

STUDIA METODOLOGICZNE

Mirosław SZREDER

Nowe źródła informacji i ich wykorzystywanie w podejmowaniu decyzji

Streszczenie. *Analizie poddano zagadnienia związane z możliwościami wykorzystywania tradycyjnych i nowych źródeł informacji (w tym statystycznych) w podejmowaniu decyzji w warunkach niepewności. Zwrócono uwagę na relację między dostępnością informacji o danej sytuacji decyzyjnej a interpretacją i estymacją prawdopodobieństwa, będącego miarą niepewności. Następnie skoncentrowano się na ocenach wiarygodności i jakości informacji uzyskiwanych z badań statystycznych oraz Big Data wykorzystywanych w podejmowaniu decyzji. Oceny te dają dodatkową perspektywę analizy sytuacji decyzyjnych, zwaną wymiarem wiedzy.*

Słowa kluczowe: podejmowanie decyzji, niepewność, ryzyko, prawdopodobieństwo, badania statystyczne, Big Data.

JEL: C18

Im mniej przewidywalna staje się rzeczywistość, w której żyjemy i funkcjonujemy — zarówno w jej wymiarze społecznym, jak i ekonomicznym — tym więcej uwagi poświęcamy kwestiom niepewności i ryzyka. Przede wszystkim zaś sposobom identyfikacji zagrożeń i zarządzania ryzykiem. Artykuł nie zawiera dyskusji na temat natury ryzyka i rozumienia samego pojęcia „ryzyko”. Literatura opisująca to zagadnienie jest obszerna, a ustalenie wspólnego stanowiska autorów precyzującego charakter ryzyka lub jednoznacznie je definiującego jest trudne. Niezależnie jednak od dziedziny naukowej, którą autorzy reprezentują, i od wywodów na temat ryzyka, jakie prowadzą, istnieje u nich wspólny ele-

ment, który dostrzegają w omawianym pojęciu. Elementem tym jest niepewność. Niepewność albo miary jej natężenia stanowią rzadko kwestionowaną składową kategorii „ryzyko”.

U niektórych autorów ryzyko jest rozumiane jako funkcja prawdopodobieństwa wystąpienia pewnego zdarzenia oraz ważności (wartości) wynikającej z tego straty. W ten sposób ujmują je Kroik, Malara, Malara i Sobol-Wojciechowska (2013). U innych z kolei jest funkcją trzech zmiennych: określonego scenariusza zdarzeń lub działań, prawdopodobieństwa tego scenariusza oraz mierzalnych konsekwencji zrealizowania się tego scenariusza (Kaplan, 1991)¹. Część autorów w podobny sposób wyróżnia trzy zmienne określające wielkość ryzyka, ale kładą oni nacisk na stan wiedzy osoby, której sytuacja ryzyka dotyczy. Do zbioru wspomnianych trzech zmiennych zaliczają: określone konsekwencje (C), miarę niepewności (Q) odnoszącą się do C (zwykle prawdopodobieństwo) oraz wiedzę (K) wspierającą Q i C , z przypisanymi tej wiedzy stopniami wiarygodności (siły) (Aven, 2009).

Słuszne jest postulowanie, aby komponentem ryzyka była niepewność, a nie jej konkretna miara — prawdopodobieństwo (Aven, 2009, s. 70). Ryzyko jest bowiem funkcją niepewności (nieprzewidywalności), a nie funkcją konkretnej miary tej niepewności, która może, lecz nie musi być w każdej sytuacji określona. Kluczową kwestią w analizie ryzyka są zasoby informacji, jakimi dysponuje decydent, a także specyficzne dla niego sposoby ich przetwarzania. Stopień niepewności decydenta, często rozumiany jako stan jego umysłu odnoszący się do danej sytuacji, zmienia się zarówno pod wpływem nowo uzyskanych informacji, jak i zdolności ich połączenia z wcześniej posiadaną wiedzą. Ważną rolę odgrywają w tych działaniach źródła informacji statystycznej zarówno tradycyjne, jak i najnowsze — Big Data.

Celem tego opracowania jest scharakteryzowanie, w jaki sposób nowe zasoby informacji, a także łatwiejsze niż w przeszłości sposoby ich pozyskiwania i przetwarzania wpływają na szacowanie tego elementu ryzyka, który określiliśmy jako niepewność. Kiedy używa się terminu „szacowanie” lub „estymacja” w odniesieniu do niepewności, to zwykle ma się na myśli najbardziej oczywistą i najpopularniejszą miarę niepewności, jaką jest prawdopodobieństwo. I chociaż na prawdopodobieństwie większość prac badawczych z zakresu pomiaru stopnia niepewności się koncentruje, to w praktyce podejmowania decyzji nierzadko odwołujemy się do podejścia jakościowego, a nie ilościowego. Wówczas celem staje się nie precyzyjny, ilościowy pomiar niepewności, lecz redukcja stopnia niepewności poprzez zdobywanie wiedzy z wykorzystaniem różnych technik jakościowych, m.in. obserwacji, wywiadów pogłębionych i grupowych.

W artykule rozważymy możliwości i zagrożenia wynikające z wykorzystania informacji pochodzących z badań próbkowych, a także zasobów informacji

¹ Popularnie te trzy elementy ująć można w formie pytań: co może się zdarzyć?, na ile prawdopodobne jest, że się zdarzy?, jakie będą konsekwencje, jeżeli się to zdarzy?

określanych mianem Big Data, w celu redukcji stopnia niepewności w podejmowaniu decyzji w warunkach ryzyka.

INTERPRETACJE PRAWDOPODOBIENSTWA I ICH PRAKTYCZNA UŻYTECZNOŚĆ

Prawdopodobieństwo — mimo że w wielu sytuacjach bywa podstawową, a najczęściej także jedyną miarą wyrażającą stopień niepewności lub przekonania decydenta o możliwości zajścia danego zdarzenia — nie jest kategorią łatwo poddającą się kwantyfikacji. Aksjomatyka Kołmogorowa, charakteryzująca trzy podstawowe właściwości funkcji zwanej prawdopodobieństwem (P), nie rozstrzyga jednoznacznie o sposobie przyporządkowania zdarzeniom losowym odpowiednich wartości funkcji P . To z kolei otwiera przestrzeń do tworzenia, zarówno w sferze analitycznej jak i praktycznej, różnych interpretacji prawdopodobieństwa oraz wynikających z nich sposobów odwzorowywania zdarzeń losowych w zbiór wartości funkcji P . Do najbardziej popularnych interpretacji zalicza się: klasyczną, częstościową (statystyczną) i personalistyczną (subiektywną)². Każda z nich odegrała w przeszłości, a często nadal odgrywa istotną rolę w teorii i praktyce zarządzania ryzykiem.

Interpretacja klasyczna, stworzona przez de Laplace'a w XVIII w., mimo że najstarsza, bywa często uważana za najbardziej odpowiadającą potocznemu rozumieniu prawdopodobieństwa i najbliższą intuicji zwykłego człowieka. Interpretacja ta jest oparta na założeniu jednakowo możliwych zdarzeń, czyli jednakowych szans realizacji podstawowych zdarzeń elementarnych. Prawdopodobieństwo zdarzenia definiuje się w niej jako stosunek liczby przypadków sprzyjających danemu zdarzeniu do liczby wszystkich, jednakowo możliwych, przypadków. Jest to koncepcja, która w praktyce może być rozszerzona na zbiory zdarzeń nieskończonych i nieprzeliczonych, ale jej ograniczeniem pozostaje założenie o jednakowych szansach realizacji każdego elementarnego zdarzenia. W praktycznych sytuacjach decyzyjnych założenie to jest trudne do spełnienia. Okoliczności zdarzeń w sferze gospodarczej czy społecznej rzadko bywają identyczne, a ponadto naturalna dynamika zmian rzeczywistości w różny sposób kształtuje i zmienia poszczególne jej elementy. Na ogół mamy do czynienia z sytuacjami bardziej złożonymi i mniej jednorodnymi niż modele rzutu monetą czy kostką do gry, urny z kulami lub talii kart. Trudno jest np. założyć, że istnieją identyczne pod względem ryzyka ubezpieczeniowego osoby, nieruchomości czy samochody. Mimo że ubezpieczający się kierowcy grupowani są w zróżnicowane klasy ryzyka w systemach bonus-malus, to w danej klasie nie znajdują się kierowcy o identycznym, lecz jedynie podobnym stopniu ryzyka. Te właśnie praktyczne ograniczenia klasycznej interpretacji prawdopodobieństwa doprowadziły do powstania interpretacji nazywanej częstościową lub statystyczną.

² Część autorów wyróżnia też interpretację logiczną, odwołującą się do prawdziwości logicznej zdań na temat określonych zdarzeń (Aven, 2014; Szreder, 1994).

Podczas gdy interpretacja klasyczna odwołuje się do znanych właściwości mechanizmu generującego zdarzenia losowe (symetryczność monety lub kostki), interpretacja częstościowa prawdopodobieństwa nawiązuje do powtarzalności zdarzeń. Warunkiem określenia (oszacowania) prawdopodobieństwa wyróżnionego zdarzenia jest możliwość obserwowania długiego ciągu zdarzeń losowych, z których jednym jest wyróżnione zdarzenie. Twórca tej koncepcji Richard von Mises (brat słynnego ekonomisty Ludwiga von Misesa) stwierdzał: *Racjonalną koncepcję prawdopodobieństwa, która stanowi jedyną podstawę do obliczeń prawdopodobieństwa, stosuje się tylko do tych problemów, w których albo te same zdarzenia powtarzają się wiele razy, albo duża liczba identycznych zdarzeń jest rozpatrywana w tym samym czasie. (...) Aby móc stosować rachunek prawdopodobieństwa, musimy posiadać nieskończony ciąg jednostkowych obserwacji* (von Mises, 1957, s. 11 — tłum. własne). Prawdopodobieństwo danego zdarzenia — zgodnie z interpretacją częstościową — jest granicą (lub asymptotyczną wartością) względnej częstości zdarzeń elementarnych sprzyjających temu zdarzeniu w długim ciągu powtarzalnych zdarzeń losowych. Takie ujęcie prawdopodobieństwa zawiera w sobie pewien element abstrakcyjny, jakim jest długi, dążący do nieskończoności ciąg powtarzalnych i obserwowalnych zdarzeń. W praktyce pojęcie tej granicy zostaje zastąpione estymowaną na podstawie dostatecznie licznej próby wartością częstości względnej zdarzeń.

Częstościowa interpretacja prawdopodobieństwa stosowana jest nie tylko w klasycznej teorii wnioskowania statystycznego (stąd jej alternatywna nazwa — interpretacja statystyczna), ale także w wielu innych dziedzinach nauki i praktyki, w tym w ubezpieczeniach. Konstrukcja tablic wymieralności i obliczenia prawdopodobieństwa dożycia określonego wieku przez osobę o ustalonych cechach odbywa się na podstawie dużej liczby danych demograficznych z przeszłości. Podobnie wiele innych zagadnień aktuarialnych w ubezpieczeniach życiowych i majątkowych rozwiązywanych jest przy wykorzystaniu informacji o powtarzalnych zdarzeniach z przeszłości. Popularności tej interpretacji prawdopodobieństwa przysporzył rozwój technologii informatycznej, w tym zwłaszcza ogromne możliwości technik symulacji komputerowej. Komputerowe modele symulacji deterministycznej bądź stochastycznej są w stanie coraz wierniej odwzorowywać rzeczywisty przebieg analizowanych zjawisk i zdarzeń. Na tej podstawie zaś coraz trafniej oceniać można prawdopodobieństwo scenariuszy lub zdarzeń interesujących badaczy.

Ograniczeniem pozostaje jednak założenie o identycznych warunkach, w jakich realizują się powtarzalne zdarzenia. Krytycy tej interpretacji słusznie zwracają uwagę na to, że powtórzenie nawet jeden raz doświadczenia w identycznych okolicznościach jest zwykle bardzo trudne lub wręcz niemożliwe. Traktowanie ciągu rzeczywistych zdarzeń ekonomicznych jako powstałych w tych samych warunkach może więc być założeniem zbyt daleko odbiegającym od rzeczywistości. To, co jest możliwe w symulacjach komputerowych, nie zawsze daje się w prosty sposób przełożyć na rzeczywistość. W stosunku jednak do

klasycznej interpretacji prawdopodobieństwa, omawiana koncepcja stanowi od dawna uznany i w praktyce najczęściej stosowany sposób szacowania prawdopodobieństwa, którego prawdziwych wartości nigdy nie poznamy. W dobie zwiększających się zasobów informacji na świecie interpretacja ta wydaje się szczególnie dobrze dostosowana do absorbowania wiedzy zawartej w dużych zbiorach danych.

Najbardziej kłopotliwe są jednak sytuacje, w których ryzyko pojawia się w zdarzeniach niemających żadnej historii. Brak możliwości obserwowania i szacowania częstości względnej zdarzeń jednostkowych czyni interpretację częstościową beзуżyteczną. W jej miejsce pojawiła się już ponad 250 lat temu, przede wszystkim w wyniku badań Bayesa (1701—1761), interpretacja personalistyczna prawdopodobieństwa, zwana inaczej subiektywną. Punktem wyjścia do stworzenia tej interpretacji było przekonanie jej twórców o potrzebie wykorzystania każdej użytecznej wiedzy posiadanej przez decydenta (wiedzy *a priori*), gdy nie ma możliwości obserwowania lub eksperymentowania w zbiorach adekwatnych zdarzeń losowych. Bernoulli (1713) w pracy pt. *Ars Conjectandi* pisał o prawdopodobieństwie jako o stopniu zaufania do realizacji danego zdarzenia na podstawie posiadanej wiedzy o ogólnych okolicznościach tego zdarzenia, a Leibniz (1646—1716) dodawał: *Ale to, co prawdopodobne, sięga dalej; trzeba je wydobyć z natury rzeczy, a mniemanie osób o wielkim autorytecie jest jedną z rzeczy, które mogą się przyczynić do uprawdopodobnienia jakiegoś mniemania* (Leibniz, 1955, s. 307). Przez subiektywne prawdopodobieństwo tego, że jakiś sąd na temat zdarzenia *A* jest prawdziwy, rozumie się stopień pewności lub przekonania danej osoby o prawdziwości tego sądu. W przeciwieństwie do interpretacji częstościowej, prawdopodobieństwo subiektywne jest traktowane personalistycznie i warunkowo ze względu na przeszłość i na całe otoczenie analizowanego zdarzenia. Jedynym ograniczeniem w stosowaniu tej interpretacji jest to, aby zbiór prawdopodobieństw był dla danej osoby logicznie zgodny i spójny³.

W praktyce interpretację subiektywną prawdopodobieństwa odnosi się najczęściej do kwantyfikacji opinii wybranych ekspertów z danej dziedziny. W literaturze z zakresu statystyki oraz podejmowania decyzji znanych jest wiele technik wydobywania od ekspertów ocen na temat określonych zdarzeń w formie rozkładów prawdopodobieństwa lub ich charakterystyki⁴. Podobnie za dobrze rozwinięte należy uznać metody agregacji opinii ekspertów, zarówno w rozkładach skokowych jak i ciągłych. Jednym z celów agregacji lub uśredniania tych opinii jest ich obiektywizacja, czyli dążenie do ograniczenia w ostatecznej ocenie prawdopodobieństwa czynników subiektywnych. Dla wielu kry-

³ Szerzej o postulatach spójności i zgodności w odniesieniu do personalistycznego prawdopodobieństwa pisał m.in. Winkler (1967).

⁴ Znaczne zasługi w tym zakresie ma Daniel Kahneman, psycholog i laureat Nagrody Nobla w dziedzinie ekonomii z 2002 r. oraz współautor wielu jego prac Amos Tversky. O tym także Tyszka (2010).

tyków interpretacji personalistycznej nie jest to jednak wystarczające. Powołując się na obiektywizm i racjonalność poznania naukowego, jako podstawowe jego atrybuty, nie godzą się oni na włączanie do procesu badawczego lub procesów podejmowania decyzji elementów subiektywnych (personalistycznych). Zauważyć jednak warto, że od dawna w wielu dziedzinach z powodu braku innych możliwości pomiaru niepewności interpretację tę stosuje się z powodzeniem⁵.

Poza tym aktualizacja wiedzy decydenta na podstawie nowych informacji, także tych o charakterze subiektywnym — jeżeli odbywa się zgodnie z twierdzeniem Bayesa — następuje z uwzględnieniem stopnia niepewności (niespójności) tych nowych informacji. Mniej pewne informacje (rozkłady prawdopodobieństwa o większym rozproszeniu) uzyskują mniejsze wagi w stosunku do tych, które charakteryzuje mały stopień niepewności (rozproszenia). Ważnymi jednak kwestiami, które nie powinny być lekceważone zwłaszcza przy szybko rosnących zasobach dostępnych informacji, są jakość informacji i wiarygodność wiedzy uzyskiwanej z nowych wcześniej nieznanymi źródłami. Nie odnosi się to wyłącznie do personalistycznej interpretacji prawdopodobieństwa, ale i do pomiaru stopnia niepewności w ogóle.

BADANIA STATYSTYCZNE A MOŻLIWOŚCI POMIARU I REDUKCJI NIEPEWNOŚCI

Współcześnie w ocenie niepewności obserwuje się tendencję do wykraczania poza wymiar wyłącznie probabilistyczny. Źródłem takiego podejścia jest coraz bardziej odczuwalna potrzeba nie tylko efektywnego pomiaru i kwantyfikacji niepewności w formie prawdopodobieństwa, ale także uwzględnienia „siły wiedzy”, na której oparte są oceny prawdopodobieństwa. Niektórzy autorzy postulują wręcz włączenie do analizy ryzyka kolejnej perspektywy badawczej, mianowicie tzw. wymiaru wiedzy⁶. Odnosi się on do szczegółowego badania wiarygodności i aktualności wiedzy, na której oparto oceny probabilistyczne, ale też do kumulowania wiedzy poprzez odbieranie i przetwarzanie sygnałów oraz ostrzeżeń, a także wykorzystania gromadzonych doświadczeń. Taki nacisk na ocenę jakości wiedzy jest spowodowany szybko rosnącymi zasobami informacji, rejestrowanymi w różnych formach, o różnym stopniu uporządkowania i wewnętrznej spójności.

Rewolucja technologiczna ostatnich dekad sprawiła, że znacznie łatwiej niż w przeszłości uzyskuje się obecnie dane ze źródeł wtórnych, a dodatkowo bardziej przystępne dla badaczy staje się projektowanie i realizacja badań statystycznych (głównie próbkowych), czyli zdobywanie wiedzy ze źródeł pierwotnych. Wraz z upowszechnieniem się badań statystycznych, wykorzystujących łatwo dostępne i przyjazne pakiety obliczeniowe (SPSS, Statistica), a przede wszystkim komunikację internetową, pojawiły się istotne zagrożenia związane

⁵ Na ten temat również Kowalczyk (2010) i Szreder (2011).

⁶ Traktują o tym zwłaszcza prace norweskiego badacza Avena (2014).

z jakością uzyskiwanych informacji, dlatego postulowaną w analizie ryzyka dodatkową perspektywę — „siłę wiedzy” — uznać należy za ważną.

Pierwszym zagrożeniem, na które warto zwrócić uwagę, są daleko idące uproszczenia metodyczne o różnym charakterze, wpływające na jakość gromadzonej wiedzy. Jednym z przejawów takiego uproszczenia jest zastępowanie badań jakościowych w problemach dotyczących motywacji i postaw jednostek w populacji szybszymi w projektowaniu i realizacji badaniami ilościowymi. Redukcja niepewności, odnosząca się do zachowań osób w określonych sytuacjach, wymaga najczęściej pogłębionych badań jakościowych, łączących elementy psychologii lub socjologii. Nie jest ich w stanie zastąpić prosta deklaracja badanych osób w sprawie tego, jak zachowałyby się w danych warunkach. Dotyczy to zarówno zachowań konsumentów na rynku, jak i różnych postaw i działań ludzi w życiu społecznym. W szczególności nie może wyrazić złożoności tych postaw zwykłe wskazanie jednego wariantu odpowiedzi w kwestionariuszu badania ilościowego. Nawet wówczas, gdy badanie ilościowe uznać można za adekwatne w danej sytuacji decyzyjnej, niektórzy badacze uciekają się do zbyt dużych uproszczeń czy to w formułowaniu pytań, czy w skalowaniu odpowiedzi. Taki przykład można znaleźć m.in. w pracy Sokołowskiej (2014), gdzie zastrzeżenia w badaniu ankietowym (w rozdz. 5) budzi zarówno niewystarczająca liczba pytań w analizowanym problemie, jak i ich skalowanie (połowa pytań ma prostą skalę dychotomiczną: tak/nie). Upowszechnienie się badań próbkowych (ankietowych, sondażowych) nie stanowi samo w sobie żadnego zagrożenia. Jest to zresztą zrozumiała reakcja badaczy na coraz łatwiej dostępne i doskonalsze instrumentarium badawcze. Badania te stają się zagrożeniem dopiero wówczas, gdy zamiast stanowić uzupełnienie pogłębionego badania jakościowego, a więc spełniać funkcję komplementarną, są traktowane jako substytut tych ostatnich⁷.

Innym dość powszechnym zagrożeniem metodycznym w dobie szerokiego dostępu do Internetu jest mało krytyczna ocena zamieszczonych tam wyników badań statystycznych, w tym pochodzących z badań o charakterze próbkowym. Część badaczy korzystających z tego źródła nie zwraca wystarczającej uwagi na charakter próby, której dotyczą wyniki. Tymczasem ma to zasadnicze znaczenie dla rozstrzygnięcia o możliwościach uogólnienia danych próbkowych na całą populację interesującą badacza. Dla młodych naukowców szczególną pokusę stanowią sondy na portalach internetowych, które najczęściej nie dają żadnej możliwości uogólnień. Próby dobiera się tam na zasadzie autoselekcji respondentów, czyli dalekie są one od spełnienia wymogu reprezentatywności. Próba oszacowania na ich podstawie prawdopodobieństwa określonych zachowań

⁷ Interesujące, że zjawisko to dostrzegali już w 1981 r. Antoni Sułek, profesor, który w swoim wystąpieniu na VI Ogólnopolskim Zjeździe Socjologicznym w Łodzi mówił: *Tym, co naganne, nie jest — rzecz jasna — samo korzystanie z danych ankietowych, lecz to, że socjologia empiryczna jest przez nie zdominowana, że w konkretnych badaniach są one prawie zawsze jedynym rodzajem zbieranych danych* (Sułek, 2011, s. 238).

w większej populacji musi być obciążona błędem i to nieznaną wielkością. Nie wiele znajdziemy w Internecie badań, w których rzetelnie wyjaśnia się ograniczenia takiej metodyki badawczej. Jednym z wyjątków jest portal CNN QuickVote informujący odbiorców, że podawane wyniki sondy internetowej nie stanowią badania naukowego i odzwierciedlają opinie jedynie tych użytkowników Internetu, którzy zdecydowali się wziąć w niej udział. A dalej ostrzega, że nie można zakładać, iż wyniki te reprezentują opinie wszystkich użytkowników Internetu, a tym bardziej całego społeczeństwa. Takim zastrzeżeniem powinny być opatrzone wszystkie sondy internetowe, których liczbę trudno byłoby nawet zliczyć. Bezkrytyczne podejście do nich, a w szczególności przypisywanie im waloru reprezentatywności w odniesieniu do określonej zbiorowości, nie znajduje racjonalnego uzasadnienia. „Siła wiedzy” wynikająca z wyników takiej sondy jest niewielka, a sama wiedza mało użyteczna w konkretnych sytuacjach zarządzania ryzykiem.

Problem uogólniania danych próbkowych, w kontekście wykorzystanej techniki doboru próby, jest oczywiście szerszy i odnosi się do kwestii umiejętności oceny przez użytkownika zalet i ograniczeń schematu próbkowania. Nie można go bowiem traktować w sposób nazbyt arbitralny, czyli odmawiać próbom nieprobabilistycznym (nielosowym) jakichkolwiek cech reprezentatywności⁸. W grupie technik nieprobabilistycznych znajdują się bowiem takie, których podstawowym postulatem jest postępowanie w celu uzyskania próby reprezentatywnej ze względu na najważniejsze cechy populacji w konkretnym problemie badawczym. Jedną z tych technik jest dobór kwotowy próby, popularny w wielu badaniach społecznych i rynkowych. Właściwości próby kwotowej czy stopień jej reprezentatywności zależą przede wszystkim od stopnia poznania badanej populacji. Im więcej wiedzy o tej populacji ma badacz przed rozpoczęciem badania, tym większe szanse, że stanowić ona będzie dobrej jakości reprezentację interesującej go zbiorowości. W Wielkiej Brytanii np. próby kwotowe są częściej stosowane w badaniach społecznych od prób losowych (w tym opinii publicznej).

Jakość informacji uzyskanych z badania próbkowego wyrażana jest najczęściej w kategoriach błędów, jakimi są obciążone jego wyniki. Występują tutaj także zagrożenia dla badaczy, którzy niekiedy zbyt powierzchownie podchodzą do tego zagadnienia. Zakładają, że dominującym albo jedynym błędem badania próbkowego jest eksponowany w komunikatach błąd próbkowania (losowy). Skłania ich do tego fakt, że w konkretnym schemacie próbkowania można określić jego wielkość, a także to, że znana jest jego natura i istota. Jest to błąd, który powstaje na skutek stosowania niedoskonałej techniki próbkowania, czyli takiej, która nie daje gwarancji uzyskania struktury próby w pełni zgodnej ze strukturą populacji. W rzeczywistości bywa on często najmniejszym składnikiem całkowitego błędu badania próbkowego, a do rangi najważniejszego składnika tego ostatniego urasta współcześnie błąd braków odpowiedzi. Nie zwracając uwagi na ten błąd, a także na inne błędy o charakterze nielosowym, badacz zbyt opty-

⁸ Szerzej na ten temat — Szreder (2012).

mistycznie może ocenić jakość uzyskanych wyników, a w konsekwencji także „siłę wiedzy”, opartej na tych informacjach. Jest to tym bardziej istotne, że znaczenie błędów o charakterze nielosowym znacznie się w ostatnich latach zwiększyło⁹.

BIG DATA JAKO NOWE ŹRÓDŁO INFORMACJI STATYSTYCZNYCH

W nieodległej przeszłości o możliwościach redukcji niepewności w podejmowaniu decyzji w warunkach ryzyka decydowały przede wszystkim możliwości uzyskania użytecznych informacji. Nacisk kładło się przede wszystkim na ilość i zakres dostępnych informacji. Przełom dokonał się w ostatnich dwóch dekadach pod wpływem nowych technologii. Każę to spojrzeć na problem nieco inaczej. Big Data, czyli sposób zdobywania nowej wiedzy i poznawania otaczającej nas rzeczywistości, który może być zrealizowany w dużej skali, dzięki najnowszym możliwościom gromadzenia i przetwarzania wielkich zbiorów danych zmienił podejście decydentów do kwestii oceny przydatności i użyteczności zasobów informacyjnych. Zasadniczym problemem przestaje być brak informacji, a wyzwaniem stają się ich selekcja, ocena jakości oraz agregacja. Rozważmy Big Data nie tyle jako nowe zjawisko współczesnej cywilizacji, lecz także jako nowe źródło informacji, ze swoimi zaletami i ograniczeniami w podejmowaniu decyzji.

Do najistotniejszych zalet Big Data jako źródła danych zaliczyć trzeba nieznane w przeszłości ułatwienia w dotarciu do dużych baz, w tym do danych administracyjnych. Zaleta ta nie ogranicza się jedynie do możliwości uzyskania tych danych — nierzadko obejmuje także sposoby i narzędzia ich przetwarzania, a także śledzenie aktualizacji w czasie rzeczywistym (bez opóźnień). Dane administracyjne, ze względu na różne formalne wymogi i rygory, które obowiązują ich „producentów”, stanowią najczęściej wiarygodne i dobrej jakości źródło informacji. O ich rosnącej roli w poznawaniu rzeczywistości może świadczyć miejsce, jakie zajmują w programach dużych badań statystyki publicznej. Otóż, coraz częściej dane z rejestrów urzędowych są w stanie skutecznie wypełnić przestrzeń informacji statystycznej, która przez dziesięciolecia zarezerwowana była dla badań pełnych (spisów powszechnych) lub badań reprezentacyjnych. Nawet w tradycyjnych ogólnokrajowych spisach ludności i mieszkań znaczna część informacji, która w przeszłości zbierana była przez rachmistrzów, obecnie uzyskiwana jest z rejestrów urzędowych. Dostępność danych z tych rejestrów sprawia, że w wielu sytuacjach decyzyjnych stanowią one główne źródło informacji albo przynajmniej podstawowe źródło danych wtórnych — niestety często niedoceniane¹⁰. Warto zdać sobie sprawę z tego, że część rejestrów urzędowych, charakteryzując określone zbiory jednostek i ich rozwój w czasie, tworzy wartościową dla decydentów i badaczy strukturę danych panelowych.

⁹ Szerzej na ten temat — Szreder (2015); Stefanowicz i Cierpień-Wolan (2015).

¹⁰ *Rządowe dane administracyjne są niemal na pewno niewystarczająco wykorzystywane* (Einav i Levin, 2013, s. 9). Możliwości zastosowań danych administracyjnych w badaniach statystycznych zostały też przedstawione w monografii — Wallgren i Wallgren (2007).

Big Data ma także inne zalety, dzięki którym zyskuje przewagę nad tradycyjnymi zbiorami danych. Jedną z nich jest możliwość gromadzenia informacji o prowadzonej w Internecie aktywności danej zbiorowości. Dzięki temu, że wszystkie operacje przeprowadzane on-line, a praktycznie każde kliknięcie jest rejestrowane, powstają nowe zbiory potencjalnie użytecznych danych. Gdy doda się do tego informacje o lokalizacji geograficznej jednostek czy relacjach łączących osoby aktywne na portalach społecznościowych, to okaże się, że decydent działa często w znacznie bardziej komfortowych warunkach niż kiedyś.

W wielu zagadnieniach społecznych i ekonomicznych główne źródło niepewności decydentów obejmuje przyszłe zachowania interesujących ich zbiorowości. I w tej roli właśnie, czyli w przewidywaniu postępowania jednostek tych zbiorowości, najlepiej sprawdzają się techniki Big Data. Na podstawie bardzo dużych zbiorów danych statystycznych, a także pomiaru współzależności i korelacji konstruowane są tzw. modele lub algorytmy predyktywne. Istotą tych modeli nie jest wyjaśnianie przyczynowo-skutkowe działań lub zjawisk (tak jak w typowych modelach ekonometrycznych), lecz wskazywanie najbardziej prawdopodobnych scenariuszy przyszłych działań lub rozwoju zjawisk¹¹. Modele takie są z powodzeniem stosowane m.in. w kryminalistyce do zapobiegania przestępstwom, gdy zachodzi największe prawdopodobieństwo ich wystąpienia, a także w ubezpieczeniach i bankowości, gdzie stanowią podstawę do określenia wielkości przyszłego ryzyka klienta. Ich prawdziwy walor ujawnia się wówczas, kiedy są w stanie aktualizować i przetwarzać duże ilości nieustrukturyzowanych danych w określone stopnie ryzyka w pełni automatyczny sposób. Dość często jest to możliwe w czasie rzeczywistym (np. samouczące się sztuczne sieci neuronowe), co dla decydenta ma szczególne znaczenie.

Możliwości Big Data sprawiają, że zapomina się niekiedy o konieczności krytycznej ich oceny. Niektórzy badacze lub decydenci, pozostając pod wrażeniem dużej ilości danych, skłonni są odstąpić od ich ewaluacji, sądząc, że ilość może zrekompensować wątpliwą lub słabą jakość. Tymczasem naturą danych Big Data jest najczęściej ich nieuporządkowanie, pewien bezład, a do tego bardziej złożona struktura od tradycyjnej, występującej w ekonomii i ekonometrii. Rzadko Big Data daje się wyrazić w formie macierzy o określonej liczbie zmiennych i obserwacji. Wymiarów tych jest najczęściej więcej, a zmienne rzadko spełniają jeden z najważniejszych we wnioskowaniu statystycznym postulatów — niezależność. Ocena jakości Big Data musi brać pod uwagę te aspekty. Uwzględniać powinna również to, że nie zawsze dane o prawie całej populacji charakteryzują ją lepiej niż poprawnie wykonane badanie reprezentacyjne. Dzieje się tak dlatego, że wszystkie kategorie błędów nielosowych, jakie mogą obciążać wyniki badania statystycznego, dotyczą zarówno badań wyczerpujących, jak i próbkowych.

¹¹ *Dumna przyczynowość musi na scenie ustąpić miejsca uniżonej korelacji* (Mayer-Schönberger i Cukier, 2014, s. 186).

Poza tym nawet mała frakcja niezbadanych jednostek populacji może istotnie zniekształcić ogólny wynik, o czym nieraz przekonali się wyborcy w naszym kraju, gdy wynik z obliczeń głosów w 90% obwodów po zakończeniu głosowania znacznie odbiegał od wyniku ze wszystkich obwodów¹². Nawet bardzo duże próby nie dają gwarancji, że otrzymane na ich podstawie wyniki lepiej charakteryzują populację aniżeli małe, ale starannie dobrane próby reprezentacyjne. Sądzę, że gdy frakcja transakcji zakupowych dokonywanych przez Internet zbliży się do 90%, nie stanie się to dla urzędów statystycznych wystarczającą przesłanką do całkowitego porzucenia tradycyjnego badania koszyka inflacyjnego. Informacje o transakcjach internetowych mogą być natomiast wykorzystane do wnioskowania opartego na połączonych zbiorach danych.

BIG DATA WSPARCIEM DLA TRADYCYJNYCH BADAŃ STATYSTYCZNYCH

W sytuacjach niepewności korzystanie ze wszelkich użytecznych i wiarygodnych informacji należy uznać za racjonalne. Wkraczania Big Data w dziedziny, w których wcześniej dominowały badania statystyczne (pełne lub próbkowe) nie trzeba koniecznie rozpatrywać w perspektywie rywalizacji tych dwu źródeł danych¹³. Bardziej naturalna jest bowiem relacja komplementarności między nimi. Big Data, podobnie jak wspomniane wcześniej rejestry administracyjne, mogą stanowić — i w praktyce już stanowią — wartościowe dopełnienie badań próbkowych. W szczególności mogą one dostarczyć ważnych informacji w sytuacjach zagrożeń badania próbkowego dużymi błędami nielosowymi, np. błędami pokrycia lub braków odpowiedzi. Innymi słowy, dodatkowe informacje o populacji, potrzebne do efektywnego zastosowania mechanizmów ważenia danych z próby lub kalibracji danych, mogą mieć swoje źródło w Big Data.

Nie tylko w podejmowaniu decyzji, ale także w empirycznych badaniach ekonomicznych Big Data ma do spełnienia funkcję raczej komplementarną, a nie substytucyjną. W badaniach ekonomicznych bowiem istotne jest przede wszystkim wyjaśnianie mechanizmów zjawisk obserwowanych na rynku i wokół niego, a więc ich przyczyn i logicznych przesłanek. Big Data z kolei koncentruje się głównie na powiązaniach między zmiennymi, rejestrując duże liczby cech, które są w stanie opisywać najistotniejsze współzależności między nimi. Te współzależności mogą mieć charakter sztuczny (pozorny) lub przyczynowo-skutkowy. Nie zawsze wnoszą one wartościową wiedzę do poznania mechanizmu przyczynowego danego zjawiska. Dobrze natomiast potrafią się przysłużyć przewidywaniom stanów lub zachowań, na podstawie algorytmów predykcyjnych.

¹² Za kontrowersyjne należy uznać stwierdzenie z monografii Mayer-Schönbergera i Cukiera (2014, s. 145), odnoszące się do jednego z opisywanych badań, brzmiące: *Najważniejsze było, że badania nie opierały się na próbie losowej, ale na czymś bliskim N=całość.*

¹³ Takie cechy nadają im w swojej książce Mayer-Schönberger i Cukier (2014).

Rozwój Big Data przebiega dynamicznie i dlatego rola wspomagająca badania statystyczne, którą obecnie dostrzega się w tym zjawisku, będzie zapewne ewoluować. Jak szybko — zależy zarówno od tempa upowszechniania się oprogramowania do analiz bardzo dużych zbiorów danych, jak i wypracowania przez naukę nowego podejścia do syntezy wiedzy z połączonych zbiorów danych o różnym charakterze i rozmaitej specyfice (w tym mało uporządkowanych), ale bardzo dużych bazach danych.

Podsumowanie

Stan niepewności w podejmowaniu decyzji jest często charakteryzowany jako stan niepełnej informacji. Współcześnie, gdy niemal w każdym problemie decyzyjnym łatwiej zdobywa się informacje niż bywało to w przeszłości, ważna staje się umiejętność oceny jakości informacji zmniejszających stan niepewności decydenta. Niektórzy autorzy, m.in. Aven (2014), postulują włączenie do analizy ryzyka jeszcze jednej perspektywy badawczej, mianowicie tzw. wymiaru wiedzy. Chodzi o określenie wiarygodności i jakości informacji („siły wiedzy”), które wykorzystane zostaną w podejmowaniu decyzji w warunkach ryzyka. Ocena użyteczności i jakości informacji obejmuje zarówno opis probabilistyczny stopnia niepewności, jak i wszelkie inne elementy podejmowania decyzji, w których wykorzystuje się informacje o charakterze ilościowym. Ważne jest właściwe ocenienie jakości danych, które pochodzą z badań statystycznych (wyczerpujących i próbkowych), a także tych, które określa się mianem Big Data. Przedstawiona w artykule analiza cech obu tych źródeł danych prowadzi do wniosku, że obecnie powinny one pełnić względem siebie funkcje komplementarne.

prof. dr hab. Mirosław Szreder — Uniwersytet Gdański

LITERATURA

- Aven, T. (2009). *Risk analysis and risk management. Basic concepts and principles. vol. 2, Hefte 1*, s. 57—73.
- Aven, T. (2014). *Risk, Surprises and Black Swans: Fundamental Ideas and Concepts in Risk*. Nowy Jork: Routledge.
- Einav, L., Levin, J.D. (2013). The Data Revolution and Economic Analysis. *NBER Working Paper, no. 19035*. Pobrane z <http://www.nber.org/papers/w19035>, dostęp 22.04.2017 r.
- Kaplan, S. (1991). Risk Assessment and Risk Management — Basic Concepts and Terminology. W: R.A. Knief, V. Broadhead Briant (red.), *Risk Management: Expanding Horizons in Nuclear Power and Other Industries*, s. 11—28. Boston: Hemisphere Publishing Corporation.
- Kowalczyk, H. (2010). O eksperckich ocenach niepewności w ankietach makroekonomicznych. *Bank i Kredyt, nr 41*, s. 101—122
- Kroik, J., Malara, Z., Malara, M.J., Sobol-Wojciechowska, J. (2013). *Ryzyko. Perspektywa jakościowa*. Oficyna Wydawnicza Politechniki Wrocławskiej.

- Leibniz, W.G. (1955). *Nowe rozważania dotyczące rozumu ludzkiego. t. 2*. Warszawa: PWN.
- Mayer-Schönberger, V., Cukier, K. (2014). *Rewolucja, która zmienia nasze myślenie, pracę i życie*. Warszawa: MT Biznes.
- Sokołowska, E. (2014). *Alternative Investments in Wealth Management. A comprehensive Study of the Central and East European Market*. Springer International Publishing.
- Stefanowicz, B., Cierpiał-Wolan, M. (2015). Błędy przetwarzania danych. *Wiadomości Statystyczne, nr 9*, s. 23—29. Warszawa: GUS i PTS.
- Sulek, A. (2011). *Obrazy z życia socjologii w Polsce*. Warszawa: Oficyna Naukowa.
- Szreder, M. (1994). *Informacje a priori w klasycznej i bayesowskiej estymacji modeli regresji*. Wydawnictwo Uniwersytetu Gdańskiego.
- Szreder, M. (2011). Prawdopodobieństwo subiektywne. *Miesięcznik Ubezpieczeniowy, nr 12*, s. 42 i 43.
- Szreder, M. (2012). Wybór próby badawczej. *Marketing i Rynek, nr 11*, s. 24—27.
- Szreder, M. (2015). Zmiany w strukturze całkowitego błędu badania próbkowego. *Wiadomości Statystyczne, nr 1*, s. 4—12. Warszawa: GUS i PTS.
- Tyszka, T. (2010). *Decyzje. Perspektywa psychologiczna i ekonomiczna*. Warszawa: Wydawnictwo Naukowe Scholar.
- von Mises, R. (1957). *Probability, Statistics and Truth*. Londyn: Allen & Unwin.
- Wallgren, A., Wallgren, B. (2007). *Register-based Statistics. Administrative Data for Statistical Purposes*. Nowy Jork: John Wiley & Sons.
- Winkler, R. (1967). The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association, vol. 62*, s. 1105—1120.

Summary. *The opportunities of using traditional and new sources of information, including statistical data, in decision-making under uncertainty were analysed. Special attention was paid to the relationship between the availability of information about the particular decision-making situation and the interpretation as well as estimation of probability as a measure of uncertainty. Subsequently, an assessment of information credibility and quality obtained from statistical research and Big Data used in decision-making was highlighted. This kind of evaluation offers additional perspective on the analysis of decision-making situation, called the knowledge dimension.*

Keywords: decision-making, uncertainty, risk, probability, statistical surveys, Big Data.