AUTOREFERAT

ROZPRAWY DOKTORSKIEJ

Wykorzystanie technik imputacyjnych w szacowaniu informacji wynikowych oraz w analizie struktury danych w statystyce przedsiębiorstw

Autor:

Paweł Lańduch

Promotor:

dr hab. Andrzej Młodak, prof. Akademii Kaliskiej im. Prezydenta Stanisława Wojciechowskiego

Kalisz, maj 2023 roku

1. Cele pracy

Problematyka wykorzystania innych źródeł danych niż informacje pochodzące od respondentów, w celu wykorzystania ich w uzyskiwaniu wynikowych danych z badań statystycznych, stanowi jeden z celów statystyki publicznej. W przypadku statystyki przedsiębiorstw problemy z tym związane dotyczą między innymi obciążeń przedsiębiorstw. Chociaż obowiązek sprawozdawczy firm wynika z mocy prawa, jednak problem braków odpowiedzi narasta. Dodatkowo, z uwagi na asymetrię rozkładu cech w populacji przedsiębiorstw, duże firmy włączane są właściwie do każdego badania, zwiększając jego obciążenia administracyjne. Z drugiej strony pojawiają się nowe źródła danych, jak chociażby zbiory typu big-data, które mogłyby być alternatywą dla tradycyjnych źródeł danych, tzn. zbiorów opartych na odpowiedziach respondentów. W przypadku statystyki krótkookresowej przedsiębiorstw dodatkowymi problemami są dynamika populacji przedsiębiorstw oraz koszty prowadzenia badań (z powodu których pomija się mikro–firmy, tzn. te z liczbą osób pracujących poniżej 10). Istotnym elementem badań statystycznych jest ich jakość. Jakość zaś to pojęcie wielowymiarowe. Jeden z najbardziej istotnych aspektów jakościowych stanowi precyzja wyników. Zastosowanie niestatystycznych źródeł danych w otrzymywaniu wyników statystycznych wymaga więc wypracowania metod, które odpowiadałyby kryteriom jakościowym, jeżeli chodzi o wyniki badań.

Integracja badań oraz różnych źródeł danych – w tym rejestrów administracyjnych oraz zbiorów typu big-data – w celu opracowania wyników jest obszarem badawczym, który leży od dawna w polu zainteresowań Europejskiego Systemu Statystycznego. Jednak złotym standardem w badaniach statystycznych pozostaje próba losowa. W niniejszej pracy zajęto się problemem zastosowania imputacji masowej do estymacji wartości globalnych cech w statystycznych badaniach przedsiębiorstw. Imputacja masowa zastosowana została jako metoda integracji danych, która łączy próbę losową z niestatystycznym, nielosowym źródłem danych. Integracja danych z próby probabilistycznej z danymi z próby nielosowej jest nową, wyłaniającą się dziedziną badań w badaniach statystycznych. Chodzi tutaj o przypadek uzyskania danych dla zmiennych będących przedmiotem zainteresowania tylko w próbie nielosowej. Są one natomiast przenoszone do próby losowej za pomocą wspólnych zmiennych występujących w obu źródłach danych. Natomiast dalsza analiza odbywa się w próbie losowej z wykorzystaniem dostępnych

prawdopodobieństw inkluzji pierwszego rzędu. Imputacja masowa, chociaż była już wcześniej stosowana, w kontekście losowania dwustopniowego, nie została jeszcze wystarczająco przebadana jako metoda łączenia próby losowej i nielosowej, tzn. stosowanej w kontekście niniejszej pracy.

Głównym celem pracy jest ocena efektywności zastosowania metod imputacji masowej, bazującej na integracji próby losowej i nielosowej, w badaniu statystycznym przedsiębiorstw.

Główny cel analizy został osiągnięty przez realizację następujących celów szczegółowych:

1. Wyznaczenie wartości globalnych obu analizowanych cech w oparciu o estymator Horvitza-Thompsona przy założeniu, że ich wartości są obserwowane w próbie losowej.
2. Wyznaczenie wartości globalnych obu analizowanych cech w oparciu o estymator Horvitza-Thompsona po wcześniejszej transmisji wartości cech do próby losowej za pomocą imputacji masowej opartej na najbliższym sąsiedztwie. Do zdefiniowania odległości wykorzystana została zmienna pomocnicza, a mianowicie liczba osób pracujących w przedsiębiorstwie. Dawca wybierany jest w ramach danych klas, gdzie klasę stanowi zbiór jednostek, których rodzaj działalności należy do tego samego działu Polskiej Klasyfikacji Działalności.
3. Wyznaczenie wartości globalnych obu analizowanych cech w oparciu o estymator Horvitza-Thompsona po wcześniejszej transmisji wartości cech do próby losowej za pomocą imputacji masowej średnią opartą na najbliższych sąsiadach. Jest to przypadek różniący się od omówionego w punkcie pierwszym, gdyż wartość imputowana stanowi tutaj średnią wartość cechy dla wyznaczonego podzbioru jednostek najbliższych danej. Podobnie jak w pkt. 2, najbliższe sąsiedztwo ograniczone jest podzbioru jednostek o tym samym rodzaju działalności względem działu PKD.
4. Wyznaczenie wartości globalnych obu analizowanych cech w oparciu o estymator Horvitza-Thompsona po wcześniejszej transmisji wartości cech do próby losowej za pomocą imputacji masowej przeprowadzonej z wykorzystaniem modelu lokalnej regresji w zbiorze nielosowym. Modele regresyjne zostały wyznaczone w ramach klas dla danego rodzaju działalności jednostki.
5. Wyznaczenie wartości globalnych obu analizowanych cech w oparciu o estymator Horvitza-Thompsona po wcześniejszej transmisji wartości cech do próby losowej za

pomocą imputacji masowej, przeprowadzonej z wykorzystaniem modelu krzywych składanych. Stosowne modele zostały wyznaczone oddzielnie dla każdej klasy jednostek, przy czym klasę będą stanowić jednostki z tym samym rodzajem działalności, tj. działem klasyfikacji PKD.

Stosownie do postawionych celów sformułowano następującą główną hipotezę badawczą:

Imputacja masowa jako metoda integracji danych, w ujęciu łączenia próby losowej i nielosowej, może być użytecznym narzędziem – wspomagającym lub podstawowym – służącym uzyskaniu podstawowych, szybkich szacunków porównywalnej jakości w porównaniu z badaniami przedsiębiorstw prowadzonymi w oparciu o losową i warstwową próbę, kiedy zakładamy, że wartości analizowanych cech są obserwowane tylko w próbie nielosowej i stamtąd mogą być zaczerpnięte.

Uwzględniając powyższą hipotezę sformułowano hipotezę dodatkową:

Integracja danych z użyciem imputacji masowej może być również wykorzystana w porównywalnym wymiarze dotyczącym jakości szacunków, w schemacie losowania jednostek proporcjonalnym do ich wielkości przy mniejszym rozmiarze próby jednak z zapewnieniem większej reprezentacji jednostek mniejszych. Innymi słowy, w oparciu o metody imputacji oraz zastosowanie schematu losowania proporcjonalnego można by docelowo zmniejszyć próbę lub bez jej zwiększania przejść na niższe poziomy agregacji. Obecnie w badaniu DG-1 prowadzonym przez statystykę publiczną w próbie znajdują się wszystkie duże jednostki (zatrudniające 50 i więcej osób) oraz 10% podmiotów średnich (10-49 pracowników) w każdej warstwie. Używając imputacji masowej można by dla przykładu podnieść próg podziału jednostek na średnie i duże (np. z 50 do 250 pracowników), co prowadziłoby do zmniejszenia wielkości próby przy utrzymaniu wysokiej jakości wyników.

2. Postawiony problem oraz założenia przeprowadzonej analizy

Problemy z nielosowym źródłem danych polegają między innymi na tym, że mechanizm doboru jednostek do niego zazwyczaj nie jest znany. Potraktowanie takiego źródła danych jako losowego może obciążać wyniki. Po drugie, wyznaczenie wskaźnika skłonności udziału jednostki w nim wymaga informacji pomocniczych na poziomie populacji. Imputacja masowa zastosowana w niniejszej pracy będzie przypadkiem integracji próby losowej i nielosowego źródła danych w

ten sposób, że zmienna stanowiąca przedmiot analizy podlega w całości imputacji, tzn. dane dla tej zmiennej są imputowane dla wszystkich rekordów. Próba nielosowa stanowi tutaj zbiór-dawcę (ang. *training set*) z wykorzystaniem którego imputacja masowa zostaje wykonana.

Niech $\mathcal{F}_N = \{(X_i, Y_i): i \in U\}$, gdzie $U = \{1, \dots, N\}$, natomiast $X_i = (X_i^1, \dots, X_i^p)$ oznacza wektor p zmiennych predyktorów (tzn. zmiennych objaśniających), zaś Y – zmienną stanowiącą przedmiot zainteresowania. Zmienna ta będzie podlegać estymacji. Dodatkowo, $\mathcal{F}_N$ jest losową próbą z modelu superpopulacji ξ, a N jest znane. Cel działań stanowi estymacja parametru $Y = \sum_{i=1}^{N} g(Y_i)$ dla pewnej znanej funkcji $g(\cdot)$. Zakłada się, że są dostępne dwa źródła danych: próba probabilistyczna, oznaczona jako próba A oraz próba nielosowa – czyli np. zbiór big data lub „duży" zbiór innego typu, w którym nie można określić prawdopodobieństwa znalezienia się w próbie, jako próba B. W tabl.1 przedstawiono schemat sytuacji wyjściowej. Zbiór O_A zawiera obserwacje z próby A:

$$O_A = \{(d_i = \pi_i^{-1}, X_i): i \in A\},$$

gdzie liczność próby A wynosi $n = |A|$, a także znane są w nim prawdopodobieństwa przynależności do próby $\pi_i = P(i \in A)$.

Zbiór O_B zawiera z kolei informacje z nielosowej próby B:

$$O_B = \{(X_i, Y_i): i \in B\},$$

gdzie liczność próby wynosi $N_B = |B|$.

Tabl. 1. Schematyczne przedstawienie dwóch źródeł danych, które zostały poddane integracji danych.

Typ próby	Numer jednostki w próbie	Wagi $d_i = \pi_i^{-1}$	Zmienne X (ang. *covariates*)	Zmienna Y (ang. *study variable*)
Próba probabilistyczna O_A	1	✓	✓	?
	.	.	.	.
	.	.	.	.
	.	.	.	.

Typ próby	Numer jednostki w próbie	Wagi $d_i = \pi_i^{-1}$	Zmienne X (ang. *covariates*)	Zmienna Y (ang. *study variable*)
	n	✓	✓	?
Próba Big data O_B	1	?	✓	✓
	.	.	.	.
	.	.	.	.
	.	.	.	.
	n	?	✓	✓

Znak ✓ oznacza występowanie danych, a znak ? –brak danych

Źródło: Yang, S., Kim, J. K., Hwang, Y., (2021) Integration of data from probability surveys and big found data for finite population inference using mass imputation, *Survey Methodology, Vol. 47, No. 1, pp. 29-58, Statistics Canada, Catalogue No. 12-001-X*

Założenia do imputacji masowej:

Niech $f(Y|X)$ będzie warunkową funkcją rozkładu zmiennej Y względem X w superpopulacji ζ. Niech $f(X)$ oraz $f(X|\delta_B = 1)$ oznacza funkcję gęstości X w populacji skończonej i w próbie B, odpowiednio, gdzie δ_B jest wskaźnikiem selekcji jednostki do próby. Przyjmuje się następujące dwa założenia.

Założenie nr 1:

Gęstość warunkowego rozkład Y względem X w próbie B jest równa tej w modelu superpopulacji, tzn.:

$$f(Y|X; \delta_B = 1) = f(Y|X)$$

W założeniu nr 1 przyjmuje się, że mechanizm selekcji do próby B można zignorować warunkowo względem wektora zmiennych X. Założenie nr 1 prowadzi także do wniosku, że $P(\delta_B = 1|X, Y) = P(\delta_B = 1|X)$. Równość ta będzie natomiast skutkować tym, że braki w zmiennej objaśnianej są określone mechanizmem MAR (ang. *missing at random*).

Założenie nr 2:

Dla wektora zmiennych objaśniających $X \in \mathbb{R}^p$ istnieją stałe C_l i C_u takie, że

$$C_l \leq f(X)/f(X|\delta_B = 1) \leq C_u$$

z prawdopodobieństwem bliskim jedności (ang. *almost surely*).

Założenie nr 2 może być sformułowane również jako $P(\delta_B = 1|X) > 0$ dla wszystkich wartości X. Założenie nr 2 nie byłoby spełnione, gdyby istniały jednostki, które nigdy nie będą wybrane do próby B. Faktycznie, ocena tego faktu podlega merytorycznej, branżowej ewaluacji w danej dziedzinie badania. Do porównań wyników w poszczególnych, zastosowanych metodach imputacji masowej zostanie użyty estymator bezpośredni Horvitza-Thompsona. Oznacza to, że przyjęto jeżeli w próbie A występuje wartość zmiennej Y, wtedy wartość globalna w populacji wynosi

$$\hat{Y} = \sum_{i \in A} \pi_i^{-1} Y_i.$$

Do oceny wariancji estymatora została użyta formuła Devilla. Jest to użyteczny w praktyce estymator wariancji estymatora, szczególnie w przypadku doboru próby o różnych prawdopodobieństwach inkluzji. Do wyliczenia oszacowań nie są wymagane prawdopodobieństwa inkluzji drugiego rzędu. Wariancja estymatora wyrażona jest wzorem:

$$\widehat{var}(\hat{Y}) = \sum_{i \in s} \frac{c_i}{\pi_i^2} (Y_i - \hat{Y}_i^*)^2,$$

gdzie $\hat{Y}_i^* = \pi_i \frac{\sum_{j \in s} \frac{c_j Y_j}{\pi_j}}{\sum_{j \in s} c_j}$, gdzie z kolei $c_i = (1 - \pi_i) \frac{n}{n-1}$.

3. Analiza empiryczna

Dla potrzeb analizy wygenerowano sztuczną populację przedsiębiorstw. Jako wzorcowa struktura do wygenerowania zbioru został użyty operat ciągłego badania statystycznego prowadzonego od wielu lat przez Główny Urząd Statystyczny. Jest to badanie – Meldunek o działalności gospodarczej, przeprowadzane w cyklu miesięcznym. Stosowne dane gromadzone są od przedsiębiorstw na formularzu o symbolu DG-1. Podmiotowy zakres badania stanowi sektor przedsiębiorstw. Ponieważ sektor przedsiębiorstw jest bardzo szeroki (przebiega w zasadzie całą Polską Klasyfikację Działalności 2007), do analiz wybrano jego bardzo istotną

część, a mianowicie przemysł (do kategorii przemysłu należą następujące rodzaje działalności: górnictwo i wydobywanie, przetwórstwo przemysłowe, wytwarzanie i zaopatrywanie w energię elektryczną, gaz, parę wodną, gorącą wodę i powietrze do układów klimatyzacyjnych, dostawa wody, gospodarowanie ściekami i odpadami oraz działalność związana z rekultywacją). Z wygenerowanej w ten sposób populacji wybrano dwa zbiory danych: próba losowa oraz zbiór niebędący próbą losową w tym sensie, że dobór jednostek w nim zawartych nie jest oparty na schemacie losowania. Próba losowa stanowi warstwową próbę losową z różnymi prawdopodobieństwami włączenia jednostki do próby. Z uwagi na asymetrię rozkładu cech w populacji przedsiębiorstw, został zastosowany schemat losowania proporcjonalny do wielkości jednostki. Warstwami losowania były jednostki drugiego poziomu klasyfikacji działalności gospodarczej PKD 2007, nazwane w niej działami. Jako warstwę losowania przyjmuje się zatem dział PKD. Próba nielosowa zostanie również wybrana warstwowo względem działów klasyfikacji PKD, z tym że wybór jednostki do próby oparty będzie o logistyczny model liniowy w jednej wersji, natomiast w drugiej – także o model logistyczny, ale w wersji nieliniowej. W pracy wybrano do analizy dwie zmienne, które są zawarte na formularzu badania DG-1, mianowicie:

- Sw_1b – Przychody netto ze sprzedaży produktów (wyrobów i usług własnej produkcji) (wiersz pierwszy formularza DG-1),
- Wb_1b – wynagrodzenia brutto osób wykazanych w wierszu 07 (tzn. dla przeciętnej liczby zatrudnionych) w tys. zł (wiersz dziewiąty formularza DG-1).

Do parowania jednostek użyto liczbę pracujących w podmiocie gospodarczym. Liczba ta jest dostępna na poziomie operatu badania DG-1. Po przeniesieniu do próby losowej wartości cech z próby nielosowej za pomocą imputacji masowej wyznaczone zostały wartości globalne w próbie losowej. Uczyniono to w oparciu o estymator Horvitza-Thompsona z wykorzystaniem prawdopodobieństw inkluzji wynikających z przyjętego schematu losowania oraz wyznaczona została jego wariancji. Analizie zostały poddane cztery metody imputacji masowej: dwie nieparametryczne oraz dwie częściowo oparte na parametrach. Użyto następujących metod imputacji masowej (dwie pierwsze są nieparametryczne):

- metodę imputacji masowej opartą na najbliższym sąsiedzie wybranym losowo: dla każdej jednostki w próbie losowej – przy wykorzystaniu wspólnej zmiennej, która występuje w

obu źródłach danych, tzn. liczby pracujących – w próbie nielosowej znajdowany jest podzbiór jednostek, które mają najmniejszy do niej dystans. Do obliczenia odległości między jednostkami wykorzystano metrykę euklidesową. W przypadku, gdy jednostek o najmniejszym dystansie do danej jest więcej niż jedna, finalny wybór jednostki do imputacji następuje losowo. Następnie wartość rozpatrywanej cechy dla owej najbliższej jednostki jest imputowana dla docelowej jednostki z próby losowej. Na podstawie próby losowej dokonywana jest estymacja wartości globalnej estymatorem bezpośrednim Horvitza-Thompsona.

- metodę imputacji średnią opartą na najbliższych sąsiadach: różni się ona od opisanej w poprzednim podpunkcie w ten sposób, że zamiast wybierać losowo daną jednostkę z podzbioru jednostek, dla których odległość od jednostki $i \in A$ jest najmniejsza, wylicza się średnią wartość badanej cechy dla tego zbioru. Następnie ta wyliczona wartość zostaje imputowana we właściwe miejsce.
- metodę imputacji masowej opartą na lokalnej regresji: do wykonania imputacji użyto podejścia określanego jako lokalna regresja lub lokalna regresja wielomianowa. Stanowi to uogólnienie średniej kroczącej i regresji wielomianowej. Wykorzystuje się tutaj ważony model wielomianowy, gdzie waga przydzielana jest na podstawie parametru α $(0 < \alpha \leq 1)$, tzn. w danym punkcie dopasowywania modelu bierze się pod uwagę tylko informacje oddalone od tegoż punktu o mniej niż α. Do wyznaczenia parametru α użyto metody iteracyjnej. Przyjęto pierwszy stopień modelu, tzn. korzystano z modelu liniowego. Zmienną objaśniającą jest zawsze liczba osób pracujących. Do próby losowej przeniesiona zostaje wartość uzyskana z modelu.
- metoda imputacji masowej oparta na krzywych składanych: zakłada się tutaj, że celem podejmowanych działań jest znalezienie takiej funkcji f, która przedstawiałaby zależność zależnej zmiennej Y od zmiennej niezależnej X. Dziedzinę X dzieli się na K rozłącznych przedziałów za pomocą uporządkowanego zbioru punktów nazywanych węzłami. W każdym przedziale szuka się funkcji f za pomocą klasycznych metod regresji. Najczęściej wykorzystywane są tutaj funkcje składane trzeciego rzędu. W pracy przyjęto takie właśnie rząd krzywych, tzn. krzywe trzeciego rzędu bez węzłów wewnętrznych.

Na podstawie 500 powtórzeń symulacji dla obu zmiennych poddanych estymacji oraz dla każdej z warstw – tzn. działów działalności gospodarczej według klasyfikacji PKD 2007 – a także

dla każdej z metod odrębnie dla każdej zmiennej, wyznaczono wartości współczynnika zmienności CV (ang. *coefficient of variation*) wedle wzoru:

$$CV_h = \frac{1}{500} \sum_{i=1}^{500} \frac{\sqrt{var(\hat{Y}_{h,i})}}{\hat{Y}_{h,i}} * 100,$$

gdzie $h = 1, \ldots, 25$.

4. Wyniki analizy

W analizie jakościowej wszystkich metod opartych na imputacji masowej można zauważyć, że wyniki mierzone wskaźnikiem zmienności estymatora kształtują się na poziomie zbliżonym do złotego standardu, tzn. przypadku, kiedy zakładamy, że analizowane zmienne są obserwowane w próbie losowej. Wartości współczynników zmienności dla zastosowanych metod kształtują się w przedziale od 0,2% do 14,9% jeżeli weźmie się pod uwagę wszystkie zakresy dla wszystkich zastosowanych metod oraz dla obu zmiennych w obu wersjach integracji (tzn. generowania próby nielosowej w oparciu o model liniowy i nieliniowy). Natomiast dla złotego standardu rozpiętość, o której tutaj mowa, wynosiła od 1,1% do 11,6%. Na podstawie tych wyników można zauważyć, że przedziały te są porównywalne. W przypadku estymacji cech w oparciu o imputację masową zakres osiągniętych wartości współczynników zaczyna się na poziomie, który w złotym standardzie był nieosiągalny, chociaż w zakresie górnym nieznacznie go przekroczył. Niskie liczności warstw miały wpływ na wyższe wartości współczynnika zmienności. Liczności dla próby losowej przyjęto na podstawie stosowanych w faktycznym badaniu DG-1, tzn. 10% liczności średnich jednostek w warstwie operatu.. Podsumowując przeprowadzone rozważania można zauważyć, że trudno jest wskazać przewagę jednej spośród analizowanych metod imputacji masowej w kontekście poziomu ich jakości wyrażonego poprzez względne współczynniki zmienności estymacji wartości globalnych analizowanych zmiennych. W tabl. 2 przedstawiono wyniki współczynników dla zmiennej Sw_1b w przypadku integracji próby losowej i nielosowej generowanej w wersji liniowej.

Tabl. 2. Wartości współczynników zmienności osiągniętych przy zastosowaniu różnych metod imputacji masowejh w przypadku integracji próby losowej i nielosowego źródła danych generowanego w modelu liniowym – przychody netto ze sprzedaży produktów.

L.p.	Dział PKD (warstwa)	HZ	KNS	NS	LREG	SPL
1	5	3,9	6,6	6,6	6,5	6,3
2	10	1,2	0,9	1,2	0,5	0,6
3	11	7,7	7,4	7,6	4,7	5,4
4	13	3,0	2,5	3,2	1,2	1,8
5	14	3,4	1,9	3,5	0,5	0,5
6	15	5,4	4,1	5,4	1,7	2,9
7	16	2,2	1,6	2,2	0,9	0,9
8	17	4,7	4,7	4,9	3,3	3,1
9	18	3,6	2,9	3,6	1,6	2,4
10	19	4,1	7,7	7,7	8,3	6,8
11	20	3,3	3,0	3,3	2,2	2,4
12	21	11,6	12,6	12,6	5,3	4,0
13	22	1,7	1,4	1,8	1,0	1,1
14	23	2,1	1,7	2,1	0,8	1,6
15	24	2,7	3,1	3,2	2,2	4,1
16	25	1,6	1,1	1,6	0,7	0,5
17	26	2,8	2,7	2,9	1,8	1,2
18	27	2,4	2,3	2,4	1,7	2,0
19	28	2,4	2,1	2,4	1,8	1,4
20	29	3,6	3,7	3,8	2,9	2,1
21	30	5,2	5,2	5,3	2,9	4,1
22	31	1,9	1,4	1,9	1,0	1,1
23	35	1,1	1,1	1,1	1,0	2,2
24	36	5,8	5,1	6,0	2,0	2,8
25	38	3,6	2,6	3,6	0,6	1,0

Uwaga: kolumna HT – zmienne obserwowane w próbie losowej; kolumna KNS – metoda imputacji opartej na średniej opartej na najbliższych sąsiadach; kolumna NS – metoda imputacji masowej oparta na najbliższym sąsiedztwie,; kolumna LREG – metod imputacji masowej opart na modelu lokalnej regresji, oraz kolumna SPL – metoda imputacji masowej oparta na krzywych składanych. Analizowaną zmienną jest zmienna Sw_1b, tj. jej wartość globalna dla poszczególnych domen, gdzie domenę stanowi dział (drugi poziom) klasyfikacji PKD.

5. Podsumowanie

Badania przedsiębiorstw mają swoje unikalne cechy – by wspomnieć chociażby asymetrię rozkładu cech, złożoność jednostki, kompleksowe, techniczne definicje cech, trudno dostępne w innych źródłach danych lub oczekiwanie na szybko dostępne wyniki, który to aspekt może nawet

mieć swoje pierwszeństwo względem jakości. Pomimo tego możliwość wykorzystania źródeł niestatystycznych – chociażby pośrednio i pomocniczo – dawałoby nowe możliwości. Wobec rosnącej liczby i zakresu nowych, niestatystycznych źródeł danych a także wzrastających oczekiwań użytkowników nowe metody integracji danych – w tym oparte na imputacji masowej – powinny być wzięte pod uwagę jako użyteczne narzędzie wspomagania szacunków wartości wynikowych wobec wzrastających oczekiwań użytkowników.

W niniejszej pracy podjęty został problem zastosowania imputacji masowej do estymacji wartości globalnych cech w badaniach statystycznych przedsiębiorstw. Imputację masową zastosowano jako metodę integracji danych. Podejście takie nie jest obecnie używane w badaniach przedsiębiorstw polskiej statystyki publicznej. Zastosowanie imputacji masowej w kontekście integracji próby losowej i nielosowej stanowi nowo wyłaniający się obszar badań. Skorzystanie z takich metod mogłoby wzbogacić możliwości wykorzystania chociażby rejestrów administracyjnych w badaniach statystycznych, analiza których pod tym kątem jest prowadzona.

Niezmiernie istotną cechą populacji przedsiębiorstw małych i mikro firm jest jej liczność i zmienność. Gdyby udało się zaadaptować metody imputacji masowej do integracji niestatycznych źródeł danych, które zawierają takie jednostki, ze zbiorami badań statystycznych, wówczas mogłoby to wzbogacić badania sektora przedsiębiorstw o jednostki małe, których – z uwagi na koszty – w krótkookresowych badaniach przedsiębiorstw się nie uwzględnia.