Imputacja i kalibracja w badaniu DG-1

Alina Szkop, Ośrodek Statystyki Małych Obszarów, Urząd Statystyczny w Poznaniu
Adam Ambroziak

Praca metodologiczna 3.207

Imputacja i kalibracja w badaniu DG-1 jest jedną z prac metodologicznych prowadzonych w OSMO w Poznaniu. Głównym celem tej pracy jest ocena przydatności imputacji i kalibracji w badaniu działalności gospodarczej przedsiębiorstw, ocena jakości uzyskanych wyników z zastosowaniem wymienionych technik, a także sformułowanie wytycznych co do możliwości ich praktycznego wykorzystania.

Badanie DG-1

Badanie DG-1 jest głównym źródłem uzyskiwania szybkich informacji o podstawowych miernikach działalności gospodarczej w przedsiębiorstwach, jak i o sytuacji społeczno-gospodarczej kraju i województw na potrzeby m.in. informacji publicznej, Narodowego Banku Polskiego czy organizacji międzynarodowych (Eurostat, OECD, Międzynarodowy Fundusz Walutowy, ONZ). Ponadto badanie to jest podstawą do opracowywania komunikatów i obwieszczeń Prezesa Głównego Urzędu Statystycznego na temat przeciętnego miesięcznego wynagrodzenia w sektorze przedsiębiorstw.
Jakość i spójność uzyskiwanych wyników są niezwykle ważnymi elementami w ramach całego procesu prowadzenia badania. Dlatego też wybór optymalnej metody z zakresu imputacji i kalibracji danych jest istotnym problemem metodologicznym.

Uogólnianie danych w badaniu

Kartoteka DG-1 + Baza B1 -> Baza B3
W bazie B3 dane uogólniane są na poziomie:

1 grupy PKD dla sekcji F, G, H, I, M, N (w bazie B1 musi istnieć co najmniej jeden rekord dla każdej grupy w ramach każdej sekcji).

2 działu PKD dla pozostałych sekcji (w bazie B1 musi istnieć co najmniej jeden rekord dla każdego działu w ramach każdej sekcji).

3 sekcji dla sektorów własności (w bazie B1 musi istnieć co najmniej jeden rekord dla sektora prywatnego i publicznego w ramach każdej sekcji)

Dane uogólnione w bazie B3 dla wyższych szczebli agregacji są sumą agregatów niższego szczebla.
Przykład: Do uogólniania w bazie B3 przychodów ze sprzedaży wyrobów i usług w cenach bazowych *Sa* stosowany jest wskaźnik ***Wu***:

$$Wu = \frac{\sum_{i=1}^{n} LPZ_i * \sum_{j=1}^{m} LPS_j}{\sum_{i=1}^{n} LPZ_i} \qquad (1)$$

wyliczany dla każdego działu lub grupy PKD. Uogólniając mamy:

$$Sa\_1b = Sa\_1bA * Wu \qquad (2)$$

gdzie: *Sa_ 1bA*-zagregowane przychody ze sprzedaży wyrobów i usług w cenach bazowych w danym miesiącu.

Zaproponowane metody imputacji i kalibracji

Zaproponowano następujące metody imputacji:
Metoda I - imputacja na szczeblu grupy/działu wg PKD (autor: Konik M.);
Metoda II - imputacja dla klasy jednostek dużych i średnich (autor: Konik M.) wg wzorów:

$$PZ\_mb = LPZ * \frac{\sum_{i=1}^{n} Pz\_mb_i}{\sum_{i=1}^{n} Pp\_3b_i} \qquad (3)$$

$$Wm = PZ\_mb * \frac{\sum_{i=1}^{n} wm_i}{\sum_{i=1}^{n} Pz\_mB_i}; \qquad (4)$$

Metoda III - imputacja na szczeblu grupy/działu wg PKD;
Metoda IV - imputacja na podstawie jednostek średnich wg wzoru:

$$Wm = LPS * \frac{\sum_{i=1}^{n} wm_i}{\sum_{i=1}^{n} Pz\_mb_i} \qquad (5)$$

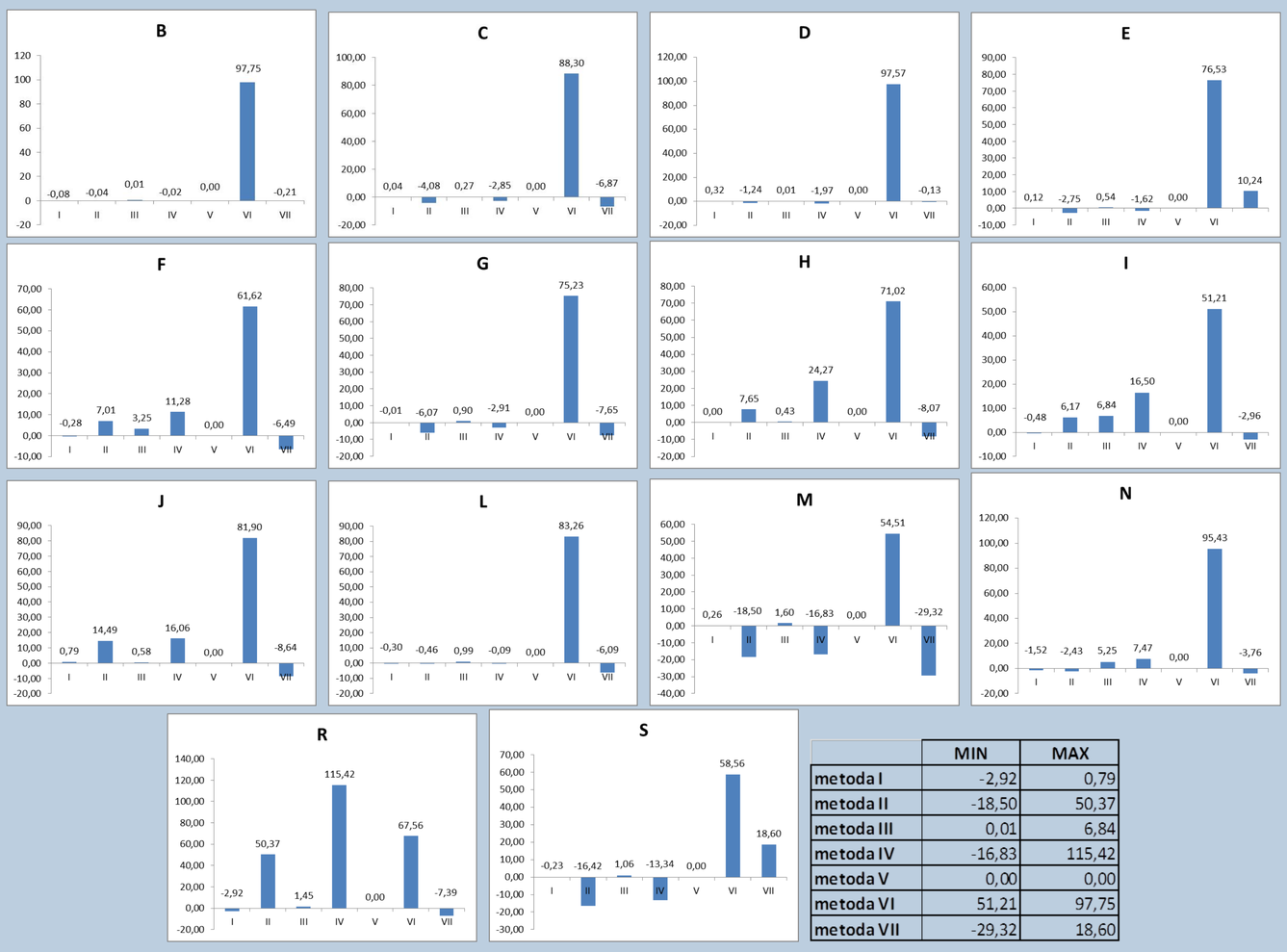
Metoda V - imputacja ilorazowa:

$$Wm = \sum_{i=1}^{n} wm_i * \frac{LPS}{\sum_{j=1}^{k} LPS_j}; \qquad (6)$$

Metoda VI - metoda imputacji średniej grupowej;
Metoda VII - random hot-deck;
Metoda VIII - kalibracja.

Rezultaty

Poniżej zaprezentowano wybrane wyniki prowadzonych prac - porównania uogólnionych siedmioma metodami wartości zmiennej Sa_1b w poszczególnych sekcjach. Zamieszczone jest również zestawienie zakresów zmian dla tych metod.



	MIN	MAX
metoda I	-2,92	0,79
metoda II	-18,50	50,37
metoda III	0,01	6,84
metoda IV	-16,83	115,42
metoda V	0,00	0,00
metoda VI	51,21	97,75
metoda VII	-29,32	18,60

Na podstawie analizy uzyskanych wyników przypuszcza się, że podejście zastosowane w metodzie V jest optymalne z punktu widzenia szybkiej i praktycznej implementacji do procesu badania DG-1. W porównaniu do uogólnianych wartości zmiennej Sa_1b zgodnie z dotychczasową metodologią badania DG-1 zmiany występują tylko na poziomie sektorów własności.

Oznaczenia

PZ_mB–przeciętna liczba zatrudnionych, LPZ–liczba pracujących w kartotece, Pz_mb-przeciętna liczba zatrudnionych w jednostce, która złożyła sprawozdanie, Pp_3b-liczba pracujących w jednostce, która złożyła sprawozdanie, n-liczba jednostek z obowiązkiem sprawozdawczym w agregacie, Wm-imputowana wartość w jednostce imputowanej, wm-wartość w jednostce, która złożyła sprawozdanie, LPS–liczba pracujących stała, k–liczba jednostek bez obowiązku sprawozdawczego w agregacie.

Literatura

1.Estevao V.M., Särndal C-E., (2006), Survey Estimates by Calibration on Complex Auxiliary Information, "International Statistical Review", Vol. 74, 127-147.
2.Założenia i wytyczne do systemu informacji o jednostkach prowadzących działalność gospodarczą na podstawie zawartych w sprawozdaniu miesięcznym DG1 w 2012 r., GUS, Warszawa, 2012.