Andrzej MŁODAK

Imputacja danych w spisach powszechnych[1]

Powszechny Spis Rolny przeprowadzany w br. oraz Narodowy Spis Powszechny Ludności i Mieszkań zaplanowany na rok następny należą do największych przedsięwzięć badawczych i organizacyjnych. Ich realizacja wiąże

[1] Autor wyraża serdeczne podziękowanie specjalistom z Pomocy Technicznej SAS Institute Sp. z o.o. w Warszawie: Łukaszowi Wołonciejowi, Mariuszowi Dzieciątku, Pawłowi Plewce, Piotrowi Burzyńskiemu oraz Katarzynie Biesialskiej, a także Tomaszowi Józefowskiemu z Ośrodka Statystyki Małych Obszarów Urzędu Statystycznego w Poznaniu za cenną pomoc okazaną podczas realizacji opisywanego eksperymentu.

się z unikalnymi wyzwaniami w zakresie metodologii, będącymi konsekwencją nowatorskich założeń i środków, które planuje się zastosować. W ten sposób statystyka publiczna wychodzi naprzeciw oczekiwaniom odbiorców na możliwie wszechstronne i wysokojakościowe dane statystyczne, wiernie odzwierciedlające obecną rzeczywistość społeczno-gospodarczą.

Wyjątkowość metodologiczna obu spisów polega głównie na zastosowaniu nowoczesnej technologii gromadzenia danych. W odróżnieniu od przedsięwzięć tego rodzaju przeprowadzanych w latach ubiegłych, obecnie główny nacisk położony będzie na uzyskanie danych ze źródeł administracyjnych. Dane te staną się podstawą identyfikacji oraz charakterystyki podstawowych cech badanych osób, mieszkań czy gospodarstw rolnych. W przypadku występowania jednostek niefigurujących w stosownych rejestrach lub dla których informacje tam zgromadzone okażą się niewystarczające, konieczne będzie przeprowadzenie celowanego badania uzupełniającego.

Drugie źródło informacji to uzupełniający spis reprezentacyjny, który planuje się przeprowadzić na 15% reprezentatywnych próbach z populacji. Pozyskiwane będą wówczas informacje niewystępujące w urzędowych rejestrach i ewidencjach. W tego typu badaniach należy się poważnie liczyć z możliwością wystąpienia braku danych dla niektórych badanych jednostek. Przyczyn takiego stanu rzeczy bywa wiele, jak np. odmowa udzielenia odpowiedzi rachmistrzowi, nieobecność wszystkich mieszkańców danego lokalu, niekompletne lub nieprecyzyjne odpowiedzi (spowodowane m.in. kłopotami z pamięcią) itp. Dlatego też jednym z celów, które postawiła sobie kierowana przez prof. dra hab. Jana Paradysza podgrupa ds. metod statystyczno-matematycznych jest wypracowanie możliwie efektywnych metod imputacji danych. Pod tym pojęciem rozumie się uzupełnienie brakujących informacji na podstawie innej pomocniczej wiedzy z wykorzystaniem nowoczesnych technik statystycznych. Wprowadzona w ten sposób informacja nazywa się *implantem statystycznym*.

Klasyczna teoria imputacji danych (Kalton, Kasprzyk, 1982, 1986; Rubin, 1987) wyróżnia kilka kategorii operacji imputacyjnych. Ze względu na zakres występowania braków danych wyróżnia się:

— imputację pozycyjną (sytuację, gdy braki danych dotyczą pojedynczych zmiennych w rekordzie);
— imputację kompleksową, gdy trzeba zaimputować wszystkie dane dotyczące konkretnej jednostki statystycznej (np. osoby, gospodarstwa domowego czy gospodarstwa rolnego);
— imputację masową, stosowaną przede wszystkim w badaniach reprezentacyjnych, gdy poprzez imputację dane uzyskane z próby przypisywane są wszystkim jednostkom zbiorowości, które z założenia nie zostały poddane badaniu (a więc nie znalazły się w wylosowanej próbie). Szczegółową charakterystykę typologii metod imputacyjnych zawiera opracowanie T. Piaseckiego i in. (2009).

W artykule przedstawimy koncepcję, rezultaty i problemy powstałe podczas realizacji symulacyjnego eksperymentu obliczeniowego, polegającego na impu-

tacji brakujących danych zbieranych w spisach powszechnych. Pod uwagę wzięto model metodologii spisowej, według którego mają zostać przeprowadzone dwa najbliższe spisy powszechne. Testowano najbardziej prawdopodobną wersję, czyli imputację pozycyjną. Będzie ona dotyczyć tych cech, które nie są dostępne w rejestrach, a dla danej osoby nie udało się zebrać stosownych odpowiedzi (np. z powodu odmowy ich udzielenia lub nieobecności respondenta w momencie badania). Istotne znaczenie w rozpatrywanej metodzie mają dwa elementy — taksonomiczne grupowanie rekordów z dostępnymi danymi na jednorodne skupienia oraz metoda „ruletki statystycznej" opracowana przez prof. dra hab. Bogdana Stefanowicza (2009). W kolejnych paragrafach omówimy poszczególne etapy zastosowanego podejścia.

OGÓLNA CHARAKTERYSTYKA METODY I GRUPOWANIE REKORDÓW ZE ZNANYMI DANYMI

Załóżmy, że przedmiotem analizy jest pewna populacja $U = \{1, 2, ..., n\}$ licząca *n* jednostek (gdzie *n* to liczba naturalna), opisana za pomocą *m* (gdzie *m* też jest liczbą naturalną) zmiennych statystycznych $X_1, X_2, ..., X_m$. Zmienne te mogą być obserwowane na różnych skalach pomiarowych — od nominalnej po ilorazową.

W spisach powszechnych najczęściej spotykane skale to nominalna i porządkowa (np. kraj urodzenia zakodowany jest zazwyczaj w skali nominalnej, zaś grupy wieku — w skali porządkowej). Załóżmy, że dla podzbioru $D \subset U$ liczącego n_1 ($n_1 \in N$, $n_1 < n$) jednostek dostępne są wszystkie gromadzone dane, natomiast w przypadku pozostałych $n_2 = n - n_1$ jednostek należących do zbioru $B = U \setminus D$ informacje są niekompletne, tzn. istnieje podzbiór m_1 ($m_1 \in N$, $m_1 < m$) zmiennych, dla którego nie są dostępne żadne informacje odnośnie rekordów należących do zbioru *B*. Dla rozważań teoretycznych, bez straty ogólności można założyć, że występuje tu brak danych z zakresu zmiennych $X_{m_2+1}, X_{m_2+2}, ..., X_m$, przy czym $m_2 = m - m_1$. Zbiór *D* nazwiemy *zbiorem dawców* danych, zaś zbiór *B* — *zbiorem biorców*.

W praktyce imputacja może być uznana za efektywną, gdy liczba biorców (n_2) jest znacznie mniejsza od całkowitej liczebności populacji (*n*). Długoletnie doświadczenie w zakresie przeprowadzania różnorodnych badań może skłaniać do oceny, że przeciętny odsetek rekordów z brakującymi danymi kształtuje się na poziomie ok. 20%, choć okazjonalnie może być mniejszy. Spostrzeżenie to zostanie wykorzystane podczas symulacji empirycznej.

Pierwszy etap eksperymentu stanowiło pogrupowanie zbioru dawców na wewnętrznie jednorodne podzbiory, które odzwierciedlać będą określone grupy informacyjne. Metodologia analizy skupień (Młodak, 2006 a) daje do wyboru dwa możliwe fundamenty algorytmu grupowania: macierz danych (poddanych ewentualnie normalizacji) bądź też macierz odległości pomiędzy badanymi obiektami skonstruowaną na podstawie tychże danych.

Ze względu na to, że w naszym modelu zmienne obserwowane są głównie na skalach nominalnej i porządkowej — gdzie wykonywanie operacji arytmetycznych, takich jak uśrednianie, jest z praktycznego punktu widzenia bezcelowe — zdecydowano się na wybór drugiej opcji. W takim przypadku istotny staje się także sposób obliczania owej odległości, uwzględniający ów charakter zgromadzonych obserwacji. Efektywnym rozwiązaniem wydaje się być tu odległość Gowera. Przyjmując, że rekord *i* ze zbioru *D* można przedstawić jako $\gamma_i = (x_{i1}, x_{i2,} ..., x_{im_2})$, odległość tę definiujemy jako:

$$d_G(\gamma_i, \gamma_k) = 1 - \delta_G(\gamma_i, \gamma_k) \tag{1}$$

gdzie $\delta_G(\gamma_i, \gamma_k) = \sum_{j=1}^{m} w_j \rho_{ikj} / \sum_{j=1}^{m} w_j$, przy czym w_j to waga przyporządkowana zmiennej X_j, natomiast ρ_{ikj} oznacza miarę podobieństwa Gowera określoną następująco:
— jeśli zmienna X_j jest mierzona na skali nominalnej, to:

$$\rho_{ikj} = \begin{cases} 1, & \text{gdy } x_{ij} = x_{kj} \\ 0, & \text{gdy } x_{ij} \neq x_{kj} \end{cases}$$

— jeżeli zaś X_j jest mierzona na skali porządkowej, przedziałowej lub ilorazowej, to:

$$\rho_{ikj} = 1 - |x_{ij} - x_{kj}|$$

dla każdego $j = 1, 2, ..., m$ oraz $i, k \in D,\ i \neq k$.

W ten sposób każda zmienna traktowana jest zgodnie ze swym charakterem, a uzyskana miara odzwierciedla praktyczny sens odległości. Nie wprowadzamy natomiast żadnego specjalnego ważenia zmiennych, przyjmując że wszystkie wagi są równe 1.

Nie mniejsze znaczenie od sposobu pomiaru odległości ma także metoda grupowania. Przyjmujemy tutaj algorytm elastycznego beta zaproponowany przez G. N. Lance'a oraz W. T. Williamsa (1967), będący przykładem rekurencyjnej metody hierarchicznej przebiegającej na drodze taksonomii wrocławskiej (Florek i in., 1951), w których odległość skupień rzędu *u* definiuje się za pomocą odległości skupień poziomu $u - 1$, $u = 2, 3, ..., n_1$. W tej konkretnej sytuacji (oznaczając przez $d_{p_u}(P_{uh}, P_{ug})$ odległość skupień P_{uh} i P_{ug} na poziomie *u*)

procedurę rozpoczynamy od zestawu trywialnych skupień jednoelementowych (tzn. na poziomie $u = 1$ każdy rekord traktujemy jak odrębne skupienie, a ich odległość dana jest wzorem (1), a następnie na każdym poziomie $u = 2, 3, ..., n_1$ łączymy skupienia P_{ug} i P_{uk}, które minimalizują dystans określony wzorem:

$$d_{p_{u+1}}(P_{[u+1]h}, P_{ug} \cup P_{uk}) = (d_{p_u}(P_{uh}, P_{ug}) + d_{p_u}(P_{uh}, P_{uk})) \cdot \frac{1-b}{2} + b \cdot d_{p_u}(P_{ug}, P_{uk})$$

gdzie $g, h, k = 1, 2, ..., p_u$, p_u dla $h \neq g, k$, jest liczbą skupień na poziomie u, zaś $u = 1, 2, 3, \ldots, n_1$.

Parametr b jest ustalany rozmaicie, najczęściej $b := -0{,}25$. G. Milligan (1989) proponuje, aby w sytuacji gdy wśród danych występują obserwacje odstające stosować mniejszy współczynnik b, np. $b := -0{,}5$. Pozwala to zwiększyć odporność na scalanie skupień zawierających takie właśnie obserwacje. W naszej sytuacji nie można wykluczyć występowania nietypowych informacji, ale — zważywszy na przedmiot dociekań — nie należy się raczej spodziewać zbyt dużej ich liczby, dlatego też przyjmujemy $b := -0{,}3$. Szczególnie użyteczną z punktu widzenia naszych rozważań zaletą metody elastycznego beta jest to, iż — w przeciwieństwie do wielu innych — na żadnym etapie nie sięga ona do niedozwolonych na niektórych skalach pomiarowych operacji liczbowych.

Ostatni krok grupowania stanowi wskazanie progu łączenia, to znaczy wartości odległości skupień łączonych na każdym etapie, po przekroczeniu której realizacja algorytmu zostanie zakończona, a struktura skupień uzyskana bezpośrednio przed tym momentem będzie uznana za ostateczną. W przeprowadzonym badaniu za granicę taką przyjęto:

$$q = \text{med}(d^*) + 2{,}5 \cdot \text{mad}(d^*) \tag{2}$$

przy czym $d^* = (d_1^*, d_2^*, ..., d_{n_1}^*)$ to wektor minimalnych odległości skupień na kolejnych etapach grupowania, med — to jego mediana, zaś $\text{mad}(d^*) = \text{med}_{u=1, 2, ..., n_1}(|d_u^* - \text{med}(d^*)|)$ — jej medianowe odchylenie bezwzględne. Za wyborem progu (2) przemawiała jego odporność na obserwacje odstające oraz tendencja do redukcji liczby skupień, co jest ważne dla sprawności obliczeń.

PRZEBIEG IMPUTACJI Z WYKORZYSTANIEM RULETKI STATYSTYCZNEJ

Podzielony na wewnętrznie jednorodne skupienia zbiór dawców danych stanowi efektywne źródło informacji niezbędnej do imputacji dla biorców. Efek-

tywność ta polega na tym, że zamiast analizować cały (częstokroć bardzo obszerny) zbiór dawców wystarczy wskazać obecnie tylko tę ich grupę, która jest najbliższa danemu biorcy. Skraca to znacznie czas i koszty obliczeń. Aby tak się stało, dla każdej grupy dawców warto wyznaczyć jej „reprezentanta", tzn. rekord, który okazuje się być najbardziej „typowy" dla tejże grupy. Bardzo dobrym rozwiązaniem mogłaby być tutaj tzw. mediana Webera, czyli wielowymiarowe uogólnienie klasycznego pojęcia mediany (Młodak, 2006 a, 2006 b, 2009). Chodzi tu o wektor, który minimalizuje sumę euklidesowych odległości od danych punktów reprezentujących rozpatrywane obiekty, a więc znajduje się niejako „pośrodku" nich, ale jest jednocześnie uodporniony na występowanie obserwacji odstających. W naszej konkretnej sytuacji jej bezpośrednie zastosowanie okazuje się jednak niemożliwe z dwóch powodów:

— konstrukcja mediany Webera opiera się na odległości euklidesowej, która ze względu na nominalny lub porządkowy charakter wielu zmiennych jest nieodpowiednia;
— generowany (i sztuczny skądinąd) wektor ma z reguły współrzędne mierzone na skali ilorazowej, tymczasem w rozpatrywanej sytuacji chodzi o informację odzwierciedlającą zakres wartości poszczególnych zmiennych.

Biorąc pod uwagę owe stwierdzenia uznano, że najbliższym, a jednocześnie efektywnym rozwiązaniem w tych warunkach będzie odnalezienie w każdej grupie dawców $P \subset D$ rekordu, którego suma odległości Gowera od pozostałych rekordów z tejże grupy jest najmniejsza, tzn. takiego wektora γ_p, że $\sum_{i \in p} \sum d_G(\gamma_i, \gamma_p) = \min_{k \in p} \sum_{i \in p} d_G(\gamma_i, \gamma_k)$. Wektor taki zostanie uznany za reprezentanta grupy.

Następnie dla każdego biorcy określamy reprezentanta grupy, który jest mu najbliższy, z uwzględnieniem braków danych, tzn. wyznaczamy reprezentanta takiego, że jego odległość Gowera (ograniczona do znanych danych) od danego rekordu — biorcy jest najmniejsza. Ujmując rzecz bardziej formalnie, dla każdego $i \in B$ znajdujemy takie $P^* \subset D$ i należące do uzyskanego układu skupień, że $\tilde{d}_G(\gamma_i, \gamma_{p^*}) = \min_p \tilde{d}_G(\gamma_p, \gamma_i)$, gdzie $\tilde{d}_G(\gamma_p, \gamma_i) = 1 - \tilde{\delta}_G(\gamma_i, \gamma_k)$, zaś $\tilde{\delta}_G(\gamma_p, \gamma_k) = \sum_{j=1}^{m_2} w_j \rho_{pkj} / \sum_{j=1}^{m_2} w_j$. Pozostałe założenia i oznaczenia są takie same, jak w formule (1).

Mając ustaloną grupę najbliższą rozpatrywanemu biorcy, trzeba obecnie uzupełnić brakujące dane informacjami pochodzącymi od członków tejże grupy. Wykonuje się to stosując metodę „ruletki statystycznej", opartej na losowym wyborze imputowanych danych. Jej schemat dla rekordu $i \in B$ przedstawia się następująco:

1) zbudowanie koła ruletki:
— przyjmujemy, że hipotetyczne koło ruletki ma długość 1,

— zakładamy, że imputowana zmienna X_j przyjmuje r różnych wartości $a_{j1}, a_{j2}, ..., a_{jr}$; obwód koła ruletki dzielimy na r ($r \in N$) odcinków — po jednym dla każdej wartości tejże zmiennej. Długość każdego odcinka (t_{js}) ustala się jako częstość pojawiania się obserwacji a_{js} w rozkładzie zmiennej X_j wśród członków grupy P_i optymalnej dla rekordu *i*, czyli $t_{js} = \frac{f_{jP_i}(a_{js})}{|P_i|}$, przy czym $f_{jP_i}(a_{js})$ oznacza liczbę obserwacji a_{js} dla X_j w grupie P_i, zaś $|P_i|$ to liczebność owej grupy. Mamy oczywiście $0 \le t_{js} \le 1$ dla $s = 1, 2, ..., r$ oraz $\sum_{s=1}^{r} t_{js} = 1$,

— początek ruletki ustalamy jako punkt 0, a następnie wyznaczamy początek q_{js} s-tego odcinka. Czynimy to w sposób następujący: $q_{j1} = 0$, $q_{js} = \sum_{z=2}^{s} t_{j(z-1)}$, $s = 2, 3, ..., r$;

2) uruchomienie koła ruletki:

— ze zbioru liczb losowych wybieramy liczbę losową λ należącą do przedziału [0,1]. Od strony informatycznej najprościej można to uczynić uruchamiając generator liczb losowych z rozkładu jednostajnego na [0,1]. Liczba λ odzwierciedla odległość od punktu 0 na obwodzie ruletki, a jednocześnie jednoznacznie wskazuje odcinek na tym „kole", z którego pochodzić będzie imputowana dana,

— niech więc $s \in \{1, 2, ..., r\}$ będzie takie, że $q_{js} \le \lambda \le q_{j(s+1)}$; wówczas gdy $\lambda < (q_{j(s+1)} - q_{js})/2$ podstawiamy implant $x_{ij} := a_{js}$, zaś w przeciwnym razie $x_{ij} := a_{j(s+1)}$.

Operację konstrukcji i uruchamiania ruletki powtarzamy dla każdego $j = m_2 + 1, m_2 + 2, ..., m$ oraz każdego $i \in B$.

EKSPERYMENT SYMULACYJNY

Celem zweryfikowania skuteczności metodologii przeprowadzono eksperyment, wykorzystując dane o osobach zgromadzone podczas Narodowego Spisu Powszechnego Ludności i Mieszkań przeprowadzonego w 2002 r. Jedną z najważniejszych kwestii, przed jaką stanął wykonawca badania, był wybór przestrzennego poziomu rozpatrywanych danych. Poczynione próby wykazały, że odfiltrowanie potencjalnych biorców ze zbioru głównego oraz konstrukcja macierzy odległości dla potencjalnych dawców przy większych rozmiarowo bazach napotyka na poważne trudności, związane z niską wydolnością obliczeniową

środowiska, którym się posługiwano (SAS Enterprise Guide 4.1, później 4.2)[2]. Dlatego też zdecydowano się na analizę na poziomie gminy, gdzie mogła ona zostać przeprowadzona w miarę sprawnie.

W tym celu z dostępnych w Hurtowni Danych SAS (funkcjonującej w sieci GUS) baz spisowych dla województw mazowieckiego i wielkopolskiego wybrano gminę Gołuchów w powiecie pleszewskim w Wielkopolsce. Powodem tego kroku był fakt, że należało się tam spodziewać dużej różnorodności mieszkańców, co wynika zarówno z położenia geograficznego gminy (sąsiedztwo z dużym Kaliszem) oraz jej walorów turystycznych (obiekty muzealne, parkowo-leśne i rekreacyjne), które mogą na stałe lub na dłuższy okres przyciągać rozmaite grupy ludności, a wyraźniejsze zróżnicowanie bazy danych pozwala lepiej wychwycić ewentualne ułomności zastosowanej metodologii imputacyjnej.

Z bazy danych gminy, liczącej łącznie 9630 rekordów, wybrano metodą losowania prostego bez powtórzeń, z jednakowym prawdopodobieństwem wyboru, trzy próbki: 5%, 10%, 20%, które w dalszej części doświadczenia uznano za zbiory biorców. Następnie próbki te odfiltrowano ze zbioru głównego, uzyskując trzy odpowiednie zestawy dawców. Zestaw zmiennych, które obejmowała owa baza danych, składał się z następujących informacji otrzymywanych od respondentów:

a) ogólne kategorie ludności:
   — charakter przebywania/nieobecności,
   — czy respondent jest zameldowany na pobyt stały,
   — czy badana osoba jest rezydentem,
   — płeć,
   — wiek (w latach),
   — pięcioletnie grupy wieku,
   — ekonomiczne grupy wieku,
   — grupy wieku (z wyszczególnieniem niektórych pojedynczych roczników),
   — data urodzenia — półrocze,
   — data urodzenia — dzień,
   — data urodzenia — miesiąc,
   — data urodzenia — rok;
b) stan cywilny:
   — stan cywilny formalnoprawny,

[2] Największy problem stanowi przetwarzanie macierzy odległości. Już bowiem na poziomie gminy — z powodu wymogów obliczeniowych i właściwości środowiska informatycznego — konieczne było zmagazynowanie na dysku lub w pamięci tablicy rozmiaru np. 9147×9147, czyli mającej 83667609 elementów. W przypadku powiatu, który przeciętnie liczy ok. 100 tys. rekordów, miejsca potrzeba nawet kilkunastokrotnie więcej, co podczas naszego eksperymentu okazało się niewykonalne. Oczywiście, w praktyce ideałem byłaby możliwość przetworzenia macierzy tego rodzaju dla całej Polski (ok. 38 mln×38 mln elementów).

— faktyczny stan cywilny,
— data zawarcia związku małżeńskiego — miesiąc,
— data zawarcia związku małżeńskiego — rok,
— wiek w chwili zawarcia związku małżeńskiego;

c) wykształcenie:
— poziom wykształcenia,
— kontynuowanie nauki;

d) niepełnosprawni:
— ograniczenie wykonywania czynności,
— orzeczenie ustalające niezdolność do pracy,
— kwalifikacja niezdolności — grupa inwalidzka,
— identyfikator niepełnosprawności;

e) kraj urodzenia, obywatelstwo:
— kraj urodzenia,
— obywatelstwo,
— kraj obywatelstwa I,
— grupa kraju obywatelstwa I,
— kraj obywatelstwa II,
— grupa kraju obywatelstwa II,
— rodzaje obywatelstwa (polskie/niepolskie, jedno/podwójne),
— kategorie ludności według obywatelstwa i kraju urodzenia;

f) narodowość, język:
— narodowość (polska/niepolska),
— język używany w domu (polski/inny),
— symbol języka niepolskiego I,
— symbol języka niepolskiego II,
— narodowość — dokładnie,
— język domowy — ogólnie (polski i niepolski — jeden, dwa lub więcej),
— język domowy — dokładnie,
— narodowość ojca,
— narodowość matki.

Wiek respondentów w latach badania i wiek w chwili zawarcia związku małżeńskiego to zmienne mierzone na skali ilorazowej. Pięcioletnie oraz specjalne grupy wieku i poziom wykształcenia charakteryzują się porządkową skalą pomiaru. Pozostałe zmienne mają charakter nominalny (czasem dychotomiczny, w innym razie polichotomiczny). Ten fakt został uwzględniony w obliczeniach (wzór 1). Łącznie rozpatrujemy więc 40 zmiennych.

W drodze analizy informacji o źródłach administracyjnych, szczególnie bazy Powszechnego Elektronicznego Systemu Ewidencji Ludności (PESEL) oraz zasobów ZUS — System Emerytalno-Rentowy — przyjęto, że podczas spisu dostępne będą następujące informacje: zmienne z grupy „ogólne kategorie ludności" — wszystkie, „stan cywilny" — wszystkie, z wyjątkiem faktycznego stanu cywilnego, „niepełnosprawni" — orzeczenie ustalające niezdolność do pracy oraz kwalifikacja niezdolności — grupa inwalidzka. Z grupy „kraj urodzenia, obywatelstwo" można spodziewać się informacji o obywatelstwie, kraju obywatelstwa (I i II) oraz jego grupie, a także o rodzajach obywatelstwa. Zakłada się, że pozostałe informacje (14 zmiennych) będą niedostępne i należy dokonać ich imputacji. Dlatego też usunięto je ze zbiorów hipotetycznych biorców[3].

W wyniku grupowania zbiorów dawców, pozostałych po odfiltrowaniu biorców z prób 5%, 10% i 20% (liczących odpowiednio: 9147, 8666 oraz 7708 rekordów), otrzymano następujące liczby skupień: 1114, 1086 i 1009. Wydają się one dość duże, ale są to najmniejsze liczby, jakie dało się uzyskać w sposób endogeniczny, tzn. ustalając próg grupowania jako statystykę elementów macierzy odległości. Najlepszy rezultat uzyskano stosując w tym zakresie podejście (2). Podczas wyznaczania najbliższej grupy dawców dla danego rekordu — biorcy — szukano dawcy z minimalną odległością od owego biorcy przede wszystkim wśród tych dawców — reprezentantów, dla których formalnoprawny stan cywilny, orzeczenie ustalające niezdolność do pracy oraz kwalifikacja niepełnosprawności były identyczne z odpowiednią informacją dla owego biorcy. Ważność owej tożsamości przyjęto w takiej właśnie kolejności.

Innymi słowy, w pierwszym rzędzie szukano najbliższego reprezentanta wśród takich, dla których wszystkie trzy wspomniane cechy były zgodne z cechami biorcy. Jeśli takich nie było, to ograniczano zgodność do dwóch pierwszych, a jeśli i to zawiodło — to tylko do pierwszej. Dzięki temu uzyskano znaczną poprawę jakości imputacji. Aby ocenić jakość finalnie otrzymanych rezultatów, dokonano stosownej analizy porównawczej danych uzyskanych na drodze imputacji dla biorców z informacjami faktycznie figurującymi w istniejącej bazie. Analiza ta przebiegała w dwu kierunkach badawczych:

1) dla każdego rekordu — biorcy wyznaczono odległość pomiędzy jego wersją zawierającą faktyczne dane dla imputowanych zmiennych zgromadzone podczas spisu a opcją z implantami; odległość ta jest w istocie rzeczy nieuśrednioną odległością Gowera, tzn. zastosowanie ma tu formuła (1), gdzie $\delta_G(\gamma_i, \gamma_{i'}) = \sum_{j=m_2+1}^{m} \rho_{ii'j,}$ gdzie: γ_i to wektor z faktycznymi danymi dla rekordu i, zaś $\gamma_{i'}$ — wektor z danymi imputowanymi dla tego rekordu, $i \in B$;

[3] Postępowanie takie nazywa się czasem *amputacją* danych.

2) dla imputowanych zmiennych wyznaczono statystykę agregacyjną (tzn. odpowiednie wielkości ogółem dla całej populacji), a następnie różnice pomiędzy uzyskanymi strukturami porównano przy pomocy testu *t*-Studenta, weryfikującego hipotezę o ich nieistotności. W celu zwiększenia efektywności porównania użyto w tym kontekście także testu znaków oraz testu znakowanych rang Wilcoxona.

Tabl. 1 zawiera wartości podstawowej statystyki opisowej dla wektora odległości biorców z implantami od stanu faktycznego we wszystkich trzech rozpatrywanych wariantach zbiorów biorców i dawców.

TABL. 1. STATYSTYKA OPISOWA DLA DYSTANSU POMIĘDZY IMPUTACJĄ A STANEM FAKTYCZNYM

Wielkość próby w %	Średnia arytmetyczna	Odchylenie standardowe	Pierwszy kwartyl	Mediana	Trzeci kwartyl	Minimum	Maksimum	Współczynnik zmienności w %
5	2,7490	2,7557	1,0000	2,0000	4,0000	0,0000	14,0000	100,2446
10	2,6501	2,6235	1,0000	2,0000	4,0000	0,0000	12,0000	98,9970
20	2,5457	2,6419	0,0000	2,0000	4,0000	0,0000	12,0000	103,7781

Ź r ó d ł o: opracowanie własne z zastosowaniem programu SAS Enterprise Guide 4.1 i 4.2.

Tabl. 1 uwidacznia nam, że zróżnicowanie rezultatów jest dość wyraźne. Warto jednak zauważyć kilka ciekawych zjawisk:

— przeciętne odległości są dość niskie, a zatem precyzję imputacji można uznać za zadowalającą;

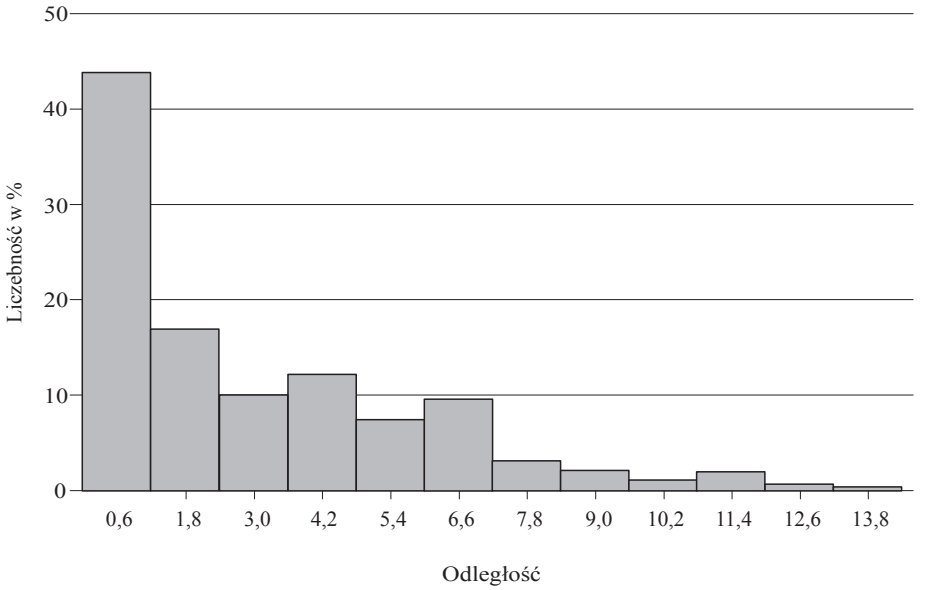
— w miarę wzrostu liczebności zbioru biorców średnia odległość maleje; może wydawać się to trochę wbrew logice, która podpowiada, że im więcej braków danych, tym większą trudność winno sprawiać ich uzupełnienie. Wydaje się to świadczyć o tym, że mniejsze zbiory dawców stają się bardziej jednorodne wewnętrznie, co potwierdzałoby hipotezę o wpływie uwarunkowań istniejących w rozpatrywanej gminie na strukturę jej ludności;

— zróżnicowanie wyników okazuje się dość wyraźne (przy czym współczynnik zmienności nie wykazuje monotoniczności — najniższy okazał się dla próby 10%), aczkolwiek oba nieparzyste kwartyle są dość podobne; może to sugerować, iż największy wpływ na zmienność wywierają pojedyncze obserwacje odstające (tzn. skrajnie duże odległości pomiędzy implantami a rzeczywistymi wartościami imputowanych zmiennych), zlokalizowane w pobliżu wartości maksymalnych.

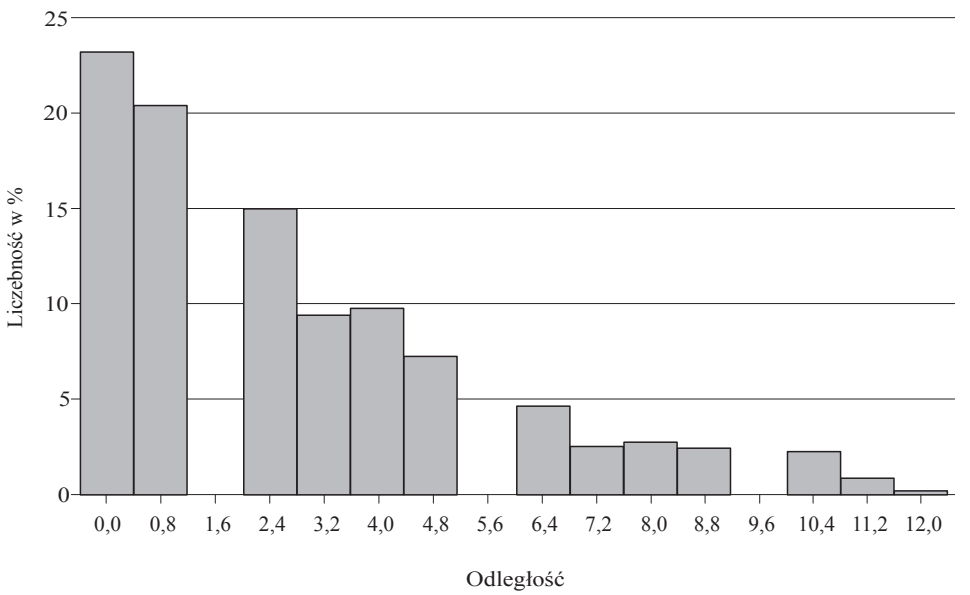
Trzy histogramy ukazują rozkład analizowanych wektorów odległości.

HISTOGRAMY CZĘSTOŚCI ODLEGŁOŚCI DANYCH IMPUTOWANYCH I SPISOWYCH

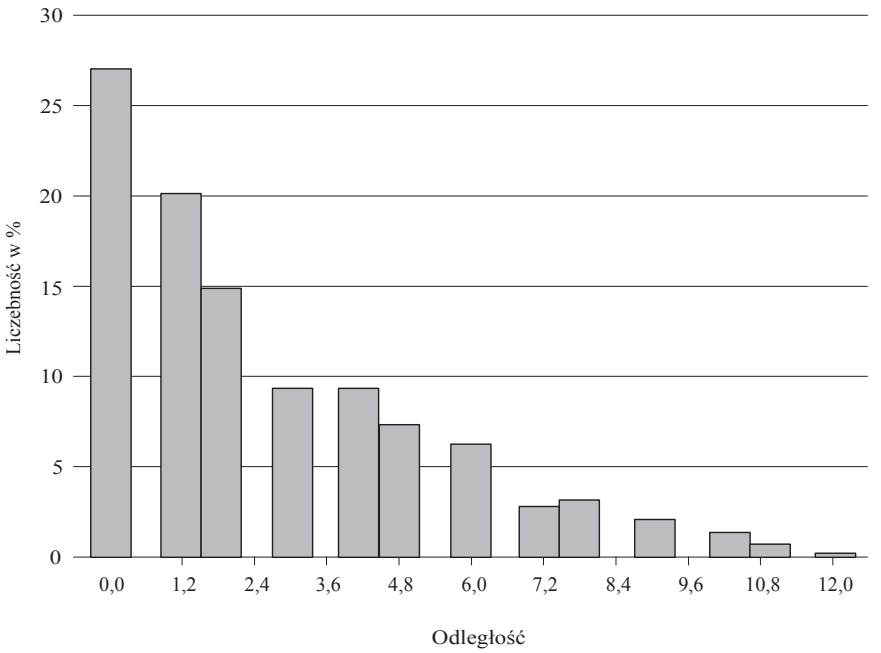
a) próba 5%



b) próba 10%



c) próba 20%



We wszystkich obserwacjach większość wyników koncentruje się wokół zera, dając wyraźną prawostronną asymetrię rozkładów. Rezultaty imputacji można uznać więc za wysoce poprawne.

Drugi etap analizy porównawczej stanowi badanie podobieństwa struktur agregacyjnych niektórych zmiennych. Przedstawimy przykład takiej analizy dla najciekawszych zmiennych polichotomicznych (tabl. 2).

TABL. 2. PORÓWNANIE STRUKTUR W ZAKRESIE FAKTYCZNEGO STANU CYWILNEGO W %

Stan cywilny	Wielkość próby					
	5 %		10 %		20 %	
	spis	imputacja	spis	imputacja	spis	imputacja
Nie dotyczy (osoby w wieku 0—14 lat)	22,20	22,20	20,87	20,87	22,64	22,64
Kawaler, panna	24,48	25,10	22,12	23,68	22,07	21,24
Żonaty, zamężna pozostający w małżeństwie	42,95	43,36	48,60	47,98	46,68	46,00
Żonaty, zamężna pozostający w związku partnerskim	0,00	0,62	0,10	0,31	0,16	1,30
Wdowiec, wdowa	8,51	8,30	6,33	5,92	7,11	7,42
Rozwiedziony, rozwiedziona	1,24	0,41	1,45	1,14	1,14	0,93
Separowany, separowana prawnie	0,00	0,00	0,00	0,00	0,00	0,00
Separowany, separowana — żonaty, zamężna niepozostający w małżeństwie	0,62	0,00	0,52	0,10	0,21	0,47
Nieustalony	0,00	0,00	0,00	0,00	0,00	0,00

Ź r ó d ł o: opracowanie własne z zastosowaniem programu SAS Enterprise Guide 4.1 i 4.2.

Struktura we wszystkich trzech przypadkach okazała się zatem bardzo zbliżona, co potwierdza dodatkowa analiza. Wartość testu *t*-Studenta przyjęła tutaj każdorazowo wartość 0, co dało poziom istotności *ex post* rzędu 1,0000. A zatem hipoteza o braku istotności różnic (a formalnie rzecz ujmując — o zerowej średniej wartości bezwzględnych różnic poszczególnych ich elementów) staje się właściwie pewną tezą. Podobnie pozytywne wyniki (dla typowych rozsądnych poziomów istotności *ex ante* — np. 0,02, 0,05, 0,10) dają testy znaków (poziom istotności *ex post* wynosi: próba 5% — 0,4531, próba 10% — 0,6875, próba 20% — 1,0000) oraz test Wilcoxona (odpowiednio: 0,3750, 0,5625 i 1,0000).

Struktury ludności według wykształcenia uzyskane drogą imputacji oraz z badania spisowego uwidacznia tabl. 3.

TABL. 3. PORÓWNANIE STRUKTUR W ZAKRESIE WYKSZTAŁCENIA W %

Wykształcenie	Wielkość próby					
	5%		10%		20%	
	spis	imputacja	spis	imputacja	spis	imputacja
Nie dotyczy (osoby w wieku 0—12 lat)	18,26	18,05	16,61	16,61	18,43	18,43
Wyższe ze stopniem naukowym co najmniej doktora	0,21	0,00	0,00	0,00	0,05	0,00
Wyższe z tytułem magistra, lekarza lub równorzędnym	3,32	3,94	2,70	4,05	2,49	3,63
Wyższe z tytułem inżyniera, licencjata	0,83	1,04	0,93	1,87	0,93	1,82
Policealne z maturą	1,24	0,41	1,35	2,28	1,30	1,77
Policealne bez matury	0,41	0,00	0,10	0,10	0,10	0,16
Średnie zawodowe z maturą	11,00	7,68	8,20	10,28	9,55	10,02
Średnie zawodowe bez matury	3,32	4,77	3,01	6,13	3,58	4,15
Średnie ogólnokształcące z maturą	1,45	4,36	3,01	2,28	2,18	3,22
Średnie ogólnokształcące bez matury	1,04	1,04	1,35	0,52	0,93	0,99
Zasadnicze zawodowe	26,76	26,56	30,32	24,51	27,57	25,23
Podstawowe ukończone	27,18	24,69	28,66	25,96	28,92	23,78
Podstawowe nieukończone i bez wykształcenia szkolnego	4,98	7,47	3,74	5,40	3,95	6,80
Nieustalone	0,00	0,00	0,00	0,00	0,00	0,00

Ź r ó d ł o: opracowanie własne z zastosowaniem programu SAS Enterprise Guide 4.1 i 4.2.

I tutaj, optycznie rzecz ujmując, struktury także są bardzo podobne, co znajduje potwierdzenie w kształtowaniu się odpowiedniej statystyki testowej. Wartość testu *t*-Studenta wyniosła każdorazowo zero (z poziomem istotności *ex post* równym 1,0000). Pozostałe testy również dały wysoce pozytywne efekty. Poziom istotności *ex post* dla testu znaków osiągnął wielkości 0,7744, 0,7539 i 0,1460, zaś dla statystyki Wilcoxona ta charakterystyka jej rozkładu przyjęła wartości odpowiednio: 0,8350, 0,5566 oraz 0,2412. I jeszcze tabl. 4.

TABL. 4. PORÓWNANIE STRUKTUR W ZAKRESIE OGRANICZENIA WYKONYWANIA CZYNNOŚCI ŻYCIOWYCH W %

Wyszczególnienie	Wielkość próby					
	5%		10%		20%	
	spis	imputacja	spis	imputacja	spis	imputacja
Całkowicie ograniczona zdolność do wykonywania czynności podstawowych	1,45	2,90	0,83	1,87	1,04	1,77
Poważnie ograniczona zdolność do wykonywania czynności podstawowych	6,85	7,26	6,33	6,75	5,09	5,61
Brak ograniczeń (zdolni do wykonywania czynności podstawowych)	91,70	89,83	92,83	91,38	93,87	92,63
Nie ustalono	0,00	0,00	0,00	0,00	0,00	0,00

Ź r ó d ł o: opracowanie własne z zastosowaniem programu SAS Enterprise Guide 4.1 i 4.2.

Widzimy tu dużą zgodność uzyskanych struktur agregacyjnych. Każdy z rozpatrywanych testów we wszystkich opcjach daje poziom istotności *ex post* w optymalnej wysokości (1,0000). Zgodność struktur jest więc statystycznie pełna. W przypadku pozostałych imputowanych zmiennych konkluzje okazały się identyczne.

Wnioski

Przedstawiona metoda w opisywanych uwarunkowaniach okazała się bardzo skutecznym narzędziem imputacji brakujących danych. Cząstkowe błędy imputacji w większości przypadków są niewielkie, choć zdarzają się nieliczne rekordy, dla których odchylenie niektórych implantów od stanu faktycznego można uznać za odstające. Wydaje się to jednak być zjawiskiem raczej marginalnym, które dałoby się w praktyce jeszcze bardziej zniwelować stosując imputację wstępną, czyli tzw. preimputację, tzn. dedukcyjnie wykluczając pewne zakresy wartości imputowanych danych, które teoretycznie są raczej niemożliwe (np. wyższe wykształcenie osoby w wieku poniżej 16 lat). Metoda ta nie wykazuje także tendencji do tworzenia dużej liczby jednakowych rekordów dla biorców, czego najczęściej obawiają się implantatorzy informacji statystycznych. Wynika to z dwóch zasadniczych elementów zastosowanego algorytmu: po pierwsze, „ruletka statystyczna" jest „uruchamiana" dla każdej zmiennej z osobna, nie ma więc zależności stochastycznej pomiędzy kolejnymi implantami. Po drugie, nałożone w trakcie przyporządkowywania biorcom optymalnych grup dawców określone warunki preferencyjne wykluczyły w znacznym stopniu „nadprodukcję" pewnych wartości kosztem innych (np. bez ich zastosowania implanty faktycznego stanu cywilnego dla osób pozostających formalnie w związkach małżeńskich zbyt często lokowały te osoby w stanie wolnym).

Na zakończenie warto wreszcie zauważyć, że opisane podejście jest uniwersalne i może być zastosowane w różnorodnych badaniach statystycznych (np. również w Powszechnym Spisie Rolnym), gdyż korzysta z ogólnych procedur,

a ewentualne warunki preferencyjne mogą być zawsze dostosowane do charakteru i metodologii prowadzonego badania.

dr hab. Andrzej Młodak — *US w Poznaniu*

LITERATURA

Florek K., Łukaszewicz J., Perkal J., Steinhaus H., Zubrzycki S. (1951), *Taksonomia wrocławska*, „Przegląd Antropologiczny", t. XVII

Kalton G., Kasprzyk D. (1982), *Imputing for Missing Survey Responses*, Proceedings of the Survey Research Methods Section, American Statistical Association (http://www.amstat.org/sections/SRMS/Proceedings/papers/1982_004.pdf)

Kalton G., Kasprzyk D. (1986), *The treatment of missing survey data*, „Survey Methodology", vol. 12, nr 1

Lance G. N., Williams W. T. (1967), *A General Theory of Classificatory Sorting Strategies. I. Hierachical Systems*, „Computer Journal", vol. 9

Milligan G. (1989), *A study of the beta-flexible clustering method*, „Multivariate Behavioral Research", vol. 24

Młodak A. (2006 a), *Analiza taksonomiczna w statystyce regionalnej*, Centrum Doradztwa i Informacji DIFIN, Warszawa

Młodak A. (2006 b), *Multilateral normalisations of diagnostic features*, „Statistics in Transition", vol. 7, No. 5

Młodak A. (2009), *Historia problemu Webera*, „Matematyka Stosowana", vol. 10/51

Piasecki T., Cybart D., Kubacki J. (2009), *Metodologiczne problemy imputacji danych w PSR 2010*, Urząd Statystyczny w Łodzi, maszynopis

Rubin D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York, John Wiley & Sons

Stefanowicz B. (2009), *Imputacja danych statystycznych*, maszynopis

SUMMARY

Results of a simulation experiment aimed at an appraisal of utility of some original model of data imputation in censuses are presented in the paper. It is based on clustering of records-donors according to their similarity and on the method of statistical roulette, i.e. a rotational algorithm arranging to records receives the lacking data in a random way from the nearest homogeneous donor clusters. The exercise, which showed high efficiency of the applied attempt, was performed using data for the gmina Gołuchów in the Wielkopolska region collected during the National Population and Housing Census conducted in 2002.

РЕЗЮМЕ

В статье представляются результаты симуляционного эксперимента, целью которого была оценка полезности одной оригинальной модели восстановления данных во всеобщих переписях. Она основывается на группировке записей-доноров согласно их сходству и на методе так

называемой «статистической рулетки», то есть ротационного алгоритма, который присваивает записям-получателям отсутствующие данные выборочным способом из ближайших им однородных кластеров доноров. Обследование показало большую эффективность используемого подхода. В нем были использованы данные для велькопольской гмины Голухув полученные во время Всеобщей переписи населения и квартир в 2002 г.