PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Jan Paradysz, Marcin Szymkowiak
Akademia Ekonomiczna w Poznaniu

TAKSONOMETRYCZNE PODSTAWY KALIBRACJI W STATYSTYCE MAŁYCH OBSZARÓW

1. Wstęp

Statystyka małych obszarów (SMO) obejmuje klasę estymatorów wykorzystujących większe zasoby informacji niż w tradycyjnej metodzie reprezentacyjnej (por. [Särndal i in. 1992; Bracha 1996]). Przede wszystkim są to informacje spoza próby w postaci różnego rodzaju rejestrów administracyjnych, spisów, wykazów itd. Takie podejście jest szczególnie ważne wówczas, gdy liczebność próby jest niewystarczająca do uzyskania wiarygodnych ocen szacowanych parametrów. Jednakże w ostatnich latach (por. [Särndal; Lundström 2005]) w estymacji dla małych obszarów dużo uwagi poświęca się brakom odpowiedzi (*nonresponse*[1]). Szczególne znaczenie odgrywa tutaj kalibracja, która jest nowoczesną metodą szacowania parametrów w badaniach reprezentacyjnych z brakami odpowiedzi (por. [Deville, Särndal 1992; Estevao, Särndal 2006; Paradysz, Szymkowiak 2007]). Idea kalibracji polega na skorygowaniu, czyli odpowiednim wyważeniu, wyjściowych wag wynikających ze schematu losowania jednostek do próby w taki sposób, aby skompensować utratę informacji związaną z brakiem danych. Korekty wag dokonuje się z uwzględnieniem zmiennych pomocniczych, które mogą pochodzić z tej samej próby co zmienna szacowana bądź z różnych innych źródeł.

Celem artykułu jest zastosowanie kalibracji i estymacji pośredniej na potrzeby oszacowania wydatków gospodarstw domowych na wybrane towary i usługi konsumpcyjne w przekroju powiatów województwa wielkopolskiego w 2002 r. na podstawie danych pochodzących z Badania Budżetów Gospodarstw Domowych. Ze względu na założenia estymacji syntetycznej, która jest jednym z możliwych podejść w SMO, dokonano – z wykorzystaniem metod taksonomicznych – wyod-

[1] Osoby uczestniczące w badaniu będziemy nazywać respondentami a nieuczestniczące – nierespondentami.

rębnienia powiatów podobnych ze względu na wybrane cechy społeczno-ekonomiczne skorelowane z rynkiem pracy.

2. Przebieg badania

Głównym celem przeprowadzonych badań było oszacowanie średnich wydatków gospodarstw domowych na wybrane kategorie towarów i usług konsumpcyjnych w przekroju powiatów województwa wielkopolskiego. W tym celu wykorzystano trzy typy estymatorów: estymator Horvitza-Thompsona, estymator kalibracyjny oraz estymator syntetyczny ilorazowy. W pierwszej kolejności oszacowano wartości globalne wydatków na wybrane towary i usługi konsumpcyjne gospodarstw domowych powiatów województwa wielkopolskiego. Następnie, opierając się na informacjach o liczbie gospodarstw domowych dla poszczególnych powiatów pochodzących z NSP 2002, dokonano oszacowania średnich wydatków przypadających na gospodarstwo domowe.

3. Opis estymatorów

Na potrzeby badania wykorzystano trzy estymatory wartości globalnych. Pierwszy z nich to znany z klasycznej metody reprezentacyjnej estymator Horvitza-Thompsona, który wyraża się wzorem

$$\hat{Y}_{HT,d} = \sum_{i \in d} d_i y_i \,, \tag{1}$$

gdzie: d_i – waga wynikająca ze schematu losowania próby; $d_i = \frac{1}{\pi_i}$ gdzie π_i oznacza prawdopodobieństwo dostania się do próby i-tej jednostki,

y_i – wartość cechy y dla i-tej jednostki badania,

d – domena.

Gdy w badaniu występują braki odpowiedzi, ważona suma $\sum_{i \in d} d_i y_i$ jest z reguły niedoszacowana w stosunku do prawdziwej, aczkolwiek nieznanej wartości globalnej. Stanowi to punkt wyjścia do konstrukcji estymatorów kalibracyjnych, które, uwzględniając dodatkowe informacje w postaci wektora zmiennych pomocniczych, wykorzystują skorygowane – w stosunku do wyjściowych – wagi mające niwelować ujemny wpływ brakujących danych.

Na potrzeby badania wykorzystano estymator kalibracyjny

$$\hat{Y}_{cal,d} = \sum_{i \in d} w_i y_i \,, \tag{2}$$

gdzie: w_i – wagi kalibracyjne o postaci: $w_i = d_i + d_i\left(\breve{\mathbf{X}} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{m} d_i \underline{x}_i \underline{x}_i^T\right)^{-1} \underline{x}_i$,

$\breve{\mathbf{X}} = \left[\sum_{i\in d} d_i x_{i1}, \sum_{i\in d} d_i x_{i2}, \ldots, \sum_{i\in d} d_i x_{ik}\right]^T$ – wektor złożony z oszacowanych wartości globalnych zmiennych pomocniczych dla domeny d,

$\hat{\mathbf{X}} = \left[\sum_{i\in r_d} d_i x_{i1}, \sum_{i\in r_d} d_i x_{i2}, \ldots, \sum_{i\in r_d} d_i x_{ik}\right]^T$ – wektor złożony z oszacowanych wartości globalnych zmiennych pomocniczych dla domeny d wyznaczony na podstawie wag przypisanych do jednostek tworzących zbiór respondentów r_d dla zmiennej y,

$\underline{\mathbf{x}}_i = \left[x_{i1}, x_{i2}, \ldots, x_{ik}\right]^T$ – wektor złożony z wartości wszystkich zmiennych pomocniczych dla i-tego respondenta,

x_{ik} – wartość k-tej zmiennej pomocniczej dla i-tej jednostki badania.

Trzecim z wykorzystywanych estymatorów był estymator syntetyczny ilorazowy dany wzorem (3).

$$\hat{Y}_{syn,d} = \frac{\sum_{i\in d} d_i x_i \sum_{i\in Sk} d_i y_i}{\sum_{i\in Sk} d_i x_i}, \tag{3}$$

gdzie: x_i – wartość zmiennej pomocniczej dla i-tej jednostki badania,

Sk – skupisko obejmujące podobne domeny ze względu na pewien zestaw cech.

W estymacji syntetycznej przyjmuje się założenie, że mały obszar (domena) jest podobny do większego obszaru – zazwyczaj zawierającego w sobie ten pierwszy. Takie założenie pozwala na podział oszacowanej wartości globalnej dla obszaru większego na części odpowiadające poszczególnym małym obszarom (domenom) [Dehnel 2003]. W podejściu zaprezentowanym w pracy połączono w celu wyznaczenia ocen estymatora syntetycznego ilorazowego podobne domeny (powiaty województwa wielkopolskiego) w większe obszary (skupiska).

4. Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację panującą na rynku pracy

Głównym celem przeprowadzonej klasyfikacji powiatów województwa wielkopolskiego była próba zastosowania estymatora syntetycznego ilorazowego na potrzeby oszacowania wybranych kategorii wydatków gospodarstw domowych. Ponieważ założenia estymacji syntetycznej są często w praktyce niemożliwe do

zrealizowania, stosowne wydaje się łączenie podobnych powiatów w grupy, co może pozwolić na uniknięcie sztywnych założeń o identycznej relacji między zmienną szacowaną Y a zmienną pomocniczą X. Na potrzeby pracy dokonano więc typizacji powiatów ze względu na sytuację panującą na rynku pracy, gdyż ma ona wpływ na strukturę wydatków i dochodów gospodarstw domowych.

Celem wykorzystania estymatora syntetycznego ilorazowego na potrzeby klasyfikacji powiatów województwa wielkopolskiego posłużono się hierarchiczną metodą Warda. Jako zmienne diagnostyczne opisujące rynek pracy w powiatach w 2002 r. przyjęto (por. [Witkowska, Witkowski 2006]).

- udział procentowy zatrudnionych w sektorze prywatnym w liczbie zatrudnionych ogółem (stymulanta),
- przeciętne miesięczne wynagrodzenie – brutto (stymulanta),
- stopę bezrobocia rejestrowanego (destymulanta),
- udział procentowy bezrobotnych do 25 roku życia w ogólnej liczbie bezrobotnych (destymulanta),
- udział procentowy długotrwale bezrobotnych w ogólnej liczbie bezrobotnych (destymulanta),
- udział procentowy bezrobotnych bez stażu lub ze stażem do 1 roku w ogólnej liczbie bezrobotnych (destymulanta),
- udział procentowy bezrobotnych z wykształceniem wyższym w ogólnej liczbie bezrobotnych (destymulanta).

W wyniku zastosowania hierarchicznej metody Warda wyodrębniono trzy skupiska zawierające powiaty o podobnej sytuacji panującej na rynku pracy w województwie wielkopolskim (rys. 1).

Pierwsze skupisko obejmuje powiaty, w których sytuacja na rynku pracy była najkorzystniejsza. Oprócz powiatu poznańskiego w jego skład weszły miasta na prawach powiatu, tj. Kalisz, Leszno, Konin i Poznań. Drugie skupisko obejmowało w zasadzie powiaty leżące w południowej części województwa wielkopolskiego – w sumie 12 powiatów. Ostanie, trzecie skupisko, tworzyły powiaty o najtrudniejszej sytuacji na rynku pracy. Swoim zasięgiem obejmowało ono powiaty wschodniej i północnej części województwa (łącznie 18 powiatów). Podobne rezultaty badawcze można zauważyć w innych opracowaniach dotyczących badanego terenu (por. np. [Gołata 2004]).

Wyniki analizy taksonometrycznej w dalszej części pracy będą nam służyć do oceny jakości estymatora kalibracyjnego.

5. Wyniki zastosowania estymatora kalibracyjnego

Na rys. 2 i 3 zestawiliśmy wyniki estymacji wydatków gospodarstw domowych na określony typ dóbr konsumpcyjnych. Do porównań wybraliśmy tylko jedną grupę wydatków na napoje alkoholowe, wyroby tytoniowe i narkotyki w przekroju powiatów województwa wielkopolskiego w 2002 r. Wybór grupy był spowodowa-

ny tym, że występujące tutaj odmowy charakteryzowały się znaczną dyspersją: od 12,5% w nowotomyskim i leszczyńskim do 50% braków odpowiedzi w powiecie ostrzeszowskim. Pod względem braku odpowiedzi nie widać jednak większych różnic między poszczególnymi częściami Wielkopolski. Zasadniczo nie różnicuje powiatów ani wielkość powiatów, ani ich charakter, ani historia (zabór pruski, zabór rosyjski).

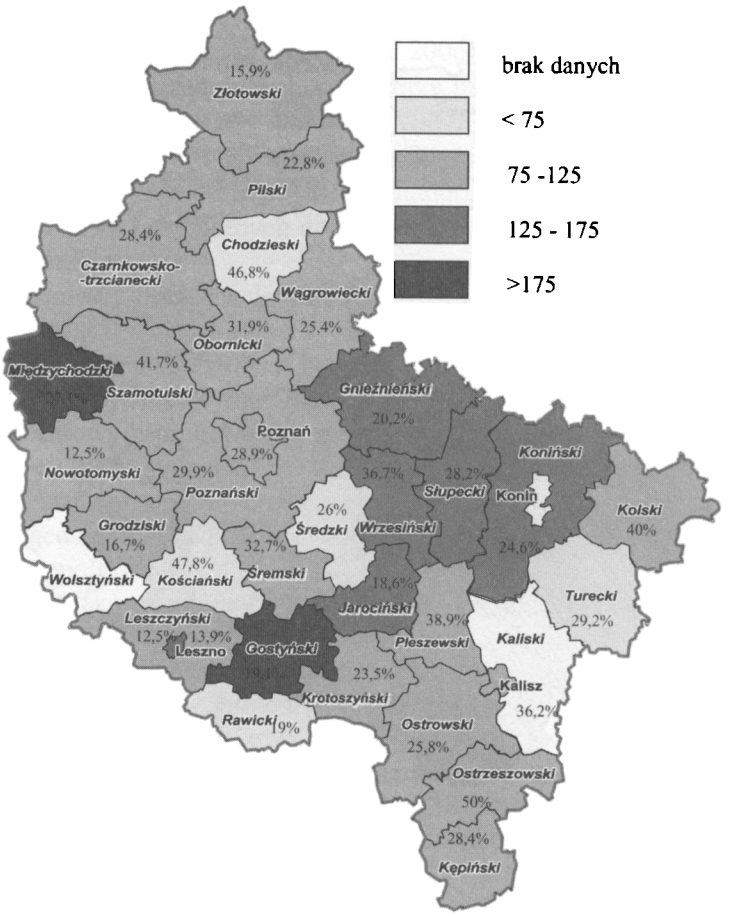
Rys. 1. Powiaty podobne pod względem sytuacji panującej na rynku pracy w województwie wielkopolskim w 2002 r.

Źródło: opracowanie własne.

Oprócz frakcji odmów, a raczej nieobecności w badaniu, rys. 2 przedstawia natężenie wydatków na napoje alkoholowe, wyroby tytoniowe i narkotyki (w zł).

Jeśli zaufać szacunkom na podstawie estymatora kalibracyjnego, to najwięcej się wydaje na ten rodzaj dóbr w powiecie gostyńskim i międzychodzkim – ponad 175 zł, a na rys. 2 widać, że jest to nawet ponad 200 zł.

Estymator kalibracyjny znacznie pogłębia różnice między powiatami (por. rys. 2). Tylko w nielicznych powiatach estymator kalibracyjny pokazuje zbliżone lub takie same oceny jak estymator bezpośredni i syntetyczny.



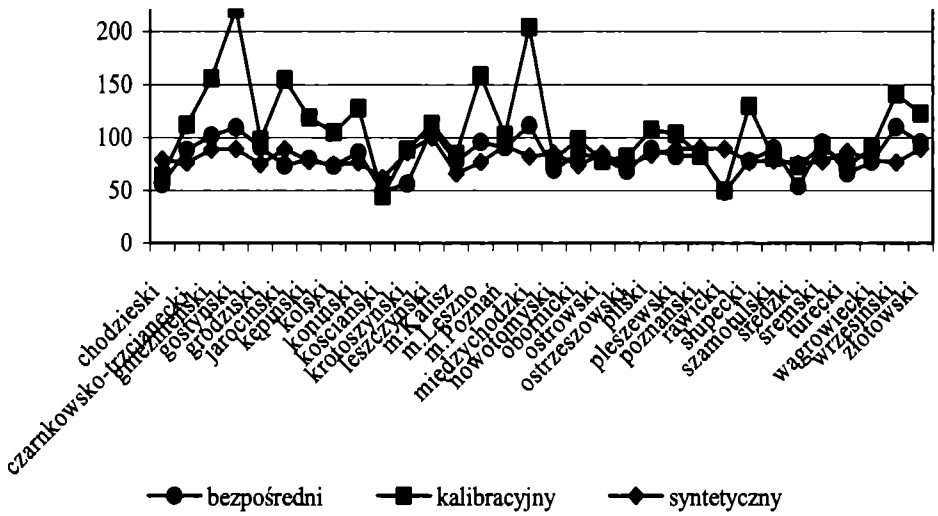
Rys. 2. Frakcja braków odpowiedzi i oszacowania średnich wydatków gospodarstw domowych na napoje alkoholowe, wyroby tytoniowe i narkotyki (w zł) z wykorzystaniem estymatora kalibracyjnego w przekroju powiatów województwa wielkopolskiego w 2002 r.
Źródło: opracowanie własne.

Na rys. 3 widać wyraźnie dość dużą zbieżność ocen estymatora syntetycznego i bezpośredniego. Oceny estymatora kalibracyjnego znacznie od nich odbiegają. Na

podstawie tego, co wiemy z analizy taksonometrycznej, większym zaufaniem bylibyśmy skłonni obdarzyć oceny estymatora kalibracyjnego. Przy tym te różnice byłyby jeszcze większe, gdyby można było uwzględnić w estymatorze kalibracyjnym różnorodność zachowań konsumpcyjnych respondentów i nierespondentów. Można bowiem przypuszczać, że wśród nierespondentów więcej wydaje się na napoje alkoholowe, wyroby tytoniowe i narkotyki.

6. Podsumowanie

Estymator kalibracyjny wydaje się przedstawiać bardziej przekonywająco różnice między powiatami województwa wielkopolskiego pod względem badanej cechy. Tradycyjne estymatory klasy SMO zazwyczaj niwelują różnice między małymi obszarami. Estymator kalibracyjny te różnice uwypukla, przez co wydaje nam się to bliższe prawdy. W każdym razie jest to bliższe temu, co można zaobserwować na podstawie analizy taksonometrycznej z wykorzystaniem wskaźników syntetycznych silnie skorelowanej z badaną zmienną.



Rys. 3. Oszacowania średnich wydatków na napoje alkoholowe, wyroby tytoniowe i narkotyki gospodarstw domowych z wykorzystaniem estymatora bezpośredniego, kalibracyjnego i syntetycznego w przekroju powiatów województwa wielkopolskiego w 2002 r.
Źródło: opracowanie własne.

Przy ostatecznej ocenie naszych wyników należy jednak mieć na uwadze, że wyniki estymacji mogą ulec zmianie. Na początku założyliśmy, że respondenci i nierespondenci charakteryzują się takimi samymi postawami konsumpcyjnymi. Jest to zapewne zbyt daleko idące uproszczenie, ale – jak na razie – brak nam pod-

staw do zmiany tego założenia. Pogłębiając problem oceny jakości estymatorów kalibracyjnych, analiza taksonometryczna mogłaby z powodzeniem spełniać funkcję kontrolną w stosunku do wyników estymacji pośredniej, a szczególnie z uwzględnieniem braków odpowiedzi.

Literatura

Bracha C. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, PWN, Warszawa.

Dehnel G. (2003), *Statystyka małych obszarów jako narzędzie oceny rozwoju ekonomicznego regionów*, AE, Poznań.

Deville J.-C., Särndal C.-E. (1992), *Calibration Estimators in Survey Sampling*, „Journal of the American Statistical Association", vol. 87, s. 376-382.

Estevao V.M., Särndal C.-E. (2006), *Survey Estimates by Calibration on Complex Auxiliary Information*, „International Statistical Review", vol. 74, s. 127-147.

Gołata E. (2004), *Estymacja pośrednia bezrobocia na lokalnym rynku pracy*, Prace habilitacyjne nr 11, AE, Poznań.

Paradysz J., Szymkowiak M. (2007), *Imputacja i kalibracja jako remedium na braki odpowiedzi w badaniu budżetów gospodarstw domowych*, [w:] Taksonomia 14, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1169, AE, Wrocław.

Särndal C.-E., Swensson B., Wretman J.H. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Särndal C.-E., Lundström S. (2005), *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Ltd.

Szymkowiak M. (2007), *Przyczynek do kalibracji w badaniach statystycznych z brakami odpowiedzi*, [w:] *Kapitał ludzki i wiedza w gospodarce – wyzwania XXI wieku*, red. E. Panek, Zeszyty Naukowe nr 96, AE, Poznań.

Witkowska A., Witkowski M. (2006), *Taksonometryczna analiza rynku pracy w województwie wielkopolskim w latach 2000-2003*, AE, Wrocław.

CLUSTERING METHODS OF CALIBRATION IN SMALL AREA ESTIMATION

Summary

The main goal of the paper is to apply calibration estimators and small area estimation methodology in Household Budget Survey (HBS) in Poland. In Polish HBS non-participation of the households selected for HBS is a very significant problem. In the paper the authors showes how new methodology of calibration and small area estimation in the context of clustering can be used in order to avoid negative impact of nonresponse. The results using calibration and synthetic estimators were also compared with results obtained with using Horvitz-Thompson estimator.