Zeszyty Naukowe Uniwersytet Ekonomiczny w Krakowie

11 (971)

ISSN 1898-6447
Zesz. Nauk. UEK, 2017; 11 (971): 5–25
DOI: 10.15678/ZNUEK.2017.0971.1101

Grażyna Dehnel

Dobór modelu a obciążenie szacunku na przykładzie estymatora GREG w badaniu małych przedsiębiorstw

Streszczenie

Estymacja dotycząca populacji charakteryzujących się silną asymetrią i obecnością obserwacji odstających jest zagadnieniem trudnym, zwłaszcza gdy prowadzona jest na niskim poziomie agregacji. Zastosowanie klasycznych, bezpośrednich metod estymacji nie pozwala na otrzymanie wiarygodnych szacunków. Potrzeba uzyskania szczegółowych informacji oraz szerszych możliwości wykorzystania danych pochodzących z rejestrów administracyjnych skłania do poszukiwania innych, nieklasycznych metod szacunku. Przykładem może być estymacja typu GREG. W artykule podjęto próbę zbadania wpływu wyboru modelu uwzględnionego w ramach estymatora GREG na jakość szacunku parametru populacji przedsiębiorstw. Analizę przeprowadzono na podstawie danych pochodzących z badania małych przedsiębiorstw. Badaną zmienną był przeciętny przychód przedsiębiorstwa. Jako zmienne pomocnicze wykorzystano zmienne opóźnione pochodzące z rejestrów administracyjnych. Badanie prowadzono w przekroju województw z uwzględnieniem rodzaju prowadzonej działalności gospodarczej.

Słowa kluczowe: estymacja GREG, statystyka gospodarcza, estymacja typu *model-assisted*, obserwacje odstające.
Klasyfikacja JEL: C40, C51.

Grażyna Dehnel, Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Statystyki, al. Niepodległości 10, 61-875 Poznań, e-mail: g.dehnel@ue.poznan.pl

1. Wprowadzenie

Rosnące potrzeby informacyjne w zakresie statystyki gospodarczej powodują konieczność prowadzenia badań w kierunku wzbogacenia i rozszerzania zakresu dostarczanych danych dotyczących przedsiębiorczości. Zadanie to jest o tyle trudne, że badaniom prowadzonym na podstawie populacji przedsiębiorstw towarzyszy zwiększający się z roku na rok odsetek braków odpowiedzi. Dodatkowo zakres zmian w systemie statystyki gospodarczej, które brane są pod uwagę, jest ograniczony przez takie czynniki, jak koszty badania oraz obciążenie respondentów wynikające ze sprawozdawczości statystycznej. Zaspokojenie potrzeby uzyskania informacji wymusza zatem poszukiwanie metod szacunku zmierzających do zwiększenia stopnia wykorzystania źródeł administracyjnych. Adaptacja nowych rozwiązań ma przyczynić się do poprawy efektywności prowadzonych szacunków, a przede wszystkim do zwiększenia liczby przekrojów, w których publikowane są dane. Próby adaptacji nieklasycznych metod estymacji w odniesieniu do podmiotów gospodarczych zostały podjęte m.in. w pracach: [Chambers i in. 2014, Clark, Kokic i Smith 2017, Dehnel 2014, 2016]. Szukając nowych podejść do estymacji parametrów dotyczących przedsiębiorstw, należy uwzględnić specyfikę badanej populacji. Populacja przedsiębiorstw charakteryzuje się m.in. obecnością obserwacji odstających. Mając to na względzie, w niniejszym artykule poddano analizie metodę szacunku stosowaną w ramach statystyki małych obszarów zaliczaną do grupy *model-assisted*. Celem badania była ocena wpływu doboru modelu uwzględnionego w ramach estymatora typu GREG na jakość szacunku przeciętnego przychodu małych przedsiębiorstw. W estymacji zaproponowano wykorzystanie opóźnionych zmiennych pomocniczych pochodzących z zasobów administracyjnych. Oceny estymatorów dokonano na podstawie badania empirycznego, w którym wykorzystano dane dotyczące małych przedsiębiorstw działających w ramach sekcji: przemysł, budownictwo, handel i transport.

2. Metoda estymacji

W badaniach prowadzonych w zakresie statystyki przedsiębiorstw opartych na podejściu modelowym często zdarza się, że warunek dotyczący homoskedastyczności nie jest zachowany. Prowadzi to do nieefektywnych ocen parametrów regresji. Stąd też poszukuje się metod, które pozwolą na złagodzenie naruszenia założeń liniowego modelu regresji. Przykładem może być modyfikacja estymatora GREG zaproponowana w pracy R. Chambersa, H. Falveya, D. Hedlina i P. Kokica [2001], zwana dalej modelem Chambersa.

W modelu Chambersa proponuje się włączenie do modelu dodatkowej zmiennej pomocniczej z_i^r. W ramach klasycznej postaci estymatora GREG wartości globalnej zmiennej Y [Lehtonen, Särndal i Veijanen 2016]

$$\hat{Y}_{GREG,d} = \sum_{i \in U_d} \hat{y}_i + \sum_{i \in s_d} w_i e_i, \tag{1}$$

gdzie $\hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}_d$, wektor parametrów modelu $\hat{\boldsymbol{\beta}}_d$ szacowany jest na podstawie zmodyfikowanego wzoru uwzględniającego dodatkową zmienną z [Chambers i in. 2001]:

$$\hat{\boldsymbol{\beta}}_d = \left(\sum_{i \in s_d} \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{z_i^\gamma} \right)^{-1} \left(\sum_{i \in s_d} \frac{w_i \mathbf{x}_i y_i}{z_i^\gamma} \right), \tag{2}$$

gdzie:

w_i – wagi wynikające ze schematu losowania,

$\mathbf{x}$ – wektor zmiennych pomocniczych, w zależności od podejścia – liczba osób pracujących lub przychód,

z_i^γ – zmienna pomocnicza w modelu regresji y względem x zakładającym heteroskedastyczność, w zależności od podejścia – liczba osób pracujących lub przychód,

γ – parametr określający stopień heteroskedastyczności,

U_d – część populacji generalnej o elementach należących do domeny d,

s_d – część próby wylosowanej z populacji o elementach należących do domeny d.

Estymator modelu Chambersa wyrażony za pomocą wzoru (1) można przedstawić w postaci, która jest tożsama z formułą klasycznego estymatora GREG:

$$\hat{Y}_{GREG,d} = \sum_{i \in s_d} w_i g_i y_i. \tag{3}$$

Różnica dotyczy jedynie definicji wagi g_i zależnej od wartości cechy dodatkowej x u jednostek wylosowanych do próby, zdefiniowanej jako:

$$g_i = 1 + (X_d - \hat{X}_{HT,d})' \left(\sum_{i \in s_d} \frac{w_i \mathbf{x}_i \mathbf{x}_i'}{z_i^\gamma} \right)^{-1} \left(\frac{\mathbf{x}_i}{z_i^\gamma} \right), \tag{4}$$

gdzie:

d – domena,

g_i – wagi zależne od wartości cechy dodatkowej u i-tej jednostki wylosowanej do próby,

$\hat{Y}_{GREG,d}$ – ocena wartości globalnej w domenie d na podstawie estymatora GREG,

$\hat{X}_{HT,d}$ – estymator bezpośredni Horvitza-Thompsona wartości globalnej zmiennej pomocniczej x w domenie d,

X – wartość globalna zmiennej pomocniczej x,

γ – parametr określający stopień heteroskedastyczności, dla $\gamma = 0$ estymator modelu Chambersa przyjmuje klasyczną postać estymatora GREG.

Na podstawie wcześniej prowadzonych badań statystycznych wiadomo, że wartość parametru γ zawiera się w przedziale ⟨1, 2⟩ [Särndal, Swensson i Wretman 1992], stąd też w przeprowadzonym badaniu przeanalizowano trzy podejścia (dla γ równego odpowiednio 1, 1,5 i 2) reprezentujące estymatory $\hat{Y}^{1}_{GREG}$, $\hat{Y}^{1,5}_{GREG}$, $\hat{Y}^{2}_{GREG}$. W estymatorach tych w liniowym modelu regresji dopuszcza się niezachowanie warunku homoskedastyczności, przy czym poziom heteroskedastyczności określany jest przez parametr γ.

3. Założenia badania

Badaniem empirycznym objęto małe przedsiębiorstwa (10–49 pracujących) prowadzące działalność gospodarczą w ramach sekcji: przemysł, budownictwo, handel i transport. Analizie poddano model, w którym zmienną zależną stanowił przychód uzyskany przez przedsiębiorstwa w czerwcu 2012 r. Za zmienne niezależne przyjęto przychód, koszt oraz liczbę pracujących według stanu zanotowanego w grudniu 2011 r. Decyzję o wykorzystaniu zmiennych opóźnionych podjęto, biorąc pod uwagę ograniczenia czasowe, z jakimi musi się liczyć GUS, prowadząc badania statystyczne. Chodzi przede wszystkim o opóźnienie, jakie ma miejsce przy przekazywaniu statystyce publicznej zasobów administracyjnych [*Wykorzystanie danych…* 2016].

Dane dotyczące zmiennej zależnej pochodziły z badania DG-1 [Dehnel 2014]. Z kolei źródłem informacji o zmiennych niezależnych były rejestry administracyjne. Szacunku dokonano w przekroju regionalnym, z uwzględnieniem rodzaju prowadzonej działalności gospodarczej. Przekrój regionalny obejmował jednostki na poziomie województw, zaś rodzajowi prowadzonej działalności odpowiadały sekcje PKD.

W ocenie jakości estymacji za punkt odniesienia przyjęto oszacowania otrzymane na podstawie estymatorów bezpośrednich – Horvitza-Thompsona (HT) i typu GREG, włączając w to jego postać ilorazową (oznaczoną w artykule jako RAT).

Estymator HT

$$\hat{Y}_{HT,d} = \frac{N_d}{n_d} \sum_{i \in s_d} y_i, \tag{5}$$

gdzie:

$\hat{Y}_{HT,d}$ – estymator bezpośredni HT wartości globalnej zmiennej y w domenie d,

N_d – liczebność populacji generalnej w domenie d,

n_d – liczebność próby w domenie d,

y_i – wartość zmiennej badanej u i-tej jednostki.

Estymator GREG

$$\hat{Y}^{0}_{GREG,d} = \hat{Y}_{GREG,d} = \sum_{i \in U_d} \hat{y}_i + \sum_{i \in s_d} w_i e_i , \tag{6}$$

gdzie $\hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}_d$, wektor parametrów modelu $\hat{\boldsymbol{\beta}}_d$ szacowany jest na podstawie wzoru [Rao, Molina 2015]:

$$\hat{\boldsymbol{\beta}}_d = \left(\sum_{i \in s_d} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i \in s_d} w_i \mathbf{x}_i y_i \right). \tag{7}$$

Estymator $\hat{Y}^{0}_{GREG}$ przyjmuje postać estymatora typu GREG, jednak jego wartości różnią się od szacunków otrzymanych na podstawie klasycznego podejścia GREG. Różnice wynikają z tego, że w przypadku estymatora modelu Chambersa $\hat{Y}^{0}_{GREG}$ w szacunku nie są uwzględniane wszystkie jednostki wylosowane do próby z domeny *d*. Pomijane są bowiem te, dla których zmienna pomocnicza '*z*' przyjmie wartość zero.

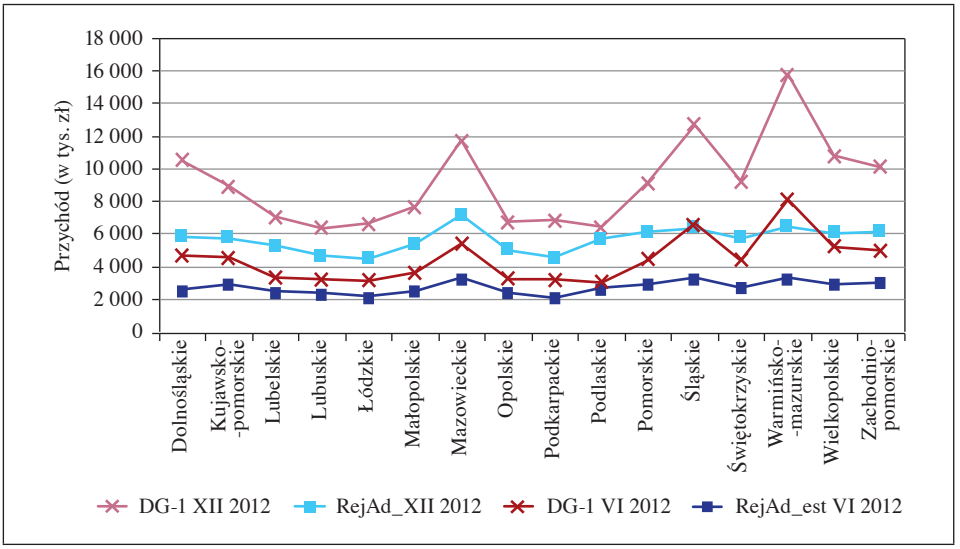
Jako zmienne pomocnicze typu '*x*' i '*z*' wykorzystano liczbę osób pracujących (dane pochodzące z ZUS) oraz przychody (dane pochodzące z rejestru podatkowego). W modelu Chambersa zakłada się, że każda ze zmiennych pomocniczych może być wykorzystana zarówno jako zmienna '*x*', jak i '*z*' (z zastrzeżeniem, że zmienną '*z*' nie może być taka zmienna, która przyjmuje wartości zerowe). W związku z powyższym ostatecznie głębszej analizie poddano podejście, w którym za zmienną '*z*' przyjęto liczbę osób pracujących.

4. Metoda oceny precyzji

Do oceny precyzji i dokładności otrzymanych szacunków wykorzystano metodę bootstrapową. Wykonano 1000 repetycji losowania podprób, na podstawie których oszacowano wartość przychodu przedsiębiorstw dla czerwca 2012 r. w przekroju przyjętych domen studiów. Efektywność estymacji oceniono na podstawie współczynnika zmienności estymatora [Bracha 2004]:

$$CV(\hat{Y}_d) = \frac{\sqrt{Var(\hat{Y}_d)}}{E(\hat{Y}_d)} = \frac{\sqrt{\frac{1}{999} \sum_{b=1}^{1000} (\hat{Y}_{b,d} - \hat{Y}_d)}}{E(\hat{Y}_d)}. \tag{8}$$

Wskaźnik ten określa udział błędu estymacji w wartości szacowanej zmiennej na poziomie domeny. W badaniach prowadzonych przez GUS oraz badaniach empirycznych przyjmuje się, że wyniki szacunków mogą być uznane za wiarygodne, jeżeli wartość *CV* nie przekracza 10%. Jeśli przyjmuje on wartości z przedziału 10–20%, szacunki należy interpretować ostrożnie. Jeżeli natomiast poziom *CV* jest wyższy od 20%, oceny estymatorów nie są uznawane za wiarygodne [*Ludność…* 2013].



Rys. 1. Wartości przychodu przedsiębiorstw w czerwcu i grudniu 2012 r. zarejestrowane na podstawie badania DG-1 oraz zeznań podatkowych w sekcji „przemysł"

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.

Ocena obciążenia wymaga znajomości wartości szacowanego parametru w populacji generalnej. Ze względu na brak dostępu w badaniu do informacji o tej wielkości oszacowano ją w sposób pośredni, na podstawie danych pochodzących z zeznań podatkowych z grudnia 2012 r. Przyjęto, że relacja przychodu zarejestrowanego w zeznaniach podatkowych dla badanych przedsiębiorstw na poziomie domeny studiów do przychodu określonego na podstawie badania DG-1 jest stała. W celu graficznej prezentacji zależności między wielkościami wykorzystano dane dotyczące sekcji „przemysł" (por. rys. 1). W pozostałych sekcjach PKD relacja między zmiennymi przedstawia się podobnie.

$$\frac{Przychód\_RejAd_{XII\,2012}}{Przychód\_DG1_{XII\,2012}} = \frac{Przychód\_RejAd_{VI\,2012}}{Przychód\_DG1_{VI\,2012}}. \tag{9}$$

Takie podejście pozwoliło na wyznaczenie przybliżonej wartości przychodu przedsiębiorstw dla czerwca 2012 r.

5. Wyniki szacunków i ocena ich jakości*

Analizę rozpoczęto od oceny rozkładów przedsiębiorstw według badanych zmiennych. Wartości współczynnika zmienności wahały się w granicach 47–649%. Zanotowano również silną asymetrię, współczynnik asymetrii przyjmował wartości w przedziale 0,6–17,1. Biorąc pod uwagę własności rozkładów podmiotów gospodarczych, za pomocą testów White'a oraz Breuscha-Pagana weryfikacji poddano hipotezę zakładającą homoskedastyczność. Wyniki testów w przypadku zdecydowanej większości wyróżnionych domen potwierdziły słuszność tezy o zmienności składnika losowego (por. tabela 1). To z kolei uzasadniało wykorzystanie opisanych wyżej estymatorów GREG uwzględniających dodatkową cechę '*z*'.

Ocenie poddano zarówno efektywność, jak i dokładność estymacji. W ocenie precyzji estymacji za punkt odniesienia przyjęto szacunek otrzymany za pomocą klasycznych, bezpośrednich estymatorów HT oraz typu GREG, włączając w to jego postać ilorazową. Biorąc pod uwagę wartości miary efektywności *CV*, można zauważyć, że największym zróżnicowaniem charakteryzuje się estymator HT (por. rys. 2–5 i tabela 2). Mniejszą zmienność obserwujemy w przypadku klasycznych estymatorów GREG, w których wykorzystuje się zmienne pomocnicze pochodzące z rejestrów administracyjnych. Analizując wyniki można zauważyć, że precyzja szacunku estymacji typu GREG zależy od liczebności próby. Jest ona z reguły wyższa w sekcjach liczniej reprezentowanych w próbie, takich jak handel i przemysł. W większości wyróżnionych domen współczynniki zmienności estymatorów modelu Chambersa kształtują się na podobnym poziomie. Wyjątek stanowi sekcja „transport", w której liczebność próby jest najmniejsza i wynosi w niektórych domenach 5 jednostek.

W ocenie dokładności szacunku przychodu przedsiębiorstw wykorzystano wartości referencyjne wyznaczone na podstawie zależności ilorazowej opisanej powyżej (por. wzór (9)). Dodatkowo, w celu pełniejszej oceny, model Chambersa porównano z estymacją HT oraz GREG (por. rys. 6–9, tabela 3). Otrzymane wyniki wskazują, że zastosowanie modelu Chambersa w znacznym stopniu poprawiło dokładność szacunku w przypadku estymacji HT i GREG. Zastosowanie estymacji HT w prawie wszystkich badanych domenach doprowadziło do znacznego przeszacowania wartości przychodu. Z kolei w przypadku wszystkich estymatorów typu GREG widoczne jest niedoszacowanie badanego parametru, jednak wielkość odchylenia od wartości referencyjnej jest zdecydowanie mniejsza niż w przypadku estymatora HT. Zastosowanie estymatora GREG, a w szczególności jego zmodyfikowanej wersji, przyniosło największą poprawę dokładności szacunku dla domen, dla których zanotowano największą dyspersję zmiennych uwzględnionych w modelu.

* Wybrane wyniki badań zostały opublikowane w pracy [Dehnel 2017].

Tabela 1. Wartości statystyk testów White'a i Breuscha-Pagana na heteroskedastyczność w przekroju województw

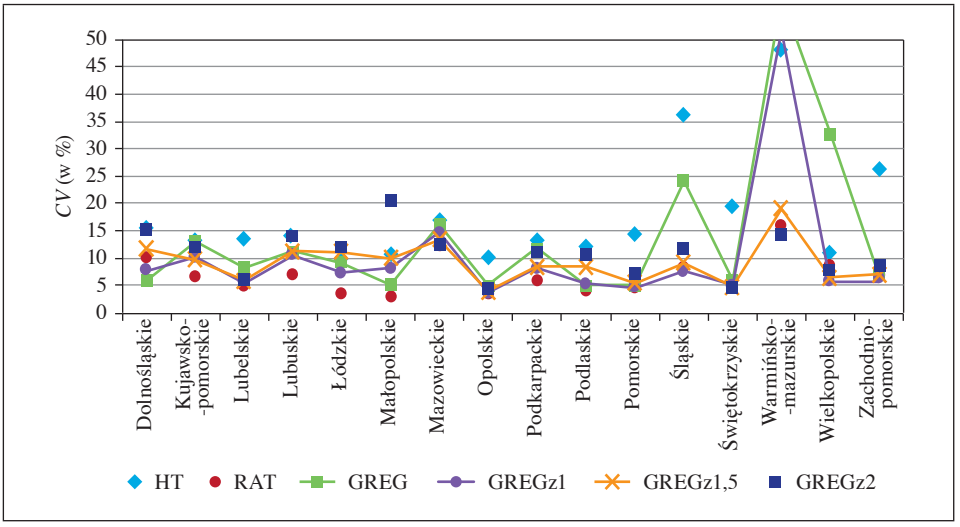
Województwo	Test White'a		Test Breuscha-Pagana		Test White'a		Test Breuscha-Pagana	
	statystyka	*p*-wartość	statystyka	*p*-wartość	statystyka	*p*-wartość	statystyka	*p*-wartość
	Przemysł				Budownictwo			
Dolnośląskie	15,28	0,0092	7,75	0,0207	39,93	*	30,4	*
Kujawsko-pomorskie	102	*	97,4	*	6,99	0,2217	0,75	0,6863
Lubelskie	48,04	*	25,31	*	28,92	*	12,14	0,0023
Lubuskie	20,5	0,001	8,87	0,0118	20,7	0,0009	16,63	0,0002
Łódzkie	179,3	*	97,86	*	35,21	*	21,22	*
Małopolskie	90,51	*	89,76	*	14,55	0,0125	9,77	0,0076
Mazowieckie	276,5	*	24,85	*	0,69	0,9835	0,02	0,9922
Opolskie	57,06	*	33,44	*	20,31	0,0011	17,71	0,0001
Podkarpackie	171,3	*	131	*	16,89	0,0047	16,01	0,0003
Podlaskie	41,21	*	28,77	*	11,27	0,0462	4,75	0,0928
Pomorskie	112,4	*	60,2	*	14,69	0,0118	12,28	0,0022
Śląskie	267,8	*	19,2	*	77,15	*	53,72	*
Świętokrzyskie	34,8	*	24,3	*	36,33	*	29,04	*
Warmińsko-mazurskie	125,8	*	14,91	0,0006	5,29	0,3818	3,73	0,1552
Wielkopolskie	101,9	*	89,19	*	81,53	*	6,22	0,0445
Zachodnio-pomorskie	25,87	*	3,83	0,1473	56,92	*	42,71	*

cd. tabeli 1

Województwo	Test White'a		Test Breuscha-Pagana		Test White'a		Test Breuscha-Pagana	
	statystyka	*p*-wartość	statystyka	*p*-wartość	statystyka	*p*-wartość	statystyka	*p*-wartość
	Handel				Transport			
Dolnośląskie	121,9	*	86,7	*	14,63	0,0121	12,9	0,0016
Kujawsko-pomorskie	46,13	*	35,05	*	16,2	0,0063	10,48	0,0053
Lubelskie	49,7	*	1,95	0,3771	2,34	0,7998	0,21	0,9013
Lubuskie	80,34	*	37,33	*	9,48	0,0913	4,16	0,1248
Łódzkie	82,18	*	34,9	*	11,98	0,035	10,06	0,0065
Małopolskie	136,1	*	6,36	0,0417	18,27	0,0026	15,31	0,0005
Mazowieckie	169,7	*	142,6	*	2,11	0,8333	0,5	0,7782
Opolskie	42,65	*	36,43	*	13,18	0,0217	1,86	0,3945
Podkarpackie	33,91	*	18,09	0,0001	8,71	0,1213	5,39	0,0677
Podlaskie	33,33	*	16,64	0,0002	5,63	0,3439	0,72	0,698
Pomorskie	40,15	*	36,32	*	24,63	0,0002	16,54	0,0003
Śląskie	97,41	*	82,98	*	23,18	0,0003	20,51	*
Świętokrzyskie	23,45	0,0003	6,18	0,0456	16,32	0,006	9,04	0,0109
Warmińsko-mazurskie	48,97	*	25,96	*	14,15	0,0147	0,44	0,8013
Wielkopolskie	121,5	*	74,83	*	45,67	*	12,75	0,0017
Zachodniopomorskie	121,1	*	113,9	*	20,77	0,0009	11,41	0,0033

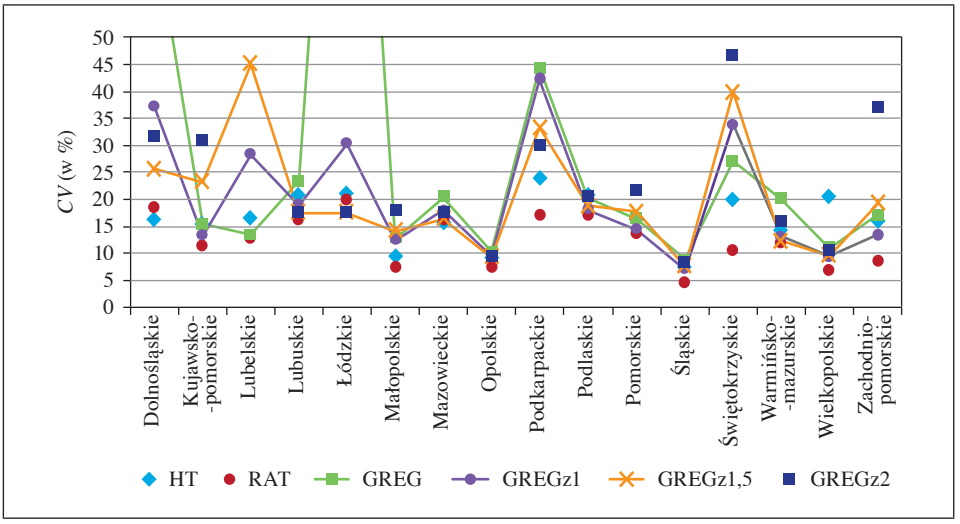
* $p < 0{,}0001$

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1.



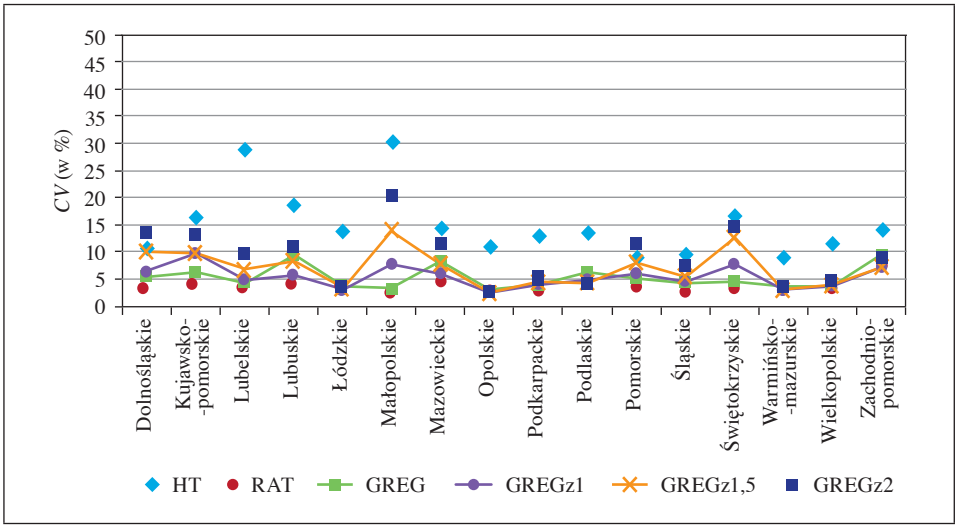
Rys. 2. Precyzja szacunku *CV* w przekroju województw – przemysł

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.



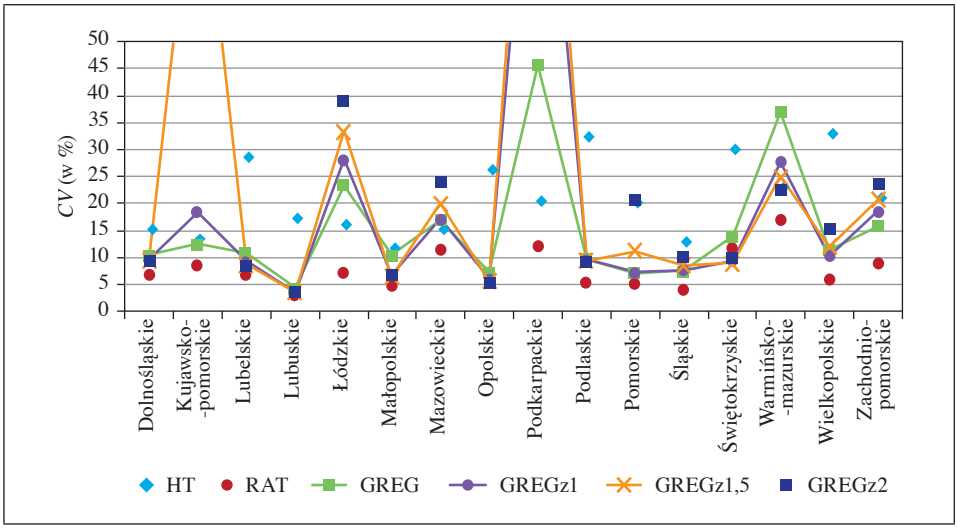
Rys. 3. Precyzja szacunku *CV* w przekroju województw – budownictwo

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.



Rys. 4. Precyzja szacunku *CV* w przekroju województw – handel

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.



Rys. 5. Precyzja szacunku *CV* w przekroju województw – transport

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.

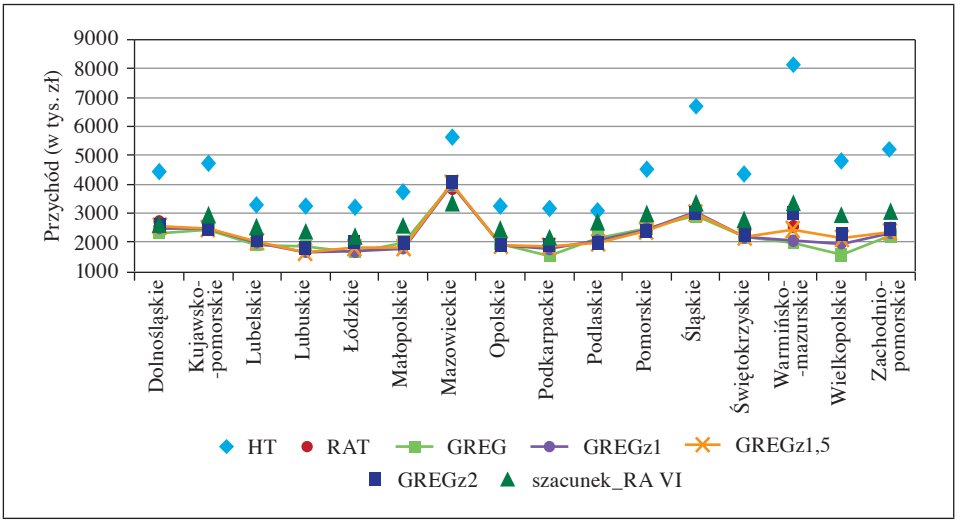
Tabela 2. Precyzja szacunku *CV* w przekroju województw i czterech sekcji PKD

Województwo	Estymator											
	HT	RAT	GREG	GREGz1	GREGz1,5	GREGz2	HT	RAT	GREG	GREGz1	GREGz1,5	GREGz2
	Przemysł						Budownictwo					
Dolnośląskie	15,5	9,8	5,8	7,7	11,7	15,1	16,1	18,6	70,3	37,1	25,7	31,6
Kujawsko-pomorskie	13,2	6,6	13,1	10,2	9,7	11,8	15,3	11,4	15,3	13,2	23,2	31,0
Lubelskie	13,6	5,0	8,3	5,4	5,9	6,5	16,5	12,7	13,3	28,3	45,1	71,2
Lubuskie	14,0	6,9	11,3	10,6	11,2	14,0	20,7	16,2	23,4	19,0	17,5	17,7
Łódzkie	9,8	3,6	9,1	7,3	11,1	11,9	21,0	19,8	151,1	30,2	17,6	17,5
Małopolskie	10,5	3,0	5,2	8,1	9,9	20,4	9,4	7,5	12,7	12,4	14,1	17,9
Mazowieckie	16,8	12,4	16,2	14,6	13,3	12,5	15,5	15,8	20,3	17,9	16,3	17,5
Opolskie	10,1	3,6	5,0	3,6	3,8	4,3	9,2	7,3	10,3	9,4	9,3	9,4
Podkarpackie	13,0	5,7	11,8	8,2	8,4	11,0	23,9	16,9	44,3	42,2	33,3	30,0
Podlaskie	12,2	4,4	5,1	5,2	8,3	10,7	20,8	17,1	20,3	18,0	18,8	20,4
Pomorskie	14,3	4,3	5,0	4,6	5,4	7,1	16,1	13,6	16,4	14,4	17,7	21,4
Śląskie	36,1	8,2	24,2	7,6	9,1	11,4	7,3	4,6	8,9	7,0	7,5	8,3
Świętokrzyskie	19,3	4,7	5,8	5,1	4,7	4,5	19,9	10,4	27,0	33,9	39,8	46,7
Warmińsko-mazurskie	48,1	15,9	57,7	51,8	19,0	14,3	14,0	11,8	20,1	13,2	12,3	16,0
Wielkopolskie	11,0	8,7	32,8	5,7	6,4	7,8	20,4	6,7	10,9	9,4	9,5	10,5
Zachodnioporskie	26,1	7,0	6,9	5,8	6,9	8,5	15,7	8,5	17,1	13,3	19,2	36,8

cd. tabeli 2

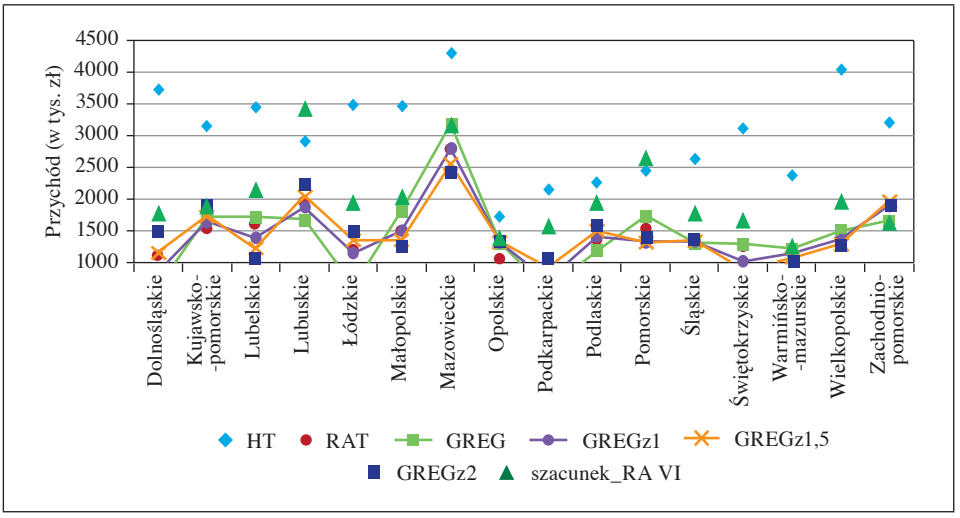
Województwo	Estymator											
	HT	RAT	GREG	GREGz1	GREGz1,5	GREGz2	HT	RAT	GREG	GREGz1	GREGz1,5	GREGz2
	Handel						Transport					
Dolnośląskie	10,4	3,2	5,4	6,3	10,1	13,3	15,4	6,9	10,5	9,6	9,6	9,7
Kujawsko-pomorskie	16,1	4,1	6,2	9,6	9,7	13,1	13,6	8,6	12,5	18,5	90,7	179,5
Lubelskie	28,8	3,5	4,3	4,8	6,7	9,6	28,7	6,8	11,0	9,3	8,7	8,4
Lubuskie	18,5	3,9	9,5	5,6	8,3	10,9	17,5	3,0	4,4	3,7	3,6	3,7
Łódzkie	13,6	3,3	3,7	3,0	3,2	3,4	16,2	7,4	23,4	28,2	33,4	39,2
Małopolskie	30,2	2,4	3,2	7,6	13,9	19,8	11,9	5,0	10,5	6,8	6,5	7,0
Mazowieckie	14,2	4,3	8,1	6,0	7,6	11,3	15,4	11,5	17,1	17,1	20,1	24,1
Opolskie	10,7	2,5	2,9	2,4	2,4	2,5	26,4	5,4	7,2	6,1	5,8	5,6
Podkarpackie	12,7	2,7	3,8	4,0	4,4	5,3	20,6	12,2	45,8	108,5	127,5	132,5
Podlaskie	13,4	4,8	6,2	4,7	4,2	4,1	32,6	5,5	9,8	9,6	9,4	9,2
Pomorskie	8,8	3,2	5,1	5,9	7,9	11,5	20,2	5,5	7,1	7,3	11,1	20,6
Śląskie	9,5	2,7	4,2	4,5	5,4	7,8	13,1	4,2	7,6	7,8	8,5	10,0
Świętokrzyskie	16,5	3,1	4,4	7,7	12,5	14,6	30,2	11,9	13,9	9,5	9,1	10,2
Warmińsko-mazurskie	9,0	2,9	3,5	2,9	3,0	3,3	23,2	17,0	37,0	27,8	25,0	22,7
Wielkopolskie	11,3	3,2	3,6	3,6	3,8	4,4	33,2	6,0	11,5	10,4	12,1	15,3
Zachodnioporskie	14,0	7,3	9,4	7,1	7,1	8,7	21,2	9,2	15,9	18,6	20,8	23,8

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.



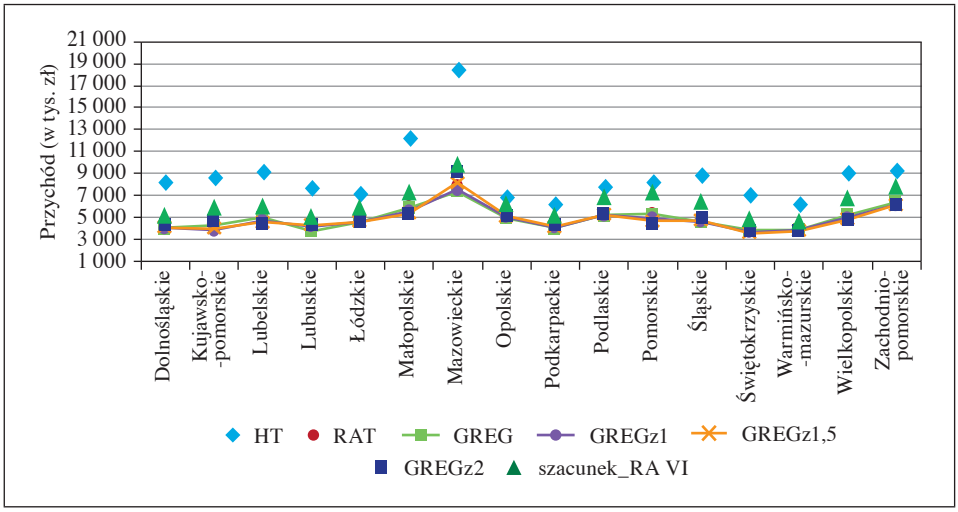
Rys. 6. Szacunek przychodu przedsiębiorstw dla czerwca 2012 r. w przekroju województw – przemysł

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.



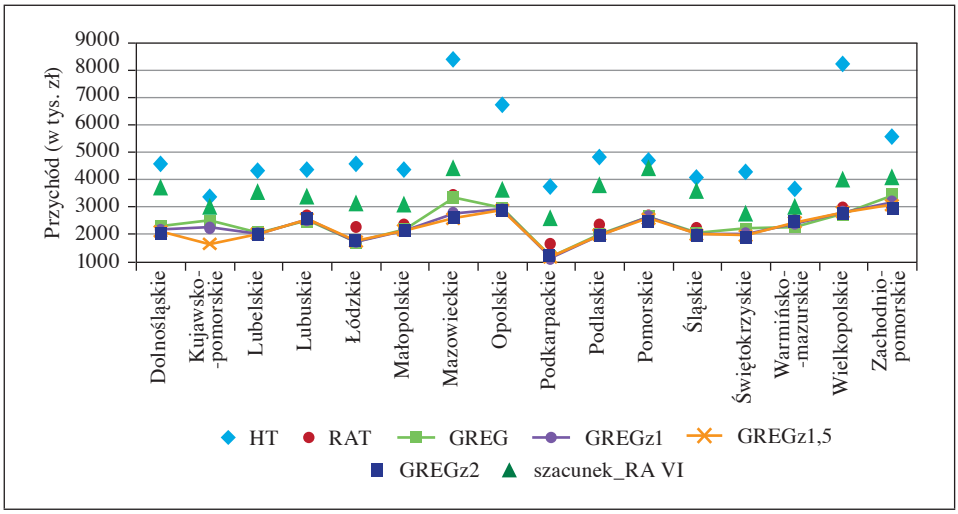
Rys. 7. Szacunek przychodu przedsiębiorstw dla czerwca 2012 r. w przekroju województw – budownictwo

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.



Rys. 8. Szacunek przychodu przedsiębiorstw dla czerwca 2012 r. w przekroju województw – handel

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.



Rys. 9. Szacunek przychodu przedsiębiorstw dla czerwca 2012 r. w przekroju województw – transport

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.

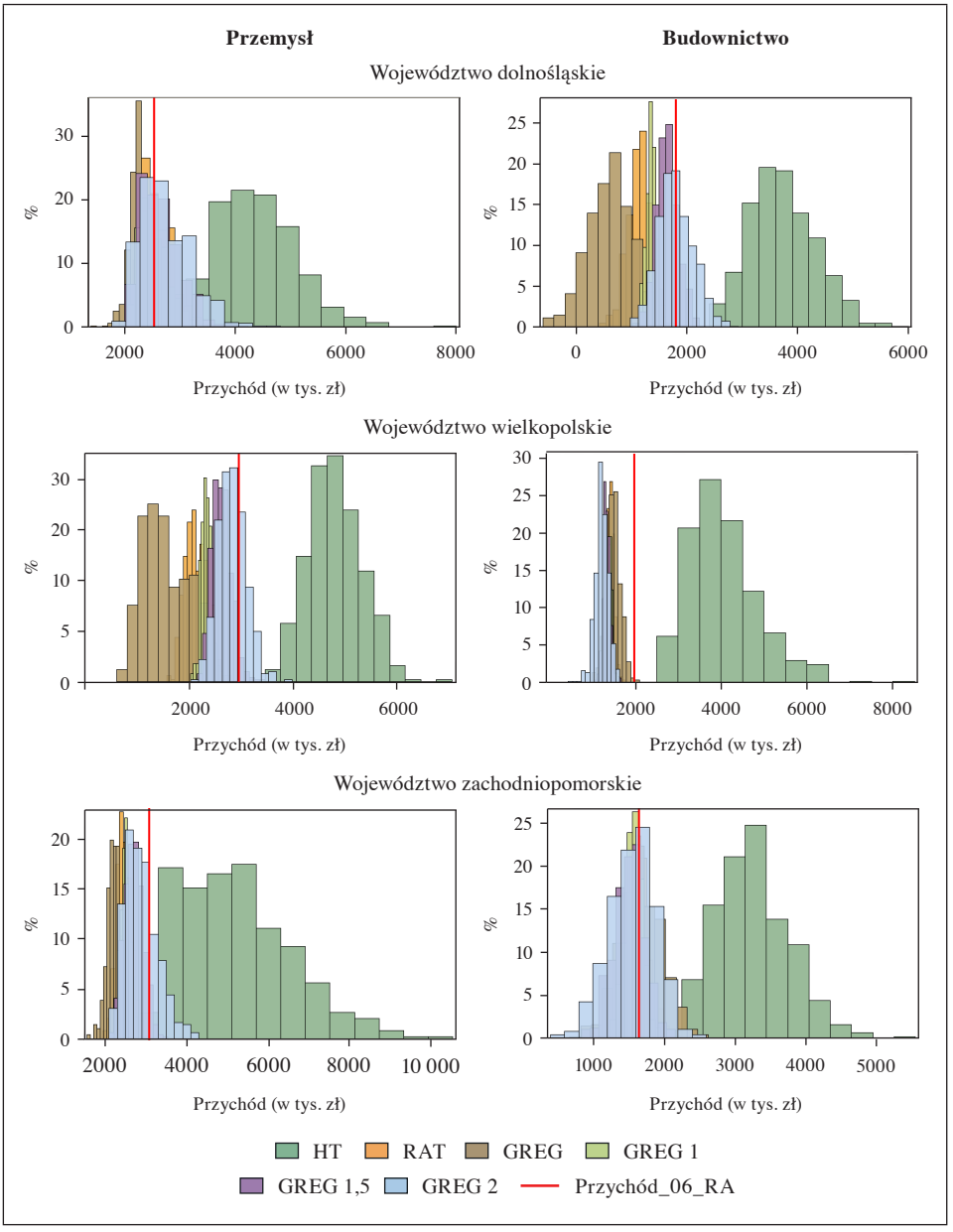
Tabela 3. Szacunek przychodu przedsiębiorstw dla czerwca 2012 r. w przekroju województw i czterech sekcji PKD (w tys. zł)

Województwo	Estymator											
	HT	RAT	GREG	GREGz1	GREGz1,5	GREGz2	HT	RAT	GREG	GREGz1	GREGz1,5	GREGz2
	Przemysł						Budownictwo					
Dolnośląskie	4437	2680	2300	2489	2570	2621	3731	1110	561	841	1200	1536
Kujawsko-pomorskie	4706	2510	2430	2477	2538	2584	3169	1653	1860	1736	2152	2811
Lubelskie	3291	2064	1901	1973	2022	2075	3451	1676	1748	1366	1147	928
Lubuskie	3227	1829	1846	1631	1668	1884	2953	2041	1657	1899	2126	2373
Łódzkie	3194	1778	1632	1694	1937	2116	3518	1199	512	1202	1454	1575
Małopolskie	3694	2024	1980	1748	1785	2024	3488	1564	1848	1484	1299	1145
Mazowieckie	5584	3911	4232	4243	4248	4178	4296	2918	3272	2771	2511	2390
Opolskie	3199	1925	1946	1903	1894	1893	1735	1087	1350	1371	1381	1386
Podkarpackie	3126	1683	1503	1780	1875	1859	2173	958	664	907	1104	1283
Podlaskie	3056	2173	2164	2021	1952	1955	2284	1407	1222	1445	1547	1640
Pomorskie	4511	2513	2475	2430	2371	2316	2457	1591	1786	1310	1333	1481
Śląskie	6814	2923	2533	3004	3064	3097	2654	1392	1311	1381	1424	1468
Świętokrzyskie	4318	2215	2168	2177	2188	2202	3142	1263	1267	931	792	715
Warmińsko-mazurskie	8222	2364	1580	1609	2247	3218	2387	1217	1314	1162	1063	1011
Wielkopolskie	4784	2028	1372	1933	2121	2268	4015	1433	1508	1359	1272	1220
Zachodnioporskie	5129	2452	2186	2280	2308	2318	3251	1651	1772	2024	2033	1679

cd. tabeli 3

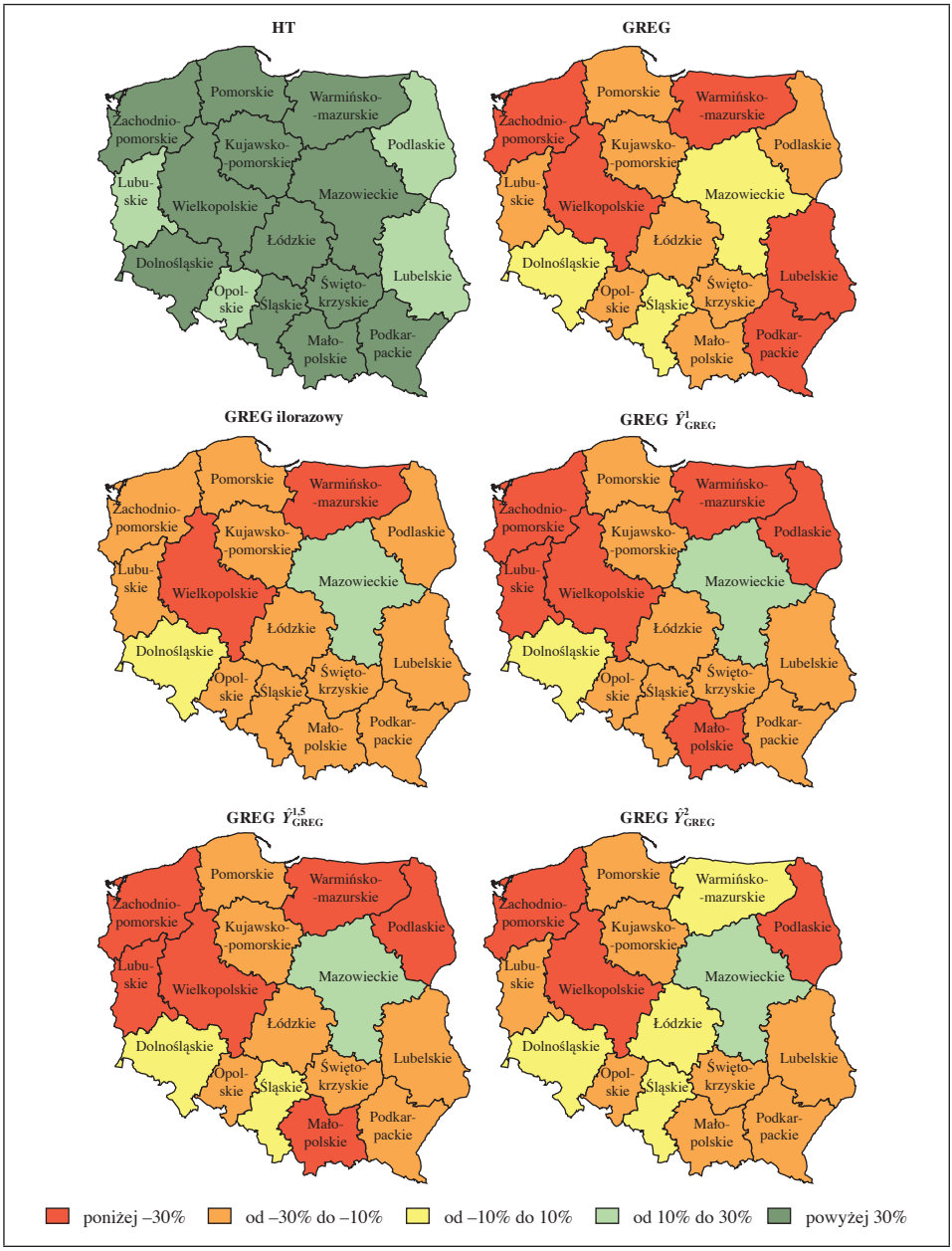
Województwo	Estymator											
	HT	RAT	GREG	GREGz1	GREGz1,5	GREGz2	HT	RAT	GREG	GREGz1	GREGz1,5	GREGz2
	Handel						Transport					
Dolnośląskie	8285	4295	3999	3961	4032	4240	4742	2207	2295	2149	2096	2083
Kujawsko-pomorskie	8761	4477	4217	3711	3942	4945	3509	2396	2498	2180	1249	1804
Lubelskie	9316	4942	4974	4711	4495	4287	4418	2079	2090	2038	2031	2035
Lubuskie	7896	4257	3658	4216	4357	4564	4527	2656	2498	2530	2546	2560
Łódzkie	7273	4685	4582	4562	4547	4562	4705	2267	1731	1744	1788	1859
Małopolskie	12448	5940	5907	5576	5389	5286	4502	2405	2231	2168	2156	2147
Mazowieckie	18637	7953	7451	7511	8245	9422	8559	3434	3349	2704	2524	2660
Opolskie	6949	5056	4898	5020	5070	5101	6915	2918	2916	2886	2876	2871
Podkarpackie	6277	4147	4019	4085	4181	4297	3868	1623	1059	637	586	589
Podlaskie	7959	5237	5136	5201	5252	5328	5005	2358	2001	1978	1978	1987
Pomorskie	8339	5325	5280	4856	4544	4237	4891	2702	2621	2638	2557	2394
Śląskie	8979	4779	4566	4550	4763	5289	4216	2242	2014	1979	1979	2025
Świętokrzyskie	7094	3913	3790	3528	3480	3847	4563	1988	2273	2034	1956	1899
Warmińsko-mazurskie	6320	3884	3863	3805	3755	3691	3782	2498	2142	2383	2439	2472
Wielkopolskie	9175	5282	5240	4960	4800	4750	8274	3080	2832	2848	2769	2635
Zachodnioporskie	9354	6365	6551	6121	6123	6202	5705	2982	3523	3192	3032	2887

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.



Rys. 10. Rozkład szacunków dla wybranych województw – przemysł i budownictwo

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.



Rys. 11. Porównanie względnego obciążenia szacunku przychodów przedsiębiorstw na podstawie estymatorów HT oraz typu GREG w województwach – przemysł

Źródło: opracowanie własne na podstawie danych pochodzących z badania DG-1 oraz rejestrów administracyjnych.

Na rys. 10 przedstawiono histogramy prezentujące rozkłady szacunków otrzymanych na podstawie badanych estymatorów metodą bootstrap dla sekcji „przemysł" i „budownictwo" dla wybranych województw. Estymatory GREG, w których wykorzystano zmienne pomocnicze pochodzące z rejestrów, charakteryzują się zdecydowanie mniejszym obciążeniem szacunków. W ich przypadku widoczna jest wyraźna koncentracja rozkładów wokół wartości referencyjnej oznaczonej czerwoną linią.

W celu przestrzennego zobrazowania rozbieżności pomiędzy oszacowaniami i wartościami rzeczywistymi sporządzono wykresy mapowe ukazujące nasilenie obciążenia w przekroju wszystkich województw (por. rys. 11). Wyniki przeprowadzonego badania wskazują, że w zdecydowanej większości domen oceny estymatorów otrzymane na podstawie estymatora HT są przeszacowane w porównaniu z wartościami zawartymi w rejestrach administracyjnych. Największe rozbieżności widoczne są w przypadku estymatora HT. Zastosowanie estymacji typu GREG wpłynęło na zmniejszenie obciążenia, jednak zakres poprawy był zróżnicowany. Na różnice miał wpływ nie tylko rodzaj estymatora GREG, ale również badana domena.

6. Wnioski

Otrzymane wyniki badań pozwalają na sformułowanie następujących wniosków:

– włączenie do szacunku zmiennych pomocniczych w ramach estymacji GREG wpłynęło na znaczną poprawę precyzji w porównaniu z estymacją HT,

– w przypadku zastosowania metody nieklasycznej większą poprawę obserwujemy dla domen charakteryzujących się znacznym zróżnicowaniem i silną asymetrią zarówno jeśli chodzi o precyzję, jak i dokładność szacunku,

– w badaniu biorą udział jedynie jednostki, dla których zmienna 'z' jest różna od 0; oznacza to, że w badaniu za zmienną 'z' można przyjąć jedynie cechę, która nie przyjmuje wartości zerowych,

– stopień poprawy precyzji estymacji w przypadku estymatorów uwzględniających transformację uzależniony jest od odpowiedniego doboru wartości parametru γ. Znaczną poprawę można osiągnąć, dobierając do danej domeny odpowiedni model. Postępowanie takie powoduje jednak, że stosowanie przedstawionych w artykule zmodyfikowanych estymatorów GREG w przypadku dużej liczby małych domen jest czasochłonne i znacznie utrudnione.

Literatura

Bracha C. [2004], *Estymacja danych z badania aktywności ekonomicznej ludności na poziomie powiatów dla lat 1995–2002*, GUS, Warszawa.

Chambers R., Chandra H., Salvati N., Tzavidis N. [2014], *Outlier Robust Small Area Estimation*, „Journal of the Royal Statistical Society: Series B (Statistical Methodology)", vol. 76, no 1, https://doi.org/10.1111/rssb.12019.
Chambers R.L., Falvey H., Hedlin D., Kokic P. [2001], *Does the Model Matter for GREG Estimation? A Business Survey Example*, „Journal of Official Statistics", vol. 17, nr 4.
Clark R.G., Kokic P., Smith P.A. [2017], *A Comparison of Two Robust Estimation Methods for Business Surveys*, „International Statistical Review", vol. 85, nr 2, https://doi.org/10.1111/insr.12177.
Dehnel G. [2014], *Winsorization Methods in Polish Business Survey*, „Statistics in Transition – New Series", vol. 15, nr 1.
Dehnel G. [2016], *M-estimators in Business Statistics*, „Statistics in Transition – New Series", vol. 17, nr 4.
Dehnel G. [2017], *GREG Estimation with Reciprocal Transformation for a Polish Business Survey* [w:] *Proceedings of the 11th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*, eds M. Papież, S. Śmiech, Foundation of the Cracow University of Economics, Cracow.
Lehtonen R., Särndal C.E., Veijanen A. [2016], *Generalized Regression and Model-calibration Estimation for Domains: Accuracy Comparison*, https://www.researchgate.net/publication/228665672_Generalized_regression_and_model-calibration_estimation_for_domains_Accuracy_comparison (data dostępu: 17.12.2017).
Ludność. Stan i struktura demograficzno-społeczna. Narodowy Spis Powszechny Ludności i Mieszkań 2011 [2013], GUS, Warszawa.
Rao J.N.K., Molina I. [2015], *Small Area Estimation*, 2nd ed., Wiley Series in Survey Methodology, Wiley, Hoboken, New Jersey.
Särdnal C.E., Swensson B., Wretman J. [1992], *Model Assisted Survey Sampling*, Springer Verlag, New York.
Wykorzystanie danych administracyjnych w badaniu: Ocena bieżącej działalności gospodarczej przedsiębiorstw [2016], GUS, Warszawa.

Model Selection and the GREG Estimator Bias in a Small Business Survey
(Abstract)

Estimation for a very skewed population containing extreme values is problematic, especially at a low level of aggregation. Traditional direct estimation methods do not provide satisfactory results. The growing demand for detailed information and the wider possibility of using data from administration registers has increased the importance of recognising more sophisticated estimation methods. Generalised Regression (GREG) estimation is an example of one such type. The paper examines the importance of the model chosen in GREG estimation in dealing with highly variable and outlier-prone populations. The model-assisted GREG estimator is applied to a real business survey. Lagged variables from administrative registers were used as the auxiliary variables. The variable of interest – mean revenue of small companies – was estimated for provinces cross-classified by categories of economic activity.

Keywords: GREG, business statistics, model-assisted estimation, outliers.