Wprowadzenie

Projektowanie i wdrażanie działań rozwojowych w różnych aspektach życia społeczno-gospodarczego oraz monitorowanie ich efektów, jak też poszerzanie się zakresu i doskonalenie narzędzi badań naukowych, wymaga coraz wszechstronniejszych i precyzyjniejszych informacji statystycznych. Rośnie zatem zapotrzebowanie odbiorców na szczegółowe i obszerne dane statystyczne obrazujące interesujący ich wycinek otaczającej rzeczywistości oraz umożliwiające efektywną analizę jego stanu i procesów w nim zachodzących.

Dane gromadzone w trakcie badań statystycznych czy ujmowane w rejestrach administracyjnych i stamtąd pozyskiwane zawierają jednak liczne informacje dotyczące indywidualnych cech jednostek, w tym ich bezpośrednich identyfikatorów. Cechy te podlegają ochronie prawnej i są objęte bezwzględną tajemnicą statystyczną, określoną w naszym kraju w art. 38 ustawy z dnia 29 czerwca 1995 r. o statystyce publicznej. W wypadku danych osobowych ich ochrona stała się kwestią szczególnie donośną społecznie na skutek wejścia w życie w dniu 25 maja 2018 r. nowej regulacji Unii Europejskiej – Rozporządzenia Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych), w środkach społecznego przekazu, a także w tej pracy, określanego skrótem RODO[1]. Rozporządzenie to spowodowało także uchwalenie ustawy z dnia 10 maja 2018 r. o ochronie danych osobowych, która jednakże – mocą art. 175 – zachowała wiele kluczowych zapisów swej poprzedniczki (tzn. ustawy z dnia 29 sierpnia 1997 r. o ochronie danych osobowych). Dlatego też przed udostępnieniem czy opublikowaniem wynikowych informacji statystycznych należy zgromadzone dane poddać weryfikacji w celu zminimalizowania w największym możliwym stopniu ryzyka ujawnienia bądź odtworzenia przez użytkowników udostępnianych zasobów wrażliwych danych identyfikujących jednostkę statystyczną (np. respondenta). Równocześnie trzeba zadbać o jak najlepszą użyteczność finalnie udostępnianych informacji dla ich użytkownika. Postępowanie takie nazywa się **kontrolą ujawniania danych** (ang. *statistical dis-*

[1] Pochodzącym od słów: rozporządzenie w sprawie ochrony danych osobowych.

closure control – SDC). Oczywiście pierwszą i najprostszą czynnością wykonywaną w ramach SDC jest usunięcie kluczowych identyfikatorów jednostek (takich jak imię, nazwisko oraz numer PESEL w wypadku osób czy nazwa oraz numer REGON i numer identyfikacji podatkowej (NIP) dla podmiotu gospodarczego).

W dzisiejszych realiach to jednak o wiele za mało, a utrwalone w wieloletniej praktyce proste reguły ukrywania okazują się dalece niewystarczające. Dzieje się tak ze względu na – zazwyczaj znaczną – liczbę zmiennych opisujących dane jednostki, co pociąga za sobą bardzo dużą liczbę możliwych kombinacji wariantów zmiennych wyrażonych na skali nominalnej lub porządkowej. Istnieje zatem spore ryzyko, że wystąpią wśród nich kombinacje unikatowe, zawierające w skrajnych sytuacjach pojedyncze przypadki, co w konsekwencji pozwoli na ich identyfikację. Szczególnie istotne jest to w przypadku mikrodanych (tzn. odpersonalizowanych danych jednostkowych) czy wielowymiarowych kostek danych OLAP[2], na które zapotrzebowanie – zwłaszcza ze strony środowisk naukowo-badawczych – rośnie i należy się spodziewać, że nadal będzie rosło. Ryzyko to może być jeszcze większe, jeśli w badaniu lub rejestrze gromadzone są informacje mierzone na skalach różnicowej (zwanej także przedziałową) lub ilorazowej. Mamy tutaj więc do czynienia ze zmiennymi ilościowymi, zatem można mówić m.in. o udziale komponentów struktur w odpowiednich wielkościach ogółem (np. o udziale liczby zatrudnionych w sektorze publicznym w liczbie zatrudnionych ogółem). Ponadto należy wziąć pod uwagę także to, że potencjalny użytkownik może dysponować również innymi, niezależnymi zasobami danych, które mogą mu tę identyfikację ułatwić. Co więcej, ze względu na obowiązujące regulacje prawne dane teleadresowe podmiotów gospodarczych (także osób fizycznych prowadzących działalność gospodarczą[3]) są jawne, a zatem w ich wypadku ochrona danych wrażliwych musi być prowadzona inaczej.

Wszystko to sprawiło, że kontrola ujawniania danych wypracowała liczne zaawansowane metody (takie jak nakładanie szumu, postrandomizacja, kontrolowane dopasowanie tablic itp.) oparte na statystyce matematycznej i stosownych rozwiązaniach informatycznych oraz stała się istotną składową metodologii badań statystycznych. Kontrola ta wyodrębnia trzy główne obszary ochrony danych, do których specyfiki dostosowane są tworzone i doskonalone jej narzędzia:

– mikrodane,
– dane tabelaryczne,
– wyniki analiz.

Rozważania przedstawione w niniejszym opracowaniu podążają właśnie w tych trzech kierunkach – charakteryzują specyficzne cechy każdego z nich, źró-

[2] OLAP – ang. *online analytical processing* (przetwarzanie analityczne on-line).

[3] W wypadku Centralnej Ewidencji i Informacji o Działalności Gospodarczej (CEIDG) – o ile osoby te nie wyrażą wobec tego sprzeciwu podczas rejestracji działalności – na mocy art. 37 ust. 1 i art. 38 ustawy z dnia 2 lipca 2004 r. o swobodzie działalności gospodarczej.

dła i fundamenty ochrony stosownych danych (w tym formalne), narzędzia teoretyczne i informatyczne służące do tego celu oraz organizacyjne i technologiczne aspekty udostępniania danych stwarzających ryzyko identyfikacji jednostek lub odtworzenia danych wrażliwych. Nie zapomniano w tym kontekście także o danych wynikowych publikowanych w formie statystyk opisowych, rezultatów estymacji modeli ekonometrycznych czy ilustracji graficznych. Autorzy starali się ukazać złożoność zagadnienia i różnorodność możliwych reguł ujawniania. W kompleksowej ocenie efektów SDC istotną rolę odgrywa szacowanie ryzyka ujawnienia z jednej strony i oczekiwanej straty informacji na skutek ukrycia bądź zniekształcenia danych wrażliwych – z drugiej. Wypracowanie rozwiązań umożliwiających optymalizację kontroli ujawniania danych z punktu widzenia minimalizacji obu tych wielkości jest podstawowym i najważniejszym celem SDC.

Ważność wspomnianego balansu znalazła odzwierciedlenie także w strukturze niniejszego opracowania. Składa się ono z sześciu rozdziałów. W pierwszym omówiono ogólne koncepcje SDC oraz definicje z tym związane, a także regulacje prawne dotyczące ochrony danych wrażliwych stosowane w różnych krajach, również w Polsce. Zaprezentowano tutaj także najważniejsze typy udostępnianych danych wynikowych, a także rolę metadanych, paradanych oraz danych dodatkowych w SDC. Tematem rozdziału drugiego jest istota i ważność ryzyka ujawnienia informacji wrażliwych oraz ocena jego poziomu. Ukazano w nim także odmienności występujące w tym zakresie między mikrodanymi a danymi zagregowanymi w postaci tablic częstości i wielkości czy wyników analiz. Rozdział trzeci poświęcono z kolei szczegółowej charakterystyce metod i technik kontroli ujawniania danych wynikowych w przedstawionych powyżej trzech typach. Wskazano w nim ponadto na zagrożenia dla poufności danych mogące wystąpić w wyniku publikowania statystyk opisowych, ilustracji i wyników analiz w opracowaniach statystycznych oraz sposoby przeciwdziałania możliwościom odtworzenia danych wrażliwych. Zagadnienia dotyczące straty informacji znalazły się w rozdziale czwartym. Scharakteryzowano tutaj istotę tego problemu oraz najistotniejsze rodzaje miar oceny owej straty, z pewnymi oryginalnymi propozycjami, a także wpływ straty informacji na jakość estymacji dokonywanej na podstawie danych poddanych SDC. W rozdziale piątym można znaleźć szczegółowe omówienie – wraz ze stosowną egzemplifikacją – narzędzi informatycznych stosowanych w SDC, przede wszystkim programów τ–Argus i μ–Argus oraz pakietów środowiska R: `sdcTable` i `sdcMicro`. W rozdziale szóstym z kolei uwaga została skoncentrowana na organizacji kontroli dostępu do danych i zasadach jej realizacji. W szczególności scharakteryzowano typy udostępnianych mikrodanych, sposoby organizacji funkcjonowania punktów dostępu oraz stosownych zabezpieczeń, przebieg efektywnej kontroli dostępu i zakres odpowiednich uprawnień. Całość wieńczy stosowne podsumowanie, w którym wskazano najistotniejsze konkluzje oraz postulaty dotyczące stosowania SDC. Dla wygody

czytelnika na końcu opracowania zamieszczono słownik występujących w nim pojęć.

Ze względu na wieloaspektowość zagadnienia z jednej strony i jednocześnie ograniczenia objętości pracy z drugiej strony publikacja została opracowana jako swoisty przewodnik dla metodologów projektujących badania statystyczne oraz odpowiedzialnych za ich jakość i bezpieczeństwo zgromadzonych w ich wyniku informacji. Mamy też nadzieję, że będzie ona ważnym źródłem poznawczym dla użytkowników informacji statystycznych w zakresie wiedzy o celach i konsekwencjach ochrony poufności danych. Stąd duży wybór przykładów i praktycznych omówień przedstawianych zagadnień, które pozwalają lepiej zrozumieć przekazywane treści. Czytelników zainteresowanych bardziej szczegółowym opisem zagadnień omawianych w monografii autorzy zachęcają do wykorzystania literatury wymienionej w spisie bibliograficznym, której spora część jest dostępna w internecie.

Autorzy wyrażają podziękowanie członkom Komisji Metodologicznej Głównego Urzędu Statystycznego za cenne uwagi, które w znacznym stopniu przyczyniły się do podniesienia jakości niniejszej publikacji, oraz współtwórcy programu τ–Argus, Peterowi-Paulowi de Wolfowi ze Statistics Netherlands (Centraal Bureau voor de Statistiek, Centralne Biuro Statystyczne Holandii) za udzielenie wsparcia technicznego, a także Bernhardowi Meindlowi ze Statistics Austria za umożliwienie dołączenia proponowanych przez nas metod oceny straty informacji do zasobów pakietu `sdcMicro` oraz Janowi Kubackiemu reprezentującemu środowisko statystyków łódzkich za cenne sugestie i opinie, które przyczyniły się do ulepszenia jakości tej książki.

Szczególne podziękowania należą się także prof. dr hab. Elżbiecie Gołacie oraz dr hab. Grażynie Dehnel, prof. UEP, z Katedry Statystyki Uniwersytetu Ekonomicznego w Poznaniu za naukowe i instytucjonalne wspieranie działań, które doprowadziły do powstania niniejszej publikacji.

Wyrazy wdzięczności należą się również innym, niewymienionym z imienia i nazwiska osobom, reprezentującym zarówno środowisko akademickie, Główny Urząd Statystyczny, jak i krajowe urzędy statystyczne, bez wiedzy i doświadczenia których niniejsza publikacja byłaby uboższa.

Autorzy żywią także nadzieję, że oddawane do rąk Czytelników opracowanie okaże się istotną pomocą w dziele ochrony informacji oraz przyczyni się do efektywnego rozwoju mechanizmów bezpiecznego upowszechniania danych statystycznych.

Andrzej Młodak, Michał Pietrzak, Tomasz Klimanek,
Tomasz Józefowski, Paweł Lańduch