PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS **nr 468 • 2017**

Taksonomia 28
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Grażyna Dehnel
Uniwersytet Ekonomiczny w Poznaniu
e-mail: g.dehnel@ue.poznan.pl

WINSORYZACJA W OCENIE MAŁYCH PRZEDSIĘBIORSTW[1]

WINSORIZATION FOR SMALL ENTERPRISES

DOI: 10.15611/pn.2017.468.06
JEL Classification: C40, C51

Streszczenie: Rozwój gospodarczy określił nowe zadania dla statystyki przedsiębiorstw. Można zaobserwować wzrost zapotrzebowania na dane statystyczne, które dostarczane są w krótkich odstępach czasu oraz charakteryzują się większą dokładnością i spójnością. Sprostanie temu wyzwaniu wymusza poszukiwanie metod szacunku zmierzających w kierunku zwiększenia stopnia wykorzystania źródeł administracyjnych. Adaptacja nowych rozwiązań ma przyczynić się do zwiększenia efektywności prowadzonych szacunków, rozszerzenia zakresu informacji zarówno co do liczby cech, jak i rodzajów przekrojów, w których dane są publikowane. Celem niniejszego badania jest próba uwzględnienia winsoryzacji w estymacji typu GREG proponowanej przez statystykę małych obszarów do szacunku charakterystyk dotyczących małych przedsiębiorstw. W estymacji jako zmienne pomocnicze uwzględniono zmienne opóźnione w czasie. Badanie prowadzono na niskim poziomie agregacji, stanowiącym przekrój województw z uwzględnieniem rodzaju PKD.

Słowa kluczowe: estymacja odporna, statystyka gospodarcza, obserwacje odstające, estymator Winsora.

Summary: The economic development assigned new tasks for business statistics. Statistical data are expected to be delivered in short intervals, with improved accuracy and coherence. It determines the demand for seeking the methods of estimation aimed at increasing the use of administrative systems. The actions undertaken are aimed at increasing the effectiveness of estimates and the extension of the scope of information both to the number of variables and the types of cross-sections in which data are published. The paper presents an attempt to estimate basic economic information about small enterprises by applying *winsorization* to one of the small area statistics methods − GREG estimation. Lagged variables from the administrative registers will be used as auxiliary variables. The study is conducted at a low level of aggregation in the joint cross-section of the province and economic activity classification.

Keywords: robust estimation, business statistics, outliers, Winsorized estimator.

[1] Projekt finansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2015/17/B/HS4/00905.

1. Wstęp

Zasadniczą część statystyki gospodarczej stanowi statystyka krótkookresowa. Obejmuje ona szeroki zakres informacji o działalności przedsiębiorstw dostarczanej w trakcie trwania roku sprawozdawczego z częstotliwością kwartalną, a nawet miesięczną. Rola i udział statystyki krótkookresowej w procesie gospodarowania z roku na rok wzrasta. Coraz bardziej liczy się bowiem informacja nie tylko dostarczana małym jednostkom, ale również dostępna w coraz krótszych odstępach czasowych. To z kolei powoduje, iż coraz więcej uwagi poświęca się próbom modyfikacji metod stosowanych w badaniach statystycznych przedsiębiorstw. Proponowane w ramach tych badań nowe rozwiązania powinny uwzględniać co najmniej dwa wymiary. Jednym z nich jest specyfika populacji przedsiębiorstw, którą charakteryzuje obecność obserwacji odstających. Drugi element ma charakter bardziej ogólny. Chodzi bowiem o trwającą aktualnie modernizację całego systemu statystyki publicznej w kierunku wykorzystania dostępnych źródeł administracyjnych. Stąd też obserwujemy rozwój takich metod, jak integracja danych, kalibracja, imputacja czy estymacja pośrednia. Mając to na względzie, w niniejszym artykule poddano analizie podejście uwzględniające w pewnym stopniu wszystkie wyżej wymienione elementy. Zaproponowano wykorzystanie zmiennych pomocniczych opóźnionych w czasie, pochodzących z zasobów administracyjnych w estymacji odpornej. Celem badania była ocena możliwości zastosowania winsoryzacji w ramach estymacji GREG, uwzględniającej zmienne dodatkowe pochodzące z rejestrów administracyjnych do szacowania informacji o działalności gospodarczej przedsiębiorstw. Oceny estymatorów dokonano na podstawie badania empirycznego, w którym wykorzystano dane dotyczące małych przedsiębiorstw działających w ramach sekcji *Przemysł* oraz *Budownictwo*.

2. Estymator Winsora

Szacunek charakterystyk dotyczących podmiotów gospodarczych ze względu na własności rozkładu jednostek według badanych cech stanowi szczególny rodzaj wyzwania. Stąd też proponowane są podejścia polegające na dokonaniu pewnej modyfikacji próby, która ma na celu „uodpornienie" estymatora na duże reszty. Jednostki wylosowane do próby, u których wartości cechy wykraczają poza pewne ustalone wartości graniczne, zostają zmienione. Ta zmiana obejmuje modyfikację wartości wag związanych z obserwacjami odstającymi lub modyfikację wartości badanych zmiennych dla tych jednostek. Przykładem tego drugiego podejścia jest estymacja Winsora [Chambers i in. 2000]. Polega ona na dokonaniu podziału jednostek wylosowanych do próby w oparciu o punkty graniczne na dwie grupy. Jedną z nich stanowią dane wykorzystane do budowy modelu w postaci niezmienionej. Druga grupa to obserwacje odstające, które są włączone do próby w postaci zmodyfikowanej. Szacunku parametrów dokonuje się na podstawie tak zmienionej próby, w oparciu o

wybrany rodzaj estymacji stosowanej w badaniach reprezentacyjnych. W przeprowadzonym badaniu empirycznym zastosowano estymację GREG wykorzystującą informacje o zmiennych pomocniczych [Rao, Molina 2015]. Ich źródłem były rejestry administracyjne.

Estymator Winsora można zapisać za pomocą wzoru:

$$\hat{Y}_{win} = \sum_{i \in s_d} \tilde{w}_i y_i^* = \sum_{i \in s_d} w_i g_i y_i^*, \tag{1}$$

gdzie zmodyfikowane wartości zmiennej badanej y_i^* są wyznaczane w następujący sposób [Preston, Mackin 2002]:

$$y_i^* = \begin{cases} \left(\frac{1}{\tilde{w}_i}\right) y_i + \left(1 - \frac{1}{\tilde{w}_i}\right) K_{Ui} & \text{jeśli} \quad y_i > K_{Ui}, \\ y_i & \text{jeśli} \quad K_{Li} \le y_i \le K_{Ui}, \\ \left(\frac{1}{\tilde{w}_i}\right) y_i + \left(1 - \frac{1}{\tilde{w}_i}\right) K_{Li} & \text{jeśli} \quad y_i < K_{Li}, \end{cases} \tag{2}$$

$$g_i = \left(1 + x_i' \left(\sum_{i \in s_d} w_i x_i x_i' \right)^{-1} \left(t_x - \sum_{i \in s_d} w_i x_i \right)' \right), \tag{3}$$

$$\hat{K}_{Ui} = \hat{\mu}_i - \frac{B_U}{(\tilde{w}_i - 1)} \qquad \hat{K}_{Li} = \hat{\mu}_i - \frac{B_L}{(\tilde{w}_i - 1)}, \tag{4}$$

gdzie: $U = \{1, ... i, ...N\}$ – populacja generalna *N*-elementowa, $s \left(s \subseteq N \right)$ – próba, s_d – część próby należąca do domeny *d*, $\tilde{w}_i = w_i g_i \quad w_i = 1/\pi_i$ – wagi wynikające ze schematu losowania, g_i – wagi zależne od wartości zmiennej pomocniczej, $x_i = (x_{1i}, ..., x_{ki}, ..., x_{Ki})'$ – wektor zmiennych pomocniczych, $t_x = \sum_{i \notin U} x_i$ – wartość globalna zmiennej *x* w populacji generalnej, K_{Ui} – górny punkt graniczny, K_{Li} – dolny punkt graniczny, $B_U = E\left[\hat{Y}_{winU} - \hat{Y}_{DIR} \right]$ – obciążenie estymatora $\hat{Y}_{winU}$, $B_L = E\left[\hat{Y}_{winL} - \hat{Y}_{DIR} \right]$ – obciążenie estymatora $\hat{Y}_{winL}$, $\hat{Y}_{winU}$ / $\hat{Y}_{winL}$ – estymator Winsora wartości globalnej uwzględniający tylko górny / tylko dolny punkt graniczny.

Oszacowanie punktów granicznych $\hat{K}_{Ui}$ i $\hat{K}_{Li}$ wymaga wyznaczenia wartości obciążenia B_U i B_L oraz parametru $\hat{\mu}_i$. Metody pozwalające na szacunek obciążenia

zostały przedstawione w publikacjach Preston, Mackin [2002] oraz Dehnel [2014]. W niniejszym badaniu głębszej analizie poddano estymację parametru $\hat{\mu}_i = \hat{\beta} x_i$. Zastosowanie znajdują tu metody regresji odpornej, takie jak: *Trimmed least squares* (TLS), *Trimmed least absolute value* (LAV), *Sample Splitting* (SPLIT).

Metoda *Trimmed least squares* (TLS) polega na minimalizacji funkcji:

$$F = \sum_{i \in s} \left(y_i - \beta^T x_i \right)^2. \tag{5}$$

Na podstawie modelu obliczane są wartości teoretyczne, a następnie reszty. W kolejnym kroku z próby usuwane są te jednostki, dla których otrzymano największe dodatnie i ujemne wartości reszt. Z reguły próba zmniejszana jest o około 5% jednostek. Dla tak zredukowanej próby wyznaczane są nowe parametry modelu, na podstawie którego wyznacza się wartości $\hat{\mu}_i = \hat{\beta} x_i$. Technika TLS ma tę zaletę, że jest szybka i prosta w użyciu.

Metoda *Trimmed least absolute value* (LAV*)* jest bardzo zbliżona do metody TLS. Różnica dotyczy jedynie rodzaju funkcji, której wartości są minimalizowane:

$$F = \sum_{i \in s} \left| y_i - \beta^T x_i \right|. \tag{6}$$

Przyjmuje się założenie, że technika LAV stanowi bardziej odporny model regresji niż TLS, gdyż duże wartości reszt w mniejszym stopniu wpływają na parametry regresji.

Trzecia z badanych metod *Sample Splitting* (SPLIT) oparta jest na KMNK. Stosuje się ją do danych, które uprzednio w sposób losowy zostały podzielone na dwie połowy. Dla każdej wyznaczany jest model regresji. Reszty dla jednostek należących do jednej połowy danych obliczane są na podstawie modelu otrzymanego dla drugiej połowy. Następnie po połączeniu próby usuwa się jednostkę, dla której zanotowano największą wartość reszty. Ten proces jest powtarzany do momentu, aż określona część danych zostanie usunięta. Zakłada się, że technika SPLIT charakteryzuje się większą odpornością niż TLS, gdyż reszty wykorzystywane do wskazania jednostek, które mają być usunięte, nie są obliczane na podstawie modelu regresji, który był dla nich wyznaczony.

3. Charakterystyka badania

Badaniem empirycznym objęto przedsiębiorstwa małe (10-49 pracujących), prowadzące działalność gospodarczą w ramach sekcji *Przemysł* i *Budownictwo*. Analizie poddano model, w którym zmienną zależną stanowił *przychód* uzyskany przez przedsiębiorstwa w czerwcu 2012 roku. Jako zmienne niezależne przyjęto *przychód*,

koszt oraz *liczbę pracujących* według stanu zanotowanego w grudniu 2011 roku. Decyzję o wykorzystaniu zmiennych opóźnionych w czasie podjęto, biorąc pod uwagę techniczne ograniczenia, z jakimi musi się liczyć GUS, prowadząc badania statystyczne. Chodzi tu przede wszystkim o przesunięcie czasowe, jakie ma miejsce przy udostępnianiu statystyce publicznej informacji zawartych w zasobach administracyjnych.

Dane dotyczące zmiennej zależnej pochodziły z badania DG1 [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.

4. Ocena szacunków otrzymanych w badaniu empirycznym

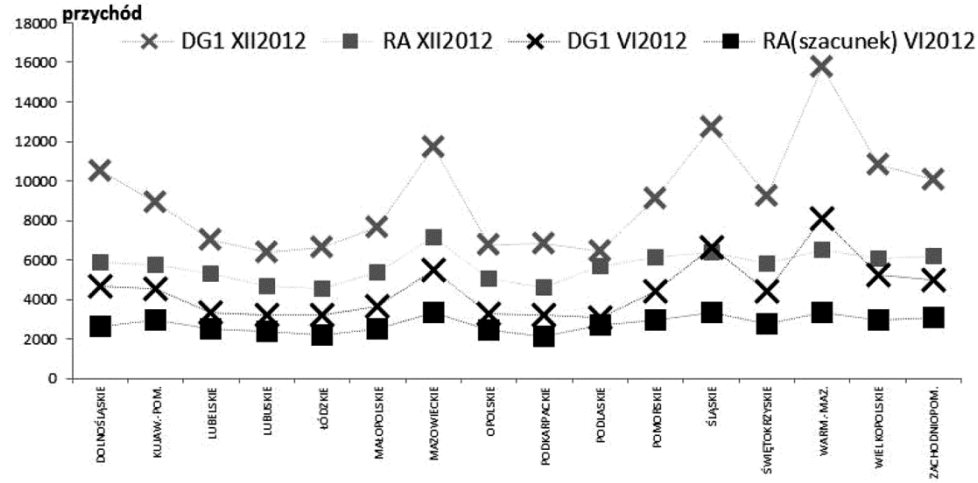
Do oceny precyzji i dokładności otrzymanych szacunków wykorzystano metodę bootstrapową. Wykonano 1000 repetycji losowania podprób, na podstawie których oszacowano wartość *przychodu* przedsiębiorstw dla czerwca 2012 roku w przekroju przyjętych domen studiów. Efektywność estymacji oceniono na podstawie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

$$CV\left(\hat{Y}_d\right) = \frac{\sqrt{Var\left(\hat{Y}_d\right)}}{\hat{Y}_d}. \tag{7}$$

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 roku. Przyjęto prawdziwość relacji: stosunek *przychodu* zarejestrowanego w zeznaniach podatkowych dla badanych przedsiębiorstw na poziomie domeny studiów do wartości *przychodu* z badania DG1 pozostaje stały (por. rys. 1).

$$\frac{\text{Przychód}_{\text{XII2012}}}{\text{Przychód\_próba}_{\text{XII2012}}} = \frac{\text{Przychód}_{\text{VI2012}}}{\text{Przychód\_próba}_{\text{VI2012}}}. \tag{8}$$

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



Rys. 1. Wartości przychodu w czerwcu i grudniu 2012 roku zarejestrowane na podstawie badania DG1 oraz zeznań podatkowych w sekcji *Przemysł*

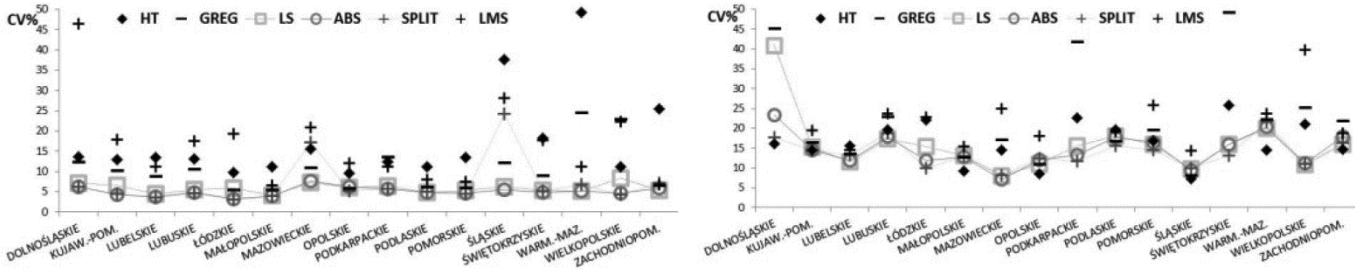
Źródło: opracowanie własne na podstawie badania Raportu [GUS 2016].

5. Wyniki empiryczne badania

W ocenie precyzji szacunku jako punkt odniesienia przyjęto szacunek otrzymany za pomocą klasycznej metody estymacji Horvitza-Thompsona (HT). Na podstawie wartości miary efektywności CV można zauważyć, że to właśnie ten estymator charakteryzuje się największym zróżnicowaniem (por. rys. 2). Mniejszą zmienność obserwujemy w przypadku estymatora GREG wykorzystującego zmienne pomocnicze pochodzące z rejestrów administracyjnych. W znacznym stopniu na zmniejszenie się wartości wariancji estymatora wpływa zastosowanie winsoryzacji, co jest zgodne z założeniami tej metody. W większości wyróżnionych domen współczynniki zmienności estymatorów Winsora, wykorzystujących techniki regresji odpornej TLS, LAV oraz SPLIT, kształtują się na podobnym poziomie. Jakość precyzji zależy od stopnia odporności zastosowanej techniki regresji.

W ocenie dokładności szacunku przychodu przedsiębiorstw posłużono się wartościami referencyjnymi wyznaczonymi na podstawie zależności ilorazowej opisanej powyżej (por. pkt 4). Dodatkowo, w celu pełniejszej oceny, estymację Winsora porównano z estymacją HT oraz GREG (por. rys. 3).

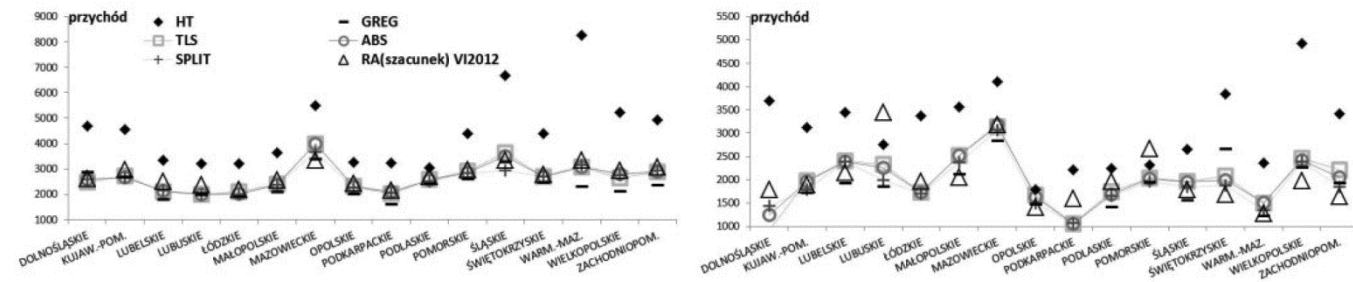
Otrzymane wyniki wskazują na to, że zastosowanie winsoryzacji w znacznym stopniu poprawiło dokładność szacunku w odniesieniu estymacji HT czy GREG. Estymacja HT w prawie wszystkich badanych domenach doprowadziła do znacznego przeszacowania wartości *przychodu*, z kolei w przypadku estymatora GREG widoczne jest niedoszacowanie badanego parametru. Największe rozbieżności zaobserwować można dla domen, dla których zanotowano największą dyspersję zmiennych uwzględnionych w modelu.



Przemysł *Budownictwo*

Rys. 2. Precyzja szacunku dla sekcji *Przemysł* i *Budownictwo*

Źródło: opracowanie własne na podstawie badania Raportu [GUS 2016].

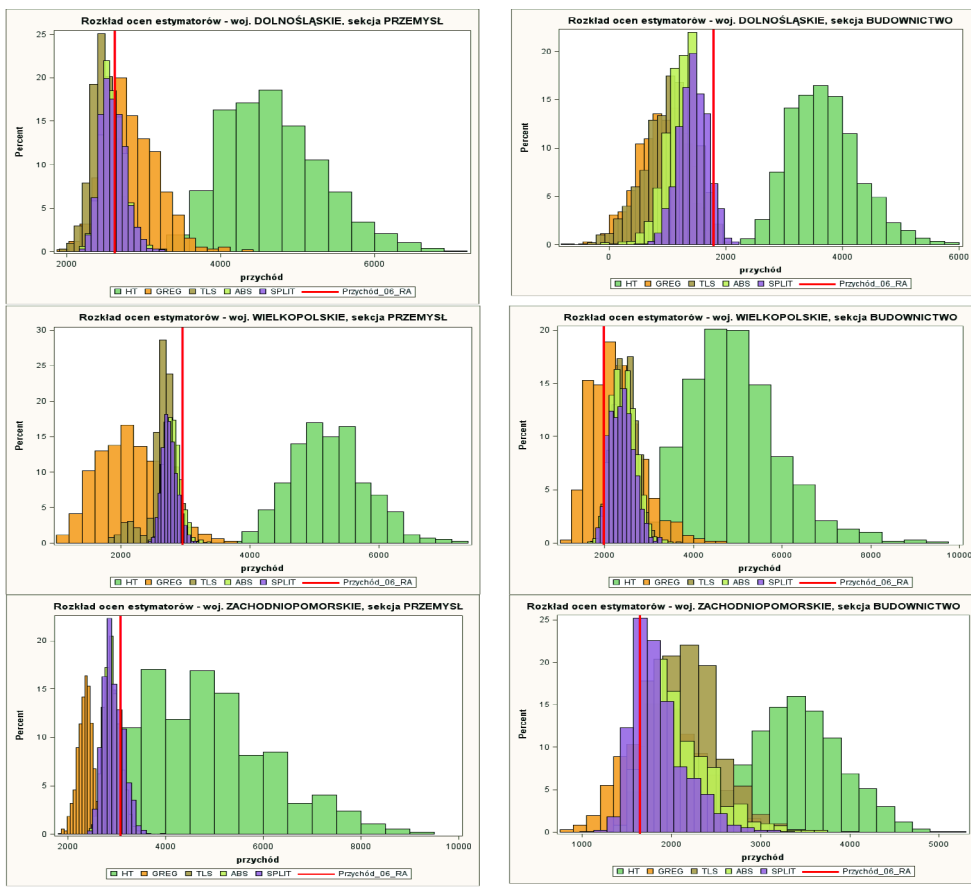


Przemysł *Budownictwo*

Rys. 3. Szacunek przychodu w czerwcu 2012 roku w sekcji *Przemysł* i *Budownictwo*

Źródło: opracowanie własne na podstawie badania Raportu [GUS 2016].

Na uwagę zasługują histogramy prezentujące rozkłady szacunków otrzymanych metodą bootstrap (por. rys. 4). Zgodnie z teorią metody reprezentacyjnej estymatory HT i GREG charakteryzuje nieobciążoność. Wyniki badania empirycznego wskazują jednak na ich obciążenie. Wśród przyczyn można wskazać, poza tzw. obciążeniem próby, wpływ obecności obserwacji odstających. Co prawda uwzględnienie w estymacji GREG zmiennych pomocniczych poprawiło dokładność oszacowań, jednak to winsoryzacja najsilniej przybliżyła szacunki do wartości referencyjnej.



Przemysł *Budownictwo*

Rys. 4. Rozkład szacunków dla wybranych województw dla sekcji *Przemysł* i *Budownictwo*

Źródło: opracowanie własne na podstawie badania Raportu [GUS 2016].

6. Zakończenie

Wykorzystanie zmiennych pomocniczych opóźnionych w czasie, pochodzących z rejestrów administracyjnych, wpłynęło na polepszenie jakości szacunku, zarówno w przypadku klasycznego estymatora GREG, jak i jego modyfikacji uwzględniającej winsoryzację. Proces winsoryzacji przyniósł zdecydowaną poprawę nie tylko pod względem precyzji, ale i dokładności estymacji. Zastosowanie każdej z trzech badanych metod regresji odpornej charakteryzowało się podobną jakością szacunku. Różnice wynikały z „odporności" metody regresji na wartości odstające. Techniki regresji bardziej odporne w większym stopniu wpływały na poprawę efektywności. Ocena jakości szacunku pod kątem dokładności potwierdziła, że modyfikacja próby przy użyciu estymatora Winsora pozwala uniknąć obciążenia spowodowanego obecnością obserwacji odstających.

Literatura

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

Chambers R., Kokic P., Smigh P., Cruddas M., 2000, *Winsorization for Identifying and Treating Outliers in Business Surveys, Proceedings of the Second International Conference on Establishment Surveys*, Alexandria, Virginia, American Statistical Association, s. 717-726.

Cox B.G., Binder A., Chinnappa N.B., Christianson A., Colledge M.J., Kott P.S., 1995, *Business Survey Methods*, John Wiley & Sons, New York.

Dehnel G., 2014, *Winsorization methods in Polish business survey*, Statistics in Transition – new series, vol. 15, no. 1, s. 97-110.

GUS, 2016, Raport „Wykorzystanie danych administracyjnych w badaniu: Ocena bieżącej działalności gospodarczej przedsiębiorstw".

Kokic P.N., Bell P.A., 1994, *Optimal winsorizing cutoffs for a stratified finite population estimator*, Journal of Official Statistics, no. 10, s. 419-435.

Preston J., Mackin, C., 2002, *Winsorization for Generalised Regression Estimation*, Paper for The Methodological Advisory Committee, November, Australian Bureau of Statistics.

Rao J.N.K., Molina I., 2015, *Small Area Estimation. Wiley Series in Survey Methodology*, 2nd ed., Hoboken, Wiley, New Jersey.