

Zastosowanie hierarchicznej estymacji bayesowskiej w szacowaniu wartości dochodów ludności dla powiatów

Jan Kubacki

Ośrodek Statystyki Matematycznej,
Urząd Statystyczny w Łodzi



Hierarchiczna estymacja bayesowska - wprowadzenie

Hierarchiczna estymacja bayesowska - jedną z częściej stosowanych współcześnie metod estymacji dla małych obszarów.

Liczne publikacje w ostatnich latach, w tym prace doktorskie autorstwa Martina Vogta (2010) i Benmei Liu (2009) oraz pracę Milany Karaganis (2009).

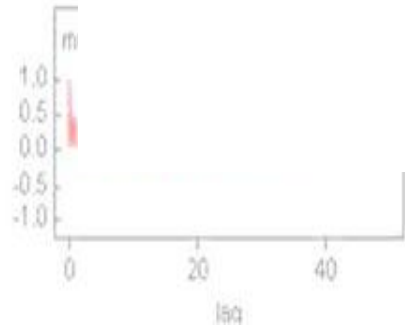
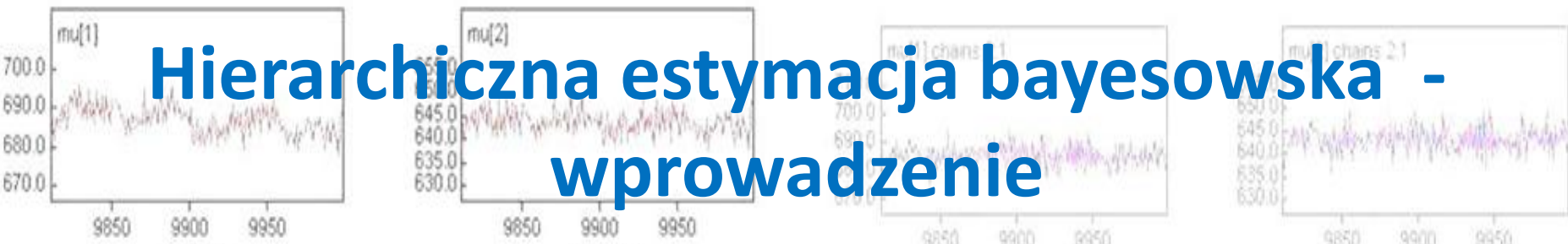
Wymagana jest znajomość rozkładów a priori $f(\lambda)$ zarówno dla parametrów rozważanego modelu, jak i rozkładów warunkowych $f(\mu, y/\lambda)$ parametrów małych obszarów μ (przy zadanych wartościach parametrów modelu) z uwzględnieniem danych y pochodzących z badania.

Zastosowanie twierdzenie Bayesa pozwala na uzyskanie rozkładu a posteriori $f(\mu/y)$.

W prostych przypadkach – rozkład można wyznaczyć analitycznie

Bardziej złożone przypadki wymagają zastosowania specjalnych metod obliczeniowych z użyciem metod MCMC (Markov Chain Monte Carlo – metody Monte Carlo dla łańcuchów Markowa),

Najczęściej realizowane jest to numerycznie za pomocą tzw. próbnika Gibbsa (Gibbs sampler).



Hierarchiczna estymacja bayesowska (HB) – zastosowanie dla małych obszarów

Założmy, że musimy otrzymać następujący rozkład a posteriori:

$$f(\mu | y) = \int f(\mu, \lambda | y) d\lambda$$

Używając wnioskowania bayesowskiego dostajemy następującą zależność

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

gdzie, $f_1(y)$ jest rozkładem brzegowym i ma postać

$$f_1(y) = \int f(y, \mu | \lambda) f(\lambda) d\mu d\lambda$$

Jak to już wspomniano we wprowadzeniu w konkretnych przypadkach do przeprowadzenia takich obliczeń potrzebna jest znajomość rozkładów a priori, które można uwzględnić przy konstruowaniu konkretnych modeli dla małych obszarów.

Rozważany model – model typu A

W rozważanym tutaj przypadku bierzemy pod uwagę model dla małych obszarów typu A, czyli tzw. podstawowy model poziomu obszaru (ang. Basic area level model), który jest postaci

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

Gdzie $\hat{\theta}_i$ jest estymatorem badanej cechy dla małego obszaru i , z_i jest wektorem zmiennych objaśniających, β jest wektorem współczynników regresji, b_i są znanymi dodatnimi stałymi, v_i określa błąd modelu, zaś e_i określa błąd wynikający ze schematu losowania.

Rozważany model

– założenia dla parameterów

Zakłada się ponadto często, że wartość składnika v_i tworzą zmienne o niezależnym rozkładzie (independent and identically distributed – iid) o następujących własnościach

$$E_m(v_m) = 0, V_m(v_i) = \sigma_v^2$$

gdzie E_m oznacza wartość oczekiwaną składnika v dla modelu, zaś V_m wariancję dla modelu. Z kolei o błędach wynikających ze schematu losowania (dla ocen bezpośrednich) zakłada się, że

$$E_p(e_i | \theta_i) = 0, V_P(e_i | \theta_i) = \psi_i$$

Przyjmuje się również, że błędy szacunku dla ocen bezpośrednich ψ_i są również znane.

Model hierarchiczny – główne założenia

Na podstawie poprzednich założeń oraz przyjmując, że znany jest oprócz rozkładów ocen bezpośrednich również rozkład błędu modelu σ_v^2 , który ma postać odwrotnego rozkładu Gamma $G^{-1}(a, b)$ (gdzie a jest współczynnikiem kształtu, zaś b współczynnikiem skali) możemy zapisać model hierarchiczny w następującej postaci

- (i) $\hat{\theta}_i | \theta_i, \beta, \sigma_v^2 \sim \text{ind } N(\theta_i, \psi_i), i=1, \dots, m$
- (ii) $\theta_i | \beta, \sigma_v^2 \sim \text{ind } N(z_i^T \beta, b_i^2 \sigma_v^2) i=1, \dots, m$
- (iii) $f(\beta) \propto 1$
- (iv) $\sigma_v^2 | \beta, \theta, \hat{\theta} \sim G^{-1}(a, b)$

Jest to przypadek znanego σ_v^2 oraz „płaskiego” rozkładu a priori dla β , dany przez $f(\beta) \propto 1$.

Zakładamy, też że (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 z punktu 10.3.3 podręcznika Rao.

Model hierarchiczny – przyjęte wartości dla rozkładu Gamma

Wartości tych parametrów - wyznaczone na podstawie empirycznego rozkładu ocen błędów modelu pochodzących z modeli regresji liniowej.

Założenie o identycznych co do definicji zmiennych objaśniających

Zbliżone co do zmienności wartości ocen zarówno ocen bezpośrednich jak i parametrów regresji liniowej β dla modelu

Przybliżenie takie może prowadzić do poprawnych ocen a posteriori dla rozważanego modelu. Zgodnie z sugestią Rao (str. 237) „gdy zakłada się o σ_v^2 , że jest znane oraz $f(\beta) \propto 1$, podejścia HB oraz BLUP w warunkach normalności prowadzą do identycznych estymacji punktowych oraz miar zmienności”.

Model (10.3.1) u Rao nie uwzględnia w pełni zmienności σ_v^2 , co skutkuje zgodnością ale dla bardziej uproszczonych miar wariancji (por wyrażenie (7.1.6) u Rao).

$$MSE(\tilde{\theta}_i^H) = E(\tilde{\theta}_i^H - \theta_i)^2 = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)$$

Uwzględnienie tej zmienności prowadzi naszym zdaniem do ocen zgodnych z szacunkami EBLUP (a więc uwzględniających pełną zmienność modelu). Dalszą dyskusję na ten temat przedstawimy w części eksperymentalnej.

Metody Monte Carlo dla łańcuchów Markowa (Markov Chain Monte Carlo - MCMC)

Przyjmijmy, że $\eta = (\mu^T, \lambda^T)^T$ jest wektorem parametrów dla małych obszarów μ oraz parametrów modelu λ

Dla bardziej złożonych modeli, wylosowanie próby z łącznego rozkładu, może być trudne ze względu na złożoną postać mianownika $f_1(y)$.

Zastosowanie metody MCMC pozwala na uniknięcie takich trudności.

Konstruuje się tutaj łańcuch Markowa $\{\eta^{(k)}, k=0,1,2,\dots\}$, taki, że rozkład $\eta^{(k)}$ jest zbieżny do rozkładu jednoznacznie stacjonarnego (lub niezmienniczego – invariant) danego przez $f(\eta|y)$, określonego jako $\pi(\eta)$.

Pomijając początkowo wylosowane wartości (w ramach tzw. symulacji wstępnej – burn in, której długość wynosi d), otrzymujemy D niezależnych prób $\eta^{(d)}, \dots, \eta^{(d+D)}$ wylosowanych na podstawie rozkładu celu $f(\eta|y)$. Próba taka jest niezależna od punktu startowego $\eta^{(0)}$



Metody Monte Carlo dla łańcuchów Markowa

Taka konstrukcja łańcucha Markowa wymaga, aby jednoetapowe (one-step) prawdopodobieństwo przejścia (transition probability) $P(\eta^{(k+1)} | \eta^{(k)})$ zależało tylko od bieżącego stanu $\eta^{(k)}$.

Prowadzi to do stwierdzenia, że prawdopodobieństwo warunkowe $\eta^{(k+1)}$, przy określonym $\eta^{(0)}, \dots, \eta^{(k)}$ nie jest zależne od historii łańcucha $\{\eta^{(0)}, \dots, \eta^{(k-1)}\}$. Spełniony tutaj musi być warunek stacjonarności dla jądra przejścia

$$\int \pi(\eta^{(k)}) P(\eta^{(k+1)} | \eta^{(k)}) d\eta^{(k)} = \pi(\eta^{(k+1)})$$

Równanie powyższe 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 rozkład $\eta^{(k)}$ dla danego $\eta^{(0)}$, oznaczony przez $P^{(k)}(\eta^{(k)} | \eta^{(0)})$ zbiega do $\pi(\eta^{(k)})$, niezależnie od tego jakiego dokonamy wyboru dla $\eta^{(0)}$.

Próbnik Gibbsa - wprowadzenie

Realizację obliczeniową metod MCMC można przeprowadzić przy pomocy tzw. próbnika Gibbsa.

Próbnik Gibbsa zakłada, że ciąg próbek $\eta^{(k)}$ uzyskujemy dzieląc wektor η na bloki η_1, \dots, η_r . Bloki te mogą zawierać jeden lub więcej elementów. Przykładowo, dla podstawowego modelu poziomu obszaru, mamy $\mu = (\vartheta_1, \dots, \vartheta_m)^T = \vartheta$ oraz $\lambda = (\beta^T, \sigma_v^2)^T$. W takim przypadku η może składać się z następujących bloków: $\eta_1 = \beta$, $\eta_2 = \vartheta_1, \dots, \eta_{(m+1)} = \vartheta_m$, $\eta_{(m+2)} = \sigma_v^2$, przy czym $r = m + 2$. Wymagane jest aby zachodził następujący zbiór rozkładów warunkowych Gibbsa $f(\eta_1 | \eta_2, \dots, \eta_r, y), f(\eta_2 | (\eta_1, \eta_3, \dots, \eta_r, y)), \dots, f(\eta_r | \eta_1, \dots, \eta_{(r-1)}, y)$.

Próbnik Gibbsa wykorzystuje wspomniane prawdopodobieństwa warunkowe do skonstruowania jądra przejścia, $P(\cdot | \cdot)$, takiego, że rozkład stacjonarny otrzymanego łańcucha Markowa jest równy $\pi(\eta) = f(\eta | y)$. Wynik ten jest konsekwencją faktu, iż $f(\eta | y)$ jest jednoznacznie określone przez zbiór warunków Gibbsa.

Próbnik Gibbsa - algorytm

Krok 0. Wybierz punkt startowy $\eta^{(0)}$ dla komponentów $\eta_1^{(0)}, \dots, \eta_r^{(0)}$, przyjmując k jako równe 0. Można na przykład wybrać jako punkty początkowe szacunki metodą REML dla parametrów modelu λ oraz szacunki EB dla parametru μ . Mogą to być jednak dowolnie dobrane punkty.

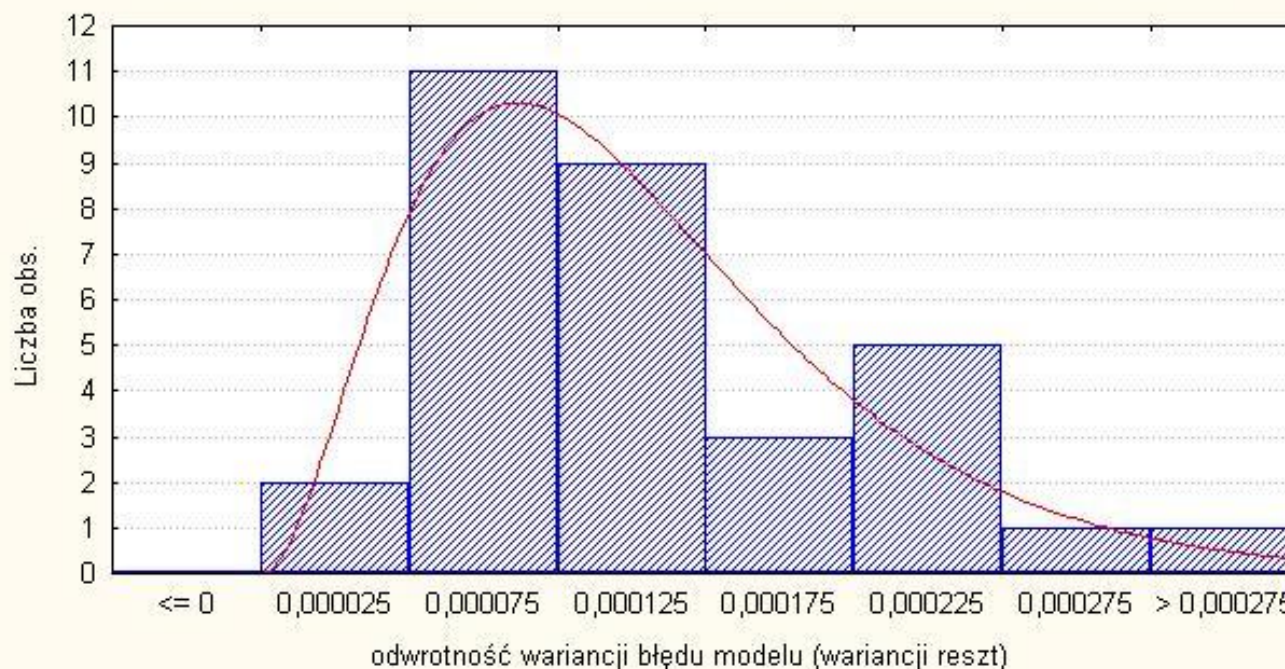
Krok 1. Wygenerować $\eta^{(k+1)} = (\eta_1^{(k+1)}, \dots, \eta_r^{(k+1)})$ w następujący sposób: Wylosować $\eta_1^{(k+1)}$ korzystając z rozkładu $f(\eta_1 | \eta_2^{(k)}, \dots, \eta_r^{(k)}, y)$, następnie wylosować $\eta_2^{(k+1)}$ z rozkładu $f(\eta_2 | \eta_1^{(k+1)}, \eta_3^{(k)}, \dots, \eta_r^{(k)}, y)$..., na koniec wylosować $\eta_r^{(k+1)}$ z rozkładu $f(\eta_r | \eta_1^{(k+1)}, \dots, \eta_r^{(k+1)}, y)$

Krok 2. Przyjąć $k=k+1$ oraz przejdź do kroku 1.

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

Rys.1. Rozkład empiryczny błędu modelu otrzymany dla regresji liniowej

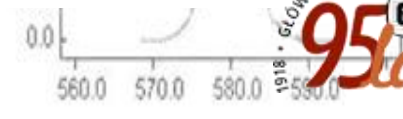
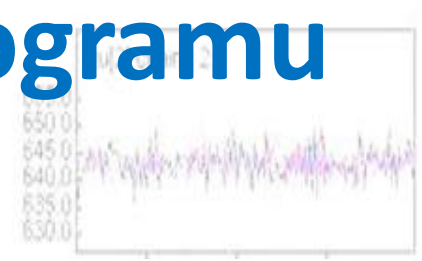
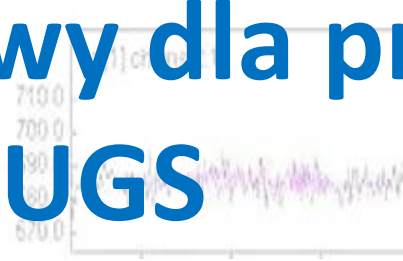
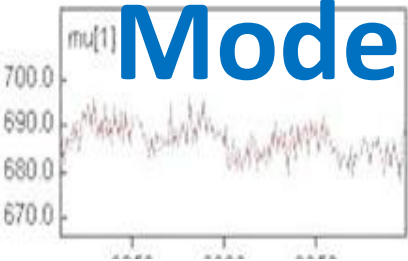
Rozkład empiryczny błędu modelu uzyskany na podstawie
modeli regresji liniowej dla dochodu rozporządzalnego w powiatach
z użyciem danych z Badania Budżetów Gospodarstw Domowych
oraz zmiennych z rejestru POLTAX dla lat 2003 i 2004 (dopasowanie-rozkład Gamma)



Model obliczeniowy dla programu WinBUGS

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] + alpha[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
}
```



Makro w R-project– określenie parametrów modelu i przeprowadzenie symulacji

determining the model parameters

```
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")
```

simulations - WinBUGS call

```
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)
```

Szacunki dla dochodu rozporządzalnego oraz ich precyzji i redukcji względnego błędu estymacji-matopolskie cz. 1

Powiat (NUTS-4)	Dochód rozporządzalny										
	Szacunki bezpośrednie			Szacunki dla metody EBLUP			Szacunki dla metody HB			Redukcja REE	
	Ocena parametru	Błąd szacunku	REE (%)	Ocena parametru	Błąd szacunku	REE (%)	Ocena parametru	Błąd szac.	REE (%)	REML	HB
bocheński	518,43	33,13	6,39	525,14	31,79	6,05	523,74	31,28	5,97	1,056	1,070
brzeski	418,25	32,80	7,84	436,81	31,77	7,27	432,23	30,96	7,16	1,078	1,095
chrzanowski	734,24	22,09	3,01	734,63	22,26	3,03	734,15	21,90	2,98	0,993	1,009
dąbrowski	553,59	106,13	19,17	480,11	66,72	13,90	490,32	71,80	14,64	1,379	1,309
gorlicki	460,38	14,97	3,25	464,59	15,06	3,24	463,79	15,02	3,24	1,003	1,004
krakowski	617,31	22,41	3,63	623,05	22,17	3,56	622,27	21,81	3,51	1,020	1,035
limanowski	576,63	69,00	11,97	527,75	53,08	10,06	536,88	54,61	10,17	1,190	1,176
miechowski	538,66	7,31	1,36	539,06	7,31	1,36	538,96	7,31	1,36	1,001	1,001
myślenicki	475,63	36,64	7,70	494,48	34,97	7,07	492,01	33,84	6,88	1,089	1,120
nowosądecki	540,52	27,76	5,14	529,10	27,26	5,15	531,27	26,97	5,08	0,997	1,012
nowotarski	525,31	34,79	6,62	514,30	33,22	6,46	518,01	32,16	6,21	1,025	1,067
olkuski	675,92	84,13	12,45	675,93	66,35	9,82	676,95	67,77	10,01	1,268	1,243
oświęcimski	732,83	30,13	4,11	726,44	29,82	4,10	728,56	29,01	3,98	1,002	1,032
proszowicki	611,29	56,11	9,18	582,99	49,62	8,51	593,89	49,39	8,32	1,078	1,104
suski	666,01	51,49	7,73	615,90	44,93	7,30	628,43	45,05	7,17	1,060	1,079
tarnowski	427,53	26,72	6,25	431,84	27,52	6,37	429,27	26,25	6,12	0,981	1,022



Szacunki dla dochodu rozporządzalnego oraz ich precyzji i redukcji względego błędu estymacji-małopolskie cz. 2



Dochód rozporządzalny

Powiat (NUTS-4)

Szacunki bezpośrednie

Ocena
parametru

Błąd
szacunku

REE
(%)

Szacunki dla metody EBLUP

Ocena
parametru

Błąd
szacunku

REE
(%)

Szacunki dla metody HB

Ocena
parametru

Błąd
szac.

REE
(%)

Redukcja REE

REML

HB

tatrzański

529,82

45,91

8,67

531,46

44,27

8,33

530,19

42,90

8,09

1,040

1,071

wadowicki

530,13

62,80

11,85

554,40

50,37

9,09

552,55

50,79

9,19

1,304

1,289

wielicki

736,62

68,03

9,24

700,67

56,16

8,02

713,04

55,65

7,80

1,152

1,183

m. Kraków

876,75

1,03

0,12

876,74

1,03

0,12

876,75

1,03

0,12

1,000

1,004

m. Tarnów

713,46

1,84

0,26

713,46

1,84

0,26

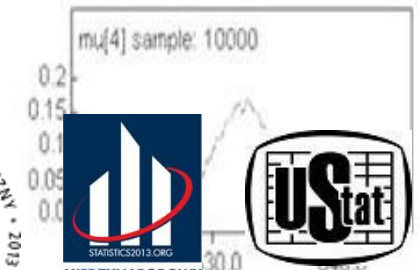
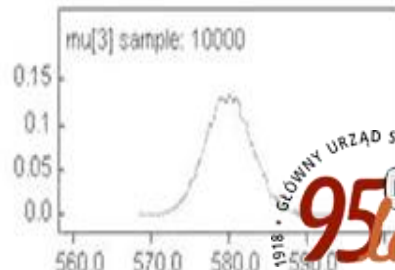
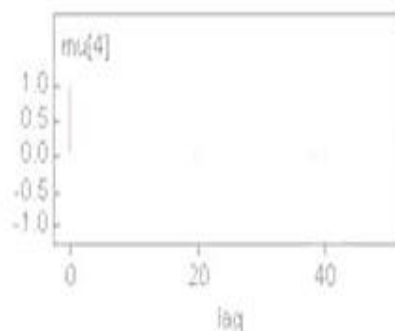
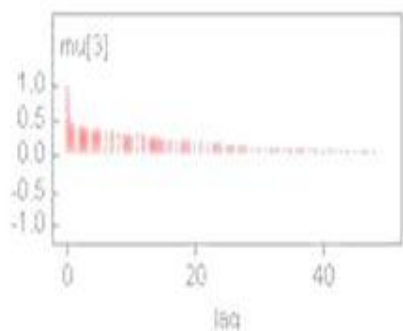
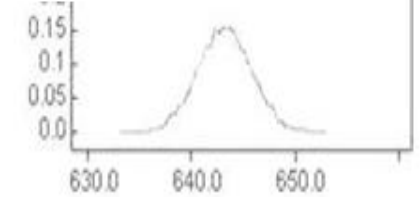
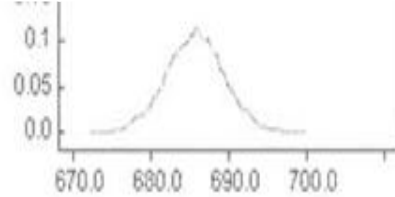
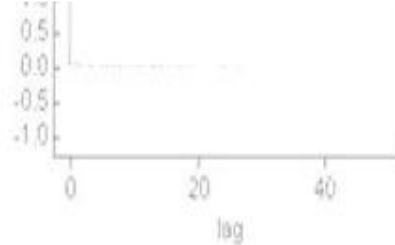
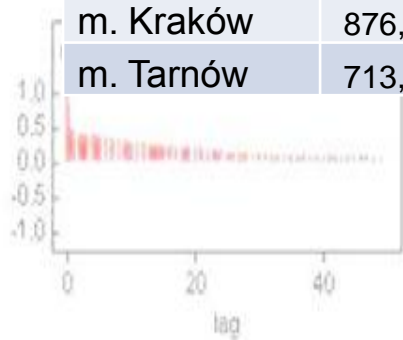
713,45

1,85

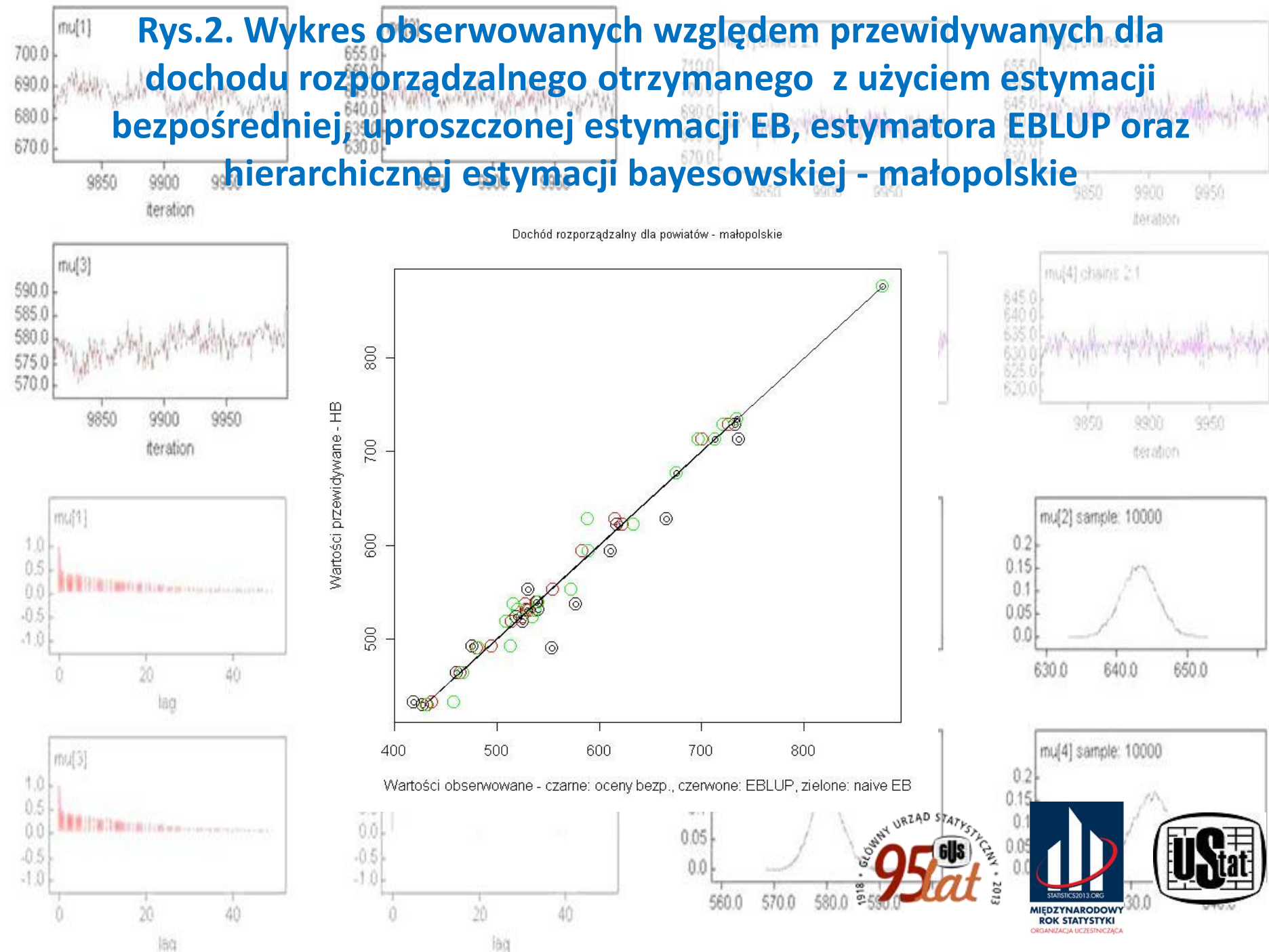
0,26

1,000

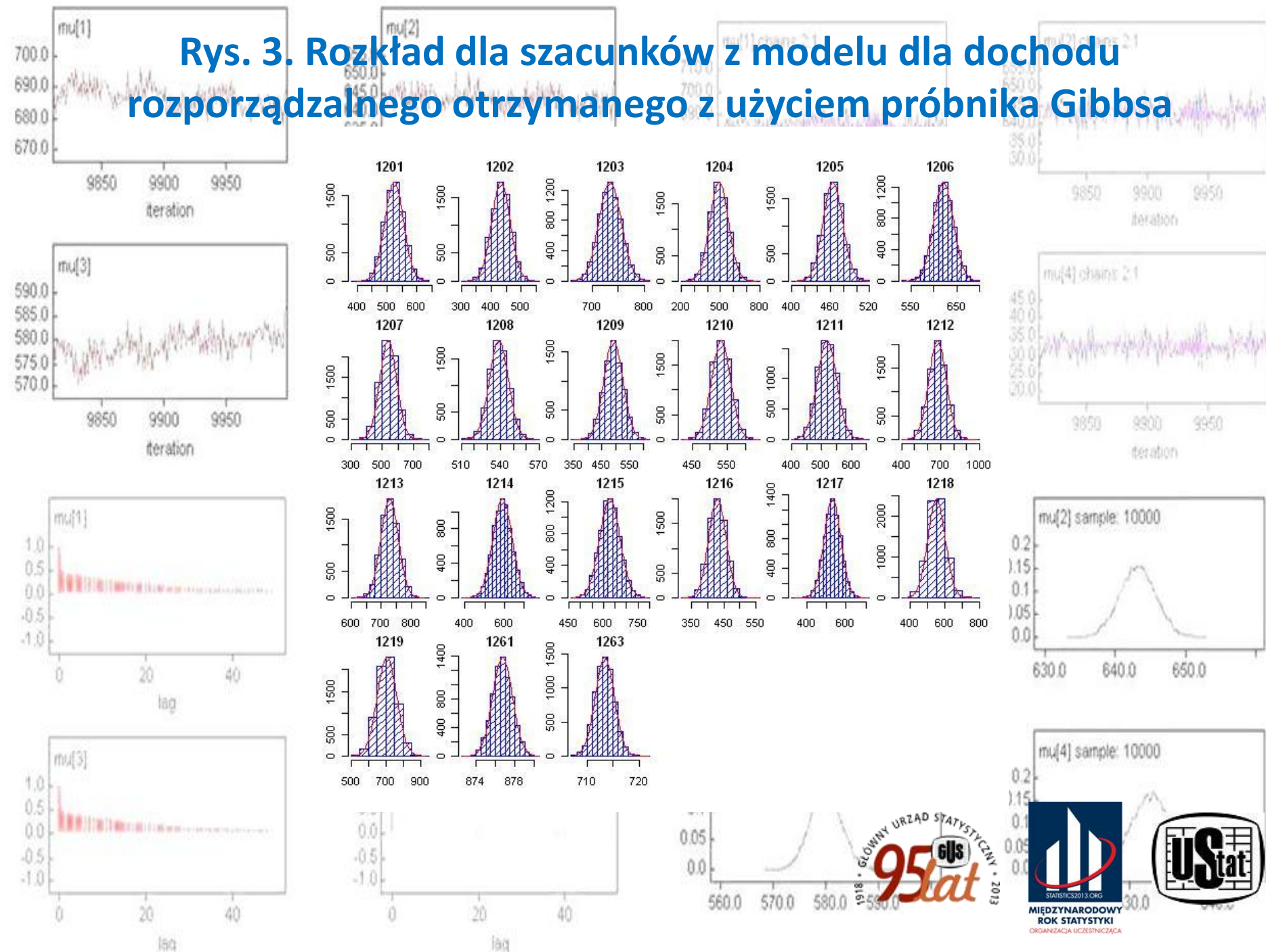
0,999



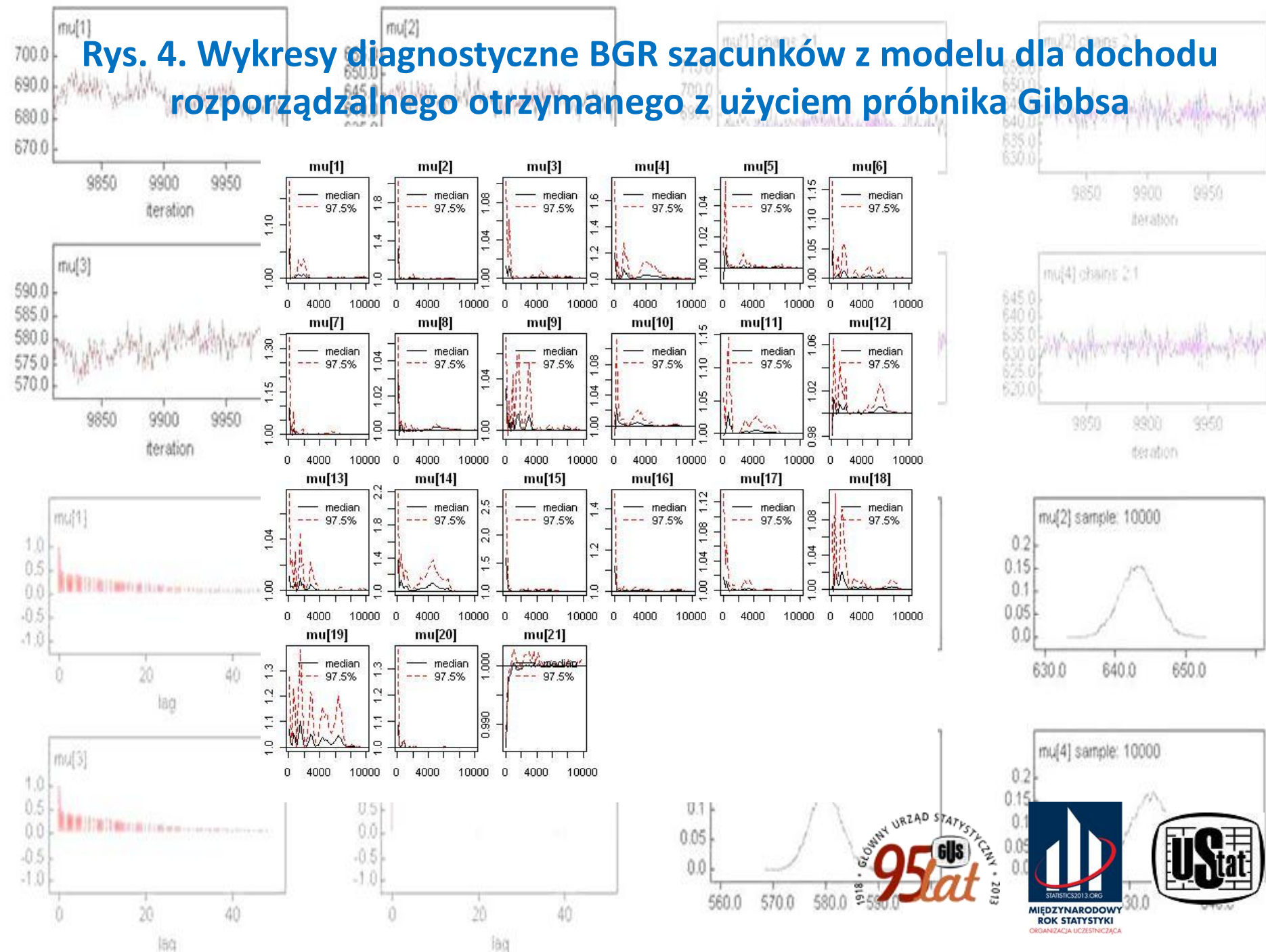
Rys.2. Wykres obserwowanych względem przewidywanych dla dochodu rozporządzalnego otrzymanego z użyciem estymacji bezpośredniej, uproszczonej estymacji EB, estymatora EBLUP oraz hierarchicznej estymacji bayesowskiej - małopolskie



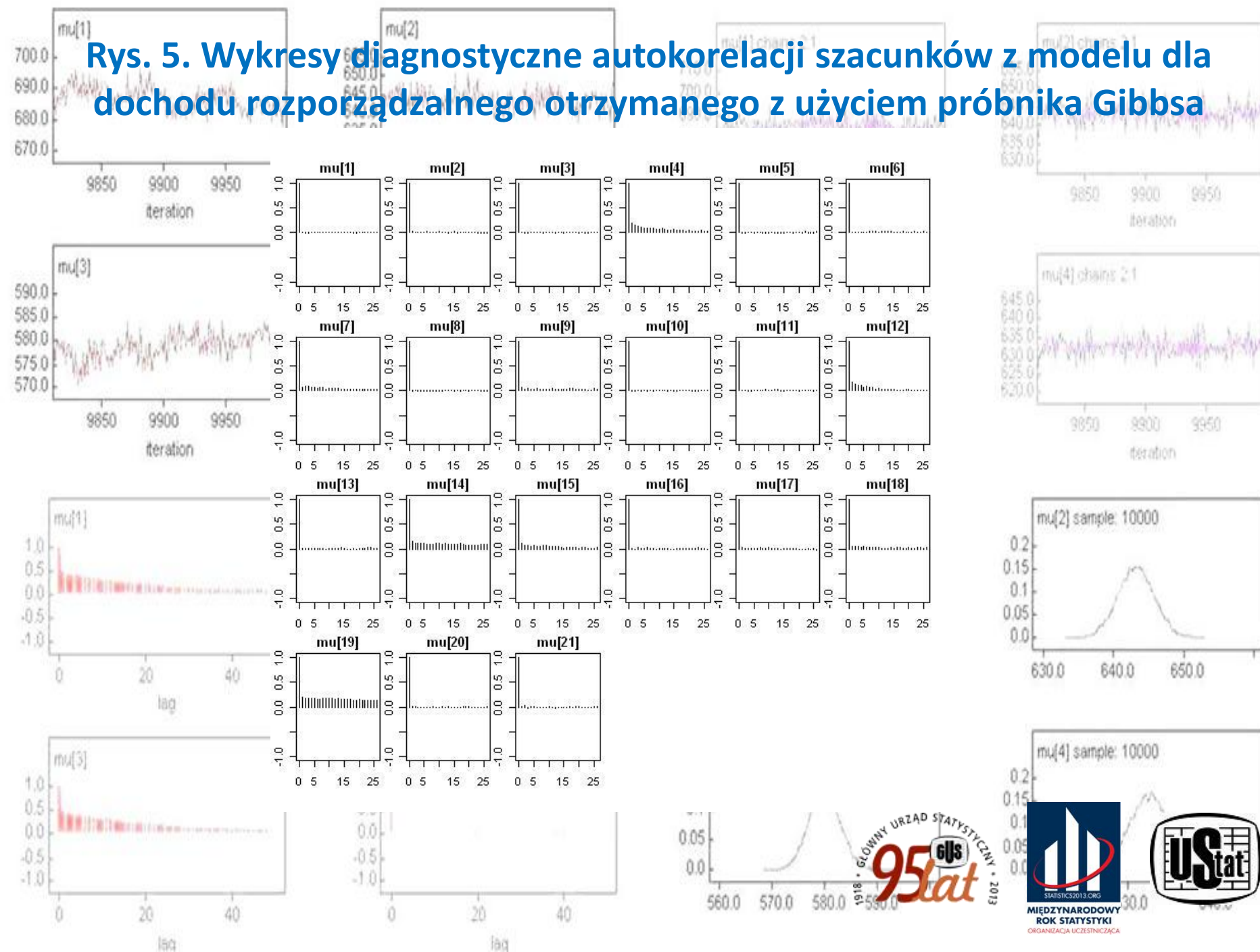
Rys. 3. Rozkład dla szacunków z modelu dla dochodu rozporządzalnego otrzymanego z użyciem próbnika Gibbsa



Rys. 4. Wykresy diagnostyczne BGR szacunków z modelu dla dochodu rozporządzalnego otrzymanego z użyciem próbnika Gibbsa

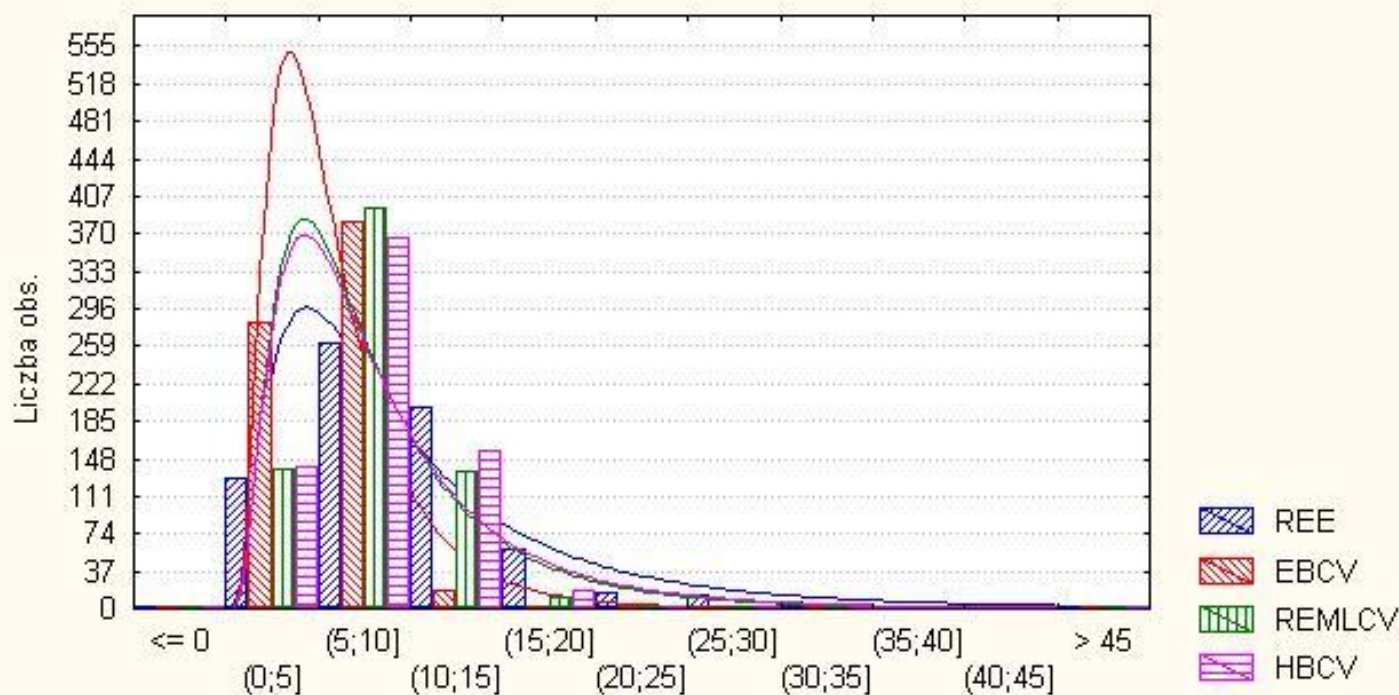


Rys. 5. Wykresy diagnostyczne autokorelacji szacunków z modelu dla dochodu rozporządzalnego otrzymanego z użyciem próbnika Gibbsa



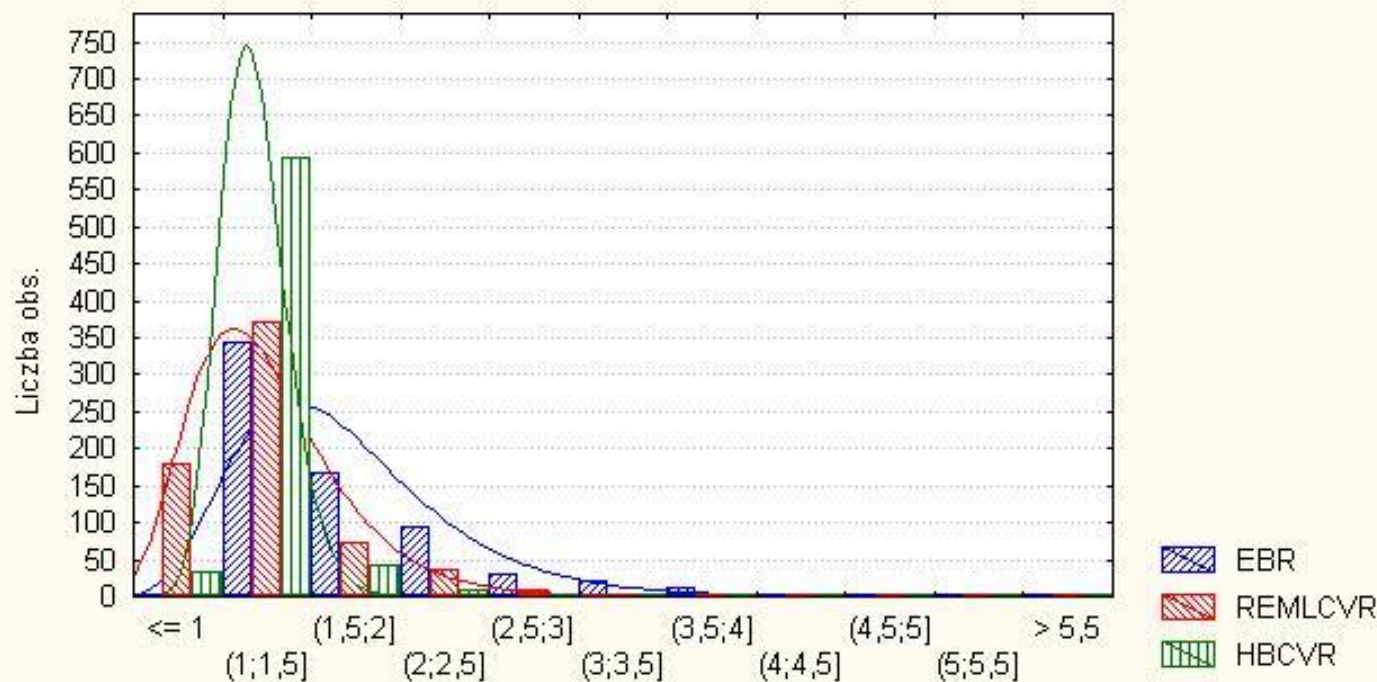
Rys. 6. Rozkład względnego błędu szacunku dla estymatora bezpośredniego, uproszczonego estymatora EB, EBLUP (wariant REML) oraz estymatora HB dla dochodu rozporządnego w powiatach (NUTS4)

Względny błąd szacunku dla estymatora bezpośredniego oraz dla estymatora EB, EBLUP (wariant REML) oraz HB dla dochodu rozporządnego w powiatach na podstawie Badania Budżetów oraz danych z systemu POLTAX dla lat 2003 i 2004



Rys. 7. Rozkład redukcji względnego błędu szacunku dla uproszczonego estymatora EB, EBLUP (wariant REML) oraz estymatora HB dla dochodu rozporządzalnego w powiatach (NUTS4)

Redukcja względnego błędu szacunku dla estymatorów EB, EBLUP (wariant REML) oraz HB dla dochodu rozporządzalnego w powiatach na podstawie Badania Budżetów oraz danych z systemu POLTAX dla lat 2003 i 2004



Wnioski

W prezentacji wykazano przydatność szacunków wykonanych z użyciem hierarchicznej estymacji bayesowskiej w przypadku znanych wartości hiperparametrów modelu.

Pokazano istnienie pewnej zgodności między szacunkami z użyciem takiej metody oraz szacunkami z użyciem innych technik dla małych obszarów, w tym metody EBLUP.

Z uwagi na dobre własności przedstawionych w pracy obliczeń symulacyjnych (brak autokorelacji i stabilność symulacji, oraz praktycznie zaniedbywalny okres symulacji wstępnej) można sądzić, iż podejście takie może znaleźć zastosowanie praktyczne.

Charakterystyczna jest tutaj też nieco większa przeciętnie efektywność obliczeń dla techniki HB niż dla EBLUP, choć dla słabiej dopasowanych modeli nie musi to być regułą.

Pewnym mankamentem jest tutaj 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.



Literatura

1. Gomez-Rubio, V., (2008), "Small Area Estimation with R Unit 5: Bayesian Small Area Estimation", useR! 2008 11 August 2008, Dortmund (Germany), http://www.bias-project.org.uk/SAE_tutorial/useR08-tutorial.tgz
2. 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>
3. Kubacki, J. (2004): Application of the Hierarchical Bayes Estimation to the Polish Labour Force Survey, Statistics in Transition, Vol. 6, No. 5, 785-796. http://www.stat.gov.pl/cps/rde/xbcr/gus/PTS_sit_6_5.pdf
4. 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, 261—278 [http://www.stat.gov.pl/cps/rde/xbcr/pts/SIT_13_2_Summer_2012nowy1\(1\).pdf](http://www.stat.gov.pl/cps/rde/xbcr/pts/SIT_13_2_Summer_2012nowy1(1).pdf)
5. Kubacki, J., Jędrzejczak A., (2012) The Comparison of Generalized Variance Function with Other Methods of Precision Estimation for Polish Household Budget Survey, Studia Ekonomiczne, 120, 58-69 http://wydawnictwo.ue.katowice.pl/uploads/media/SE_120.pdf
6. 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
7. Rao, J.N.K. (2003) Small Area Estimation, Wiley Interscience, Hoboken, New Jersey
8. Salvati, N., Gómez-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
9. Spiegelhalter, D.J., Thomas, A., Best, N., and Lunn, D. (2003), WinBUGS User Manual, Version 1.4.
10. Sturtz, S., Ligges, U., i Gelman, A. (2005), R2WinBUGS: A Package for Running WinBUGS from R., Journal of Statistical Software, 12(3), 1-16
11. 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

