Bogdan STEFANOWICZ, Marek CIERPIAŁ-WOLAN

Błędy przetwarzania danych

W artykule na temat błędów nielosowych, obciążających wyniki reprezentacyjnych badań statystycznych, o których pisze profesor Mirosław Szreder[1], Autor wymienia cztery ich rodzaje. Są to następujące błędy:

- pokrycia badanej zbiorowości przez operat stanowiący podstawę losowania jednostek do badania,
- spowodowane brakiem odpowiedzi respondentów,
- pomiaru, związane z zarejestrowaniem nieprawdziwych informacji o respondencie,
- przetwarzania danych.

Profesor Szreder ustosunkowuje się do pierwszych trzech rodzajów błędów, czwarty zaś pomija. Tymczasem wspomniane przez Profesora przetwarzanie danych jest często poważnym źródłem błędów, znacząco obniżających jakość wyników badania statystycznego.

Zachęca to do przedstawienia refleksji w tej sprawie. Wydaje się to tym bardziej uzasadnione, że zarówno błędy losowe, jak i nielosowe stanowią istotną część jednego z wymiarów jakości Europejskiego Systemu Statystycznego dotyczących badań statystycznych w zakresie: przydatności, dokładności, terminowości, punktualności, dostępności, przejrzystości, porównywalności i spójności, zidentyfikowanych i precyzyjnie opisanych w powszechnie dostępnym dokumencie[2].

W przetwarzaniu danych statystycznych można wyróżnić kilka kluczowych operacji, takich jak: wprowadzanie, symbolizacja, redagowanie czy imputacja. Każda z nich wiąże się z ryzykiem popełnienia błędów.

Celem artykułu jest zwrócenie uwagi na ten fakt i wykazanie, że jednym ze składników błędu całkowitego, o którym pisze Mirosław Szreder, może być błąd generowany przez niektóre z tych operacji.

WPROWADZANIE DANYCH

Jedną z pierwszych operacji w przetwarzaniu danych jest ich wprowadzanie — wypełnianie formularzy statystycznych i zapisywanie do baz danych. Obie operacje stają się często źródłem ewidentnych pomyłek. W szczególności powszechnie popełniany jest tzw. czeski błąd (polegający na przestawianiu znaków, np. cyfr w liczbie), który w większości przypadków jest trudny do zidenty-

[1] Szreder M. (2015), *Zmiany w strukturze całkowitego błędu badania próbkowego*, „Wiadomości Statystyczne", nr 1, s. 4—12.

[2] Na przykład w związanym z Badaniem Aktywności Ekonomicznej Ludności (BAEL) w krajach Unii Europejskiej http://ec.europa.eu/eurostat/cache/metadata/EN/employ_esms.htm#accuracy 1423673804060.

fikowania nawet przez drobiazgowe kontrole. Skutki tego typu błędów są często rozpoznawane dopiero po upływie pewnego czasu, zwykle w wyniku bardziej zaawansowanych porównań. Zdarza się, że następuje to dopiero na etapie porównań międzynarodowych prowadzonych przez inne instytucje, np. Eurostat. Konsekwencje tego stają się szczególnie dotkliwe w sytuacjach, gdy badana zmienna wyrażana jest przez duże rzędy wartości (w mln lub mld). Często zdarza się, że owa błędnie wprowadzona wartość stanowi istotną część wyliczanych na jej podstawie wskaźników i wówczas taka sytuacja jest bardzo trudna do skorygowania. Sytuacja komplikuje się zwłaszcza wtedy, gdy jest to wskaźnik ogłaszany w różnego rodzaju aktach prawnych i stanowi podstawę do podejmowania decyzji na poziomie mikro-, mezo- czy makroekonomicznym. Warto zauważyć, że statystyka publiczna opracowuje ok. 200 wskaźników zawartych w różnego rodzaju rozporządzeniach, obwieszczeniach i komunikatach, które bezpośrednio wpływają na wiele istotnych rozstrzygnięć.

Podobne skutki mają błędy popełniane przy wprowadzaniu danych, kiedy osoba wypełniająca formularz nie zwraca uwagi na jednostki miary, w jakich wyrażone są wartości danej zmiennej, np. wprowadzamy domyślnie wartości w jednostkach, podczas gdy w rzeczywistości są one wyrażone w tysiącach.

Warto nieustannie podkreślać, że duże znaczenie ma przy tym konstrukcja formularza, im jest on bardziej skomplikowany i nieprzejrzysty, tym więcej tego rodzaju błędów powstaje na etapie jego wypełniania. Nadmierna szczegółowość stanowi bowiem źródło błędów. W niektórych sprawozdaniach statystycznych wymaga się od respondenta wypełniania danych za bieżący okres i narastająco, co stanowi dodatkową trudność.

Kolejne błędy, trudne do zidentyfikowania, mogą być popełniane podczas symbolizacji. Przykładem usterek tego typu są błędy popełniane przy kodowaniu chorób przez lekarzy orzeczników lub przyporządkowywanie badanym zjawiskom symboli pochodzących z różnych klasyfikacji, np. dla wydatków konsumpcyjnych według klasyfikacji COICOP/HBS[3] w badaniu budżetów gospodarstw domowych czy dla zawodów według klasyfikacji ISCO[4] w BAEL.

Pojawiają się też niekiedy błędy techniczne. Pomimo niemal doskonałych algorytmów przesyłania danych, w praktyce wykorzystywania urządzeń przenośnych zdarza się, że na etapie transmisji występują nieprzewidywane awarie. Mamy wówczas do czynienia z niekompletnie wypełnionym formularzem, a jest to błąd trudny do wykrycia.

REDAGOWANIE DANYCH

Na szczególną uwagę zasługuje redagowanie danych, czyli sprawdzanie ich poprawności (kontrola) oraz poprawianie wykrytych błędów (korekta). Często dokonuje się też wyprowadzenia pewnych danych pochodnych jako uzupełnie-

[3] *Classification of Individual Consumption by Purpose* — Klasyfikacja Spożycia Indywidualnego według Celu.

[4] *Classification of Occupations and Specializations* — Klasyfikacja Zawodów i Specjalności.

nie zebranego materiału badawczego, np. wyliczenie wieku badanych osób na podstawie zarejestrowanych dat urodzenia.

W procesach tych zazwyczaj wykorzystuje się technologie informatyczne, ze względu na duże zbiory danych. W tego rodzaju okolicznościach zachodzi konieczność opracowania odpowiednich programów: kontroli danych, korekty wykrytych błędów, a także wyznaczania danych pochodnych. W każdym z tych przypadków rodzi się potrzeba sformułowania odpowiednich reguł przez specjalistę statystyka. Nawet najzdolniejszy informatyk nie jest w stanie dokonać tego ze względu na brak merytorycznej wiedzy w stosownym zakresie. Jego zadaniem jest przekształcenie tych reguł do postaci algorytmu i zakodowanie go jako programu komputerowego.

Trzeba mieć bowiem na uwadze, że człowiek często się myli, mimo swojego geniuszu, wiedzy, doświadczenia i rzetelności, np. z powodu zmęczenia, a w konsekwencji przeoczenia jakiegoś szczegółu. Nie można zatem zakładać, że najdoskonalszy nawet specjalista statystyk jest w stanie w każdych okolicznościach bezbłędnie sformułować odpowiednie reguły kontroli oraz korekty w odniesieniu do każdego przypadku.

Trzeba też brać pod uwagę zmienność zjawisk, zwłaszcza społecznych i ekonomicznych. Nawet przy założeniu, że specjalista ma bogatą wiedzę i zna rzeczywiste cechy obserwowanych zjawisk trwających w jakimś okresie, to ze względu na ich zmienność, w chwili opracowywania potrzebnego algorytmu wiedza ta może być nieaktualna. Jest to szczególnie istotne w przypadku zjawisk ulegających dynamicznym zmianom, jak np. kształtowanie się cen na rynku w okresach sezonowych.

W rezultacie algorytm kontroli danych czy korekty błędów powstaje na podstawie wiedzy nieaktualnej. W tej sytuacji, z powodu błędnych założeń, wyniki operacji kontrolnych i korygujących będą nieuchronnie obciążone błędem systematycznym. Fakt ten nabiera szczególnego znaczenia w odniesieniu do błędów rzadko występujących, ale obciążających dane opisujące kluczowe cechy i znaczące jednostki badane — choćby duże przedsiębiorstwa w badaniach statystycznych przemysłu.

Wspomniane kryteria kontrolne mogą też być niekiedy obciążone pewnymi wadami „wrodzonymi”:

- niektóre z nich są zbyt liberalne — przyjmują, że „wszystko w życiu jest możliwe”. Sprawia to, że niejednokrotnie w istocie błędne dane zostają przyjęte jako poprawne, chociaż faktycznie nimi nie są. Przyczynia się to do włączenia do zbioru przetwarzanych danych wielu błędów;
- inne kryteria mogą być bardzo restrykcyjne, „podejrzewające” niemal wszystkie dane o błędy. W tym przypadku kryterium to (a w konsekwencji także odpowiedni program) będzie odrzucać jako błędne i takie dane, które nie są obciążone wadami. Dane te — rzekomo błędne (nazwiemy je błędami pozornymi) — w dalszych operacjach są korygowane ręcznie lub za pomocą odpowiednich procedur programowych, czyli w ramach tzw. korekty automatycznej.

W przypadku stosowania takiej korekty tylko w wyjątkowych sytuacjach istnieje możliwość ponownego dotarcia do badanej jednostki i uzyskania właściwej odpowiedzi, którą należy wprowadzić na miejsce wykrytego błędu. W większości zaś przypadków następuje poszukiwanie wartości najbardziej prawdopodobnej, która powinna być wstawiona w miejsce stwierdzonego błędu. Często wymienia się tu procedurę imputacji danych, która polega na dobieraniu wartości szacunkowych spełniających zakładane kryteria poprawności i wstawianych w miejsce wykrytych błędów lub jako „plomba" zamiast brakujących danych.

Niestety, bardzo rzadko wyznaczone dane korygujące są tymi wartościami, które powinny zastąpić wykryty błąd. W konsekwencji w zbiorze danych wejściowych pojawiają się wartości, które tylko formalnie spełniają przyjęte kryteria poprawności, ale w istocie są to dane sztuczne, odbiegające od faktycznego stanu badanej rzeczywistości.

Naturalnie, w każdym takim przypadku niewykryte błędy lub błędy pozorne i wprowadzone w ich miejsce dane szacunkowe w jakimś stopniu zniekształcą zebrany materiał statystyczny, a ich włączenie do dalszego procesu przetwarzania wypacza wyniki badania. Im więcej takich uzupełnień („plomb") pojawi się w zbiorze, tym większe obciążenie wywrą one na jakość wyników. Jest to niezależne od stosowanych technologii informatycznych.

W szacowaniu całkowitego błędu badania wpływ ten jest w praktyce niestety pomijany.

PRZETWARZANIE DANYCH JAKO PROCES WIELOSTOPNIOWY

Jak wiadomo, przetwarzanie danych statystycznych to nie tylko wspomniane wcześniej wprowadzanie danych i ich redagowanie (chociaż te etapy przetwarzania zostały tu tylko zasygnalizowane). W szerszym ujęciu na proces ten składa się długa lista wielu innych operacji związanych z opracowaniem wyników badania, archiwizowaniem danych, udostępnianiem informacji wynikowych itp. Każda z nich — realizowana ręcznie lub z wykorzystaniem dostępnych technologii — jest zabiegiem wykonywanym na zgromadzonym materiale źródłowym, wyprowadzającym określone informacje pochodne.

Wszystkie etapy przetwarzania wiążą się z koniecznością posłużenia się odpowiednimi algorytmami, a w praktyce statystycznej na ogół z programami komputerowymi. W każdym takim przypadku wspomniane algorytmy są budowane na podstawie określonych założeń, wynikających z wiedzy empirycznej i koncepcji teoretycznych. W konsekwencji powstają wyniki „dopasowane" do owych koncepcji modelowych. Niestety, są one obciążone przyjętymi założeniami i tylko w pewnym przybliżeniu odzwierciadlają rzeczywistość, którą przedstawiają dane zebrane na początku badania.

Rodzi się pytanie, czy uzyskiwane w ten sposób wyniki mają jakąś wartość poznawczą i mogą służyć w praktyce do podejmowania decyzji, czasem o znaczeniu kluczowym, a także globalnym? Odpowiedź jest oczywista — mają. Decydent, który z nich korzysta musi jednak być świadom, że informacja staty-

styczna zawsze tylko w przybliżeniu opisuje rozpatrywaną rzeczywistość. Przyczyny tego są rozmaite. Zwróćmy uwagę na następujące spośród nich:

- uzyskiwane dane opisują przeszłość. Im jest ona bardziej odległa od momentu, w którym owe dane zostają wykorzystane, tym mniej aktualne są informacje, którymi decydent dysponuje, dlatego ważną cechą informacji statystycznych jest ich aktualność;
- wszelkie operacje wykonywane w ramach przetwarzania zgromadzonego materiału (tzw. wyrównania sezonowe, uogólnianie, analiza, interpretacja, porównania) wnoszą jakąś dozę sztuczności i przyczyniają się do pojawienia się nieco zdeformowanego obrazu obserwowanej rzeczywistości. Stąd w kontekście jakości informacji wynikowych rodzi się wymaganie rzetelności tych wyników, której warunkiem koniecznym jest właściwy dobór stosowanych metod;
- globalizacja i procesy integracyjne sprawiają, że instytucje międzynarodowe prowadzące badania i tworzące bazy danych muszą zapewnić ich porównywalność. Wymaga to zastosowania różnego rodzaju procedur i metod ekonometrycznych. Powszechnie wiadomo, że np. w zakresie wyrównań sezonowych dominują dwie metody — TRAMO/SEATS[5] i X-12 ARIMA[6]. Od wielu lat środowiska naukowe nie są w stanie jednoznacznie rekomendować jednej z nich do powszechnego stosowania. Na stronach Eurostatu możemy np. znaleźć informację, że liczba bezrobotnych w Polsce według BAEL w pierwszym kwartale 2010 r. wynosiła 1,798 mln, natomiast liczba bezrobotnych wyrównana sezonowo przy użyciu metody TRAMO/SEATS to 1,650 mln, a więc różnica wynosi 148 tys. Przy zastosowaniu drugiej metody otrzymujemy inny wynik. Zatem preferowanie tylko jednej z nich nie zapewni optymalnego rozwiązania w tym zakresie i stanowi kolejne potencjalne źródło błędów.

  Dodajmy, że z formalnego punktu widzenia za każdym razem należałoby dokonać oceny jakości wyrównania sezonowego. Porównanie wyników otrzymanych metodami TRAMO/SEATS i X-12 ARIMA powinno więc każdorazowo przebiegać na podstawie analizy różnorodnej statystyki, która pozwala na ocenę jakości dekompozycji i stabilności wygładzonych szeregów czasowych (wartość kryteriów informacyjnych, kryterium idem potencji, miary (nie)zgodności, kryterium rewizji, *sliding spans*). Podczas procedury wstępnych wyrównań sezonowych wykonywane są zwykle czynności pozwalające wykryć nietypowe zaburzenia w szeregu czasowym, m.in. takie jak wartości odstające. Występowanie takich zaburzeń, będących efektem sporadycznych, nieregularnych zdarzeń, powoduje zniekształcenia w analizie szeregów, które utrudniają lub wręcz uniemożliwiają ich modelowanie. Muszą one być zatem zidentyfikowane (wiele programów robi to automatycznie), a przede wszystkim odpowiednio potraktowane przez analityka. Może się także zdarzyć, że wartości od-

[5] *Time series Regression with ARIMA noise, Missing values and Outlieres.*
[6] *Autoregressive integrated moving average.*

stające są skutkiem błędów wprowadzania bądź symbolizacji danych, co prowadzi do powielania się (multiplikowania) błędów;
- w przypadku badań reprezentacyjnych przetwarzanie danych obejmuje także uzupełnianie bazy danych o wagi, które pozwalają na uogólnienie wyników. O ile same wagi związane są z wybraną metodą losowania, gdzie błędy wynikają ze schematu losowania i są dobrze opisane w literaturze, to często dokonuje się także kalibracji wag. W przypadku badań społecznych kalibracja wag może być związana z koniecznością dostosowania struktury demograficznej badanej próby do populacji. Często punktami odniesienia są dane pochodzące ze spisów ludności, które organizowane są średnio co dziesięć lat. Praktyka wskazuje, że na ogół każdy spis wiąże się z korektą danych. Stąd wszystkie kalibracje wag dokonywane w okresach międzyspisowych obarczone są błędem.

Wnioski

Nasze refleksje skłaniają do sformułowania następujących wniosków:

1. Fakt, że przetwarzanie danych statystycznych jest związane z ryzykiem pojawiania się pewnych błędów w zbiorze danych wejściowych nie może być interpretowany jako ostrzeżenie przed stosowaniem kontroli danych czy korekty błędów podczas ich redagowania i innych operacji. Nie jest też ostrzeżeniem przed stosowaniem dostępnych technologii informatycznych. Są one koniecznością w obliczu posługiwania się dużymi zbiorami danych i zazwyczaj krótkich terminów uzyskania wyników.
2. Sformułowany przez Mirosława Szredera postulat w sprawie konieczności podjęcia badań nad oszacowaniem całkowitego błędu w badaniach reprezentacyjnych zasługuje na uwagę zarówno pod względem teoretycznym, jak i praktycznym. Ocena taka powinna obejmować wymienione przez Autora cztery rodzaje błędów, jak i uwzględniać wiele innych czynników, np. błędy w przygotowaniu badania statystycznego, w organizacji i przeprowadzeniu operacji zbierania danych (w szczególności prowadzenia wywiadu) i in. Na uwagę zasługuje podjęcie takich analiz nie tylko w odniesieniu do badań reprezentacyjnych, ale także wszystkich badań statystycznych. Może to stanowić rozległy temat badań naukowych. Jeden z wątków tych dociekań to ustalanie wielkości próby w badaniu reprezentacyjnym przy planowaniu stosowania automatycznych procedur redagowania danych. Skoro wiadomo, że procedury takie mogą obniżyć w jakimś stopniu jakość wyników, to obciążenie to powinno być oszacowane już podczas planowania badania.
3. Taka rozszerzona analiza błędów nielosowych w badaniach statystycznych powinna znaleźć odzwierciedlenie w programach nauczania statystyki i w podręcznikach akademickich. Na ogół można w nich znaleźć opisy wielu programów komputerowych do analiz statystycznych. Brakuje jednak analizy skutków ich udziału w ostatecznej jakości wyników badania.

Trzeba wszakże zauważyć, że w praktyce znaczenie błędów przetwarzania i częstość ich występowania dostrzegają specjaliści statystycy, w szczególności podejmując nieustanne wysiłki nad doskonaleniem rozwiązań informatycznych, niekiedy wielokrotnie modyfikując systemy informatyczne. Praktyka wskazuje, że w niektórych przypadkach nawet niewielkie zmiany wprowadzane w badaniach realizowanych cyklicznie wymagają dokonania kilku, a nawet kilkunastu poprawek oprogramowania w części dotyczącej interfejsu użytkownika i wstępnej kontroli danych. Podczas każdej zmiany aplikacyjnej koryguje się zwykle od jednego do kilku błędów związanych z wprowadzaniem danych.

prof. dr hab. Bogdan Stefanowicz — *Wyższa Szkoła Informatyki Stosowanej i Zarządzania pod auspicjami Polskiej Akademii Nauk*
dr Marek Cierpiał-Wolan — *Urząd Statystyczny w Rzeszowie, Uniwersytet Rzeszowski*

SUMMARY

The article highlights the need to broaden the analysis of the quality of the survey results, taking into account the negative impact of certain operations of so-called editing input data, such as checking their accuracy and correction of errors. In the conclusions it underlines the need to extend the programs for academic lectures in statistics for analysis of the impact of processing operations on the quality of the results.

РЕЗЮМЕ

В статье было обращено внимание на необходимость расширения анализа качества результатов статистических обследований в отношении к отрицательному влиянию некоторых операций так называемой редакции входных данных, таких как контроль их правильности и корректировка ошибок. В статье подчеркивается необходимость расширения программы обучения в области статистики дополняя ее анализом влияния операций обработки на качество результатов.