dla innych zmiennych i modeli dla innych województw. Istnienie tej autokorelacji nie zmienia jednak w istotny sposób poziomu błędu ocen (REE), choć zmienia, ale dość rzadko w wyraźny sposób, rozkład składników losowych (parametry u w schemacie obliczeniowym). Ponadto zazwyczaj z testów diagnostycznych zbieżności symulacji (test *Gelmana-Rubina-Brooksa* — GRB) wynika również, że jej przebieg zachowuje zbieżność. Jest to szczególnie widoczne dla ocen parametrów z modelu (zmienne m, u — wykr. 5), a także dla składników losowych modelu (zmienna v odpowiadająca w schemacie obliczeniowym zmiennej u — wykr. 6), choć tutaj zależność jest słabsza i pełną zbieżność osiąga się dla dłuższego okresu. Jednak zazwyczaj i w tym przypadku wartość statystyki GRB osiąga wartości bliskie 1 już po 2—3 tys. iteracji (co wiąże się z istnieniem autokorelacji). Ostatecznie jednak po 5—7 tys. iteracji nawet przebiegi z autokorelacją dla składników losowych stabilizują się, co widoczne jest na wykr. 6.

TABL. 1. WARTOŚCI OSZACOWAŃ DOCHODU ROZPORZĄDZALNEGO NA OSOBĘ NA PODSTAWIE BBGD ORAZ WYBRANYCH ZMIENNYCH NA PODSTAWIE REJESTRU POLTAX DLA 2003 R. I POWIATÓW W WOJ. MAŁOPOLSKIM, WRAZ Z OCENĄ PRECYZJI SZACUNKÓW (względny błąd szacunku w %), A TAKŻE REDUKCJĄ WZGLĘDNEGO BŁĘDU SZACUNKU, OSZACOWANE Z UŻYCIEM ESTYMATORA BEZPOŚREDNIEGO ORAZ METODY EBLUP STOSUJĄCEJ TECHNIKĘ REML

Powiaty	Dochód rozporządzalny						
	szacunki bezpośrednie			szacunki dla metody EBLUP			redukcja REE
	ocena parametru	błąd szacunku	REE w %	ocena parametru	błąd szacunku	REE w %	
Bocheński	518,43	33,13	6,39	525,14	31,79	6,05	1,056
Brzeski	418,25	32,80	7,84	436,81	31,77	7,27	1,078
Chrzanowski	734,24	22,09	3,01	734,63	22,26	3,03	0,993
Dąbrowski	553,59	106,13	19,17	480,11	66,72	13,90	1,379
Gorlicki	460,38	14,97	3,25	464,59	15,06	3,24	1,003
Krakowski	617,31	22,41	3,63	623,05	22,17	3,56	1,020
Limanowski	576,63	69,00	11,97	527,75	53,08	10,06	1,190
Miechowski	538,66	7,31	1,36	539,06	7,31	1,36	1,001
Myślenicki	475,63	36,64	7,70	494,48	34,97	7,07	1,089
Nowosądecki	540,52	27,76	5,14	529,10	27,26	5,15	0,997
Nowotarski	525,31	34,79	6,62	514,30	33,22	6,46	1,025
Olkuski	675,92	84,13	12,45	675,93	66,35	9,82	1,268
Oświęcimski	732,83	30,13	4,11	726,44	29,82	4,10	1,002
Proszowicki	611,29	56,11	9,18	582,99	49,62	8,51	1,078
Suski	666,01	51,49	7,73	615,90	44,93	7,30	1,060
Tarnowski	427,53	26,72	6,25	431,84	27,52	6,37	0,981
Tatrzański	529,82	45,91	8,67	531,46	44,27	8,33	1,040
Wadowicki	530,13	62,80	11,85	554,40	50,37	9,09	1,304
Wielicki	736,62	68,03	9,24	700,67	56,16	8,02	1,152
M. Kraków	876,75	1,03	0,12	876,74	1,03	0,12	1,000
M. Tarnów	713,46	1,84	0,26	713,46	1,84	0,26	1,000

Ź r ó d ł o: obliczenia własne.

TABL. 2. WARTOŚCI OSZACOWAŃ DOCHODU ROZPORZĄDZALNEGO NA OSOBĘ NA PODSTAWIE BBGD ORAZ WYBRANYCH ZMIENNYCH NA PODSTAWIE REJESTRU POLTAX DLA 2003 R. I POWIATÓW W WOJ. MAŁOPOLSKIM, WRAZ Z OCENĄ PRECYZJI SZACUNKÓW (względny błąd szacunku w %), A TAKŻE REDUKCJĄ WZGLĘDNEGO BŁĘDU SZACUNKU, OSZACOWANE Z UŻYCIEM ESTYMATORA BEZPOŚREDNIEGO ORAZ METODY HIERARCHICZNEJ ESTYMACJI BAYESOWSKIEJ (HB)

Powiaty	Dochód rozporządzalny						
	szacunki bezpośrednie			szacunki dla metody HB			redukcja REE
	ocena parametru	błąd szacunku	REE w %	ocena parametru	błąd szacunku	REE w %	
Bocheński	518,43	33,13	6,39	523,74	31,28	5,97	1,070
Brzeski	418,25	32,80	7,84	432,23	30,96	7,16	1,095
Chrzanowski	734,24	22,09	3,01	734,15	21,90	2,98	1,009
Dąbrowski	553,59	106,13	19,17	490,32	71,80	14,64	1,309
Gorlicki	460,38	14,97	3,25	463,79	15,02	3,24	1,004
Krakowski	617,31	22,41	3,63	622,27	21,81	3,51	1,035
Limanowski	576,63	69,00	11,97	536,88	54,61	10,17	1,176
Miechowski	538,66	7,31	1,36	538,96	7,31	1,36	1,001
Myślenicki	475,63	36,64	7,70	492,01	33,84	6,88	1,120
Nowosądecki	540,52	27,76	5,14	531,27	26,97	5,08	1,012
Nowotarski	525,31	34,79	6,62	518,01	32,16	6,21	1,067
Olkuski	675,92	84,13	12,45	676,95	67,77	10,01	1,243
Oświęcimski	732,83	30,13	4,11	728,56	29,01	3,98	1,032
Proszowicki	611,29	56,11	9,18	593,89	49,39	8,32	1,104
Suski	666,01	51,49	7,73	628,43	45,05	7,17	1,079
Tarnowski	427,53	26,72	6,25	429,27	26,25	6,12	1,022
Tatrzański	529,82	45,91	8,67	530,19	42,90	8,09	1,071
Wadowicki	530,13	62,80	11,85	552,55	50,79	9,19	1,289
Wielicki	736,62	68,03	9,24	713,04	55,65	7,80	1,183
M. Kraków	876,75	1,03	0,12	876,75	1,03	0,12	1,004
M. Tarnów	713,46	1,84	0,26	713,45	1,85	0,26	0,999

Ź r ó d ł o: obliczenia własne.

Podane zależności dla symulacji metodą MCMC obserwowane są też w przypadku innych zmiennych niż dochód rozporządzalny, mimo że niekiedy dopasowanie modelu jest słabsze. Zachowanie normalności rozkładów z symulacji w takich przypadkach może oznaczać też, że założenia dotyczące normalności rozkładów warunkowych ocen rozważanych parametrów (w tym $\hat{\theta}_i$) oraz ich błędów mogą być w praktyce spełnione. Może to usprawiedliwiać zastosowanie modelu (7) oraz metod EBLUP dla takich szacunków, które często takiego założenia o normalności wymagają. Trudno jednak sądzić, czy hipotezę tę można by potwierdzić empirycznie, tym bardziej że w rzeczywistych badaniach obserwuje się zarówno zmianę uwarunkowań społeczno-ekonomicznych związanych z badaniem oraz zmianę poziomu samego zjawiska (np. wskutek zmian cen oraz PKB). Stąd obserwowane prawidłowości mogą dotyczyć ściśle tylko hipotetycznych populacji, często określanych jako superpopulacje, zaś ich postać teoretyczna wyrażona modelem (7) oraz jego praktyczna realizacja może być raczej przybliżeniem zależności obserwowanych w rzeczywistości.

Obliczenia przeprowadzone dla woj. małopolskiego wykazują na ogół małe różnice między wartościami błędu szacunku dla metody EBLUP (pokazano tutaj przykład obliczeń wykonanych dla metody REML z użyciem pakietu SAE[19]) oraz dla HB, podobnie dla większości pozostałych modeli, dla których oceny błędów szacunku dla metody EBLUP oraz HB są zbliżone. Świadczy o tym porównanie rozkładu względnego błędu szacunku (REE) przedstawione na wykresie 8. Istnieją jednak tutaj pewne różnice, o czym świadczy wykr. 9. Widać tutaj, że estymacja hierarchiczna daje z grubsza nieco wyższą redukcję REE niż technika EBLUP. Metodę EBLUP wyróżnia jednak nieco bardziej płaski przebieg redukcji REE, co szczególnie widoczne jest dla uproszczonej empirycznej estymacji bayesowskiej (*naive EB*).

Obserwowane dla innych województw różnice (Kubacki, 2012[20]) można też objaśnić słabszym dopasowaniem modelu do danych dla większej liczby małych obszarów (powiatów). Jest to widoczne nawet w zakresie zwykłych modeli regresji i dlatego może stanowić ograniczenie w stosowaniu metod takich, jak HB. Należy jednak dodać, że dla bardziej specyficznych zmiennych (np. dla modelu dla rent rodzinnych lub zasiłków dla bezrobotnych) modele hierarchiczne przedstawiane tutaj mają tę przewagę, że nie obserwuje się dla nich braku zbieżności, tak jak niekiedy widoczne jest to dla obliczeń z użyciem techniki EBLUP. Jest to jednak nie tyle cecha modelu hierarchicznego, co doboru parametrów, gdyż — jak sprawdzono empirycznie — inny, arbitralny zestaw parametrów rozkładu Gamma (np. taki jak zaproponowany w pracy M. Vogta (2010)[21], $a = 0{,}5$, $b = 0{,}0005$) nie sprawdza się w praktyce dla bardziej specyficznych zmiennych. W takim bowiem przypadku obserwowano częściej autokorelację oraz niekiedy brak stabilności przebiegu symulacji (np. oscylacje dla dłuższych przebiegów). Zastosowanie zatem takiego bardziej ogólnego podejścia nie zawsze jest skuteczne.

Należy dodać, że tak przedstawiony dobór parametrów modelu jest możliwy tylko wtedy, gdy dysponuje się większą liczbą modeli o podobnych własnościach, jak w przypadku modeli dla powiatów w latach 2003 i 2004. W bardziej specyficznych przypadkach (np. gdy rozważany jest model dochodu rozporządzalnego dla województw) dostępność dostatecznie dużej liczby przypadków może być ograniczona. W takiej sytuacji przyjęcie innej strategii niż opisana może być uzasadnione. Jedno z takich podejść (kiedy nie jest znane σ_v^2) pokazano w części 10.3.3 podręcznika Rao[22]. Być może porównanie obu tych metod pozwoliłoby na pełniejszą ocenę podejścia zaproponowanego w opracowaniu oraz metody przyjętej u Rao i ewentualne rozstrzygnięcie, który z rozważanych wariantów jest właściwszy.

---

[19] Salvati N., Gómez-Rubio V. (2006).
[20] Strona 269—271.
[21] Strona 13.
[22] Strona 240—243.

Podsumowanie

W artykule wykazano przydatność szacunków wykonanych z użyciem hierarchicznej estymacji bayesowskiej w przypadku znanych wartości hiperparametrów modelu. Pokazano istnienie zgodności między szacunkami z użyciem takiej metody oraz szacunkami z użyciem innej techniki dla małych obszarów, w tym metody EBLUP. Z uwagi na dobre własności przedstawionych obliczeń symulacyjnych (na ogół brak autokorelacji i stabilność symulacji) można sądzić, że podejście takie może znaleźć zastosowanie praktyczne. Charakterystyczna jest też tutaj nieco większa efektywność obliczeń dla techniki HB, choć dla słabiej dopasowanych modeli nie musi to być regułą. Mankamentem jest jednak konieczność posiadania wstępnego materiału empirycznego, dla którego wyznaczany jest rozkład parametru σ_v^2. W przypadku modeli dla powiatów jest to jednak możliwe i może być korzystne ze względów praktycznych.

---

dr Jan Kubacki — *Urząd Statystyczny w Łodzi*

LITERATURA

Bracha Cz., Lednicki B., Wieczorkowski R. (2004), *Wykorzystanie złożonych metod estymacji do dezagregacji danych z Badania Aktywności Ekonomicznej Ludności w roku 2003*, GUS, „Z prac Zakładu Badań Statystyczno-Ekonomicznych GUS i PAN", z. 299

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

Gelman A., Rubin D. B. (1992), *Inference from Iterative Simulation Using Multiple Sequences*, „Statistical Science", No. 7

Gołata E. (2004), *Estymacja pośrednia bezrobocia na lokalnym rynku pracy*, Wydawnictwo Akademii Ekonomicznej w Poznaniu

Gomez-Rubio V. (2008), *Small Area Estimation with R Unit 5: Bayesian Small Area Estimation*, useR! 2008, 11 August 2008, Dortmund, http://www.bias-project.org.uk/SAE_tutorial/useR08--tutorial.tgz

Kalton G., Kordos J., Platek R. (1993), *Small Area Statistics and Survey Designs*, Vol. I: Invited Papers: Vol. II: Contributed Papers and Panel Discussion, Central Statistical Office, Warsaw

Karaganis M. (2009), *Small Area Estimation for Survey Data: A Hierarchical Bayes Approach*, A Thesis submitted to the Faculty of Graduate Studies of The University of Manitoba, Winnipeg, http://mspace.lib.umanitoba.ca/bitstream/1993/3207/1/Final%20thesis%20-%20Aug%2020%202009.pdf

Kubacki J. (2004), *Application of the Hierarchical Bayes Estimation to the Polish Labour Force Survey*, „Statistics in Transition", Vol. 6, No. 5

Kubacki J. (2012), *Estimation of parameters for small areas using hierarchical Bayes method in the case of known model hyperparameters*, „Statistics in Transition-new series", Summer 2012, Vol. 13, No. 2

Kubacki J., Jędrzejczak A., Piasecki T. (2011), *Wykorzystanie metod statystyki małych obszarów do opracowania wyników badań statystycznych*, Raport z pracy metodologicznej 3.065, Ośrodek Statystyki Matematycznej, Urząd Statystyczny w Łodzi

Kubacki J., Jędrzejczak A. (2012), *The Comparison of Generalized Variance Function with Other Methods of Precision Estimation for Polish Household Budget Survey*, „Studia Ekonomiczne", nr 120, http://wydawnictwo.ue.katowice.pl/uploads/media/SE_120.pdf

Liu B. (2009), *Hierarchical Bayes Estimation and Empirical Best Prediction of Small Area Proportions*, Dissertation submitted to the Faculty of the Graduate School of the University of Maryland, College Park, http://drum.lib.umd.edu/bitstream/1903/9149/1/Liu_umd_0117E_10245.pdf

Longford N. T. (2005), *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*, Springer-Verlag, New York

Metodologia Badania Budżetów Gospodarstw Domowych (2011), GUS, http://stat.gov.pl/obszary--tematyczne/warunki-zycia/dochody-wydatki-i-warunki-zycia-ludnosci/metodologia-badania--budzetow-gospodarstw-domowych,10,1.html

Niemiro W., Wesołowski J. (2010), *Statystyka małych obszarów — symulacyjne badanie próbnika Gibbsa w hierarchicznym modelu bayesowskim*, „Biblioteka Wiadomości Statystycznych", tom 62, GUS

Plummer M., Best N., Cowles K., Vines K. (2006), *CODA: Convergence Diagnosis and Output Analysis for MCMC*, R News, Vol. 6

Rao J. N. K. (2003), *Small Area Estimation*, Wiley Interscience, Hoboken, New Jersey

Salvati N., Gomez-Rubio V. (2006), *SAE: Small Area Estimation with R*, R package version 0.07, http://www.bias-project.org.uk/software/SAE_0.07.zip

Spiegelhalter D. J., Thomas A., Best N., Lunn D. (2003), *WinBUGS User Manual*, Version 1.4

Sturtz S., Ligges U., Gelman A. (2005), *R2WinBUGS: A Package for Running WinBUGS from R*, „Journal of Statistical Software", No. 12(3)

Venables W. N., Ripley B. D. (2002), *Modern Applied Statistics with S*, Fourth Edition, Springer Verlag, New York

Vogt M. (2010), *Bayesian Spatial Modeling: Propriety and Applications to Small Area Estimation with Focus on the German Census 2011*, PhD Thesis, University of Trier, http://ubt.opus.hbz--nrw.de/volltexte/2010/578/pdf/Dissertation_Martin_Vogt.pdf

Wolter K. M. (2007), *Introduction to Variance Estimation*, Second Edition, Springer Verlag, New York

Żądło T. (2008), *Elementy statystyki małych obszarów z programem R*, Akademia Ekonomiczna Katowice

SUMMARY

The author presents a method of hierarchical bayesian estimation to estimate the value of the different income variables on the basis of studies of household budgets and POLTAX tax register. Calculations have been made for the case where approximately known a priori evaluation of hyperparameters used to construct a conditional probability, which is used in the model. The author compares the efficiency of the estimates obtained by using other hierarchical methods of estimation for small areas, including the EBLUP estimators type. This gave congruity in precision of the estimated parameters using both techniques.

РЕЗЮМЕ

Статья представляет метод байесовской иерархической эстимации для оценки значения разных статей дохода на основе Обследования бюджетов домашних хозяйств и регистра налогов POLTAX. Исчисления были проведены для случая, когда являются приблизительно известными оценки гиперпараметров a priori, использованные для образования условной вероятности использованной в модели.

В статье были также сопоставлены эффективность оценок полученных в результате использования иерархических методов с другими методами оценивания для малых домэн, в том числе оценок типа EBLUP получая соответствие оценок точности параметров оцениваемых с использованием обоих техник.