

Paweł LANDUCH

Pomiar statusu osób na rynku pracy w Narodowym Spisie Powszechnym Ludności i Mieszkań 2011

Streszczenie. *Wielokrotny pomiar statusu osób na rynku pracy z wykorzystaniem kwestionariusza statystycznego może posłużyć do zastosowania metod ilościowych do oceny jego jakości. W celu oceny jakości wyników badania stosuje się m.in. metody powtórne, a także wielokrotnego pomiaru na wybranej próbie losowej. Na podstawie danych uzyskanych z Narodowego Spisu Powszechnego Ludności i Mieszkań przeprowadzonego w 2011 r. (NSP 2011), zaprezentowano zastosowanie współczynnika zgodności Cohena oraz analizy klasy ukrytej do oceny jakości klasyfikowania osoby na rynku pracy. Pytania kwestionariusza spisowego pozwoliły na ocenę statusu osób na rynku pracy zgodnie z rekomendacją Międzynarodowej Organizacji Pracy (MOP) i samooceny respondenta. W artykule przedstawiono tylko kwestię zgodności wyników uzyskanych w tych ujęciach w części spisu opartej na próbie losowej, pomijając wiele innych zagadnień dotyczących błędów losowych i nielosowych.*

Słowa kluczowe: kwestionariusz statystyczny, błąd pomiaru, indeks zgodności, klasa ukryta.

JEL: C83

W badaniach rynku pracy jednym z kluczowych elementów jest określenie statusu osoby na tym rynku. GUS prowadzi w tym zakresie szereg badań obejmujących instytucje, jak i gospodarstwa domowe. Jednym z nich, przeprowadzanych cyklicznie, jest Badanie Aktywności Ekonomicznej Ludności. Dane o rynku pracy są gromadzone m.in. na podstawie sprawozdawczości przedsiębiorstw i jednostek sfery budżetowej, a także administracyjnych źródeł danych. Bezrobocie rejestrowane badane jest według danych powiatowych urzędów pracy. W publikacji, opisującej zasady metodyczne statystyki rynku pracy i wy-

nagrodzeń (GUS, 2008), omówiono szczegółowo badania rynku pracy i aktywności ekonomicznej prowadzone regularnie przez GUS. Tematykę aktywności zawodowej ludności obejmują też spisy powszechne i spisy rolne. Przyjmując zalecenia i definicje stosowane przez MOP, w NSP 2011, w części poświęconej aktywności ekonomicznej, badano m.in. status osoby na rynku pracy. Przyjęto definicję bieżącej aktywności ekonomicznej wyróżniając trzy kategorie:

- pracujący,
- bezrobotni,
- bierni zawodowo (poza rynkiem pracy).

Kategorie te w sposób zupełny i rozłączny klasyfikują osobę na rynku pracy w tym sensie, że musi ona należeć do jednej, i tylko jednej, z wymienionych kategorii. W spisie powszechnym zbierano dane różnymi metodami. Część informacji miała swoje źródło w systemach informacyjnych, a kolejną część uzyskano z wywiadów przeprowadzonych na reprezentacyjnej grupie ludności. W tej części dane zebrano przy pomocy kwestionariusza statystycznego trzema różnymi metodami, tj. na podstawie samospisu internetowego oraz wywiadów telefonicznego i bezpośredniego.

Istotnym elementem oceny przeprowadzonego badania statystycznego jest jego jakość, która jest kategorią wielowymiarową, bardzo istotną i opisaną w wielu materiałach krajowych urzędów statystycznych oraz Eurostatu. Jednym z dokumentów dotyczącym raportów jakości badań w Europejskim Systemie Statystycznym jest regularnie aktualizowany podręcznik na ten temat (Eurostat, 2014).

Celem artykułu jest ocena jakości związanej z pomiarem badanej cechy. W tym przypadku badaną zmienną jest status na rynku pracy. Poruszono problem wiarygodności poszczególnych kategorii uzyskanych z wykorzystaniem kwestionariusza statystycznego. Przeprowadzono analizę różnic w klasyfikacjach na podstawie dwukrotnego pomiaru badanej cechy. W tym celu zastosowano współczynnik zgodności, który na podstawie obserwowanych wartości mierzonej cechy w dwóch pomiarach ocenia stopień zgodności między nimi. Określono również różnice w pomiarach tej samej cechy z wykorzystaniem modelowania statystycznego. Użyto tutaj metody analizy klasy ukrytej. Zakłada ona, że obserwowana wartość jest tylko reprezentantem wartości badanej cechy, a prawdziwa jej wartość pozostaje ukryta. Często wartości obserwowanej cechy nazywa się zmienną manifestowaną. Na koniec dokonano krótkiego podsumowania badania. Użyta metoda ilościowa jest próbą znalezienia ewentualnych problemów z pomiarem tej klasyfikacji na etapie gromadzenia danych.

WYKORZYSTANIE WSPÓŁCZYNNIKA ZGODNOŚCI

Błąd pomiaru możemy zdefiniować jako różnicę między prawdziwą wartością badanej cechy a wartością otrzymaną w badaniu. W nominalnej skali pomiarowej błąd klasyfikacji można przedstawić jako różnicę liczby osób, które błędnie

przypisano do kategorii X , chociaż faktycznie nie należały do tej kategorii (*false positive*) i liczby osób, które przypisano do innej kategorii, chociaż faktycznie należały do kategorii X (*false negative*). Lohr (1999) zaznacza osiem przyczyn, które mogą prowadzić do błędu klasyfikacji:

- respondenci czasami nie mówią prawdy;
- respondenci nie zawsze rozumieją pytania zawarte w kwestionariuszu;
- pamięć respondenta bywa ograniczona (np. przez czas);
- rozmaite odpowiedzi mogą być udzielone przez respondenta różnym ankietantom;
- respondenci mogą udzielić odpowiedzi, które uznają za oczekiwane przez ankietera;
- ankietier nie zawsze wiernie odczytuje pytanie, wpływając tym samym na jakość odpowiedzi lub pomyłkowo rejestruje inną odpowiedź niż została podana;
- znaczenie słów może być różnie rozumiane przez respondentów;
- sformułowanie pytań oraz kolejność ich zadawania mogą mieć wpływ na udzielane odpowiedzi.

Biemer (2004), w artykule dotyczącym wykorzystania modelowania statystycznego do oceny pytań zawartych w kwestionariuszach statystycznych, wskazuje przykłady i omawia ogólne podejście do wykorzystania analizy statystycznej w testowaniu kwestionariuszy statystycznych. W celu użycia tych metod zakładamy, że posiadamy wielokrotny pomiar cechy, który może być dokonany poprzez:

- powtórny wywiad dotyczący danej cechy lub cech, realizując ponowny kontakt z próbką jednostek uczestniczących w wywiadzie pierwotnym;
- pojedynczy pomiar, w którym badana cecha jest zawarta w kwestionariuszu statystycznym dwukrotnie przy użyciu różnie sformułowanych pytań;
- wielokrotny pomiar, gdzie jedną grupę danych uzyskuje się za pomocą kwestionariusza, a drugą z innych źródeł, np. z danych administracyjnych;
- badanie panelowe odnoszące się do tych samych jednostek w różnych momentach czasu.

Dane uzyskane w ten sposób mogą być następnie poddane ocenie przy użyciu następujących metod statystycznych:

- analizy danych w celu identyfikacji pytań, które potencjalnie generują odpowiedzi o małej wiarygodności;
- kolejnej analizy prowadzącej do identyfikacji źródeł błędów w pytaniach, które zostały uznane jako mające małą wiarygodność;
- weryfikacji źródeł błędów poprzez dodatkowo uzyskane dane z eksperymentów, testów pilotażowych, metod kognitywnych itd.;
- wyboru strategii, która ma wyeliminować lub zmniejszyć błąd pomiaru poprzez jej implementację.

W praktyce statystycznej wykorzystywane są przede wszystkim metody statystyczne wyszczególnione w pierwszych dwóch punktach.

W artykule przedmiotem porównania jest klasyfikacja osoby na rynku pracy. Dane uzyskane do porównania pochodzą z kwestionariusza statystycznego użytego w NSP 2011. Wywiad z użyciem kwestionariusza przeprowadzono na reprezentacyjnej grupie ludności. Podstawą zastosowanych definicji i pojęć były zalecenia międzynarodowe do spisów ludności i mieszkań przyjęte przez Konferencję Statystyków Europejskich w czerwcu 2006 r. (ONZ, 2006). W części dotyczącej aktywności ekonomicznej, definicje były oparte na rezolucji dotyczącej statystyki pracujących, bezrobotnych i niepełnozatrudnionych, przyjętej na XIII Międzynarodowej Konferencji Statystyków Pracy w październiku 1982 r. (z późniejszymi zmianami), zalecane do stosowania przez MOP (ILO, 1982). Zgodnie z tymi zaleceniami zakwalifikowanie osoby do jednej z kategorii (pracujący, bezrobotni, poza rynkiem pracy) następuje po udzieleniu przez respondenta odpowiedzi na szereg pytań, zadanych nie wprost i w odpowiedniej kolejności. Pytania dotyczyły wykonywania i posiadania pracy w badanym tygodniu, aktywnego jej poszukiwania oraz bycia gotowym do jej podjęcia w krótkim okresie. Na tej podstawie następowało zakwalifikowanie osoby do jednej z wyszczególnionych kategorii (GUS, 2013b). W dalszej części artykułu klasyfikację tę określono jako „definicję MOP” lub „według MOP”. Niezależnie od opisanego sposobu definiowania osób na rynku pracy, zastosowano jednocześnie drugi sposób pomiaru subiektywnej oceny na rynku pracy dokonanej przez samego respondenta. Pytanie brzmiało jak opisałby (opisałaby) Pan(i) swoją sytuację na rynku pracy w tygodniu od 25 do 31 marca 2011 r.? Proszę wybrać jedną odpowiedź.

Pytanie to zawierało następujące warianty odpowiedzi:

- 1 — pracowałem(am) wyłącznie poza rolnictwem;
- 2 — pracowałem(am) głównie poza rolnictwem i dodatkowo w rolnictwie;
- 3 — pracowałem(am) głównie w rolnictwie i dodatkowo poza rolnictwem;
- 4 — pracowałem(am) wyłącznie w rolnictwie;
- 5 — byłem(am) bezrobotny(a);
- 6 — uczyłem(am) się, studiowałem(am);
- 7 — byłem(am) na emeryturze, wcześniejszej emeryturze;
- 8 — nie pracowałem(am) z powodu złego stanu zdrowia (niepełnosprawności);
- 9 — zajmowałem(am) się domem, rodziną;
- 10 — byłem(am) bierny(a) zawodowo z innych przyczyn niż wyżej wymienione;
- 99 — nieustalona.

Opowiedzi na pytania 1—4 pozwalały na wyłonienie kategorii pracujących, odpowiedź na pytanie 5 dotyczyła kwalifikacji osób jako bezrobotnych, a pozostałe warianty — biernych zawodowo.

Przedmiotem dokonywanego w artykule porównania jest zgodność uzyskanych wyników klasyfikacji otrzymanych dwoma metodami, zawartymi w jednym kwestionariuszu. Miarą wykorzystaną do oceny porównania wyników jest współczynnik zgodności kappa (κ) w wersji zaproponowanej przez Cohena (1960).

Autor ten poddał krytyce porównanie zgodności za pomocą statystyki χ^2 , która ocenia tylko siłę związku pomiędzy dwiema zmiennymi, ale nie uwzględnia zgodności w poszczególnych kategoriach. W celu dokonania porównania zakłada się, że:

- jednostki są niezależne,
- kategorie są niezależne, rozłączne i zupełne,
- pomiary są niezależne.

TABL. 1. TABLICA KONTYNGENCJI ZAWIERAJĄCA KLASYFIKACJE OSÓB NA RYNKU PRACY WEDŁUG DEFINICJI MOP ORAZ SAMOOCENY

Status według samooceny	Status według MOP			
	pracujący	bezrobotni	bierni zawodowo	brak danych
Pracujący	2884247	19045	149926	3
Bezrobotni	22517	318806	152239	4
Bierni zawodowo	157223	92963	2639768	2
Brak danych	339394	52486	121359	544086

Ź r ó d ł o: wyliczenia własne na podstawie danych NSP 2011¹.

W danych (tabl. 1) dla dokonywanych porównań pominięto osoby, dla których status na rynku pracy nie został ustalony. Pozostanie więc klasyfikacja składająca się z 3 podkategorii. Liczba wszystkich osób wyniosła więc 6436734. Najliczniejsze były grupy osób pracujących i biernych zawodowo. W stosunku do pozostałych nieliczną kategorię stanowili bezrobotni. Jednak duża liczba osób, mających status bezrobotnych według definicji MOP, w samoocenie została zaliczona do osób biernych zawodowo. Poza tym w otrzymanych wynikach zauważalne są różnice w brakach danych w obu kategoriach. Ponad 0,5 mln osób, dla których ustalono status według definicji MOP, według samooceny miało status nieustalony. Jeśli status według samooceny występuje, braki w statusie MOP są sporadyczne. Jednym z powodów tego braku mogła być kolejność zadawanych pytań. Pytania o status według MOP zadawane są w pierwszej kolejności, co jest kluczowe dla jakości wyników.

Porównano klasyfikację poprzez wyliczenie współczynnika κ dla całej tablicy oraz dla każdej kategorii z osobna poprzez przekodowanie tablicy z 3 kategoriami na 3 tablice zawierające zmienną dychotomiczną. W tabl. 2 podano wyliczenia wraz z wartościami teoretycznymi i brzegowymi, co pozwala zauważyć duże różnice w kategoriach pracujących i biernych zawodowo na rynku pracy. Ma to ogromny wpływ na wzrost statystyki χ^2 , co prowadzi do podkreślania związku między nimi. W tabl. 3 przedstawiono odpowiednie frakcje dla zmiennych klasyfikacyjnych.

¹ Badanie zostało przeprowadzone na próbie losowej ok. 20% mieszkań w skali kraju. Jednostką losowania było mieszkanie, a dokładniej jego adres. Szczegółowy opis badania znajduje się w publikacji GUS (2013a), rozdział 1 *Metodologia spisu ludności i mieszkań 2011 — wybrane aspekty*.

**TABL. 2. TABLICA KONTYNGENCJI ZAWIERAJĄCA KLASYFIKACJE OSÓB
NA RYNKU PRACY WEDŁUG DEFINICJI MOP ORAZ SAMOOCENY
WRAZ Z WYLICZENIAMI TEORETYCZNYMI (w nawiasach)**

Status według samooceny	Status według MOP			
	ogółem	pracujący	bezrobotni	bierni zawodowo
Ogółem	6436734	3063987	430814	2941933
Pracujący	3053218	2884247 (1453399)	19045 (204354)	149926 (1395491)
Bezrobotni	493562	22517 (234944)	318806 (33034)	152239 (225583)
Bierni zawodowo	2889954	157223 (1375670)	92963 (193425)	2639768 (1320858)

Źródło: jak przy tabl. 1.

**TABL. 3. FRAKCJE DLA OBU KLASYFIKACJI
WRAZ Z WYLICZENIAMI TEORETYCZNYMI (w nawiasach)**

Status według samooceny	Status według MOP			
	ogółem	pracujący	bezrobotni	bierni zawodowo
Ogółem	1,0000	0,4760	0,0669	0,4570
Pracujący	0,4743	0,4481 (0,2258)	0,0029	0,0232
Bezrobotni	0,0767	0,0035	0,0495 (0,0051)	0,0236
Bierni zawodowo	0,4490	0,0244	0,0144	0,4101 (0,2052)

Źródło: jak przy tabl. 1.

Cohen (1960) wprowadza współczynnik κ wyrażony wzorem:

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

gdzie:

p_0 — frakcja jednostek, dla których występuje zgodność klasyfikacji w obu pomiarach,

p_c — frakcja jednostek, dla których oczekiwana jest zgodność teoretyczna.

Współczynnik ten wynosi 0 dla idealnej zgodności rozkładu teoretycznego z obserwowanym. Maksymalnie może on przyjąć wartość 1, gdy obie klasyfikacje będą miały taką samą liczebność dla każdej kategorii. Interesujące jest, jaki maksymalny poziom mógłby przyjąć współczynnik κ . Wyznacza się go poprzez wyliczenie:

$$\kappa_m = \frac{p_m - p_c}{1 - p_c}$$

gdzie p_m jest sumą minimalnych wartości brzegowych dla każdej kategorii w obu klasyfikacjach (w naszym przypadku liczba ta wynosi $0,4743+0,0669+0,4490=0,9902$). Wtedy $1-\kappa_m$ wskazuje, jaki współczynnik mógłby być osią-

gnięty, gdyby nie różnice w wartościach brzegowych, co z kolei może wskazywać na graniczną niejasność kategorii.

Błąd standardowy może być oszacowany według wzoru:

$$\sigma_k = \sqrt{\frac{p_0(1-p_0)}{N(1-p_c)^2}}$$

gdzie N oznacza liczebność populacji. Dodatkowo, przy dużej wartości N , przedział ufności może zostać oszacowany z granicznej normalności rozkładu κ . Test dla istotności statystycznej κ może zostać wykonany z wykorzystaniem statystyki z :

$$z = \frac{\kappa_1 - \kappa_2}{\sigma_{k1}^2 - \sigma_{k2}^2}$$

dla różnych κ . Hipoteza zerowa zakłada, że κ równa się 0 i wtedy $p_c = p_0$. Cohen (1960) wskazuje, że samo testowanie istotności statystycznej ma niewielką wartość poznawczą, może jednak stanowić warunek minimum dla wymagań stawianych oszacowaniu współczynnika κ .

Jarosz-Nowak (2007) przytacza skale porównawcze, które mają praktyczne zastosowanie przy ocenie współczynnika Cohena. Chociaż progi ustalone są arbitralnie, można z ich pomocą sformułować wnioski i dokonać stosownych porównań. Przykład takiej skali według Kocha i Landisa (1977) przedstawia tabl. 4.

TABL. 4. INTERPRETACJA WSPÓŁCZYNNIKA COHENA WEDŁUG KOCHA I LANDISA

Wartość współczynnika	Interpretacja zgodności
Poniżej 0,00	brak zgodności
0,00—0,20	słaba
0,21—0,40	średnia
0,41—0,60	umiarkowana
0,61—0,80	pokaźna
0,81—1,00	prawie perfekcyjna

Źródło: Koch i Landis (1977).

Autorka wskazuje, że bardziej efektywna może być analiza współczynnika κ według poszczególnych kategorii klasyfikacji. Dowodzi, że w przypadku zmiennych klasyfikacyjnych składających się z więcej niż 2 kategorii, możliwe jest wyprowadzenie współczynnika κ z modelu binarnego dla każdej kategorii z osobna jako ich średniej ważonej. Dodatkowo wyprowadza współczynnik κ metodą największej wiarygodności. Tak uzyskane estymatory są identyczne z uzyskanymi heurystycznie.

Wracając do oceny statusu danych wynikowych z NSP 2011 (tabl. 1), wartość statystyki χ^2 jest bardzo wysoka i wynosi 7824882 przy 4 stopniach swobody i wartości $p < 0,0001$. Wartości współczynników i statystyki testującej dla tabl. 3 wynoszą:

$$p_0 = 0,4481 + 0,0495 + 0,4101 = 0,9077,$$

$$p_c = 0,2258 + 0,0051 + 0,2052 = 0,4361,$$

$$\kappa = \frac{0,9077 - 0,4361}{1 - 0,4361} = 0,8363,$$

$$\kappa_m = \frac{(0,4743 + 0,0669 + 0,4490) - 0,4361}{1 - 0,4361} = 0,9826,$$

$$\sigma_k = \sqrt{\frac{0,9077(1 - 0,9077)}{6436734(1 - 0,4361)^2}} = 2,0231 \cdot 10^{-4},$$

95% przedział ufności dla współczynnika κ — od 0,8360 do 0,8365,

statystyka $z = 2521,8375$,

$p < 0,001$.

Wartość współczynnika κ jest bardzo wysoka i według np. charakterystyki Kocha i Landisa osiąga się tu prawie perfekcyjną zgodność. Test parametryczny z wykorzystaniem przedstawionej wcześniej statystyki z wskazuje, że należy odrzucić hipotezę zerową, która zakłada wartość współczynnika κ wynoszącą 0. Właściwie zerowa wartość istotności p nie powoduje żadnych problemów, przynajmniej w zakresie przyjęcia tego minimalnego kryterium dla κ . Niska wartość błędu standardowego precyzuje współczynnik w bardzo małym przedziale. Jeżeli przyjąć miarę κ jako sposób oceny poprawności pomiaru, to mamy do czynienia z pomiarem dobrej jakości. Kryterium to powinno być oczywiście uznane za jeden z elementów jakości badania. Prosta konstrukcja współczynnika κ , która mierzy zgodność dwóch ocen, pozwala na szybkie obliczenie, natomiast skale porównawcze umożliwiają praktyczne zastosowanie do porównań. Różnice w empirycznych wartościach brzegowych oraz wyliczenie κ_m wskazują, że dla zaobserwowanych wartości brzegowych można by maksymalnie osiągnąć wartość współczynnika κ równą 0,9826, czyli wysoką wartość bliską jedności, natomiast $1 - \kappa_m$ wskazuje na proporcję zgody niemożliwą do osiągnięcia ze względu na różnicę w wartościach brzegowych. W rozpatrywanym przypadku wartość ta wynosi 0,0174, czyli jest bardzo niska i może wskazywać na brak problemów z granicami poszczególnych kategorii, na co wskazuje Cohen (1960) interpretując ten wskaźnik. Zauważa on, że wartość współczynnika może sygnalizować niejasność określenia wyraźnego podziału na poszczególne kategorie.

Przechodząc do porównania każdej kategorii z osobna, obliczono poszczególne współczynniki κ . W tabl. 5 przedstawiono wartości zmiennych po przekodowaniu na kategorie dychotomiczne oraz odpowiadające im wartości κ . W ko-

lumnach i wierszach oznaczono odpowiednio jako 1 sytuację posiadania stosownego statusu i 0 w przypadku jego braku.

TABL. 5. LICZBA ORAZ WARTOŚĆ WSPÓŁCZYNNIKÓW ZGODNOŚCI WEDŁUG KATEGORII KLASYFIKACJI NA RYNKU PRACY

Posiadanie statusu według samooceny	Posiadanie statusu według MOP	
	1	0
Pracujący, $\kappa = 0,8914$		
1	2884247	168971
0	179740	3203776
Bezrobotni, $\kappa = 0,6659$		
1	318806	174756
0	112008	5831164
Bierni zawodowo, $\kappa = 0,8269$		
1	2639768	250186
0	302165	3244615

Ź r ó d ł o: jak przy tabl. 1.

Najniższą wartość współczynnik zgodności przyjął dla kategorii osób bezrobotnych, co wskazuje na pokaźną zgodność obu pomiarów według skali Kocha i Landisa (1977). Tym samym na tle pozostałych kategorii wyróżnia się on dość zdecydowanie niższą wartością. W analizowanej klasyfikacji jest to więc kategoria najmniej wiarygodna. W definicji MOP do zakwalifikowania osoby jako bezrobotnej jest wymagana aktywność w poszukiwaniu pracy oraz gotowość do jej podjęcia. Niespełnienie któregoś z tych kryteriów kwalifikuje bowiem respondenta do grupy biernych zawodowo, chociaż w jego samoocenie może być osobą bezrobotną. Fakt bycia uczniem lub studentem w potocznym rozumieniu jest również często identyfikowany z biernością zawodową, podobnie jak przebywanie na rencie lub emeryturze. Ponadto podstawą definicji MOP nie jest fakt rejestracji lub jej brak w powiatowych urzędach pracy, tylko posiadanie lub wykonywanie pracy. Rejestracja w urzędzie pracy, a mimo to wykonywanie pracy, np. w szarej strefie, według MOP kwalifikuje osobę jako pracującą. Z kolei brak gotowości do podjęcia pracy w klasyfikacji MOP uwzględnia osobę jako bierną na rynku pracy, pomimo że jest ona ujęta w rejestrze urzędu pracy i występuje w kategorii tzw. bezrobocia rejestrowanego, co jednocześnie może być podstawą odpowiedzi na pytanie dotyczące samooceny.

ZASTOSOWANIE ANALIZY KLASY UKRYTEJ

Innym sposobem porównania wiarygodności pomiaru może być modelowanie statystyczne. Biemer (2004) omawia wykorzystanie modeli statystycznych do oceny zebranych danych. Wskazuje na możliwości osłabienia niektórych założeń, jak np. możliwość modelowania bez założenia, iż jeden z pomiarów pełni rolę punktu odniesienia. Można zatem przyjąć, że w dwukrotnym pomiarze oba jego rezultaty są obciążone błędem. Punktem wyjścia jest założenie, że prawdziwa wartość mierzonej cechy jest ukryta, a wartość zaobserwowana stanowi indyktor, czyli jej wartość manifestowaną. Celem oceny jest znalezienie związ-

ku między obiema wartościami. W przypadku danych dyskretnych stosowana jest analiza klas ukrytych (*Latent Class Analysis*), których omówienie można znaleźć np. w publikacji Brzezińskiej (2015). Analiza będzie zatem prowadziła do oszacowania klasy ukrytej dla trzech mierzonych kategorii jako prawdziwego oszacowania przynależności do danej kategorii klasyfikacji. W zakresie dwóch zmiennych, dwóch ocen przynależności do danej kategorii klasyfikacji pochodzących z dwóch pomiarów, interesuje nas prawdopodobieństwo warunkowe tych pomiarów względem klasy ukrytej. Jeśli przez X_i oznaczymy klasę ukrytą, natomiast przez A_i i B_i odpowiednio pomiar pierwszy oraz drugi ($i = 1, 2, 3$), to szczególnie interesuje nas prawdopodobieństwo warunkowe $P(A=a|X=x)$ oraz $P(B=b|X=x)$, gdzie dla $a \neq b$ szacunek będzie dotyczył błędu pomiaru. Zakłada się, że w ramach danej klasy ukrytej prawdopodobieństwo wynosi w sumie jeden, tzn. $\sum_a P(A=a|X=x) = 1$. Drugie założenie dotyczy lokalnej niezależności.

Oznacza ono, że dla danej klasy ukrytej zmienne obserwowalne są niezależne statystycznie. W przypadku dwóch niezależnych pomiarów przyjmuje się dodatkowo, że prawdopodobieństwo błędów w obydwu pomiarach jest takie samo. Z kolei dla trzech kategorii statusu na rynku pracy mierzonych dwukrotnie chcemy oszacować klasę ukrytą w odniesieniu do każdej z tych kategorii. Liczba parametrów przekracza liczbę komórek w tablicy liczebności (trzeba oszacować po 6 parametrów dla prawdopodobieństwa warunkowego oraz prawdopodobieństwo bezwarunkowe), a tablica składa się 9 komórek, model jest więc niemożliwy do oszacowania. Hui i Walter (1980) zaproponowali metodę stanowiącą jedno z rozwiązań tego problemu. Wprowadza ona zmienną grupującą, która dzieli tablicę obserwowanej częstości dla 3 mierzonych kategorii na 2 grupy ze względu na płeć, tzn. $G = 1$ dla kobiet i $G = 2$ dla mężczyzn. Założeniem jest takie samo prawdopodobieństwo błędu pomiaru w obu grupach oraz $\sum_x P(X=x|G=g) = 1$. Tablica zawiera wówczas 18 liczb i możliwy do osza-

cowania jest model nasycony (trzeba oszacować 18 parametrów). W przykładzie przedstawionym przez Biemera (2004) zmienną wprowadzającą podział dla ocenianych danych jest płeć respondenta. Tabl. 6 przedstawia wynikową tablicę liczebności w zakresie ocenianych klasyfikacji.

TABL. 6. LICZEBNOŚĆ WEDŁUG KLASYFIKACJI POMIARU STATUSU OSOBY NA RYNKU PRACY

Status według samooceny	Status według MOP					
	kobiety			mężczyźni		
	pracujące	bezrobotne	biernie zawodowo	pracujący	bezrobotni	biernie zawodowo
Pracujący	1269252	8213	80104	1614995	10832	69822
Bezrobotni	9096	140452	82258	13421	178354	69981
Biernie zawodowo	98221	58194	1612792	59002	34769	1026976

Do oceny klasy ukrytej wykorzystano program LEM (Vermunt, 1997), który umożliwia w szerokim zakresie analizę klas ukrytych, szacując parametry iteracyjną metodą maksymalizacji prawdopodobieństwa i używając algorytmu EM (*Expectation Maximization*). W celu umożliwienia oceny dopasowania, przyjęto założenie niezależności dwóch pomiarów, tzn. założono, że $P(A = a | X = x) = P(B = b | X = x)$ dla $a = b$. Tym samym zostaje uwolnionych 6 stopni swobody. Do oceny dopasowania można użyć ilorazu wiarygodności statystyki L^2 , która jest porównywana ze statystyką χ^2 przy 6 stopniach swobody. Wartość L^2 jest bardzo wysoka i wynosi 19887,95, a zatem model musi zostać odrzucony ($p < 0,0001$). Tym samym nie można uznać równoległości pomiarów. Jednak można oszacować nasycony model dla obserwowanej tablicy bez założenia o niezależności. Tabl. 7 przedstawia wyniki pomiaru statusu według MOP oraz samooceny.

TABL. 7. WARUNKOWE PRAWDOPODOBIEŃSTWO DLA STATUSU WEDŁUG MOP ORAZ SAMOOCENY

Status „prawdziwy”	Pracujący	Bezrobotni	Bierni zawodowo
Według MOP			
Pracujący	97,78 (0,01)	0,00 (nieokreślony)	2,22 (0,01)
Bezrobotni	4,42 (0,01)	79,30 (0,05)	16,28 (0,04)
Bierni zawodowo	5,43 (0,01)	3,22 (0,01)	91,35 (0,02)
Według samooceny			
Pracujący	100,00 (0,00)	0,00 (nieokreślony)	0,00 (nieokreślony)
Bezrobotni	4,89 (0,04)	95,11 (0,04)	0,00 (nieokreślony)
Bierni zawodowo	2,90 (0,02)	3,10 (0,02)	94,00 (0,03)

U w a g a. W nawiasach podano błąd standardowy.

Ź r ó d ł o: jak przy tabl. 1.

Należy podkreślić, że wyniki są pochodną przyjętego modelu. Jeśli model został błędnie określony, wyniki będą obciążone. Dla $P(A=s|X=s)$ obserwujemy poprawnie zmierzony status, a dla $P(A=a|X=s)$, gdzie $a \neq s$ błędy pomiaru poszczególnych kategorii. Z danych przedstawionych na podstawie spisu reprezentacyjnego wynika, że w przypadku pomiaru statusu według MOP, najniższy wynik ma kategoria osób uznanych za bezrobotne, co potwierdza sugestię wynikającą z indeksu κ Cohena. Całkiem duży odsetek osób (16,28%) został uznany według modelu za bierne zawodowo zamiast bezrobotne, a 4,42% — za pracujące. Różnica dotyczy zatem 20,7% osób dla tej kategorii, a przynajmniej taka sugestia płynie z przyjętego modelu. Z kolei na podstawie samooceny w zmiennej klasyfikującej nie występuje wyraźna różnica w poszczególnych kategoriach.

Wszystkie kategorie należałoby uznać za mające wysoką wiarygodność, chociaż niższą dla osób bezrobotnych i biernych zawodowo. Może to wynikać z faktu, że pomimo wielu wariantów odpowiedzi, ocena następuje na podstawie jednego pytania w kwestionariuszu spisowym. Pomimo zastrzeżeń dotyczących adekwatności modelu i ostrożności w formułowaniu wniosków można jednak stwierdzić, że rozpatrywany model statystyczny pozwala na analizę różnic w klasyfikacjach, jak również na sformułowanie przypuszczeń o możliwych źródłach różnic. Pozwala on również na analizę relacji pomiędzy obserwowaną liczebnością i nieobserwowaną, ukrytą zmienną. Model ten pozwala na głębsze analizy niż analizowany wcześniej współczynnik zgodności.

Podsumowanie

Do oceny zastosowanej klasyfikacji mogą być użyte metody statystyczne polegające na jej wielokrotnym pomiarze. Jednak warunkiem jest niezależność pomiarów. Należy zapewnić podobne czy wręcz takie same warunki pomiarów w celu uznania ich za niezależne w sensie wyciągania wniosków opartych na prawdopodobieństwie. Dane pomiarowe z NSP 2011 pozwalają na uzyskanie w jednym badaniu dwóch klasyfikacji osoby na rynku pracy. Jedną prowadzącą wprost do uzyskania statusu według definicji MOP oraz drugą uzyskaną z pytania o subiektywną samoocenę respondenta. Uzyskując podwójny pomiar można zatem dokonać porównania, ale w interpretacji wyników należy zachować ostrożność. Podstawowym tego powodem są różnice definicyjne w uzyskanych klasyfikacjach. Dodatkowo, w przypadku badań prowadzonych na podstawie wywiadu, dochodzą typowe elementy mające wpływ na wyniki, takie jak:

- odpowiedzi w zastępstwie, tzw. odpowiedzi PROXY, kiedy w gospodarstwie domowym jeden z członów udziela odpowiedzi za innego członka gospodarstwa;
- pamięć oraz efekt nazywany teleskopowym, polegający na umiejscowieniu zdarzenia w innym okresie niż rzeczywiście miało to miejsce. NSP 2011 był przeprowadzony w okresie od 1 kwietnia do 30 czerwca 2011 r. według stanu w dniu 31 marca 2011 r. o godz. 24.00, a więc wraz z upływem czasu odpowiedzi mogły stawać się coraz mniej precyzyjne;
- pytania dotyczące oceny własnej osoby (samoocena) — może wystąpić naturalna rozbieżność między respondentami.

Założenie o niezależności pomiarów nie będzie w tym przypadku prawdopodobnie wypełnione. Przyjęty do analizy model klasy ukrytej potwierdza, że założenie niezależności może nie być spełnione. Niemniej, pamiętając o wszystkich zastrzeżeniach, istotność statystyczna współczynnika zgodności czy niska wartość błędów standardowych w modelu wskazują, że można dokonać oceny zgodności wyników. Należy także pamiętać o dużej liczbie braków danych, co obciąża wyniki. Do porównania brano jednak tylko te osoby, w przypadku których uzyskano odpowiedzi dla obu zmiennych klasyfikacyjnych. Dodatkowo,

liczebność próby jest bardzo wysoka. Zbadano aktywność zawodową prawie 6,5 mln osób.

Wykorzystanie modelowania statystycznego czy współczynnika zgodności byłoby efektywniejsze, gdyby porównanie dotyczyło wyników mierzonych cech na podstawie tych samych definicji. W praktyce raczej nie będzie możliwy powtórny pomiar spełniający warunki niezależności na tak dużej próbie, przy czym wykonalne byłoby powtórne badanie na niewielkiej podpróbie w celach porównania ilościowego. Możliwość dodatkowych źródeł danych do porównań stwarzają również rejestry administracyjne.

mgr Paweł Lańduch — *Urząd Statystyczny w Poznaniu*

LITERATURA

- Biemer, P. (2004). Modeling Measurement Error to Identify Flawed Questions. W: S. Presser, et al., *Methods for Testing and Evaluating Survey Questionnaires. Chapter 12*, Wiley, New York.
- Brzezińska, J. (2015). Analiza klas ukrytych w badaniach sondażowych. *Prace Naukowe Uniwersytetu Wrocławskiego*, nr 384.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*, vol. 20, s. 37—46.
- Eurostat (2014). *ESS handbook for quality reports — 2014 edition*.
- GUS (2008). *Zasady Metodyczne Statystyki Rynku Pracy i Wynagrodzeń*. http://stat.gov.pl/download/cps/rde/xbcr/gus/Zasady_metodyczne_stat_ryнку_pracy_i_wynagrodzen.pdf.
- GUS (2013a). *Ludność. Stan i struktura demograficzno-społeczna. Narodowy Spis Powszechny Ludności i Mieszkań*. http://stat.gov.pl/cps/rde/xbcr/gus/LUD_ludnosc_stan_str_dem_spo_NSP2011.pdf.
- GUS (2013b). *Aktywność ekonomiczna ludności Polski, Narodowy Spis Powszechny Ludności i Mieszkań 2011*. <http://stat.gov.pl/spisy-powszechno/nsp-2011/nsp-2011-wyniki/aktywnosc-ekonomiczna-ludnoscipolski-nsp-2011,12,1.html>.
- Hui, S., Walter, S. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, no. 36, s. 167—171.
- ILO (1982). *Resolution concerning statistics of the economically active population, employment, unemployment, and underemployment adopted by the Thirteenth International Conference of Labour Statisticians*. International Labour Organization, October, http://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/normativeinstrument/wcms_087481.pdf.
- Jarosz-Nowak, J. (2007). Modele oceny stopnia zgody pomiędzy dwoma ekspertami z wykorzystaniem współczynników kappa. *Matematyka Stosowana: matematyka dla społeczeństwa*, nr 8, s. 126—154.
- Koch, G.G., Landis, J.R. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, vol. 33, s. 159—174.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. New York: Duxbury Press.
- ONZ (2006). *Conference of European statisticians recommendations for the 2010 censuses of population and housing*. Prepared in cooperation with the Statistical Office of the European Communities (Eurostat), Genewa, <http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2006/zip.1.e.pdf>.
- Vermunt, J. (1997). *REM: A General Program for the Analysis of Categorical Data*. Tilburg University, Holandia.

Summary. *Multimeasurement of the person's labour market status based on the statistical questionnaire may be used for the assessment of its quality with quantitative methods. In order to evaluate the measurement quality we employ among others, remeasurement or multiple measurements for a random sample of units. Based on the data from the National Population and Housing Census 2011, this article presents appliance of the Cohen's coefficient of contingency and Latent Class Analysis for the assessment of respondent's labour force classification. The questions in the census questionnaire enable the evaluation of persons' labour market status according to International Labour Organisation (ILO) recommendation as well as give an opportunity to respondent's self-assessment. The article focuses on the issue of results' consistency obtained from these two types of information in the part of the Census based on a random sample. Many other aspects relating to random and non-random errors were omitted in the research.*

Keywords: statistical questionnaire, measurement error, index of consistency, latent class.