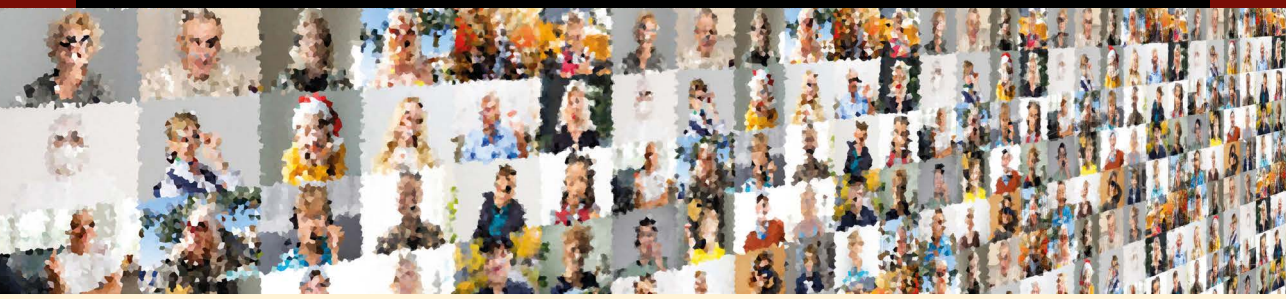
Andrzej Młodak ▪ Michał Pietrzak
Tomasz Klimanek ▪ Tomasz Józefowski ▪ Paweł Lańduch

Poufność a użyteczność informacji statystycznych

Dylematy ochrony udostępnianych danych



WYDAWNICTWO UEP
UNIWERSYTET EKONOMICZNY W POZNANIU

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

Poufność a użyteczność informacji statystycznych

Dylematy ochrony udostępnianych danych

WYDAWNICTWO UEP
UNIWERSYTET EKONOMICZNY W POZNANIU
Poznań 2023

Publikacja dofinansowana ze środków budżetu państwa w ramach programu Ministra Edukacji i Nauki pod nazwą Doskonała nauka, nr projektu DNM/SP/548365/2022, kwota dofinansowania 17 472,40 zł, całkowita wartość projektu 19 430,40 zł.

Autorzy
Andrzej Młodak
Michał Pietrzak
Tomasz Klimanek
Tomasz Józefowski
Paweł Lańduch

Sugerowane cytowanie:
Młodak, A., Pietrzak, M., Klimanek, T., Józefowski, T. i Lańduch, P. (2023). *Poufność a użyteczność informacji statystycznych. Dylematy ochrony udostępnianych danych*. Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu. https://doi.org/10.18559/978-83-8211-168-2

ISBN 978-83-8211-167-5
eISBN 978-83-8211-168-2
https://doi.org/10.18559/978-83-8211-168-2
https://doi.org/10.18559/978-83-8211-168-2-summary

WYDAWNICTWO UNIWERSYTETU EKONOMICZNEGO W POZNANIU
ul. Powstańców Wielkopolskich 16, 61–895 Poznań
tel. 61 854 31 54, 61 854 31 55
www.wydawnictwo.ue.poznan.pl, e-mail: wydawnictwo@ue.poznan.pl
adres do korespondencji: al. Niepodległości 10, 61–875 Poznań

Skład: FNCE Wydawnictwo Naukowe

Druk: Zakład Graficzny Uniwersytetu Ekonomicznego w Poznaniu
ul. Towarowa 53, 61–896 Poznań, tel. 61 854 38 06, 61 854 38 03

Spis treści

Wykaz skrótów

Skrót	Termin w języku angielskim	Znaczenie
ABS	Australian Bureau of Statistics	australijskie biuro statystyczne
ANOVA	analysis of variance	analiza wariancji
ASA	American Statistical Association	Amerykańskie Towarzystwo Statystyczne
BMP	BitMaP	mapa bitowa, format plików graficznych w grafice rastrowej
CART	classification and regression tree	drzewo klasyfikacji i regresji
CBS	Centraal Bureau voor de Statistiek	Centralne Biuro Statystyczne Holandii (znane na arenie międzynarodowej także jako Statistics Netherlands)
CDEP	The Committee on Digital Economy Policy	Komitet ds. Polityki Gospodarki Cyfrowej
CEIDG		Centralna Ewidencja i Informacja o Działalności Gospodarczej
CEREM	Centre for Research on Economic Microdata	Centrum Badań Mikrodanych Gospodarczych (Holandia)
CES	Center for Economic Research	Centrum Badań Ekonomicznych (w Stanach Zjednoczonych)
CGADP	combined general additive data perturbation	metoda GADP oparta na wielowymiarowej dystrybuancie łącznego rozkładu
CGM	computer graphics metafile	komputerowy metaplik graficzny, format wymiany graficznych danych wektorowych
CKM	cell-key method	metoda kluczy komórkowych
CPS	Centre for Policy Related Statistics	Centrum Statystyki dla celów Polityki Społeczno-Gospodarczej
CSV	comma-separated values	wartości rozdzielone przecinkiem, format danych
CTA	controlled tabular adjustment	kontrolowane dopasowanie tablic
DDI	data documentation initiative	inicjatywa na rzecz dokumentacji danych
DLL	dynamic-link library	dynamiczna biblioteka łącz
DP	differential privacy	różnicowanie prywatności

Skrót	Termin w języku angielskim	Znaczenie
EASD	Enhancing access to and sharing of data	Poprawa dostępu do danych i ich udostępniania, dokument OECD
EBIL	entropy-based information loss measure	miara straty informacji oparta na entropii
EMF	enhanced metafile	ulepszony metaplik, format wymiany graficznych danych wektorowych
EPS	encapsulated PostScript	schermetyzowany PostScipt, format wymiany graficznych danych wektorowych
ESS	European Statistical System	Europejski System Statystyczny
Euratom		Europejska Wspólnota Energii Atomowej (także EWEA)
Eurostat		Urząd Statystyczny Unii Europejskiej
EWG		Europejska Wspólnota Gospodarcza
FCS	fully conditional specification	specyfikacja w pełni warunkowa
FHNW	Fachhochschule Nordwestschweiz	Wyższa Szkoła Zawodowa Północno-Zachodniej Szwajcarii
FRIBS	Framework regulation for integrating business statistics	Ramowe regulacje w sprawie integracji statystyki działalności gospodarczej
GADP	general additive data perturbation	uogólnione addytywne zakłócenie danych
GAFAM	Google, Apple, Facebook, Amazon, Microsoft	Łączne określenie grupy pięciu firm o światowym zasięgu dysponującymi największymi zasobami dnaych: Google, Apple, Facebook, Amazon, Microsoft
GAN	generative adversarial network	generatywna sieć kontradyktoryjna
GUS		Główny Urząd Statystyczny
GUI	graphical user interface	graficzny interfejs użytkownika
HH	household	gospodarstwo domowe
HiTaS	hierachical tables suppresion, a heuristic approach to cell suppression in hierarchical tables	heurystyczne podejście do łączenia komórek w tablicach hierarchicznych
ILP	integer linear programming	programowanie całkowitoliczbowe
IPSO	information preserving statistical obfuscation	statystyczne zaciemnianie zachowujące informacje
ISI	International Statistical Institute	Międzynarodowy Instytut Statystyczny
JPEG	joint photographic experts group	wspólna grupa ekspertów fotograficznych, format kompresji grafiki rastrowej
LP	linear programming	programowanie liniowe

Skrót	Termin w języku angielskim	Znaczenie
MASSC	micro agglomeration, substitution, subsampling and calibration	mikroaglomeracja, podstawianie, podpróbkowanie i kalibracja
MCMC	Markov chain Monte Carlo	algorytm Monte Carlo łańcucha Markowa
MD	maximum-distance criterion	kryterium największej odległości
MDAV	multivariate microaggregation based on maximum distance to average vector	wielowymiarowa mikroagregacja oparta na maksymalnej odległości od wektora średnich
NIP		Numer Identyfikacji Podatkowej
NP	nondeterministic polynomial problem	niedeterministycznie wielomianowy problem
OECD	Organisation for Economic Co-operation and Development	Organizacja Współpracy Gospodarczej i Rozwoju
OLAP	online analytical processing	przetwarzanie analityczne on-line
ONS	Office for National Statistics	krajowy urząd statystyczny Wielkiej Brytanii
ONZ		Organizacja Narodów Zjednoczonych
PBSSP		Program badań statystycznych statystyki publicznej
PCA	principal component analysis	analiza głównych składowych
PESEL		Powszechny Elektroniczny System Ewidencji Ludności
PKD		Polska Klasyfikacja Działalności
PNG	portable network graphics	przenośna grafika sieciowa, format plików graficznych w grafice rastrowej.
POLTAX		polski kompleks baz danych dotyczących podatników płacących podatki do budżetu państwa obejmujący m.in. ewidencję podatników i płatników podatków, windykacji należności, mandatów karnych, transakcji w zakresie nieruchomości itp.
PRAM	the post-randomization method	metoda postrandomizacyjna
RAF	remote access facility	system zdalnego dostępu
RDC	Research Data Centre	Centrum Danych Naukowych
REGON		powszechny rejestr podmiotów gospodarki narodowej; czasem tym akronimem określa się potocznie identyfikator nadany podmiotowi w powyższym rejestrze
RMDM	robust Mahalanobis distance based microaggregation	odporna mikroagregacja oparta na odległości Mahalanobisa
RODO		Rozporządzenie o Ochronie Danych Osobowych

Skrót	Termin w języku angielskim	Znaczenie
ROMM	random orthogonal matrix masking	maskowanie losową macierzą ortogonalną
SBS	structural business statistics	strukturalna statystyka działalności gospodarczej
SDC	statistical disclosure control	kontrola ujawniania danych
SDMX	Statistical Data and Metadata Exchange	wymiana danych i metadanych statystycznych
SNZ	Statistics New Zealand	nowozelandzki urząd statystyczny
SPSS	statistical package for the social sciences	pakiet statystyczny dla nauk społecznych
SSHRC	Social Sciences and Humanities Research Council	Komitet Nauk Społecznych i Humanistycznych (w Kanadzie)
SWOT	strengths, weaknesses, opportunities, threats	analiza mocnych stron, słabości, szans i zagrożeń
TERYT		rejestr podziału terytorialnego kraju
TIFF	tagged image file format	etykietowany format pliku z obrazem, format plików graficznych w grafice rastrowej
TNO	Nederlandse Organisatie voor toegepast-natuurwetenschap-pelijk onderzoek	Holenderska Organizacja Stosowanych Badań Naukowych
TRS	targeted record swapping	celowana wymiana rekordów
UE	European Union	Unia Europejska
VPN	virtual private network	wirtualna sieć prywatna
UNECE	United Nations Economic Commission for Europe	Komisja Gospodarcza Organizacji Narodów Zjednoczonych dla Europy (znana również jako Europejska Komisja Gospodarcza ONZ)
WE		Wspólnota Europejska
WMF	Windows metafile	metaplik Windows, format wymiany graficznych danych wektorowych
ZHAW	Zürcher Hochschule für Angewandte Wissenschaften	Zurychska Szkoła Nauk Stosowanych

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

1

Koncepcja ochrony danych i definicje podstawowych pojęć

W rozdziale tym zaprezentowano koncepcje i fundamenty kontroli ujawniania danych oraz jej podstawy formalne, prawne i etyczne. W szczególności omówiono istotę zjawiska ujawnienia, rodzaje użytkowników danych z punktu widzenia SDC, typologię danych wynikowych z uwzględnieniem ich specyfiki w kontekście ochrony informacji wrażliwych oraz balans między ryzykiem a użytecznością. Dokonano przy tym przeglądu najważniejszych rozwiązań i regulacji w tym zakresie stosowanych w różnych krajach europejskich. W dalszej kolejności zaprezentowano najistotniejsze w świetle SDC rodzaje danych jednostkowych oraz rolę metadanych i paradanych w tymże procesie.

1.1. Kontrola ujawniania danych statystycznych

1.1.1. Ujawnienie i ryzyko ujawnienia

Potrzeba wypracowania zasad i metod chroniących dane ich dostawców w badaniach statystycznych wynika zarówno z przepisów prawa, jak i z zasad etycznych, na podstawie których tworzone są kodeksy dobrych praktyk oraz reguły postępowania. U jej podłoża leży stosowana powszechnie zasada ochrony danych dotyczących respondentów i przez nich przekazywanych oraz wykorzystywania pozyskanych informacji jednostkowych wyłącznie do celów statystycznych. Z drugiej jednak strony celem prowadzenia badań statystycznych jest jak najlepsze informowanie społeczeństwa oraz organów władzy o aktualnej sytuacji społeczno-gospodarczej, a także umożliwienie prowadzenia analiz badawczych instytutom i uczelniom. Stąd starania krajowych urzędów statystycznych i innych gestorów o publikowanie wyników na jak najbardziej szczegółowych poziomach agregacji oraz o zapewnienie dostępu do jak największego zakresu danych. Te dwa przeciwstawne cele leżą u podstaw wypracowania metodologicznych ram dla kontroli ujawniania danych w procesowym ujęciu badania statystycznego. Duncan i in. (2001) przedstawili zobowiązanie do poufności wobec podmiotów, od których gromadzone są dane, jako prawo respondenta do autonomii wynikające

z etyki, a także z przesłanek pragmatycznych dotyczących jakości uzyskiwanych danych. Cox (1995) zwrócił uwagę na problem poufności w badaniach przedsiębiorstw, który uznał za balansowanie między poufnością a prawem społeczności do informacji, co w konsekwencji prowadzi do zanegowania zapewnienia poufności jako absolutu.

Pojęcie *ujawnienia* określa poznanie przez podmiot (osobę lub instytucję) faktów na temat innego podmiotu (osoby lub instytucji), do których nie uprawnia go ani obowiązujące prawo, ani zasady wynikające z innych norm. Ujawnienie może nastąpić poprzez dostęp podmiotów do danych statystycznych, na przykład na skutek publikacji wyników badań. Podmiot można tu również rozumieć jako jednostkę obserwacji, co najczęściej dotyczy osób, gospodarstw domowych, rodzin, przedsiębiorstw oraz organów administracji. Rozróżnia się trzy typy ujawnienia danych:

- **identyfikację** – następuje ustalenie związku między udostępnionymi danymi a podmiotem; podmiot zostaje zidentyfikowany, co skutkuje poznaniem przez osobę nieuprawnioną danych na jego temat, które są uznane za poufne,
- **ujawnienie atrybutu** – na podstawie udostępnionych danych możliwe jest poznanie dodatkowej charakterystyki podmiotu lub grupy podmiotów oraz ich identyfikacja; Templ i in. (2014) podali następujący przykład: jeśli szpital ujawni, że wszystkie pacjentki w wieku 56–60 lat mają raka, to wystarczy znajomość wieku pacjentki – bez znajomości personaliów – żeby znać charakterystykę medyczną pacjentki tego szpitala,
- **ujawnienie poprzez wnioskowanie** – wystąpi wówczas, gdy na podstawie publikowanych danych jest możliwe oszacowanie charakterystyki podmiotu z większą dokładnością, niż zakłada to publikacja; szacunek ów może być wykonany za pomocą metod wnioskowania statystycznego o wysokim stopniu precyzji, na przykład przy użyciu metod regresji charakteryzujących się wysokim stopniem dopasowania modelu do danych empirycznych (Domingo-Ferrer, 2009).

Z pojęciem ujawnienia wiąże się nieodłącznie problem ryzyka ujawnienia, czyli prawdopodobieństwa zajścia zdarzenia polegającego na tym, że dany podmiot pozna charakterystykę innego podmiotu na podstawie udostępnionych danych, naruszając tym samym jego autonomię w zakresie prawa do prywatności. Relację między ryzykiem ujawnienia a użytecznością informacji omówiono w podrozdziale 1.3, z kolei pomiar i ocenę ryzyka ujawnienia danych szczegółowo przedstawiono w rozdziale 2 niniejszej monografii.

Warto zwrócić uwagę na trudności związane z tłumaczeniem terminu angielskiego *statistical disclosure control* na język polski przy zachowaniu krótkiej trzywyrazowej formy i jednoczesnym ukazaniu całego kontekstu pojęciowego wyrażenia. Po pierwsze należy zauważyć, że sformułowanie *disclosure* może

mieć – i nierzadko ma – nieco szersze znaczenie niż ujawnianie. Można bowiem wskazać przykłady pozycji literatury, gdzie *disclosure* zostało przetłumaczone jako udostępnienie (Litwiński, 2009, s. 180; RODO, 2016, art. 4, ust. 2), przekazanie (Jarosz, 2013), przesłanie, rozpowszechnianie (RODO, 2016, art. 4, ust. 2), upublicznienie (Czech, 2017, s. 195) bądź nawet „jakakolwiek możliwość zapoznania się z danymi osobowymi" (Bielak-Jomaa i Lubasz, 2017, s. 538). Przyznać trzeba, że w podstawowym akcie prawnym regulującym funkcjonowanie statystyki oficjalnej w Polsce – ustawie z dnia 29 czerwca 1995 r. o statystyce publicznej – nie występuje sformułowanie ujawnianie. Wyrażeniem, które w kontekście dbałości o poufność danych zostało użyte w ustawie, jest udostępnienie (por. art. 2, art. 10 oraz przede wszystkim art. 38). Tłumaczenie *disclosure control* jest już nieco łatwiejsze, wskazuje bowiem – chociaż nie w sposób bezpośredni – na kontekst ryzyka ujawnienia poufnych danych, które może wystąpić podczas udostępniania zgromadzonych danych. Gestor danych w dzisiejszych uwarunkowaniach musi mieć możliwości zarządzania tym ryzykiem. Zatem to ryzyko w pewnym zakresie może – i powinno – zostać poddane kontroli za pomocą wprowadzonych rozwiązań prawnych czy organizacyjnych, ale także metod i technik ochrony danych zastosowanych przez wykwalifikowany zespół analityków.

Po drugie zaś przymiotnik *statistical* ma obecnie także szerszy kontekst niż tylko odniesienie do funkcjonowania krajowych urzędów statystycznych. Postulowanie stosowania metod SDC dotyczy obecnie bowiem także tych podmiotów gospodarczych, które ze względu na prowadzoną działalność dysponują ogromnymi zasobami danych o osobach fizycznych czy firmach: banków, towarzystw ubezpieczeniowych, podmiotów funkcjonujących w systemie ochrony zdrowia, firm należących do tzw. GAFAM[4] czy platform streamingowych itp. W ich wypadku także można mówić o statystyce, w najszerszym rozumieniu tego słowa, jako o ilościowych metodach badania prawidłowości występujących w zbiorowościach oraz o charakteryzowaniu tych prawidłowości za pomocą liczb. Należy zatem przyznać, że na przykład sformułowanie kontrola przestrzegania tajemnicy statystycznej byłoby interesującym rozwiązaniem w kontekście tłumaczenia sformułowania *statistical disclosure control*, jednakże wyłącznie na gruncie funkcjonowania statystyki publicznej w Polsce. Biorąc natomiast pod uwagę szersze spektrum instytucji życia społeczno-gospodarczego zainteresowanych potencjalnie możliwością wykorzystania zagadnień przedstawionych w niniejszej publikacji, autorzy będą stosowali tłumaczenie *statistical disclosure control* jako kontrola ujawniania danych statystycznych – z podkreśleniem, że termin ten rozumieją jako kontrolę ryzyka ujawnienia informacji, które powinny być poufne, zawartych w udostępnionych danych statystycznych, mogących przybrać różne formy: mikrodanych, danych tabelarycznych, wykresów, wyników analiz i innych.

[4] Google, Apple, Facebook, Amazon, Microsoft.

Przebieg procesu kontroli ujawniania danych statystycznych jest uwarunkowany tym, dla jakiej postaci wynikowych danych statystycznych ma zostać zapewniona ochrona poufności danych. W związku z tym można wyróżnić następujące rodzaje kontroli ujawniania danych statystycznych:

- **kontrolę ujawniania mikrodanych**, która jest przeprowadzana na zbiorach danych jednostkowych,
- **kontrolę ujawniania tablic statystycznych** – obejmującą dane zagregowane w postaci tablic statystycznych (ale także agregaty w innym układzie, np. siatkę kilometrową czy hiperkostki),
- **kontrolę ujawniania wyników analiz** (ang. *output checking*) – polegającą na sprawdzeniu pod kątem zachowania poufności danych przetworzonych do postaci różnych wyników analiz (charakterystyk liczbowych, statystyk opisowych, komentarzy analitycznych, wykresów statystycznych, w tym kartogramów i kartodiagramów, infografik itp.).

1.1.2. Strata informacji

Proces kontroli ujawniania danych statystycznych jest zawsze kompromisem pomiędzy dbałością o zapewnienie poufności danych a ich użytecznością. Niestety, zastosowanie którejkolwiek z metod SDC powoduje w zasadzie w każdym wypadku zmniejszenie użyteczności danych poprzez ubytek zasobu informacyjnego w stosunku do oryginalnych danych. Ubytek ten nazywa się stratą informacji. W idealnym przypadku stratę informacji ocenia się pod kątem potrzeb i zastosowań końcowych użytkowników danych. Jednakże różni użytkownicy końcowi danych mogą mieć bardzo różne oczekiwania co do udostępnionych danych i zastosowania ich; ustalenie wszystkich potencjalnych zastosowań udostępnionych danych może nie być możliwe. Czasami nawet te potrzeby informacyjne mogą być krańcowo odmienne, np. jeden użytkownik będzie zainteresowany jak najbardziej szczegółowymi przekrojami podziału terytorialnego, podczas gdy drugi użytkownik będzie potrzebował szczegółowej struktury wiekowej populacji bez konieczności dysponowania przekrojami terytorialnymi. Strata informacji jest zatem pojęciem subiektywnym, gdyż w idealnym ujęciu stratę informacji ocenia się pod kątem potrzeb i zastosowań końcowych użytkowników danych. Mimo to powstały określone miary utraty informacji, zależne od rodzaju udostępnionych danych, które obiektywizują proces SDC. Mają one tutaj ogromne znaczenie, bowiem aby osiągnąć kompromis między minimalizacją ryzyka ujawnienia a maksymalizacją użyteczności danych dla użytkowników końcowych, konieczny jest obiektywny pomiar użyteczności danych poddanych SDC. Strata informacji jest odwrotnością użyteczności danych – zatem im większa użyteczność danych po przeprowadzeniu procesu kontroli ujawniania danych statystycznych, tym mniejsza strata informacji. Istota straty informacji i dylematy jej pomiarów zostały szczegółowo opisane w rozdziale 4 niniejszego opracowania.

1.1.3. Kontrola ujawniania danych a użytkownik danych

Kontrola ujawniania danych to postępowanie polegające na dokonaniu przed udostępnieniem czy opublikowaniem wynikowych informacji statystycznych (w postaci danych jednostkowych, opisu analitycznego, tablic, wykresów bądź kartogramów) ich weryfikacji w celu wyeliminowania – lub przynajmniej jak największego zminimalizowania – ryzyka ujawnienia bądź odtworzenia danych wrażliwych przez użytkowników udostępnianych zasobów. W praktyce bowiem całkowite usunięcie ryzyka jest na ogół niemożliwe lub wiązałoby się ze zbyt dużym uszczerbkiem informacyjnym. Metody SDC dotyczą operacji przeprowadzanych na danych, które mają być udostępnione. Istota tych sposobów polega na wprowadzeniu ograniczeń w ilości publikowanych danych lub publikowaniu danych poddanych modyfikacji. Rozróżnienie to prowadzi do wyodrębnienia dwóch odmiennych zespołów metod stosowanych w kontroli ujawniania danych. Jedna grupa owych metod polega na zmniejszeniu ilości publikowanych danych poprzez ukrycie pewnych danych lub agregację tych danych. Druga wprowadza zaplanowaną korektę (tzw. szum – ang. *noise*) w danych, w wyniku czego publikacja nie jest rzeczywistym wynikiem badania, a jedynie zasobem informacji, w którym występują pewne zakłócenia, wprowadzone świadomie w celu zachowania poufności.

W metodach kontroli ujawniania danych wprowadza się też rozróżnienie na dwa rodzaje użytkowników danych:

- **analityka danych** – badacz lub innego rodzaju użytkownik danych, którego celem jest analiza danych albo prowadzenie badań, nieprzejawiający intencji naruszania prywatności bądź poufności danych,
- **intruza** – użytkownik danych, którego celem jest naruszenie prywatności innych osób lub poufności danych niezgodnie z misją danego gestora danych.

W literaturze anglojęzycznej poświęconej SDC zwraca uwagę stosowanie większej liczby pejoratywnych określeń na intruza – *data spy*, *attacker*, *snooper* (Duncan i in., 2001; Hundepool i in., 2012) oraz *intruder* (Hundepool i in., 2012; Templ i in., 2021). Należy przy tym zauważyć, że obecnie sformułowanie *intruder* w literaturze anglojęzycznej całkowicie zdominowało określenie użytkownika o niecnych zamiarach. Terminy: *snooper* – szperacz, szpicel, podglądacz, *data spy* – szpieg danych, *attacker* – napastnik, atakujący – o, jak się wydaje, dużo większym ładunku negatywnych emocji, w piśmiennictwie z zakresu SDC stanowią zdecydowaną mniejszość. Pomimo funkcjonowania podanego powyżej podziału użytkowników danych, w stosowanych metodach SDC faktycznie nie ma możliwości ich rozróżnienia.

W metodach kontroli ujawniania danych przyjmuje się kryterium *użyteczność a ryzyko*. Zwiększenie użyteczności wiązałoby się z większym ryzykiem. Jedno-

czesne zwiększanie użyteczności i zmniejszanie ryzyka jest niemożliwe. Niemniej rozróżnienie operacyjne tych dwóch grup użytkowników pozwala spojrzeć modelowo na ocenę użyteczność–ryzyko w relacji do tego, jakie informacje na temat procesu ochrony poufności i stosowanych kryteriów poufności są udostępniane publicznie przez instytucje prowadzące badania statystyczne. Problem rozróżnienia rzeczonych dwóch grup użytkowników poruszają Duncan i in. (2001). We wspomnianej publikacji wskazuje się zresztą, że termin *snooper* oznacza każdego, kto dysponuje uprawnionym dostępem do danych (ang. *legitimate access to data*), jednakże intruzem czyni go to, że zastosowane metody analizy – a przede wszystkim cele wykorzystania udostępnionych danych – nie są zgodne z misją gestora danych. Stąd na przykład haker, który włamuje się do systemu komputerowego, nie jest traktowany jako intruz w kontekście SDC. Nie jest on bowiem użytkownikiem z uprawnionym dostępem do danych.

Na kanwie rozważań dotyczących miejsca i roli użytkownika, w tym intruza, w procesie kontroli ujawniania danych warto wspomnieć o rozwiązaniu, które stosuje na przykład krajowy urząd statystyczny Wielkiej Brytanii (ang. Office for National Statistics – ONS) (Spicer i in., 2014). W rozwiązaniu tym została wykorzystana koncepcja tzw. przyjaznych intruzów (ang. *friendly intruders*), opisana w publikacji Willenborga (2012), tzn. pracowników służb statystycznych albo innego gestora danych, imitujących zachowanie prawdziwych intruzów, których cele nie są zgodne z misją danego urzędu statystycznego czy gestora. Zadaniem przyjaznych intruzów jest zweryfikowanie ochrony np. danych ze spisu oraz ocena stopnia zaangażowania i wiedzy potrzebnej do pozyskania poufnych informacji.

1.2. Typy danych wynikowych

W niniejszym podrozdziale omówiono najważniejsze – z punktu widzenia kontroli ujawniania danych statystycznych – typy danych wynikowych. Są nimi:

- tablice statystyczne,
- mikrodane (dane jednostkowe),
- wyniki analiz.

Mogą mieć one – niezależnie od swej postaci – poufny charakter, jak również mogą danymi poufnymi nie być. **Dane poufne** to takie dane wynikowe, które pozwalają w sposób bezpośredni lub pośredni na identyfikację jednostek statystycznych, ujawniając tym samym informacje o nich. Nie chodzi tu jedynie o te jednostki statystyczne, którym odpowiadają np. obserwacje w zbiorach danych jednostkowych. Dotyczy to również wszelkich innych jednostek statystycznych, dla których jakiekolwiek informacje są dostępne w takich zasobach. Na przykład w mikrodanych z badania wypadków przy pracy gromadzone są dane na temat osób poszkodowanych, którym odpowiadają rekordy, ale również o wypadkach

zbiorowych (obejmujących więcej niż jedną osobę poszkodowaną) oraz o podmiotach gospodarczych (w których doszło do wypadku przy pracy). Nie jest to jednak jedyne kryterium uznania danych wynikowych za poufne bądź nie. Drugim jest to, czy dane takie podlegają ochronie prawnej. Muszą bowiem dotyczyć osób fizycznych lub prawnych, by były objęte tajemnicą statystyczną. Tylko takie dane wynikowe, które z punktu widzenia obu kryteriów należy uznać za dane poufne, powinny zostać poddane odpowiednio zaplanowanemu i przeprowadzonemu procesowi kontroli ujawniania danych statystycznych przed udzieleniem do nich dostępu użytkownikom zewnętrznym.

1.2.1. Tablice statystyczne

Wyróżniamy dwa podstawowe typy danych tabelarycznych:

- **tablice częstości** – tablice, w których każda komórka reprezentuje liczność (częstość), czyli liczbę jednostek należących do danego przekroju, tzn. do wymiaru tablicy, który dana komórka obrazuje; tablice te są typowe dla badań społecznych – ze względu na to, że w badaniach tych przeważają zmienne mierzone na skali nominalnej lub porządkowej;
- **tablice wielkości** – tablice, w których każda komórka reprezentuje wartość sumaryczną cechy ilościowej dla grupy respondentów należących do danego przekroju tablicy; tablice te są charakterystyczne dla badań gospodarczych, w których licznie występują zmienne mierzone na skali różnicowej bądź ilorazowej.

Ryzyko ujawnienia w wypadku tablic statystycznych wiąże się z występowaniem komórek o małych licznościach w tablicach częstości lub, dodatkowo, o wysokiej dominacji pojedynczych jednostek w wartości agregatu – w tablicach wielkości. Komórki te są określane jako **komórki z ryzykiem pierwotnym** lub inaczej – jako **komórki wrażliwe**. Ze względu na potrzebę zapewnienia efektywnej ochrony w wypadku zastosowania metod opartych na ukrywaniu komórek wyróżnia się także **komórki z ryzykiem wtórnym**. Są to komórki, dla których występuje wtórne ryzyko ujawnienia informacji wrażliwych, czyli odtworzenia danych jednostkowych na podstawie zawartości innych komórek.

Szczególnym przypadkiem tablic statystycznych lub zestawu kilku takich tablic rozpatrywanym w kontroli ujawniania tablic statystycznych są tablice hierarchiczne i tablice łączone. **Tablice hierarchiczne** to takie, w których jako zmiennej grupującej, wyznaczającej przekroje, użyto jednej bądź kilku zmiennych hierarchicznych (szczegółowy opis tej klasy zmiennych można znaleźć w części 1.2.2 niniejszej monografii). Przykładem takich zmiennych może być kod TERYT, klasyfikacja zawodów lub wykształcenia, a także Polska Klasyfikacja Działalności (PKD). Jeżeli w jednej tablicy statystycznej przedstawiono agregaty dla różnych poziomów klasyfikacji – na przykład na poziomie województw, powiatów i gmin

(w wypadku kodu TERYT) czy w układzie sekcji, działu, grupy, klasy i symbolu (w wypadku PKD), to wartości dla bardziej szczegółowych obszarów lub domen będą się sumować do wartości dla tych bardziej zgrubnych kategorii. Uwzględnienie występowania hierarchii w tablicy statystycznej jest kluczowe, by proces kontroli ujawniania tablic tego typu przebiegł w sposób prawidłowy. **Tablice łączone** z kolei to zestaw co najmniej dwóch tablic statystycznych, w których zachodzi pełna lub choć częściowa zgodność pomiędzy ich przekrojami, a w związku z tym występuje prawdopodobieństwo, że wskutek połączenia danych z takich tablic po pasujących przekrojach może dojść do ujawnienia informacji poufnych. Jest to drugie, po występowaniu hierarchii, kluczowe zagadnienie dla zapewnienia w sposób prawidłowy ochrony poufności udostępnianych tablic statystycznych.

1.2.2. Mikrodane

Mikrodane (dane jednostkowe) to zbiory danych składające się z rekordów (wierszy), spośród których każdy zawiera wartości zmiennych (atrybutów), czyli informacje o badanej cesze jednostki podlegającej analizie. W bardziej formalnym zapisie zbiór mikrodanych V to zbiór zawierający n rekordów, przy czym każdy rekord zawiera m zmiennych (atrybutów) charakteryzujących respondenta. Określenie *zmienna* (atrybut) oznacza cechę jednostek tworzących określoną populację, będącą przedmiotem badania statystycznego lub rejestracji, opisaną określonego rodzaju danymi statystycznymi dla tychże jednostek. Atrybuty w niechronionym zbiorze charakteryzują się na ogół różnorodnym stopniem poufności i wrażliwości (Domingo-Ferrer i Torra, 2003; 2004; 2005).

Mikrodane są podstawową formą przechowywania danych. I to na ich podstawie krajowe urzędy statystyczne tworzą wszystkie udostępniane dane wynikowe. Obecnie mają one dwojaką postać: danych tabelarycznych, w których komórki zawierają naliczone agregaty, oraz mikrodanych, czyli zbiorów rekordów, spośród których każdy zawiera informacje na poziomie jednostki statystycznej. Jak twierdzą Hundepool i in. (2010; 2012), obecnie coraz częściej mikrodane stają się pożądanymi danymi wynikowymi. Obejmują one próby z gospodarczych lub społecznych badań reprezentacyjnych, jak również zbiory ze spisów powszechnych lub ze źródeł administracyjnych. Oczekiwaniem osób pożądających mikrodanych jest ich jak największa szczegółowość. Niezbędne jednak staje się jej pogodzenie z obowiązkiem krajowych urzędów statystycznych czy innych gestorów dotyczącym ochrony danych poufnych dostarczonych przez respondenta. Ponadto Domingo-Ferrer i Torra (2003) podkreślają, że o ile doświadczenie w udostępnianiu zestawień tabelarycznych jest duże, o tyle udostępnianie zbiorów danych jednostkowych jest zjawiskiem nowym.

W zbiorach danych jednostkowych zmienne dzieli się, ze względu na funkcje pełnione przez nie w procesie kontroli ujawniania mikrodanych oraz nośność

informacji wrażliwych, na kilka niekoniecznie rozłącznych klas. Chociaż w zależności od źródła klasyfikacje te mogą się różnić, to autorzy niniejszej monografii – posiłkując się następującymi pozycjami: Eurostat i FRIBS Task Force (2017), Hundepool i in. (2012), Willenborg i de Waal (2001), Templ i in. (2014), Templ i in. (2021), Government Statistical Service (GSS, 2014) – wyróżnili następujące klasy zmiennych:

- **Identyfikatory** – w sposób jednoznaczny, bezpośredni i precyzyjny identyfikują jednostkę statystyczną. Istotą tych zmiennych, z perspektywy tworzenia oficjalnych danych statystycznych, jest możliwość łączenia danych z różnych źródeł oraz monitorowania jednostek w czasie (dane panelowe). Jeżeli użytkownik zewnętrzny ma uzyskać dostęp do mikrodanych, zakłada się ich usunięcie bądź zaszyfrowanie przed przystąpieniem do przeprowadzania procesu SDC lub przed zastosowaniem innych metod ochrony poufności. Przykłady: numer PESEL, imię i nazwisko, adres zamieszkania, numer dowodu osobistego, numer paszportu, numer prawa jazdy, numer polisy ubezpieczeniowej, nazwa podmiotu gospodarczego, adres siedziby przedsiębiorstwa czy numery ewidencyjne firmy.
- **Quasi-identyfikatory** (inaczej pseudoidentyfikatory, zmienne kluczowe lub domniemywane identyfikatory) – nie identyfikują jednostki statystycznej w sposób bezpośredni, ale w połączeniu z innymi takimi zmiennymi mogą pozwolić na jej jednoznaczną identyfikację. Zwykle trudno wskazać precyzyjnie, który zestaw zmiennych należy uznać za quasi-identyfikatory; niezbędne jest przy tym pójście na pewne kompromisy i przyjęcie określonych założeń. Najczęściej bowiem pomiędzy takimi zmiennymi występują powiązania, zależności funkcyjne i inne relacje. Są to często zmienne niezwykle cenne w analizach statystycznych czy ekonometrycznych. Ponadto potencjalnie każda zmienna, w zależności od przyjętych założeń, może się stać częścią klucza. Przykłady: płeć, wiek, wyznanie religijne, orientacja seksualna, wykonywany zawód, dochód, stan zdrowia, narodowość czy poglądy polityczne, sektor gospodarki, wielkość firmy (wyrażona liczbą pracowników) czy dochód lub strata.
- **Zmienne wrażliwe** (poufne zmienne wynikowe) – zawierają wrażliwe informacje na temat jednostki statystycznej. Ich wartości nie mogą zostać powiązane z żadnym respondentem w zbiorze danych. Najczęściej wynika to z uwarunkowań legislacyjnych lub etycznych. Nie ma co prawda jednoznacznych wytycznych, które zmienne zaliczyć do tej kategorii, a których nie, ale pewne wskazówki można znaleźć w RODO lub na stronach Eurostatu. Trzeba zaznaczyć, że wrażliwy charakter mogą mieć wszystkie kategorie lub wartości zmiennej, lecz równie dobrze może to dotyczyć jedynie jednego lub kilku jej wariantów. Może się zdarzyć, że pomimo zapewnienia ochrony związanej z ujawnieniem tożsamości udostępnienie zmiennych wrażliwych

nadal jest w stanie doprowadzić do ujawnienia określonej chronionej prawem cechy. Przykłady: dochody, stan zdrowia i niepełnosprawność, orientacja seksualna czy obroty podmiotów gospodarczych.

- **Zmienne niewrażliwe** (zmienne wynikowe niebędące poufnymi) – nie zawierają wrażliwych informacji o respondencie, jednak ze względu na to, że teoretycznie mogą się one stać quasi-identyfikatorami, nie można ich pominąć w procesie SDC. Przykłady: płeć, stan cywilny, aktywność ekonomiczna, lokalizacja siedziby przedsiębiorstwa czy powierzchnia zakładu pracy.
- **Zmienne ilościowe** (w literaturze z zakresu SDC nazywane zmiennymi ciągłymi) – przyjmują wartości ze zbioru nieskończonego wyrażone na skali różnicowej lub ilorazowej, a zatem takie, dla których określone operacje arytmetyczne są wykonalne i mają swoje interpretacje. Przykłady: wiek, staż pracy czy liczba pracujących w przedsiębiorstwie.
- **Zmienne jakościowe** (w literaturze z zakresu SDC nazywane zmiennymi kategorialnymi) – przyjmują wartości ze skończonego zbioru wariantów. Mogą być mierzone na skali nominalnej lub porządkowej. Przykłady: płeć, poziom ukończonego wykształcenia, sektor własności podmiotu gospodarczego czy wielkość firmy.
- **Zmienne ważące** – są wykorzystywane w procesie estymacji do uogólniania wyników próby na populację generalną i charakterystyczne dla badań reprezentacyjnych lub mieszanych. Są one opracowywane z wykorzystaniem schematu doboru próby, przy uwzględnieniu poprawek związanych z jednostkowymi brakami odpowiedzi. Należy rozważyć, czy udostępnienie takich wag nie ujawni informacji, których gestor danych nie chciałby upubliczniać (np. na podstawie wag możliwe może się okazać odtworzenie warstw, a te z kolei pozwolą na zlokalizowanie obszaru, w którym dany respondent się znajduje). Jeśli tak, to należałoby podjąć stosowne działania polegające np. na modyfikacji wag w sposób niewpływający istotnie na jakość finalnych szacunków dla populacji, ale redukujących do maksimum ryzyko ujawnienia informacji poufnych. Przykłady: waga z losowania.
- **Zmienne hierarchiczne** – to zmienne z wyróżnionymi obszarami bądź domenami na różnych poziomach, które mają charakter hierarchiczny, zagnieżdżony i mogą zostać uporządkowane według stopnia tego zagnieżdżenia. Najmniej szczegółowe są domeny wyróżnione na najwyższym poziomie hierarchii, najbardziej natomiast – te na poziomie najniższym. Wszystkie pozostałe poziomy znajdują się pod względem szczegółowości między tymi dwoma skrajnymi poziomami. Przykłady: klasyfikacja zawodów, klasyfikacja dziedzin wykształcenia, kod TERYT czy Polska Klasyfikacja Działalności (PKD).
- **Zmienne regionalne** – przestrzennie lokalizują jednostki statystyczne. Często zamiast tychże zmiennych – zwłaszcza dla małych jednostek podziału

terytorialnego (tj. powiatów czy gmin) lub dla małych domen wyróżnionych rzeczowo, a nie przestrzennie – udostępnia się zmienne pochodne o takiej jednostce. Przeprowadzając kontrolę ujawniania mikrodanych, należy mieć na względzie to, czy na podstawie kombinacji wartości zmiennych pochodnych nie będzie możliwe zidentyfikowanie obszaru lub domeny. Przykłady: stopień urbanizacji, gęstość zaludnienia czy klasa miejscowości zamieszkania (wyrażona liczbą mieszkańców).

- **Zmienne kartograficzne** – wykorzystywane przy tworzeniu wykresów mapowych. Przykłady: kod TERYT, współrzędne geograficzne czy siatka kilometrowa.
- **Zmienne predefiniujące nadrzędną jednostkę statystyczną** – przyjmują identyczne wartości dla każdej jednostki statystycznej podrzędnej wchodzącej w skład jednostki statystycznej nadrzędnej. Występują w zbiorach danych jednostkowych o układzie hierarchicznym, w których zbierane są informacje na temat różnych jednostek statystycznych, lub gdy agregacja danych o podrzędnych jednostkach statystycznych pozwala na uzyskanie danych na poziomie nadrzędnej jednostki statystycznej. Popularnym w badaniach społecznych przykładem takiego hierarchicznego układu jednostek statystycznych jest układ osoby (podrzędne) i gospodarstwa domowe (nadrzędne). Ich uwzględnienie w procesie SDC jest o tyle istotne, że nawet jeżeli każda podrzędna jednostka statystyczna z osobna jest trudna do identyfikacji, to wyróżniająca się pod względem pewnych charakterystyk nadrzędna jednostka statystyczna może się okazać łatwa do rozpoznania i prowadzić do ujawnienia wszystkich jednostek wchodzących w jej skład. Warto zauważyć, że szczególne potraktowanie gospodarstw domowych (np. do nazywania opisywanej klasy zmiennych) wynika z tego, że pierwsze zastosowania SDC dotyczyły właśnie tego kierunku oraz że w dalszym ciągu jest to jedno z głównych pól stosowania rozpatrywanej kontroli. Niemniej wydaje się, że szczególne eksponowanie gospodarstw domowych w tym kontekście jest obecnie mniej potrzebne. Kontrola ujawniania danych rozwinęła się bowiem na tyle, że jej narzędzia stosuje się do różnorodnych zbiorów danych jednostkowych. Stąd autorzy zaproponowali bardziej uniwersalną nazwę tej grupy zmiennych. Przykłady: typ biologiczny gospodarstwa domowego, rozmiar gospodarstwa domowego, główne źródło utrzymania gospodarstwa domowego, źródło ogrzewania mieszkania, rok wybudowania budynku, typ budynku czy numer identyfikacyjny REGON.

W nawiązaniu do powyższej klasyfikacji w niniejszym opracowaniu terminy *quasi-identyfikatory* i *zmienne kluczowe* będą stosowane zamiennie (jako synonimy), podobnie jak w niektórych pozycjach literatury międzynarodowej (Hundepool i in., 2012). Na przykład w środowisku R stosuje się podział na *kategorialne* (*jakościowe*) *zmienne kluczowe* i na *ciągłe* (*ilościowe*) *zmienne kluczowe*, choć to na-

zwa tych pierwszych sugeruje, że są one zmiennymi kluczowymi. Autorzy opracowania mają jednak świadomość tego, że termin *zmienne kluczowe* ma węższe znaczenie i powinien raczej się odnosić jedynie do zmiennych jakościowych uznanych za quasi-identyfikatory. To bowiem przede wszystkim na podstawie ich wartości tworzone są klucze przy wyznaczaniu wartości miar *a priori* ryzyka ujawnienia informacji poufnych. Miary te omówiono szerzej w rozdziale 2. Oczywiście można by rozważyć włączenie zmiennych ilościowych do zestawu zmiennych kluczowych, o ile nie przyjmują one wartości ze zbioru o dużej liczbie elementów bądź nieskończonego.

W klasyfikacji wyróżniono zmienne ilościowe i jakościowe jako klasy zmiennych w kontroli ujawniania danych statystycznych. W obcojęzycznej literaturze przedmiotu zmienne takie określa się mianem **zmiennych kategorialnych** (ang. *categorical*) – w wypadku zmiennych jakościowych – i **zmiennych ciągłych** (ang. *numerical*) – w wypadku zmiennych ilościowych (Domingo-Ferrer i Torra, 2005). Ze względu jednak na to, że podział ten nie jest zgodny z żadną funkcjonującą obecnie klasyfikacją w literaturze w języku polskim, autorzy zdecydowali się stosować w niniejszej pracy określenia **zmienne jakościowe** i **zmienne ilościowe**. Istotnym czynnikiem zaklasyfikowania zmiennej do jednego z tych dwóch typów jest skala pomiarowa – sposób pomiaru zjawisk obserwowanych przez statystykę zależnie od ich specyfiki i potrzeb badacza. Każdej obserwacji przyporządkowuje się określoną liczbę rzeczywistą zgodnie z własnością przedmiotu obserwacji oraz przyjętymi zasadami. Stevens (1946) zaproponował cztery skale pomiarowe: nominalną, porządkową, różnicową i ilorazową.

- Dane wyrażone na **skali nominalnej** mogą służyć jedynie do rozróżniania jednostek; można zatem stwierdzić, czy dwie jednostki są w zakresie danej cechy tożsame, czy różne (taka sytuacja występuje w wypadku płci, stanu cywilnego lub gminy zamieszkania). Wartości obserwacji wyrażonych na skali nominalnej nie da się uporządkować, jedyną relację pomiędzy nimi stanowi relacja równości. Liczb nie można tu zatem porównywać ani wykonywać na nich żadnych operacji algebraicznych – są one bowiem faktycznie tylko etykietami, oznaczeniami.
- Dane wyrażone na **skali porządkowej** mogą służyć do rozróżniania i uporządkowania jednostek pod określonym względem (jak poziom wykształcenia czy stopień niepełnosprawności osób). Zakłada się tutaj bowiem istnienie w zbiorze danych określonego porządku spełniającego klasyczne założenia spójności (jeżeli dwa obiekty są różne, to jeden jest mniejszy od drugiego), antysymetrii (jeżeli jeden obiekt jest mniejszy od drugiego, to nie może być równocześnie na odwrót) i przechodniości (jeśli jeden obiekt jest mniejszy od drugiego, a ów drugi od trzeciego, to pierwszy jest mniejszy od trzeciego). Podobnie jak w wypadku skali nominalnej, na wyrażonych za jej pomocą danych nie można wykonywać żadnych operacji arytmetycznych.

– Dane wyrażone na **skali różnicowej** (przedziałowej, interwałowej) mogą służyć do rozróżniania i uporządkowania jednostek pod określonym względem oraz mogą być dodawane i odejmowane, ale ich mnożenie i dzielenie jest niewykonalne. O skali różnicowej pomiaru powiemy, jeśli liczbowy opis danej cechy przewiduje występowanie obserwacji wyrażonych zarówno liczbami ujemnymi, jak i dodatnimi (oraz zera). Może to dotyczyć na przykład przyrostu naturalnego na 1000 ludności, salda migracji, wskaźnika rentowności przedsiębiorstwa itp.
– Dane wyrażone na **skali ilorazowej** mogą służyć do rozróżniania i uporządkowania jednostek pod określonym względem. Można na nich wykonywać wszystkie operacje arytmetyczne (dodawanie, odejmowanie, mnożenie i dzielenie). Przykłady tego rodzaju danych to: przeciętne wynagrodzenie miesięczne, liczba ludności przypadającej na jednego lekarza, stopa bezrobocia itp. Przyjmuje się tutaj milcząco istnienie zera bezwzględnego, które wszakże w praktyce zwykle nie występuje.

Mikrodane z badań statystycznych prowadzonych przez krajowe urzędy statystyczne lub pozyskane przez innych gestorów mają odzwierciedlać fragment otaczającej nas rzeczywistości, by umożliwić sformułowanie pewnych prawidłowości statystycznych w nim zachodzących. W związku z tym pomiędzy zmiennymi czy rekordami w zbiorach danych jednostkowych będą zachodzić zależności o naturze logicznej lub statystycznej. Istotną kwestią jest to, czy można je zignorować w procesie kontroli ujawniania mikrodanych i jaki wpływ będzie to miało na bezpieczeństwo danych wynikowych dla użytkownika zewnętrznego. Rozpatrywanie wielu zależności, które trzeba by wykryć, przeprowadzając wybrane analizy statystyczne, mogłoby się okazać trudne do wykonania i niemożliwe do uwzględnienia w założeniach procesu SDC. Niezbędne jednak jest uwzględnienie takich zależności, które są powszechnie znane i które użytkownik zasobów dążący do złamania poufności dostępnych mu danych wykorzysta, aby przywrócić oryginalne wartości zmiennych dla części lub całości rekordów, co mogłoby się okazać dla niego wręcz trywialnym wyzwaniem.

Chociaż zależności pomiędzy zmiennymi a rekordami w mikrodanych nie mogą zostać zignorowane przy zapewnianiu ochrony poufności, to często jednak nie są one ujmowane w wypracowanych podejściach, np. z powodu niemożności uwzględnienia ich występowania przy parametryzacji wykorzystywanych metod.

Najważniejszymi – nie tylko z punktu widzenia organizacji i funkcjonowania państwa, ale również dla środowiska naukowo-badawczego – źródłami zbiorów danych jednostkowych są **badania statystyczne** prowadzone przez krajowe urzędy statystyczne, a także **źródła administracyjne** (w tym rejestry). Oczywiście poza badaniami statystycznymi i źródłami administracyjnymi statystyka publiczna czerpie dane również z wielu innych źródeł. Mogą to być zasoby pozyskiwane od podmiotów gospodarczych (np. operatorów telefonii komórkowych

czy dostawców energii), jak też badania i analizy prowadzone przez instytucje naukowe, komercyjne, ośrodki badań opinii publicznej itp. Należy zasygnalizować, że coraz częściej pozyskuje się je również z innych, alternatywnych źródeł, w tym tych publicznie dostępnych, na przykład z internetu (m.in. big data, web scraping). Jednak ogólne zasady etyczne oraz metody, techniki i narzędzia w zakresie ochrony udostępnianych danych wykorzystywane w statystyce publicznej mają i tam zastosowanie. Z punktu widzenia funkcjonowania i możliwości krajowych urzędów statystycznych nadal jednak kluczową rolę odgrywają dwa źródła wymienione na początku. Do mikrodanych z takich źródeł coraz częściej dostęp chcą uzyskać użytkownicy zewnętrzni, tj. osoby ze środowiska naukowego, z sektora publicznego oraz prywatnego. Dzieje się tak przede wszystkim ze względu na bogactwo informacyjne takich zbiorów danych jednostkowych oraz szerokie spektrum nowych możliwości, jakie daje ich wykorzystanie w prowadzeniu analiz (tj. możność opracowania lub skorzystania z nowych metod badawczych, które wcześniej pozostawały poza zasięgiem badacza – np. z powodu ograniczonej i zagregowanej postaci dostępnych danych). Poniżej te dwa źródła krótko scharakteryzowano, podkreślono również istotę ich ochrony.

Badania statystyczne prowadzone przez służby statystyki publicznej, jak również funkcjonowanie tych organów i całego systemu statystyki publicznej, regulowane są ustawą z dnia 29 czerwca 1995 r. o statystyce publicznej. W świetle tej ustawy, badaniem statystycznym jest zbieranie, gromadzenie i opracowywanie danych statystycznych (dotyczących zjawisk, zdarzeń, obiektów i działalności podmiotów gospodarki narodowej oraz życia i sytuacji osób fizycznych, w tym danych osobowych, pozyskanych bezpośrednio od respondentów albo z systemów informacyjnych administracji publicznej i rejestrów urzędowych, od momentu ich zebrania na potrzeby wykonywania zadań statystyki publicznej) oraz ogłaszanie i udostępnianie wyników dokonanych obliczeń, opracowań i analiz, w tym podstawowych wielkości i wskaźników (do których ogłaszania prezes Głównego Urzędu Statystycznego jest zobowiązany każdorazowo na podstawie odrębnych przepisów). Statystyka publiczna natomiast to system zbierania danych statystycznych, gromadzenia, przechowywania i opracowywania zebranych danych oraz ogłaszania, udostępniania i rozpowszechniania wyników badań statystycznych jako oficjalnych danych statystycznych.

Celem funkcjonowania statystyki publicznej jest dostarczanie rzetelnych, obiektywnych i systematycznych informacji o sytuacji ekonomicznej, demograficznej, społecznej oraz środowisku naturalnym wszystkim zainteresowanym takimi informacjami stronom (społeczeństwu, organom państwa oraz administracji publicznej, a także podmiotom gospodarki narodowej). Prowadzone w ramach statystyki publicznej badania statystyczne (metodą obserwacji pełnej lub reprezentacyjnej na wylosowanej bądź dobranej celowo próbie) mogą dotyczyć każdej dziedziny życia społecznego i gospodarczego oraz występujących w nim

zjawisk dających się obserwować i analizować z wykorzystaniem metod statystycznych.

Wykaz wszystkich badań statystycznych, które mają zostać przeprowadzone przez prezesa Głównego Urzędu Statystycznego w ramach statystyki publicznej, zostaje ustalany corocznie w drodze rozporządzenia Rady Ministrów. Wykaz ten nazywa się Programem badań statystycznych statystyki publicznej (w skrócie PBSSP) (BIP, b.d.). Dla każdego badania określane są: temat, organ lub podmiot go prowadzący, cykliczność, cel, szczegółowy zakres podmiotowy i przedmiotowy, źródła danych, podmioty przekazujące dane, informacje dotyczące przekazywanych danych, a także rodzaje wynikowych informacji statystycznych oraz formy i terminy ich udostępnienia. Oprócz PBSSP powszechnie dostępne są również wykazy kosztów związanych z każdym badaniem, wraz ze wskazaniem źródeł ich pokrycia.

Drugim ważnym źródłem, z którego pozyskiwane są mikrodane w ramach statystyki publicznej, są źródła administracyjne. Przede wszystkim zaś są nimi – zgodnie z definicją sformułowaną przez Oleńskiego (2005; 2006) – wykazy, listy i spisy: podmiotów (osób fizycznych, osób prawnych, jednostek organizacyjnych nieposiadających osobowości prawnej); obiektów materialnych; procesów ekonomicznych lub technologicznych; zdarzeń społecznych, ekonomicznych, technicznych, ekologicznych i innych, których rejestrowanie i ewidencjonowanie jest niezbędne organom administracji publicznej, jednostkom sektora publicznego bądź innym jednostkom dla realizacji ich funkcji publicznych, do czego jednostki te są zobowiązane z mocy prawa. Nyczaj (2010) zauważył, że przytoczoną powyżej definicję uznaje się za jedną z najbardziej precyzyjnych, ponieważ uwypuklone zostało w niej kryterium funkcji publicznej, którą sprawuje gestor danych (może nim być również podmiot niepubliczny), a także jednostkowość elementu rejestrowanego.

Rejestry administracyjne posiadają rękojmię wiary publicznej. Oznacza to, że są legalnymi i wiarygodnymi podwalinami informacyjnymi, na podstawie których mogą być podejmowane decyzje lub działania ekonomiczne, społeczne bądź administracyjne. Podstawowym celem funkcjonowania rejestrów jest identyfikacja objętych nimi jednostek – dokonywana jest ona z mocy prawa na potrzeby jednostek administracji państwowej oraz innych organizacji realizujących zadania publiczne (Oleński, 2005).

Oleński (2005) przeprowadził kompleksową analizę rejestrów administracyjnych. Dokonał ich klasyfikacji ze względu na zakres (ogólnokrajowe i wewnętrzne), a także na trzy klasy: rejestry podmiotowe, przedmiotowe i zdarzeń. Omówił ich funkcje i modele oraz podał przykłady, jak również przeanalizował wpływ technologii na funkcjonowanie rejestrów administracyjnych.

Rejestry statystyczne także wlicza się do rejestrów administracyjnych. Są nimi te, w których gromadzi się jednostkowe dane pozyskiwane w ramach PBSSP (Ny-

czaj, 2010). Stąd czasami rozróżnienie pomiędzy danymi rejestrowymi a danymi z badań statystycznych bywa nieostre.

Rejestry administracyjne odgrywają istotną rolę w statystyce publicznej – ich wykorzystanie jest bowiem jedną z fundamentalnych zasad prowadzenia oficjalnej statystyki w Europie. Ponadto coraz częściej stają się one alternatywą dla przeprowadzania ankietowych badań statystycznych. Często też łączy się dane jednostkowe z różnych źródeł. Przemawia za tym jakość i kompletność rejestrów (w porównaniu z badaniami statystycznymi bardzo często nie są obarczone brakami odpowiedzi, a jeśli takie braki już występują, to są nieznaczne w stosunku do rozmiaru zasobów rejestru i nie wpływają zasadniczo na jakość danych wynikowych; luki te można zresztą łatwo wypełnić np. na drodze imputacji) oraz ich aktualność (co często jest uwarunkowane prawnie). Istotne znaczenie mają tutaj też zwykle wysokie koszty i czas prowadzenia badań statystycznych, a także związane z nimi obciążenie respondentów. Rejestry wykorzystuje się również między innymi w konstrukcji operatów dla badań statystycznych, do uzupełnienia danych zebranych w ankietach oraz do weryfikacji jakości pozyskanego materiału statystycznego.

Zarówno rejestry administracyjne, jak i dane z badań statystycznych są zasobami niezwykle bogatymi w informacje i z tego powodu zapewniającymi nowe możliwości użytkownikom zewnętrznym. Coraz częściej osoby ze świata nauki zwracają się do krajowych urzędów statystycznych, ale również do innych gestorów danych, z prośbą o udostępnienie im baz danych jednostkowych. Mikrodane pochodzące zarówno z jednego, jak i z drugiego rozpatrywanego tutaj źródła zawierają jednak zmienne pozwalające – w sposób bezpośredni lub pośredni – na identyfikację jednostki statystycznej lub respondenta. Z tego powodu niezwykle istotna jest ochrona tychże zasobów przed ujawnieniem informacji poufnych bądź wrażliwych.

Rejestry administracyjne, bogate w zmienne o charakterze identyfikatorów, prawdopodobnie zawierają więcej zmiennych pozwalających na bezpośrednią identyfikację osoby bądź innego podmiotu. Przykładem takiego rejestru jest Powszechny Elektroniczny System Ewidencji Ludności (PESEL), prowadzony na mocy ustawy o ewidencji ludności, zawierający dla wszystkich osób mu podlegających m.in. informacje o: imionach i nazwisku (także rodowym), imionach i nazwiskach rodowych rodziców, dacie, miejscu i kraju urodzenia, stanie cywilnym, płci, akcie urodzenia, numerze PESEL, obywatelstwie, małżonku, miejscu zamieszkania, dowodzie osobistym i paszporcie, wyjazdach z kraju i powrotach do kraju oraz o zgonie (dla cudzoziemców zbierane są ponadto dodatkowe informacje). Wiele spośród tych cech pozwoliłoby na bezpośrednią identyfikację osoby. Podobnie sytuacja przedstawia się w wypadku rejestru ubezpieczonych prowadzonego przez Zakład Ubezpieczeń Społecznych, rejestrów podatkowych POLTAX czy Centralnej Ewidencji Pojazdów i Kierowców itp. Warto wszakże

wspomnieć, że istnieją i takie źródła administracyjne, które bezpośrednich identyfikatorów nie zawierają (przykładem mogą tu być zasoby danych o zakupach z kas fiskalnych sklepów i punktów usługowych, gdzie nabycie towarów i usług jest rejestrowane bez danych o kliencie).

Badania statystyczne nie są tak bogate w zmienne identyfikatory, lecz zgromadzone w ich wyniku dane pozwalają na uzyskanie szerokiej i kompleksowej charakterystyki badanej jednostki statystycznej na płaszczyźnie różnych dziedzin i sfer. Wiele spośród zmiennych zebranych w kwestionariuszu – zwłaszcza o charakterze demograficznym czy społecznym – mogłoby umożliwić pośrednie zidentyfikowanie respondenta (na podstawie quasi-identyfikatorów). W zależności od zakresu tematycznego badania oraz kompleksowości kwestionariusza różne mogą być liczba i typ zmiennych pozwalających na pośrednią identyfikację respondenta. Przykładem zbioru danych z badania statystycznego, bogatym w zmienne, które potencjalnie mógłby wykorzystać intruz, jest złoty rekord zawierający dane jednostkowe z Narodowego Spisu Powszechnego Ludności i Mieszkań z 2011 roku, w którym dodatkowo część zmiennych ma pochodzenie administracyjne. Łącznie w złotym rekordzie znajduje się kilkaset cech charakteryzujących poszczególne osoby. W kontekście badań statystycznych należy również przypomnieć, że rodzaj pozyskiwanych danych osobowych oraz możliwości ich wykorzystania są uwarunkowane legislacyjnie.

Wpływ na bezpieczeństwo informacji wrażliwych bądź poufnych ma ponadto zakres jednostek statystycznych, o których informacje dostępne są w zbiorach mikrodanych. Rejestry administracyjne zawierają dane o wszystkich jednostkach z punktu widzenia spełniania ustalonego kryterium. Badania statystyczne prowadzone metodą reprezentacyjną, na wylosowanej bądź dobranej celowo próbie, obejmują informacje jedynie o tych respondentach, którzy znaleźli się w owej próbie. Badania statystyczne prowadzone metodą pełnej obserwacji zawierają informacje o wszystkich jednostkach należących do ściśle określonej zbiorowości statystycznej. Ponadto, o czym już wspomniano przy omawianiu klasyfikacji zmiennych występujących w zbiorach danych jednostkowych, w jednym takim zbiorze – niezależnie od jego źródła – zbierane mogą być liczne informacje na temat różnych jednostek statystycznych, które dodatkowo mogą tworzyć hierarchiczną strukturę, począwszy od podrzędnych jednostek statystycznych aż do tych nadrzędnych.

1.2.3. Metadane, paradane i dane dodatkowe

Mikrodane nie są jedynym produktem, który krajowe urzędy statystyczne udostępniają użytkownikom zewnętrznym w celu realizacji prac naukowo-badawczych. W badaniach statystycznych przez nie prowadzonych gromadzone są bowiem różnego rodzaju informacje, do których również może być udzielony dostęp. Wśród nich można wyróżnić następujące grupy:

– dane zebrane w kwestionariuszu badania,
– metadane badania,
– paradane badania,
– dane dodatkowe.

Dane zebrane w kwestionariuszu badania zostały już szczegółowo omówione w poprzednim punkcie – to mikrodane, które stanowią podstawę do naliczania wszelkiej postaci danych wynikowych w postaci tablic statystycznych, wyników analiz, ale również zbiorów danych jednostkowych, które podlegają opublikowaniu, udostępnieniu lub w inny sposób są rozpowszechniane.

Przez pojęcie **metadanych** należy rozumieć dane o danych (Nicolaas, 2011). Rola systemów metadanych w krajowych urzędach statystycznych stale rośnie ze względu na to, że systemy te są niezbędne do upowszechniania statystyk. Metadane są podstawami funkcjonowania hurtowni danych statystycznych i zyskują coraz większe znaczenie w wymianie danych statystycznych między różnymi organizacjami (zwłaszcza na płaszczyźnie międzynarodowej), gdyż dodatkowo są one wykorzystywane w celu ułatwienia komunikacji i wspólnego zrozumienia wymienianych danych.

Szczegółową analizę systemu metadanych można znaleźć m.in. w artykule Signore i in. (2015). Metadane dzieli się tam na **metadane strukturalne** oraz **metadane referencyjne**. Pierwsze to te metadane, które identyfikują strukturę danych (np. nazwy kolumn w mikrodanych lub wymiary w kostkach statystycznych) lub strukturę powiązanych metadanych (np. jednostki miary). Drugie z kolei to te metadane, które opisują zawartość i jakość danych statystycznych. W najlepszym przypadku powinny one obejmować:

– **metadane pojęciowe** – opisujące użyte pojęcia i ich praktyczną implementację, pozwalające użytkownikowi zrozumieć, co statystyki mierzą i ich przydatność do użycia,
– **metadane metodologiczne** – opisujące metody wykorzystane do utworzenia zbioru danych (np. metody doboru próby, pozyskiwania wyników czy ich przygotowania i obróbki),
– **metadane jakościowe** – opisujące różne wymiary jakości statystyk wynikowych (np. ich aktualność, dokładność).

Jeżeli dla badania statystycznego zostaną przygotowane procedury gromadzenia danych i dokumentowania zebranych danych (a więc odpowiednie metadane), to dokumentacja ta powinna zostać udostępniona badaczom w cyfrowej bazie danych, gdyż ułatwi to korzystanie z danych oraz ich wyszukiwanie. Do prezentacji dostępnych źródeł korzysta się z repozytoriów metadanych i metodologii semantyki w interaktywny i łatwo dostępny sposób. Metadane z badań statystycznych mogą pomóc ich użytkownikom zapoznać się z dostępnymi zbiorami danych, potwierdzić, czy niezbędny jest dostęp do plików o ograniczonym dostępie, podjąć decyzję, czy dane te będą wspierały sformułowany i pożądany

cel naukowo-badawczy, a także zapoznać się z metodologią badania, jak również pomóc im w odkrywaniu ogromnych materiałów, które nie są im znane (Poljičak i Stančić, 2014).

Oczywiste jest, że w procesie kontroli ujawniania danych statystycznych przeprowadzanym na mikrodanych należy również zweryfikować to, czy udostępnienie razem z nimi metadanych nie wpłynie negatywnie na poufność i nie spowoduje wzrostu ryzyka ujawnienia informacji poufnych. Stosując odpowiednie przekształcenia na zbiorach danych statystycznych, wynikające z zastosowania metod ochrony tajemnicy statystycznej, należy pamiętać o tym, by odpowiednie zmiany zastosować również w zbiorach metadanych (np. agregując czy w inny sposób modyfikując bądź zakłócając wartości lub warianty zmiennej, trzeba uwzględnić te same zmiany w metadanych dotyczących tej zmiennej).

Metadane odgrywają również jeszcze jedną istotną rolę w procesie udostępniania przez gestora danych różnej postaci poufnych danych wynikowych użytkownikom zewnętrznym, przede wszystkim zaś mikrodanych. Otóż Thomas i in. (2011) oraz Poljičak i Stančić (2014) w swoich pracach koncentrują się na opisie wykorzystania repozytorium metadanych w systemie autoryzacji dostępu. Będzie miał on zastosowanie przede wszystkim wówczas, gdy dostęp do poufnych zbiorów danych jednostkowych będzie zdalny (poprzez sieć komputerową, na przykład z miejsca pracy, innej wyznaczonej przez gestora danych lokalizacji, a nawet z miejsca zamieszkania użytkownika) lub stacjonarny (czyli jedynie na specjalnie wyznaczonym stanowisku w siedzibie gestora danych). W wypadku dostępu zdalnego zbiory te mogą się stać widoczne dla ich użytkownika, jak również mogą być przed nim ukryte. W drugiej sytuacji osobie korzystającej z dostępu zdalnego pozwala się jedynie na wyświetlanie metadanych, by na ich podstawie mógł prowadzić swoje prace naukowo-badawcze.

Oczywiste jest, że dostęp do poufnych danych jednostkowych powinien być możliwy tylko dla upoważnionych użytkowników. Powinien się odbywać zgodnie ze wszystkimi przepisami prawnymi i technicznymi, jak również innymi uregulowaniami branymi pod uwagę i wdrażanymi w celu zapewnienia odpowiedniego poziomu ochrony danych. Według Poljičak i Stančicia (2014) jest to możliwe, jeżeli wykorzystane będzie repozytorium metadanych użytkowników, które zostanie umieszczone w rdzeniu systemu odpowiedzialnego za dokumentowanie użytkowników, ich praw oraz dostępnej dla nich listy usług.

W repozytorium metadanych powinna być przechowywana informacja o użytkowniku, o danych z badań statystycznych, do których ma dostęp lub z których może korzystać, jak również o danych udostępnionych użytkownikowi do eksploracji bądź przeprowadzenia dodatkowych analiz. Repozytorium będzie więc zawierać całą historię praw użytkownika do zbiorów danych. Informacje takie mogłyby zostać wykorzystane w wypadku naruszenia poufności, jak również w razie konieczności poinformowania użytkownika o ważnych terminach

czy zdarzeniach (tj. przesyłania informacji o nowych usługach bądź zbiorach, o zbliżającym się wygaśnięciu subskrypcji itp.). Dodatkowo informacje te mogłyby zostać użyte do spersonalizowania zasobów systemowych zgodnie z zainteresowaniami i potrzebami użytkowników (Poljičak i Stančić, 2014). Na koniec warto jeszcze podkreślić, że przy zarządzaniu dostępem powinno być możliwe korzystanie z tego samego języka metadanych, który był używany do dostarczania opisowych i strukturalnych informacji o danych (Thomas i in., 2011).

Oprócz artykułu Thomas i in. (2011) informacje o stosowanych w praktyce standardach repozytoriów metadanych – tj. DLL (ang. *dynamic-link library*, dynamiczna biblioteka łącz) czy SDMX (ang. *statistical data and metadata exchange*, wymiana danych i metadanych statystycznych) – można znaleźć między innymi w następujących pracach: Gregory i Heus (2007), Gregory (2011), Vale (2010), OECD (2014) czy Signore i in. (2015).

Paradane to termin wprowadzony do metodologii badań reprezentacyjnych przez Coupera (1998). Jednakże ówcześnie odnosił się on do zbieranych w sposób automatyczny danych o procesie badania. Obecnie przez to pojęcie rozumie się wszystkie dane o procesie zbierania danych w badaniach reprezentacyjnych, np. dane adresowe i geograficzne respondenta, nagranie wywiadu, długość wywiadu, informacje o przyciskanych klawiszach, charakterystykę osoby ankietującej, jak również obserwacje dokonane przez nią oraz informacje z jej kwestionariusza (mimo że dane takie nie opisują samego procesu) (Nicolaas, 2011).

Wykorzystanie paradanych może pozytywnie wpłynąć na niektóre wyzwania, przed którymi stają służby statystyki publicznej – jak na przykład rosnący odsetek odmów odpowiedzi w badaniach reprezentacyjnych (a w związku z tym rosnące ryzyko obciążenia jego wyników), błędy pomiaru czy rosnące koszty zbierania danych w tego typu badaniach. Dodatkowo obserwuje się wzrost zakresu i szczegółowości paradanych, co jest skutkiem postępu technologicznego, przez który cały proces zbierania danych ankietowych staje się bardziej skomputeryzowany. Nicolaas (2011) szczegółowo przeanalizował korzyści płynące z wykorzystania paradanych – zarówno po stronie krajowych urzędów statystycznych, jak i po stronie odbiorców informacji statystycznych. Dane tego typu jednak zwykle nie są udostępniane. Autor wyżej wskazanej pracy doszukuje się przyczyn takiego stanu rzeczy w:

- kosztach ich przygotowania,
- potrzebie udzielania pomocy w zarządzaniu nimi i w ich interpretacji (różnice w paradanych pod względem formatu, struktury czy stopnia złożoności; paradane mogą być bardzo duże i nieuporządkowane, co generuje koszty i czas ich przygotowania, a ponadto poprawne z nich skorzystanie może wymagać merytorycznego wsparcia ze strony gestora danych),
- dbałości o poufność ankietera i respondenta,
- trosce o interes gestora danych (czasem udostępnienie paradanych może prowadzić do wyjawienia informacji niejawnych, np. o procesie prowadze-

nia przez niego badań statystycznych czy zbierania danych, co może zostać wykorzystane przez jego konkurencję).

Z punktu widzenia niniejszego opracowania interesujący jest trzeci wymieniony powyżej aspekt, odnoszący się do ochrony informacji poufnych. Udział w badaniach statystycznych opiera się bowiem na zasadzie społecznego zaufania do systemu statystyki publicznej i wiary w to, że tożsamość ani żadne inne osobowe dane – w tym przede wszystkim te, które uznać należy za wrażliwe – nie zostaną ujawnione. Zapewnienie poufności respondentom jest zatem kluczowe – dotyczy to nie tylko mikrodanych, ale również powiązanych z nimi zbiorów paradanych, spośród których wiele może zawierać informacje umożliwiające ich identyfikację. Oczywiste jest, że takie dane nie mogą zostać udostępnione bez dokładnego ich przetworzenia i usunięcia lub zaszyfrowania wszelkich informacji, które potencjalnie mogłyby zostać wykorzystane przez użytkownika (samodzielnie lub w połączeniu z innymi informacjami) do identyfikacji osoby ankietowanej. O ile usunięcie lub zmodyfikowanie części informacji w paradanych (np. adresu respondenta) jest proste, o tyle w wypadku informacji przechowywanych w postaci częściowo ustrukturyzowanych plików tekstowych lub nagrań audio może się to okazać czasochłonne i trudne do wykonania – zwłaszcza jeżeli zależy nam jednocześnie na zapewnieniu jak największej ich użyteczności dla celów naukowo-badawczych. Jest to szczególnie dokuczliwe, gdy procesu czyszczenia nie można zautomatyzować (jak w wypadku wspomnianych plików tekstowych o różnej strukturze czy nagrań audio). Dodatkowym wyzwaniem w zakresie przeprowadzania procesu kontroli ujawniania danych statystycznych na paradanych jest łączenie ich z różnych źródeł i dodawanie do danych z badania reprezentacyjnego, rejestru administracyjnego lub z innego źródła (Nicolaas, 2011). Niezbędna jest również ochrona danych o ankieterach, które także mogą się znajdować w zbiorach paradanych (a organy prowadzące badania statystyczne są zobligowane do ich rejestrowania).

Przy udostępnianiu paradanych należy pamiętać o konieczności zastosowania metod kontroli ujawniania danych statystycznych. Problemem może być to, że o ile dla mikrodanych, jak i danych zagregowanych w postaci tablic statystycznych czy wyników analiz, wypracowane zostały pewne podejścia w zakresie ochrony poufności, o tyle nie ma metod i narzędzi informatycznych przeznaczonych dla paradanych. Również na arenie międzynarodowej trudno jest doszukać się szczegółowych opisów prac przeprowadzonych w tym zakresie. Z tego powodu można rozważyć alternatywnie udostępnienie paradanych w chronionym i ściśle nadzorowanym środowisku, co uniemożliwiłoby nieautoryzowany dostęp do nich oraz wykluczyłoby możliwość niedozwolonego sposobu ich użycia, a zwłaszcza ich rozpowszechnienia lub utworzenia ich kopii.

Ostatnią grupą danych często powiązaną z trzema omówionymi powyżej są dane dodatkowe. Są to wszelkie dane ze źródeł zewnętrznych takich jak inne badania reprezentacyjne, spisy powszechne czy rejestry administracyjne.

1.2.4. Wyniki analiz

Wyniki analiz to wszelkiej postaci opracowania, które zostały przygotowane na podstawie mikrodanych, których zawartość jest trudna do przewidzenia. W dokumencie wypracowanym w projekcie ESSNet (Brandt i in., 2010) poświęconym temu zagadnieniu zwraca się uwagę na różnorodność analiz oraz na to, że ich zawartość nie wpisuje się w żaden określony schemat wyników. W literaturze przedmiotu z zakresu kontroli ujawniania wyników analiz porządkuje się dane wynikowe poprzez przypisanie ich do określonych klas. Po tym następuje nadanie etykiety *bezpieczne* lub *niebezpieczne*. Więcej na ten temat piszemy w podrozdziale 2.3.

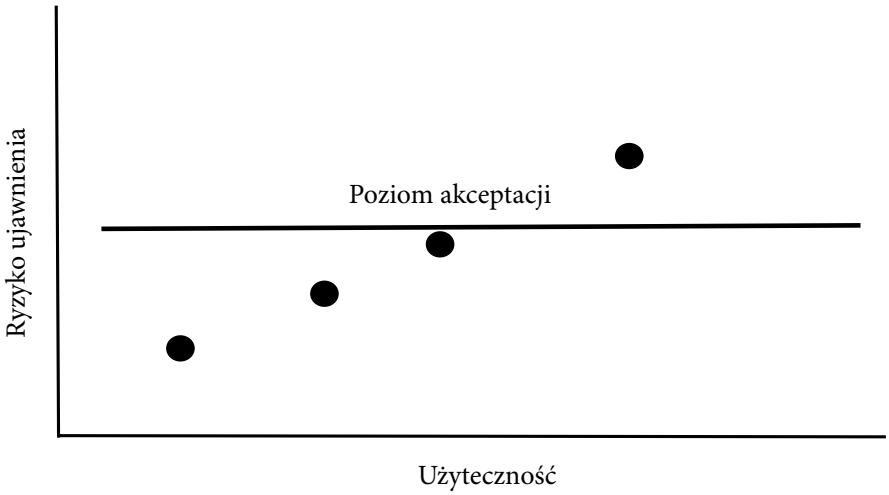
Klasami wyników analiz ze względu na ich strukturę mogą być na przykład:

- dla danych tabelarycznych: tablice częstości i wielkości,
- dla statystyk opisowych: wartość minimalna i maksymalna, rozstęp, kwantyle, w tym: kwartyle (dolny, mediana i górny), kwintyle, decyle czy percentyle, średnia, dominanta, wariancja, odchylenie standardowe, ćwiartkowe i przeciętne, kowariancja, skośność, pozycyjny współczynnik asymetrii, kurtoza i współczynniki koncentracji, a także wszelkie inne wskaźniki,
- dla rezultatów analiz statystycznych: współczynniki modelu regresji liniowej lub nieliniowej (w tym wyraz wolny), reszty z modelu regresji, statystyki podsumowujące (np. współczynnik determinacji R^2) i testowe (np. statystyka χ^2), analiza czynnikowa, analiza skupień, analiza kontyngencji czy współczynniki korelacji,
- dla ilustracji graficznych: wykresy (np. słupkowe, kołowe, punktowe, liniowe, histogramy, mozaikowe, pudełkowe itd.), kartogramy i kartodiagramy, wykresy mapowe, piktogramy.

1.3. Ryzyko ujawnienia a użyteczność – istota zapewnienia balansu

Ponieważ kontrola ujawniania danych jest wynikiem praktycznych potrzeb związanych z kontrolą publikacji wyników badań statystycznych, u jej podstaw nie leży ugruntowana jednolita teoria, która wyznaczałaby jednoznaczny sposób postępowania przed fazą publikacji wyników badania. Różnorodność struktur danych, możliwości powiązania ich z innymi dostępnymi źródłami danych i rozróżnienie użytkowników danych na tych, którzy nie stanowią zagrożenia dla poufności, oraz tych, których zamiarem jest naruszenie poufności, są ogromnym wyzwaniem dla wypracowywania metod kontroli ujawniania danych. Proponowane modele stosowane w tejże kontroli, przedstawione np. w podręczniku Hundepoola i in. (2012), to:

– Mapa ryzyko-użyteczność (R-U) – polega na wypracowaniu metod, które minimalizują ryzyko ujawnienia danych i jednocześnie maksymalizują ich użyteczność. Zarówno użyteczność, jak i ryzyko są wyrażone jako liczbowe miary, na podstawie których instytucja udostępniająca dane przyjmuje maksymalny dopuszczalny poziom prawdopodobieństwa ujawnienia danych. Cox i in. (2011) przedstawili sformułowanie tego problemu w postaci mapy (R-U) jako dobry sposób jego postrzegania, który ma jednak braki, jeśli chodzi o praktyczny, efektywny sposób postępowania. Zależność R-U przedstawiono na rysunku 1.1.



Rys. 1.1. Zależność pomiędzy użytecznością danych a ryzykiem ujawnienia po zastosowaniu kontroli ujawniania danych

Uwaga: Kropki reprezentują różne zbiory danych, w stosunku do których zastosowano metody kontroli ujawniania danych. Kropka w dolnym lewym rogu reprezentuje niskie ryzyko ujawnienia danych, natomiast kropka w prawym górnym rogu przekracza przyjętą dopuszczalną granicę ryzyka ujawniania przy wysokiej użyteczności.

Źródło: Hundepool i in. (2012).

– *k*-anonimowość – zbiór spełnia kryterium *k*-anonimowości, dla $k > 1$, jeśli dla każdej kombinacji pseudoidentyfikatorów istnieje w tym zbiorze reprezentacja przynajmniej *k* rekordów. Jeśli to kryterium zostaje uznane za wystarczające, staje się jedynym kryterium branym pod uwagę w zakresie badania stopnia poufności danych, przy jednoczesnym zaakceptowaniu ponoszonej straty informacji. Kryterium to jest jednak krytykowane jako jedynie konieczne, ale niewystarczające.
– Zróżnicowanie prywatności – rozpatrywanie problemu prywatności z innej perspektywy. Nie jest tu celem poszukiwanie pseudoidentyfikatorów lub

kombinacji pseudoidentyfikatorów. Problem zachowania prywatności rozwiązuje się jako dodawanie do zbioru danych dodatkowych rekordów, co ma zwiększyć gwarancję prywatności w sensie probabilistycznym. W praktyce dane przed publikacją są modyfikowane poprzez dodanie do rzeczywistych danych tzw. błędu, który ma pewien rozkład. Najczęściej przyjmowanym w tym celu rozkładem jest rozkład Laplace'a. Metoda ta jest krytykowana za poświęcanie zbyt małej uwagi użyteczności danych.

Jak wyżej wspomniano, w niektórych wypadkach zasada *k*-anonimowości może być niewystarczająca. Chodzi tutaj mianowicie o sytuację, gdy wartości pewnych cech wrażliwych są takie same dla zbyt wielu rekordów, wskutek czego istnieje ryzyko identyfikacji cech jednostki na podstawie tego podobieństwa i innych dostępnych danych. Stąd niektórzy badacze (np. Machanavajjhala i in., 2006) zaproponowali stosowanie także zasady *l*-różnorodności (ang. *l-diversity*). Mianowicie zbiór spełnia kryterium *l*-różnorodności, dla $l > 1$, jeśli zawiera przynajmniej l różnych wrażliwych wartości każdej cechy takich, że są one najczęściej przyjmowanymi wartościami, a ich częstości są identyczne lub prawie identyczne. Inne podejście w tym zakresie stanowi zasada *t*-bliskości (ang. *t-closeness*, sugerowana przez N. Li i in. (2007)), mówiąca, że dane są bezpieczne, gdy rozkład cechy wrażliwej w każdej badanej klasie jest bliski rozkładu tejże cechy w całej populacji (tzn. różnica pomiędzy tymi rozkładami nie przekracza ustalonego progu t). Z tych dwóch zasad zasada *l*-różnorodności wydaje się nieco łatwiejsza do spełniania w praktyce, jako że specyfika definicji i struktury określonych klas jednostek powodują, że i rozkłady badanych cech w takich klasach z natury rzeczy mogą mieć pewne niezbywalne formy. W takich wypadkach bardziej racjonalne może być spełnienie zasady *l*-różnorodności. Z drugiej strony wszakże *l*-różnorodność jest ograniczona w swym założeniu o wiedzy posiadanej przez intruza: może on bowiem pozyskać informacje o wrażliwej cesze, gdy dysponuje jej globalnym rozkładem. Poza tym metody ochrony prywatności danych zazwyczaj koncentrują się na danych jakościowych, z których łatwiej odczytać intruzowi wrażliwe informacje, zwłaszcza jeśli ma on do dyspozycji dodatkowo dane wyrażone na skali różnicowej lub ilorazowej, bliskie odpowiednich wartości w udostępnianym zbiorze.

1.4. Inne problemy związane z poufnością danych

Cox i in. (2011) w przeprowadzonej dyskusji na temat paradygmatu R-U podali takie problemy związane z metodami kontroli ujawniania danych jak: transparentność metod, problem danych panelowych i dane administracyjne. Omówiono je poniżej.

– **Transparentność metod.** Ogólną praktyką w tym zakresie jest, że instytucje udostępniające dane statystyczne odmawiają ujawnienia szczegółów na temat zastosowanych metod kontroli ujawniania danych w procesie przygotowawczym do publikacji. Transparentność oznacza zakres, w jakim

instytucja informuje o metodach zastosowanych w kontroli ujawniania danych lub parametrach użytych w tychże metodach. Ideę transparentności zaczerpnięto z rozwiązań stosowanych w kryptografii, gdzie reguła kodowania nie polega na tajności algorytmu, ale na poufności stosowanych kluczy. Wybór rodzaju ujawnianych informacji często jest trudny i arbitralny.

– **Problem danych panelowych.** Dane takie stwarzają dodatkowe ryzyko ujawnienia informacji wrażliwych, które na ogół nie jest brane pod uwagę. Przy ocenie ryzyka uwzględnia się bowiem tylko aspekt przekrojowy, który nie obejmuje czynnika panelowego. Dodatkowo nie ma wypracowanego sposobu postępowania w tym wypadku. Na ogół wprowadza się dodatkowe ograniczenia w dostępie do takich danych. Jednym ze sposobów radzenia sobie z tym problemem jest przygotowanie danych syntetycznych poprzez zamianę pierwotnych danych na szacunki uzyskane z modelu ekonometrycznego lub poprzez symulację regresji.
– **Dane administracyjne.** Przez dane administracyjne rozumie się tu zbiory danych gromadzonych w celach innych niż statystyczne. Wraz z rozwojem technologii informatycznych rośnie ilość gromadzonych danych, a w konsekwencji zainteresowanie takimi danymi ze strony statystyków, gdyż mogą być one wykorzystane jako dodatkowe źródło wiedzy do analiz i podnoszenia jakości badań statystycznych. Warto zauważyć, że nie ma wystarczających dowodów na efektywną możliwość zastosowania paradygmatu R-U do danych administracyjnych. Dodatkowo zależność R-U powinna być widziana w tym kontekście jako trójwymiarowa: koszt-ryzyko-użyteczność. Drugą kwestią jest to, że dane administracyjne zawierają jakościowe charakterystyki, które mogą być nieznane lub nierozumiane przez instytucje statystyczne. Relacja R-U jest niełatwa do zastosowania w środowisku kontrolowanym. Tym trudniej zatem oczekiwać jej efektywności w środowisku nierozpoznanym. Trzeci argument kwestionuje w ogóle możliwość zastosowania związku R-U do zmienionych danych administracyjnych. Dane administracyjne, aby mogły być wykorzystane do celów statystycznych, muszą być odpowiednio przetworzone. Metody stosowane w tym celu to przede wszystkim statystyczne łączenie rekordów. Technika łączenia rekordów zmienia zarówno ryzyko, jak i użyteczność danych, co wpływa na możliwość zastosowania paradygmatu R-U.

Należy zwrócić uwagę, że wartość analityczna zbiorów danych może być znacznie wyższa poprzez połączenie danych z różnych źródeł. Łączenie owo może się opierać na jednoznacznych identyfikatorach lub na łączeniu statystycznym (ang. *record linkage*) czy parowaniu statystycznym (ang. *statistical matching*). W niektórych krajach statystyka publiczna jest gestorem połączonych zbiorów danych, które są tworzone poza nią. Takie zbiory pozostają przedmiotem zainteresowania środowiska naukowego, mogą jednak również zwiększać ryzyko utraty poufności.

Przeprowadzone w wielu krajach badania pokazały, że dane powstałe z połączenia różnych źródeł są same w sobie postrzegane jako ryzykowne z punktu widzenia prywatności (UNECE, 2007). W tej sytuacji istotne jest przestrzeganie wcześniej zarysowanych kryteriów przy udostępnianiu mikrodanych. Na przykład w Kanadzie od połowy lat 80. funkcjonuje polityka dotycząca łączenia różnych źródeł danych jednostkowych pod nazwą *record linkage policy*. Ma ona gwarantować prywatność jednostek, umożliwiając jednak warunkowo łączenie różnych źródeł danych. Łączenie może zostać podjęte tylko w celach statystycznych, a jego wartość dodana jest ważona względem możliwości naruszenia poufności. Istnieje pewien zdefiniowany zbiór możliwych połączeń, dla którego nie trzeba uzyskiwać akceptacji kolegialnego ciała – Policy Commitee. Są to łączenia, dla których ryzyko naruszenia prywatności jest niskie. Wszystkie inne typy łączenia muszą uzyskać zgodę Komitetu, a jeśli zawierają dane osobowe, przypadki połączeń są publikowane na stronie internetowej urzędu. Łączeniem jest każde wiązanie dwóch lub więcej mikrorekordów, które utworzą rekord złożony, a które dotyczą osoby, rodziny, gospodarstwa domowego, mieszkania, firmy, instytucji itd. (Statistics Canada, 2017).

Bardzo złożoną kwestią pozostaje stosowanie narzędzi kontroli ujawniania danych w wypadku big data, czyli zbiorów danych jednostkowych o bardzo dużym rozmiarze, zmienności i różnorodności, których przetwarzanie jest zazwyczaj czaso- i kosztochłonne, gdyż wymaga wysokiej wydajności narzędzi oraz skomplikowanych algorytmów, ale dzięki temu można uzyskać niezwykle wartościowe informacje. Dyskusję na temat SDC w tym wypadku przeprowadzili m.in. de Wolf i Zeelenberg (2015). Zasugerowali, że najlepszymi rozwiązaniami w tym zakresie, chroniącymi dane wrażliwe, mogą być: opracowywanie danych syntetycznych podobnych do oryginalnych, tworzenie statystycznych informacji strumieniowych czy wprowadzenie niepewności tylko do informacji wrażliwych (w wypadku big data pełna populacja nie jest znana).

Nie można także zapominać o danych dostarczanych „spontanicznie" przez różnego rodzaju osoby na blogach lub na portalach społecznościowych. Statystycy powinni w miarę możliwości monitorować takie informacje, gdyż mogą one potencjalnie stać się pomocniczymi źródłami wiedzy w identyfikacji informacji wrażliwych na podstawie danych statystyki publicznej.

1.5. Rozwiązania prawne i zasady stosowane w praktyce międzynarodowej

Punktem wyjścia konstrukcji właściwego i efektywnego mechanizmu kontroli ujawniania danych muszą być stosowne uregulowania formalne. Na nich bowiem opiera się każdy zorganizowany system statystyczny. Również w wypadku badań statystycznych wykorzystujących dane, metody albo narzędzia statystyczne, a pro-

wadzonych poza takim systemem (np. przez ośrodki badania opinii społecznej, agencje ratingowe, studentów i doktorantów itp.), konieczne jest przestrzeganie pewnych ogólnie przyjętych reguł etycznych. W tej części opracowania przedstawiono najważniejsze przykłady takich regulacji. Uwzględniono tutaj także realia polskiej statystyki publicznej w tym kontekście. Część ta rozpoczyna się od prezentacji aktów i innych dokumentów prawnych na ten temat, by w następnej kolejności przejść do opracowań bardziej ogólnych, które określają pewne ramy etyczne działań niepodlegających formalnym przepisom, ale powszechnie realizowanych przez różne osoby czy instytucje. Spełnianie tych ramowych warunków jest bowiem oczekiwane chociażby ze względu na pryncypia współżycia i interesu społecznego.

1.5.1. Regulacje prawne

Jednym z kluczowych dokumentów w omawianym zakresie jest Rozporządzenie Parlamentu Europejskiego i Rady (WE) nr 223/2009 z dnia 11 marca 2009 r. w sprawie statystyki europejskiej oraz uchylające rozporządzenie Parlamentu Europejskiego i Rady (WE, Euratom) nr 1101/2008 w sprawie przekazywania do Urzędu Statystycznego Wspólnot Europejskich danych statystycznych objętych zasadą poufności, rozporządzenie Rady (WE) – nr 322/97 w sprawie statystyk Wspólnoty oraz decyzję Rady 89/382/EWG, Euratom w sprawie ustanowienia Komitetu ds. Programów Statystycznych Wspólnot Europejskich. Podstawowe założenia zawarte w tym rozporządzeniu, które stanowią fundamenty systemów kontroli ujawniania danych w krajowych urzędach statystycznych państw UE (i nie tylko), są – według preambuły dokumentu – następujące:

- Poufne informacje zbierane przez krajowe i wspólnotowe organy statystyczne do tworzenia statystyki europejskiej powinny być chronione w celu zdobycia i utrzymania zaufania stron odpowiedzialnych za przekazywanie tych danych. Poufność danych musi podlegać tym samym zasadom we wszystkich państwach członkowskich.
- Środowisko naukowe powinno mieć możliwość szerszego dostępu do poufnych danych wykorzystywanych do opracowywania, tworzenia i rozpowszechniania statystyki europejskiej dla analizy na rzecz postępu naukowego w Europie. Należy zatem poprawić pracownikom naukowym dostępność do poufnych danych wykorzystywanych do prowadzonych badań, bez uszczerbku dla wysokiego poziomu ochrony, jakiej wymagają poufne dane statystyczne.
- Wykorzystywanie poufnych danych do celów innych niż wyłącznie statystyczne – takich jak cele administracyjne, prawne lub podatkowe, bądź do kontroli jednostek statystycznych – powinno być surowo zabronione.

Wynika stąd, że kontrola ujawniania danych powinna uwzględniać – czasem stojące w pewnym konflikcie wobec siebie – interesy różnych grup społecznych: osób

i podmiotów dostarczających danych oraz użytkowników tychże danych. Warto zauważyć szczególne podkreślenie w rozporządzeniu roli środowiska naukowego jako motoru postępu społeczno-ekonomicznego i technologicznego współczesnego świata. Twórcy rozporządzenia niejako przyznają mu szersze niż innym użytkownikom danych prawa dostępu do informacji statystycznych, równocześnie jednak zobowiązują naukowców (na równi ze statystykami) do zapewnienia odpowiedniego poziomu ochrony poufności tychże danych.

Wśród sześciu zasad statystycznych wymienionych w omawianym rozporządzeniu do kontroli ujawniania danych można odnieść bezpośrednio zasady poufności i opłacalności. Pierwsza z nich jest tutaj rozumiana (zgodnie z art. 2) jako ochrona poufnych danych odnoszących się do każdej jednostki statystycznej, które zostały pozyskane bezpośrednio do celów statystycznych lub pośrednio ze źródeł administracyjnych lub innych. Wynika z tego zakaz wykorzystywania uzyskanych danych do celów innych niż statystyczne oraz ich bezprawnego ujawnienia.

Zatrzymajmy się obecnie przy drugiej z tych zasad – opłacalności. Pojęcie to oznacza, że koszty tworzenia danych statystycznych muszą być współmierne do wagi oczekiwanych wyników i korzyści oraz że zasoby muszą być wykorzystywane w sposób optymalny, a obciążenie respondentów musi być ograniczone do minimum. Wymagane informacje, o których przekazanie wystąpiono z wnioskiem, są, jeżeli istnieje taka możliwość, łatwe do pobrania z dostępnych rejestrów lub źródeł (art. 2).

Chociaż kwestie związane z kontrolą ujawniania danych nie zostały tutaj literalnie wymienione, to koszty tworzenia danych statystycznych obejmują wszak także koszty prowadzenia kontroli mikrodanych i danych wynikowych pod kątem zabezpieczenia poufności wrażliwych informacji przed ich ujawnieniem osobom do tego nieuprawnionym. Tym samym zastosowane metody kontroli ujawniania danych także powinny być współmierne do spodziewanego zakresu ochrony i użyteczności finalnie udostępnionych danych dla ich użytkownika. Tak więc i w tym wypadku konieczne jest zagwarantowanie optymalnej równowagi pomiędzy tymi dwoma potrzebami.

Przytoczmy jeszcze kilka innych kluczowych definicji z art. 3 tego rozporządzenia, ważnych z punktu widzenia rozważanych tutaj spraw:

- dane statystyczne: ilościowe i jakościowe, zagregowane i reprezentatywne informacje opisujące złożone zjawisko w rozpatrywanej populacji,
- opracowywanie danych: działania zmierzające do utworzenia, usprawnienia i udoskonalenia metod, standardów i procedur statystycznych stosowanych przy tworzeniu i rozpowszechnianiu danych statystycznych, jak również do projektowania nowych danych i wskaźników,
- tworzenie informacji statystycznych: wszelka działalność związana ze zbieraniem, przechowywaniem, przetwarzaniem, zestawianiem i analizą niezbędnych do zestawienia danych statystycznych,

– rozpowszechnianie danych: działanie, którego celem jest udostępnianie danych statystycznych i analiz statystycznych użytkownikom,
– zbieranie danych: badania oraz wszelkie inne formy pozyskiwania informacji z różnych źródeł, w tym ze źródeł administracyjnych,
– jednostka statystyczna: podstawowa jednostka objęta obserwacją – osoba fizyczna, gospodarstwo domowe, podmiot gospodarczy czy inny podmiot – do której odnoszą się dane,
– dane poufne: dane umożliwiające bezpośrednią lub pośrednią identyfikację jednostek statystycznych, co skutkuje ujawnieniem informacji indywidualnych; aby określić, czy możliwa jest identyfikacja danej jednostki statystycznej, należy wziąć pod uwagę wszystkie przewidywalne środki, które mogą być użyte przez osobę trzecią do jej zidentyfikowania,
– wykorzystanie do celów statystycznych: wykorzystanie wyłącznie do celów opracowywania i tworzenia wyników oraz analiz statystycznych,
– identyfikacja bezpośrednia: identyfikacja jednostki statystycznej według jej nazwy lub adresu lub według publicznie dostępnego numeru identyfikacyjnego;
– identyfikacja pośrednia: identyfikacja jednostki statystycznej z użyciem wszelkich środków innych niż środki identyfikacji bezpośredniej.

Wszystkie te definicje w większym lub mniejszym stopniu wiążą się z kontrolą ujawniania danych. Kontrola ta wpisuje się bowiem zarówno w prace dotyczące zbierania danych (tutaj dotyczy to przede wszystkim zapewnienia bezpieczeństwa gromadzonych informacji, a także zdobycia zaufania respondenta w tym zakresie z poszanowaniem jego własnej stosownej pragmatyki), jak też ich opracowania i rozpowszechniania. Możliwość bezpośredniej i pośredniej identyfikacji jednostek zaś to podstawa tejże kontroli.

Kwestii poufności danych statystycznych został poświęcony rozdział piąty rozpatrywanego rozporządzenia. Podkreślono w nim, że krajowe i międzynarodowe instytucje statystyczne powinny wykorzystywać zgromadzone poufne dane wyłącznie do celów statystycznych – chyba że zainteresowana jednostka statystyczna wyrazi jednoznaczną zgodę na wykorzystanie swych danych do innych celów. Podobne reguły ten akt prawny przewiduje dla wynikowych informacji statystycznych, pozwalających na identyfikację konkretnej jednostki. Dopuszcza się tu wszakże pewne szczególne wyjątki określone innymi przepisami, pod warunkiem że rozpowszechnianie to nie stanowi uszczerbku dla poufności informacji statystycznych.

Do kontroli ujawniania danych *sensu stricto* odnosi się zapis mówiący, że krajowe instytucje statystyczne i inne oraz organy Komisji Europejskiej (w tym Eurostat) „podejmują wszelkie niezbędne środki regulacyjne, administracyjne, techniczne i organizacyjne działania w celu zapewnienia ochrony fizycznej i logicznej poufnych danych (kontrola ujawniania danych statystycznych)". Działania te

mają polegać na wykorzystaniu wszelkich dostępnych środków niezbędnych do zapewnienia harmonizacji zasad i wytycznych dotyczących zarówno fizycznej, jak i logicznej ochrony poufnych danych.

Rozporządzenie przewiduje także możliwość przyznawania prawa dostępu do poufnych danych pracownikom naukowym prowadzącym analizy statystyczne do celów prowadzonych badań, pod warunkiem że dane te umożliwiają tylko pośrednią identyfikację jednostek statystycznych. W praktyce oznacza to, że udostępniane mogą być dane jednostkowe, z których usunięto podstawowe cechy identyfikacyjne jednostki (np. imię, nazwisko, numer identyfikacyjny ewidencji ludności w przypadku osoby). Jednak w i tej sytuacji powinny być podjęte działania zapewniające uniemożliwienie pośredniego odtworzenia innych danych wrażliwych (i tym samym identyfikacji jednostki). To właśnie zadanie kontroli ujawniania danych.

Dokument przyznaje ponadto prawo dostępu do danych zawartych w źródłach administracyjnych i ich wykorzystywania w stopniu, w jakim są one niezbędne do opracowywania, tworzenia i rozpowszechniania statystyki europejskiej. Choć nie mówi tego wprost, to jednak i tutaj kontrola ujawniania danych bywa potrzebna. Rozporządzenie uznaje natomiast prawo do nakładania sankcji za naruszenie poufności informacji statystycznych.

Warto podkreślić, że określenie „kontrola ujawniania danych statystycznych" – mimo że powyższe rozporządzenie go używa – nie zostało tam formalnie zdefiniowane. Uczyniono to dopiero w Rozporządzeniu Komisji (UE) nr 1151/2010 z dnia 8 grudnia 2010 r. w sprawie wykonania Rozporządzenia Parlamentu Europejskiego i Rady (WE) nr 763/2008 w sprawie spisów powszechnych ludności i mieszkań w odniesieniu do ustaleń dotyczących raportów jakości i ich struktury oraz formatu technicznego przekazywania danych. Według art. 2 tego rozporządzenia kontrola ujawniania danych statystycznych „oznacza metody i procesy zastosowane w celu zminimalizowania ryzyka ujawnienia informacji na temat poszczególnych jednostek statystycznych, przy jednoczesnym udostępnianiu jak największej ilości informacji statystycznych".

Kolejny impuls rozwojowy w zakresie kontroli ujawniania danych przyniosła zmiana rozporządzenia Parlamentu Europejskiego i Rady nr 223/2009, dokonana mocą Rozporządzenia Komisji (UE) nr 557/2013 z dnia 17 czerwca 2013 r. w sprawie wykonania rozporządzenia (WE) nr 223/2009 Parlamentu Europejskiego i Rady w sprawie europejskiej statystyki w zakresie dostępu do poufnych danych do celów naukowych i uchylające rozporządzenie Komisji (WE) nr 831/2002.

Dokument ten precyzuje zasady dostępu do danych statystycznych przekazywanych Komisji Europejskiej i jej organom przez krajowe instytucje statystyczne osobom, które będą te dane wykorzystywać do celów naukowych. Twórcy rozporządzenia uwypuklają więc specyfikę udostępniania danych w tym zakresie wynikającą z jednej strony z istotności badań naukowych oraz ich wyników, a z drugiej strony z tego, że uzyskanie odpowiedniej jakości rezultatów takich badań

zależy od możliwie jak najwszechstronniejszego dostępu do niezbędnych danych statystycznych. Stąd pojawiła się konieczność opracowania w takim wypadku specjalnych reguł pozwalających na stworzenie użytkownikom ze środowiska naukowego, większych niż innym korzystającym, możliwości w zakresie dostępu do informacji jednostkowych. Nakłada to jednak poważniejsze obowiązki – zarówno na stronę udostępniającą, jak i na użytkownika danych, zwłaszcza w zakresie kontroli udostępniania i wykorzystania danych. Dokument wprowadza kilka nowych definicji z tym związanych, a mianowicie:

- poufne mikrodane do celów naukowych: dane, które umożliwiają tylko pośrednią identyfikację jednostek statystycznych, przyjmujące formę plików mikrodanych do użytku chronionego lub plików mikrodanych do zastosowań naukowych,
- mikrodane do użytku chronionego: poufne dane do celów naukowych, wobec których nie zastosowano dalszych metod kontroli ujawniania danych statystycznych,
- mikrodane do zastosowań naukowych: poufne dane do celów naukowych, wobec których zastosowano metody kontroli ujawniania danych statystycznych, tak by ryzyko zidentyfikowania jednostki statystycznej ograniczyć do odpowiedniego poziomu i zgodnie z obecnymi najlepszymi praktykami,
- metody kontroli ujawniania danych statystycznych: metody zmniejszenia ryzyka ujawnienia informacji na temat jednostek statystycznych polegające zazwyczaj na ograniczaniu ilości lub modyfikacji uwalnianych danych,
- punkt dostępu: środowisko fizyczne lub wirtualne wraz z jego strukturą organizacyjną, w którym udzielany jest dostęp do poufnych danych do celów naukowych.

Rozporządzenie opisuje procedurę udzielania przez Komisję Europejską dostępu do danych. Czyni się to na wniosek – jak to określono – uznanej jednostki badawczej o przeprowadzenie badań naukowych, w którym jednostka taka wskazuje rodzaj poufnych danych do celów naukowych, o dostęp do których się ubiega. Dostęp ów jest przyznawany przez Komisję (Eurostat) lub przez inny punkt dostępu zatwierdzony przez Komisję, pod warunkiem że właściwa krajowa instytucja statystyczna, która dostarczyła danych, wyrazi zgodę na ich udostępnienie. Uznana jednostka badawcza oznacza zaś jednostkę, którą określono jako taką, biorąc pod uwagę m.in. cel jej funkcjonowania na podstawie jej statutu, misji lub innej deklaracji, to czy cel ten odnosi się do badań, udokumentowanych osiągnięć naukowych lub renomy i doświadczenia jednostki jako naukowo-badawczej, posiadanie przez jednostkę odrębnej osobowości prawnej, niezależność i samodzielność w formułowaniu wniosków naukowych, brak związków z politycznymi działaniami podmiotu, do którego należy, a także zapewnienie danym odpowiedniego zabezpieczenia (spełnianie odpowiednich wymogów techniczno-infrastrukturalnych gwarantujących bezpieczeństwo danych).

Jednostka wnioskująca o dostęp do danych musi złożyć – podpisane przez swego przedstawiciela – zobowiązanie do zachowania poufności obejmujące wszystkich naukowców w niej pracujących, którzy będą mieli dostęp do poufnych danych dla celów naukowych (przy czym każdy naukowiec powinien podpisać stosowne zobowiązanie oddzielnie) oraz określające warunki dostępu (w tym wykorzystanie punktu dostępu), obowiązki owych naukowców, środki przestrzegania poufności danych statystycznych i sankcje w przypadku naruszenia tych obowiązków[5]. Wniosek o dostęp do danych powinien zawierać także m.in. sformułowany cel i oczekiwany rezultat badań oraz uzasadnienie, dlaczego danego badania nie można zrealizować bez danych poufnych. Udostępnienie danych przez Komisję wymaga najpierw zatwierdzenia tego przez krajową instytucję statystyczną, która jest gestorem owych danych.

Artykuł 8 dokumentu reguluje też kilka kwestii organizacyjnych związanych z dostępem do danych dla celów naukowych. W tym celu można utworzyć specjalne punkty dostępu, których prawidłowość organizacji i bezpieczeństwa podlega ocenie i zatwierdzeniu przez Komisję (Eurostat). Punkt taki powinien się znajdować na terenie placówki krajowej instytucji statystycznej. Ewentualnie – o ile instytucja dostarczająca danych wyrazi na to zgodę – może on być ulokowany w innym miejscu. Formalną podstawą funkcjonowania i kontroli punktów dostępu przez Komisję jest specjalna umowa zawierana między nią a organizacją czy instytucją, w której ów punkt się znajduje.

Rozporządzenie wyznacza tym samym pewne określone wzorcowe standardy udostępniania mikrodanych dla celów naukowych nie tylko na szczeblu europejskim, ale – pośrednio – także krajowym. W tym ostatnim kontekście można ująć je w następujących punktach:

- dostęp do mikrodanych powinien być udzielany na podstawie wniosku jednostki prowadzącej działalność naukowo-badawczą, w którym wnioskodawca podaje dokładne dane o sobie oraz o zakresie wnioskowanego dostępu do danych i możliwego ich zabezpieczenia; instytucja udostępniająca dane ma zaś prawo dokonania szczegółowej weryfikacji takiego wniosku,
- wszystkie osoby z wnioskującej jednostki, które będą miały dostęp do tych danych, powinny złożyć stosowne oświadczenia o zobowiązaniu do zachowania poufności otrzymanych danych,
- podstawą udostępnienia danych jest stosowna umowa zawierana między zainteresowanymi stronami,
- jednostka udostępniająca dane lub z nich korzystająca powinna zorganizować odpowiednio wyposażony i zabezpieczony punkt dostępu do tych

[5] Przywołany akt prawny nie przewiduje bezpośrednio żadnych sankcji za takie naruszenie, gdyż legislacja UE pozostawia wprowadzenie stosownych regulacji państwom członkowskim oraz umawiającym się stronom.

danych; w tym drugim wypadku instytucja udostępniająca ma prawo do kontroli i weryfikacji prawidłowości funkcjonowania takiego punktu.

Przepisy te nie przewidują wprost możliwości występowania o dostęp do danych przez naukowców prowadzących indywidualne badania naukowe. Jednak i w tej sytuacji postępowanie to da się zastosować – z tym że w takim wypadku umowa o dostęp do danych byłaby zawierana nie z instytucją (której tutaj nie ma), a bezpośrednio z taką osobą, punkt dostępu zaś powinien być ulokowany w siedzibie instytucji udostępniającej dane.

Kwestia kontroli ujawniania danych pojawia się także w Rozporządzeniu Wykonawczym Komisji (UE) 2017/881 z dnia 23 maja 2017 r. w sprawie wykonania rozporządzenia Parlamentu Europejskiego i Rady (WE) nr 763/2008 w sprawie spisów powszechnych ludności i mieszkań w odniesieniu do ustaleń dotyczących raportów jakości i ich struktury oraz formatu technicznego przekazywania danych, zmieniającym rozporządzenie (UE) nr 1151/2010. Warto zauważyć, że użyta tutaj definicja rzeczonej kontroli różni się nieco od definicji wprowadzonej we wspomnianym rozporządzeniu nr 1151/2011. Tam bowiem podkreślano zminimalizowanie ryzyka ujawniania danych wrażliwych, jednak przy jednoczesnym zapewnieniu udostępniania jak największej ilości informacji statystycznych. Definicja w przytoczonym dokumencie z 2017 r. aspekt ilościowy pomija milczeniem, ustalając, że kontrola ujawniania danych statystycznych to „metody i procesy zastosowane w celu zminimalizowania ryzyka ujawnienia informacji na temat poszczególnych jednostek statystycznych podczas udostępniania informacji statystycznych". Wydaje się jednak, że – niezależnie od tego – kluczowym celem kontroli ujawniania danych jest wypracowanie optymalnego kompromisu pomiędzy bezpieczeństwem danych a maksymalizacją możliwego zakresu ich udostępniania.

Warto zobaczyć na kilku przykładach, jak kwestia ochrony poufności danych przedstawia się w różnych krajach europejskich.

W Wielkiej Brytanii zakres oraz zabezpieczenie poufności danych jest ustalone na podstawie zalecenia Komisji Gospodarczej Organizacji Narodów Zjednoczonych dla Europy (ang. United Nations Economic Commission for Europe – UNECE) z 1992 r.[6] Pełną – i zarazem bardzo zwięzłą – prezentację tych zasad zawiera dokument *Narodowe doradztwo statystyczne: poufność statystyki publicznej* (GSS, 2009). Autorzy dokumentu przytaczają podstawowe zasady etyki statystycznej i wskazują, jak odnoszą się do nich przepisy brytyjskie. Punktem wyjścia jest tutaj zasada mówiąca, że prywatne informacje o konkretnych osobach (w tym prawnych) pozyskane podczas badań statystycznych statystyki publicznej są poufne i powinny być wykorzystane wyłącznie dla celów statystycznych. Przypominają, że informacje legalnie umieszczone w domenach i na forach publicznych oraz publicznie dostępne nie stają się automatycznie poufne, gdy wykorzystuje się je dla

[6] W Polsce ciało to określane jest również jako Europejska Komisja Gospodarcza ONZ.

celów statystycznych. Nadmienia także o zaistniałej w ostatnim latach likwidacji klauzuli poufności dla niektórych rodzajów informacji, co wiąże się z rozwojem technologicznym przyczyniającym się do ich – niejako automatycznego – upowszechniania (np. mapy Google). Znalazło to odzwierciedlenie m.in. w podstawowym dla brytyjskiej statystyki dokumencie, którym jest ustawa w sprawie statystyki i usług ewidencyjnych z 26 lipca 2007 r. (ang. The Statistics and Registration Service Act 2007). Dokument ustanawia Zespół Statystyki (ang. *Statistics Board*), składający się z członków wykonawczych (ang. *executive*) i niewykonawczych (ang. *non-executive*). Członkowie niewykonawczy to: przewodniczący Zespołu (mianowany przez królową Elżbietę II[7]) i przynajmniej pięć osób desygnowanych przez ministra Urzędu Gabinetu (ang. Ministry for the Cabinet Office). Członkowie wykonawczy zaś to: Narodowy Statystyk (ang. National Statistician, mianowany przez królową i zatrudniony przez Zespół na warunkach wskazanych przez monarchinię) oraz dwaj inni pracownicy Zespołu wskazani przez jego członków niewykonawczych. Zespół pełni nadzór nad statystyką publiczną.

Kwestię poufności informacji osobowych uregulowano tam w art. 39. Odnosi się ona przede wszystkim do Zespołu, ale tyczy też całej statystyki, jako że statystyka publiczna jest Zespołowi podporządkowana. Informacja osobowa jest tutaj rozumiana jako informacja, która odnosi się do konkretnej osoby (w tym osoby prawnej) i ją identyfikuje, czyli identyfikacja osoby:

– jest określona w samej tej informacji,
– może zostać wydedukowana z tej informacji,
– może być wydedukowana z tej informacji w połączeniu jej z jakąkolwiek inną publicznie dostępną informacją.

Wspomniany wyżej Zespół Statystyki ustala m.in. kryteria udzielania dostępu do informacji poufnych dla naukowców oraz formalnie przewiduje udzielanie takiego właśnie dostępu. Ustawa reguluje też odpowiedzialność karną za ujawnienie chronionych danych lub dopuszczenie do takiego ujawnienia.

Przywołany wcześniej dokument *Narodowe doradztwo statystyczne: poufność statystyki publicznej* precyzuje istotę informacji osobowej (tutaj zwanej prywatną – ang. *private information*) jako informację, która spełnia jednocześnie trzy następujące warunki:

– odnosi się do identyfikowalnej osoby prawnej lub fizycznej,
– nie znajduje się w domenie publicznej ani w innych powszechnie dostępnych źródłach wiedzy,
– w razie ujawnienia spowodowałaby szkodę, krzywdę lub cierpienie.

Osoby gromadzące, przetwarzające i opracowujące dane w statystyce publicznej powinny być ponadto – według autorów dokumentu – świadome tego, że informacją prywatną jest również ta informacja dotycząca konkretnej osoby, która

[7] Od 8 września 2022 r. – przez króla Karola III.

została im przekazana, ale oczekuje się, że nie zostanie publicznie ujawniona, oraz inna informacja, która została uznana za prywatną na mocy odpowiednich przepisów prawnych.

Dokument postuluje podawanie do publicznej wiadomości ogólnej informacji, w jaki sposób chroniona jest poufność informacji w danym badaniu statystycznym. Odpowiednie zobowiązanie do takiej ochrony powinien także zawierać plan badania. Plan ów powinien być jednak odpowiednio zoptymalizowany w celu umożliwienia publikacji tak szczegółowych informacji wynikowych, jak to tylko jest rozsądnie możliwe, biorąc też pod uwagę potrzeby użytkowników. Wymóg, aby wybór stosowanych metod kontroli ujawniania danych był wystarczający, nie wydaje się nadmiernie restrykcyjny, gdyż według niego można używać w zasadzie dowolnej metody. Jednak warto pamiętać, że pewne podejścia są w tej mierze bardziej użyteczne od innych, dając lepszy poziom ochrony. Zastosowane metody kontroli ujawniania danych oraz ich użyteczność dla statystyki publicznej powinny być opisane, np. jako część metadanych. Opis ów jednak nie może zostać sporządzony z taką szczegółowością, która zwiększyłaby ryzyko ujawniania chronionych danych, stanowiąc niejako instrukcję dla osoby nieuprawnionej, jak wprowadzoną ochronę naruszyć. W wypadku przekazywania poufnych danych należy sporządzić stosowną dokumentację, która powinna się składać z umowy na takie udostępnienie oraz zbioru operacyjnego zawierającego zapisy wszystkich kroków związanych z tym przekazaniem i posługiwaniem się udostępnionymi danymi. To umożliwi dokładną kontrolę bezpieczeństwa owych danych.

W Niemczech kluczowe regulacje dotyczące statystyki są zawarte w ustawie o statystyce dla celów federalnych (niem. Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG)), uchwalonej w 1987 r. i później wielokrotnie zmienianej. Warto zauważyć, że §13a rzeczonej ustawy przewiduje możliwość łączenia różnych zbiorów danych pochodzących ze źródeł administracyjnych, a zatem:

- danych gospodarczych i z zakresu ochrony środowiska według przedsiębiorstw, zakładów i miejsc pracy, łącznie z informacjami statystycznymi zgromadzonymi przez Niemiecki Bank Federalny (niem. Deutsche Bundesbank)[8],
- danych z rejestru statystycznego,
- danych pozyskanych na mocy ustawy o wykorzystaniu danych administracyjnych,
- danych pozyskanych przez urzędy federalne oraz urzędy krajów związkowych z powszechnie dostępnych źródeł.

Według tego przepisu numery identyfikacyjne jednostek powinny być przechowywane przez 30 lat od dnia pozyskania danych. Po upływie tego czasu muszą

[8] Bank centralny Republiki Federalnej Niemiec.

być usunięte. Ustawa w §16 nakłada na osoby zatrudnione w administracji publicznej i zajmujące się gromadzeniem i przetwarzaniem takich danych obowiązek zachowania ich w tajemnicy. Nie dotyczy to danych jednostkowych:

- na ujawnienie których pisemną zgodę wyraziły zainteresowane jednostki (o ile w szczególnych okolicznościach nie wyrażono odmiennej woli w innej formie),
- pochodzących z powszechnie dostępnych źródeł,
- które Federalny Urząd Statystyczny lub urzędy statystyczne krajów związkowych pozyskały łącznie z danymi jednostkowymi innych respondentów i umieściły wśród wynikowych danych statystycznych,
- które nie zostały przyporządkowane do konkretnych respondentów czy zainteresowanych jednostek.

Ostatnia opcja dotyczy w istocie także danych odpersonalizowanych, czyli danych jednostkowych, z których usunięto podstawowe cechy identyfikacyjne jednostek oraz za pomocą metod kontroli ujawniania danych ograniczono do minimum ryzyko pośredniej identyfikacji owych jednostek.

Inne ważne z rozpatrywanego punktu widzenia przepisy omawianej ustawy dotyczą udostępniania danych organom państwowym w celach planistycznych oraz współpracy między Federalnym Urzędem Statystycznym a urzędami statystycznymi krajów związkowych w zakresie przetwarzania danych jednostkowych i opracowywania na ich podstawie informacji wynikowych. W szczególności dane jednostkowe pochodzące ze statystyki publicznej mogą być wykorzystywane do realizacji statystycznych zadań gmin i obowiązków gmin, pod warunkiem że będzie się to odbywać na specjalnych stanowiskach odpowiadających wymogom przepisów o statystyce publicznej, które są odseparowane od pozostałych stanowisk pracy w urzędach gminnych oraz których organizacja i stosowne metody gwarantują zachowanie poufności danych.

Dla celów naukowych dane jednostkowe mogą być udostępniane wyższym uczelniom oraz innym instytucjom realizującym suwerenne badania naukowe. Jednak dane te muszą spełniać dwa następujące warunki:

- przyporządkowanie zawartych w nich informacji konkretnym jednostkom mogłoby nastąpić tylko przy użyciu nieproporcjonalnie dużych nakładów czasu, kosztów i pracy (a zatem chodzi o dane faktycznie zanonimizowane),
- dostęp do nich będzie się odbywać w specjalnie ustalonych częściach Federalnego Urzędu Statystycznego lub urzędów statystycznych krajów związkowych, w których zostaną zapewnione skuteczne środki dla zapewnienia poufności owych danych.

Rzecz jasna, osobami, którym udostępniane są dane, mogą być tylko osoby pełniące określoną funkcję publiczną, z mocy której zobowiązane są do zachowania tajemnicy, jak też osoby, które zobowiązane są do zachowania poufności danych na takich samych zasadach.

Z kolei w Holandii do dnia dzisiejszego obowiązuje ustawa o statystyce gospodarczej uchwalona 28 grudnia 1936 r. (modyfikowana ustawowymi zmianami z 11 lutego 1988 r. oraz 4 czerwca 1992 r.), dotycząca miar stosowanych w celu uzyskania prawidłowych gospodarczych informacji statystycznych. Ustawa m.in. upoważnia naczelny organ statystyki publicznej w tym kraju, którym jest do dziś Centralne Biuro Statystyczne Holandii (nl. Centraal Bureau voor de Statistiek – CBS), znane na arenie międzynarodowej także jako Statistics Netherlands, do pozyskiwania i gromadzenia danych niezbędnych dla uzyskania rzetelnych informacji o gospodarce.

Artykuł 4 tejże ustawy zabrania osobom wykonującym jakiekolwiek zadania przewidziane ustawą wykorzystywać ustalenia i informacje oraz dane uzyskane w drodze inspekcji ksiąg, dokumentów i opracowań (do przeprowadzenia której pracowników CBS może upoważnić nadzorujący minister spraw gospodarczych) do innych celów niż konieczne do realizacji ich obowiązków.

Do kontroli ujawniania danych odnosi się bezpośrednio art. 5 ustawy, stanowiący, że dane zgromadzone na podstawie omawianej ustawy nie będą ujawniane w takiej formie, w której mogłyby zostać wydedukowane dane i informacje o konkretnej osobie, firmie lub instytucji – chyba że taka osoba, szef firmy lub zarząd instytucji nie miały zastrzeżeń odnośnie do takowego ujawniania. Za świadome złamanie tego przepisu art. 6 przewiduje dla winnego karę do sześciu miesięcy pozbawienia wolności lub odpowiednio wysoką grzywnę, dla osób zaś odpowiedzialnych za ujawnianie wrażliwych danych – karę pozbawienia wolności do trzech miesięcy lub niższą grzywnę.

Postanowienia te powtórzone zostały w ustawie o Centralnym Biurze Statystycznym Holandii z dnia 20 listopada 2003 r. z późniejszymi zmianami (Statistics Netherlands Act – Act of 20 November 2003 enacting a Law governing Statistics Netherlands). Można znaleźć tu pewne doprecyzowania wspomnianych wcześniej zapisów. W sekcji 37 zapisano mianowicie, że dane otrzymane i gromadzone przez dyrektora generalnego statystyki (czyli dyrektora CBS) mogą być publikowane jedynie w taki sposób, aby z tychże danych nie można było odtworzyć czy rozpoznać żadnych informacji o konkretnej osobie, gospodarstwie domowym, firmie czy instytucji. W wypadku podmiotów gospodarczych przyjmuje się, że ochrona taka jest prowadzona, o ile nie ma racjonalnych powodów, aby przyjąć, że dany podmiot nie będzie miał nic przeciwko publikacji dotyczących go danych. Zgodnie z postanowieniami sekcji 38 dyrektor generalny statystyki jest odpowiedzialny za zastosowanie środków technicznych i organizacyjnych odpowiednich dla zabezpieczenia danych przed ich utratą, zniszczeniem, a także nieuprawnionym ich przeszukiwaniem, modyfikacją czy dostarczaniem. Dotyczy to także zabezpieczeń danych przekazywanych do organów Unii Europejskiej. W sekcji 39 podkreślono, że podczas każdorazowego przekazywania tych danych dyrektor generalny powinien się upewnić, że podjęto wszystkie możliwe admini-

stracyjne, techniczne i organizacyjne środki dla zapewnienia fizycznej i logistycznej ochrony poufnych danych przed ich bezprawną publikacją lub wykorzystaniem dla celów niestatystycznych.

W wypadku kanadyjskiego urzędu statystycznego (ang. Statistics Canada) zasady udostępniania danych opierają się na ustawie o statystyce (ang. Statistics Act) z 1985 r. Warto zauważyć, że reguluje ona m.in. kwestię udostępniania informacji zgromadzonych nie tylko przez Statistics Canada, ale i przez inne organy publiczne i korporacje. Artykuł 12 określa, że udostępnianie takie odbywa się w drodze porozumienia zawieranego przez ministra (członka Tajnej Rady Królewskiej dla Kanady, desygnowanego przez Gubernatora Kanady i właściwego dla spraw regulowanych tym aktem prawnym) z odpowiednią instytucją, organizacją czy podmiotem. Respondent jest informowany o możliwości zawarcia owego porozumienia. Jeśli nie wyrazi on dyrektorowi urzędu statystycznego (ang. *Chief Statistician*) pisemnej zgody na udostępnienie takich danych, owo udostępnienie nie będzie możliwe (chyba że udostępnienie takie danemu organowi czy podmiotowi jest obligatoryjne z mocy prawa). Co więcej, art. 13 wyraźnie określa, że osoba mająca pieczę nad danymi zgromadzonymi u danego gestora ma obowiązek udostępnienia owych zasobów osobie upoważnionej do dostępu do nich przez dyrektora urzędu statystycznego. Zasady udostępniania reguluje art. 17, wskazujący rodzaje informacji, na których udostępnienie musi wyrazić zgodę dyrektor urzędu statystycznego (w wypadku danych indywidualnych i jednostkowych zakłada on uzyskanie uprzedniej pisemnej zgody zainteresowanej osoby/jednostki) oraz uściśla, że dostęp do danych wrażliwych mogą mieć tylko osoby, które złożyły specjalne, ustalone przepisami tej ustawy, przyrzeczenie. Artykuł 18 nakłada z kolei bezwzględną tajemnicę statystyczną na dane wrażliwe. Mówi on bowiem, że – z wyjątkiem celów określonych w omawianej ustawie – żadna informacja identyfikująca jednostkę dostarczona do Statistics Canada nie może być użyta w żadnym postępowaniu. Ponadto w ust. (2) jednoznacznie stwierdza się, że żadna osoba upoważniona do dostępu do danych wrażliwych nie może zostać zmuszona przez żaden sąd, trybunał czy inny organ do składania ustnego świadectwa dotyczącego takich danych. Ustępy 18.1 i 18.2 przewidują, że dane indywidualne zgromadzone podczas powszechnych spisów ludności przeprowadzonych w latach 1910–2005 oraz w i po roku 2021 nie podlegają ochronie po upływie 92 lat od dnia przeprowadzenia spisu. W wypadku spisów z lat 2006, 2011 i 2016 – a także Narodowego Badania Gospodarstw Domowych (ang. National Household Survey) z 2011 r. – możliwe jest również zastosowanie tego przepisu, o ile jednak zainteresowana osoba wyraziła na to zgodę podczas spisu/badania.

Na zakończenie tych rozważań warto się jeszcze przyjrzeć rodzimym uregulowaniom w rozpatrywanym zakresie. Podstawę funkcjonowania polskiego systemu statystyki publicznej stanowi ustawa z dnia 29 czerwca 1995 r. o statystyce

publicznej. W zakresie zagadnień kontroli ujawniania danych wprowadza ona w art. 2 pojęcia danych jednostkowych identyfikowalnych, danych jednostkowych nieidentyfikowalnych oraz danych jednostkowych. Pierwsze z nich (**dane jednostkowe identyfikowalne**) określa dane statystyczne zawierające informacje dotyczące konkretnego podmiotu gospodarki narodowej albo osoby fizycznej, identyfikujące bezpośrednio ten podmiot albo osobę według nazwy, imienia i nazwiska, adresu lub publicznie dostępnego numeru identyfikacyjnego oraz pozwalające na pośrednią identyfikację tego podmiotu albo osoby z użyciem innych środków niż środki pozwalające na bezpośrednią identyfikację, z wyłączeniem środków wymagających nadmiernych kosztów, czasu lub działań. **Dane jednostkowe nieidentyfikowalne** to z kolei dane statystyczne zawierające informacje dotyczące konkretnego podmiotu gospodarki narodowej albo osoby fizycznej, niepozwalające na bezpośrednią ani pośrednią identyfikację tego podmiotu albo osoby. Natomiast **dane jednostkowe** stanowią sumę danych jednostkowych identyfikowalnych i danych jednostkowych nieidentyfikowalnych. Określenie jednostkowych danych identyfikowalnych odpowiada więc pojęciu danych poufnych występującemu w regulacjach na poziomie Unii Europejskiej czy określeniu „informacje prywatne/osobowe" według stosownej legislacji brytyjskiej. Istota takich danych występuje też w przepisach regulujących funkcjonowanie statystyki publicznej w innych krajach, aczkolwiek nie zawsze pod konkretną, zdefiniowaną uprzednio, nazwą. Oryginalną cechą polskiej ustawy jest wprowadzenie *sensu stricto* pojęcia danych jednostkowych nieidentyfikowalnych. Oczywiście, w stosownych aktach prawnych na arenie europejskiej występują określenia wskazujące wyraźnie na ten typ danych, ale zazwyczaj odbywa się to pośrednio (w kontekście możliwości identyfikacji pośredniej lub bezpośredniej, albo też – jak w przypadku niemieckim – w odniesieniu do danych, które nie zostały przyporządkowane do konkretnych respondentów czy zainteresowanych jednostek).

Artykuł 35d ustawy nakłada na służby statystyki publicznej, jeżeli pozwala na to cel przetwarzania danych, wymóg, aby dane zostały poddane procesowi pseudonimizacji. Obowiązujące od 25 maja 2018 r. rozporządzenie RODO, o którym była mowa we wprowadzeniu, przewiduje również możliwość pseudonimizacji danych, którą definiuje się w art. 4 jako przetworzenie danych osobowych w taki sposób, by nie można ich było już przypisać konkretnej osobie, której dane dotyczą, bez użycia dodatkowych informacji, pod warunkiem że takie dodatkowe informacje są przechowywane osobno i są objęte środkami technicznymi i organizacyjnymi uniemożliwiającymi ich przypisanie zidentyfikowanej lub możliwej do zidentyfikowania osobie fizycznej.

W zakresie podstawowych reguł kontroli ujawniania danych w art. 38 – analogicznie do innych demokratycznych krajów – ustawa wyklucza możliwość publikacji oraz udostępniania uzyskanych w badaniach statystycznych danych jednostkowych identyfikowalnych. W odróżnieniu jednak od tamtych uregulo-

wań (które – jak np. w Holandii – poprzestają na ogólnym zakazie udostępniania danych, z których można odczytać lub wydedukować chronione informacje indywidualne) polski akt stanowi, że nie mogą być publikowane ani udostępniane uzyskane w badaniach statystycznych dane statystyczne możliwe do powiązania i zidentyfikowania z konkretną osobą fizyczną oraz informacje i dane statystyczne charakteryzujące wyniki ekonomiczno-finansowe podmiotów gospodarki narodowej prowadzących działalność gospodarczą, jeżeli na daną agregację składają się mniej niż trzy podmioty lub udział jednego podmiotu w określonym zestawieniu jest większy niż trzy czwarte całości. Ten ostatni wymóg dość sztywno określa ramy ujawniania danych gospodarczych, podczas gdy kontrola tego ujawniania (szczególnie uwzględniająca np. powiązania międzytablicowe czy międzyagregatowe w kontekście możliwości dedukcji informacji wrażliwych) może czasami potrzebować większej elastyczności. Ważna tutaj jest również klauzula mówiąca, że w wypadku podmiotów gospodarki narodowej wrażliwe informacje i dane statystyczne mogą być publikowane, jeżeli osoba upoważniona do reprezentowania danego podmiotu wyraziła zgodę na opublikowanie określonych danych charakteryzujących wyniki ekonomiczno-finansowe tego podmiotu. W uzasadnionych przypadkach (umotywowanych przygotowaniem określonych programów, prognoz i analiz) prezes Głównego Urzędu Statystycznego może udostępnić Prezydentowi, Sejmowi i Senatowi, organom administracji rządowej, Najwyższej Izbie Kontroli, Narodowemu Bankowi Polskiemu, organom jednostek samorządu terytorialnego oraz innym instytucjom rządowym – na ich wniosek – identyfikowalne dane jednostkowe podmiotów sektora finansów publicznych. Dane te jednak mogą być wykorzystane wyłącznie w celu wskazanym we wniosku i z zachowaniem zasad tajemnicy statystycznej oraz kontroli ujawniania danych.

Podsumowując, można stwierdzić, że – zarówno w Polsce, jak i w innych krajach czy w wymiarze międzynarodowym – ochrona danych wrażliwych jest bardzo ważnym elementem funkcjonowania systemów statystycznych. Stosowne regulacje prawne w tym zakresie różnią się jedynie poziomem szczegółowości zapisów i zakresem ograniczeń oraz wymogów formalnych. Ogólnie można zauważyć, że – zwłaszcza w krajach zachodnioeuropejskich – konkretyzację zasad kontroli ujawniania danych ceduje się milcząco na odpowiedzialnych za to metodologów, w aktach prawnych ograniczając się do zapisów absolutnie koniecznych.

1.5.2. Opracowania, wytyczne i rekomendacje organizacji międzynarodowych

Wskazane w poprzednim punkcie uregulowania prawne opierają się w znacznej mierze na ogólnych kodeksach dobrych praktyk statystycznych, zawierających najważniejsze reguły etyczne, którymi powinien się kierować współczesny statystyk. Kodeksy dobrych praktyk zaś odnoszą się – w odróżnieniu od legislacji – do

kwestii etycznych, które wynikają z ogólnie funkcjonujących w społeczeństwie reguł współżycia i wzajemnego poszanowania. Stąd też zawartość takich kodeksów zasługuje na odrębne omówienie. Warto wszakże zauważyć, że najistotniejsze reguły etyczne zazwyczaj leżą u podstaw stosownych uregulowań prawnych, stąd trudno wyraźnie rozdzielić te dwa aspekty. Przypomnijmy może jeszcze, że Główny Urząd Statystyczny (GUS) jest organizacyjnym, Polskie Towarzystwo Statystyczne – afiliowanym, a Prezes GUS – *ex officio* indywidualnym członkiem Międzynarodowego Instytutu Statystycznego (ang. International Statistical Institute– ISI). Polska należy też do Organizacji Narodów Zjednoczonych (ONZ) i jej agend, Banku Światowego, OECD i Unii Europejskiej, a zatem rekomendacje tych organizacji są dla naszego kraju szczególnie istotne.

Jednym z najważniejszych „drogowskazów" w tym zakresie jest *Deklaracja etyki zawodowej* (ang. *Declaration on professional ethics*) przyjęta przez Radę Międzynarodowego Instytutu Statystycznego podczas obrad w Reykjaviku (Islandia) w dniach 22 i 23 lipca 2010 r. (ISI Declaration, 2010).

Już na samym początku w dokumencie tym wymieniono podstawowe wartości zawodowe, którymi powinni się kierować statystycy. W części dotyczącej szacunku (ang. *respect*) pierwsze stwierdzenie orzeka, że statystycy szanują prywatność innych i dane im obietnice dotyczące poufności. Konieczność ochrony danych wrażliwych została jednak szerzej wskazana w wykazie fundamentalnych reguł etycznych statystyki. Reguła 6 – Zabezpieczenie informacji niejawnych – stanowi, że niejawne informacje powinny być poufne. Nie dotyczy to wszakże metod i procedur statystycznych wykorzystywanych do prowadzenia badań lub opracowania publikowanych danych. Reguła 8 – Utrzymywanie poufności w statystyce – odnosi się do promocji znaczenia i pewności co do zachowania poufności danych w społeczeństwie. W tym celu statystycy powinni zapewnić, że precyzyjnie i prawidłowo opisują swe wyniki, łącznie z objaśniającą mocą ich danych. Na statystykach ciąży też obowiązek informowania potencjalnych użytkowników danych o występujących ograniczeniach w zakresie rzetelności i stosowalności danych. Z kolei zgodnie z regułą 12 – Ochrona interesów podmiotów – statystycy są zobligowani do ochrony podmiotów (zarówno indywidualnie, jak i zbiorowo) na tyle, na ile to możliwe, przed potencjalnie szkodliwymi efektami uczestnictwa w systemie statystycznym. Odpowiedzialność ta nie zwalnia z owego uczestnictwa – ani w drodze zgody, ani poprzez wymogi prawne. Drażliwy potencjał niektórych form badań statystycznych wymaga, aby formy te były stosowane z wielką rozwagą, pełnym uzasadnieniem stosownych potrzeb oraz potwierdzeniem ich użycia. Zapytania te powinny być oparte, w ramach dotychczasowych praktyk w tym zakresie, na dobrowolnie wyrażonej i przekazanej zgodzie podmiotu. Dane identyfikacyjne i informacje jednostkowe dotyczące respondentów powinny być utrzymywane w poufności. Odpowiednie mierniki statystyczne powinny być używane w taki sposób, aby uchronić dane przed rozpowszechnianiem

w takiej formie, która pozwalałaby wydedukować lub wywnioskować informacje identyfikujące podmiot bądź respondenta.

Z powyższych zapisów wynika zatem, że chronione powinny być dane, ale nie procedury. Oznacza to, że każdy zainteresowany ma prawo dostępu do wiedzy na temat, w jaki sposób dane statystyczne są gromadzone, jak są przetwarzane i jak chronione przed nieuprawnionym dostępem, a także, jaką efektywność i jakość mają te działania. Postulowana ochrona obejmuje także uniemożliwienie odtworzenia informacji identyfikujących konkretny podmiot czy konkretną osobę. Ujawnienie takich danych może nastąpić tylko na podstawie wyraźnej i dobrowolnej zgody zainteresowanego.

Kolejnym istotnym w tym kontekście dokumentem są *Wskazówki etyczne w praktyce statystycznej* opracowane przez Amerykańskie Towarzystwo Statystyczne (ang. American Statistical Association) (ASA, 2016). Opisane tam reguły postępowania statystyka podzielono na osiem części. Postulaty związane z ochroną informacji poufnych i kontrolą ujawniania danych znajdują się już w części B, dotyczącej integralności danych i metod. Warto przytoczyć tutaj punkt 3, w którym stwierdza się, że etyczny statystyk w publikacjach, raportach czy oświadczeniach określa, kto jest odpowiedzialny za wykonanie danej pracy statystycznej, o ile nie jest to widoczne w inny sposób. Ten zapis podkreśla przyjęcie osobistej odpowiedzialności autora danego opracowania za zawarte w nim treści i efektywność ochrony danych wrażliwych. Z kolei punkt 10 nakłada na statystyka postępującego zgodnie z zasadami etyki – w celu wspierania procesu recenzji i replikacji – powinność dzielenia się danymi wykorzystanymi w analizie, ilekroć jest to możliwe/dopuszczalne, oraz zachowania ostrożności dla ochrony danych własnościowych i poufnych, łącznie z takimi danymi, które mogłyby w niepożądanych okolicznościach zidentyfikować respondenta. Jest to bardzo ważna sprawa, albowiem udostępnianie innym wykorzystanych danych umożliwia w nauce weryfikację uzyskanych wyników przez niezależnych badaczy.

W części C amerykańskiego dokumentu sprecyzowano odpowiedzialność statystyka wobec nauki, społeczeństwa, sponsora oraz klienta. W punkcie 5 określono, że etycznie postępujący statystyk rozumie i stosuje wymogi poufności w odniesieniu do gromadzenia, publikowania i udostępniania danych, a także każde ograniczenie ich użycia nałożone przez dostarczyciela danych (w stopniu określonym prawem), oraz odpowiednio chroni użycie i udostępnianie danych. Strzeże także poufnych informacji pracodawcy, klienta lub sponsora.

Z kolei w punkcie 3 części D (Odpowiedzialność wobec podmiotów badań) wskazano, że etyczny statystyk chroni prywatność i poufność badanych podmiotów oraz dotyczących ich danych – bez względu na to, czy dane owe uzyskał bezpośrednio od tych podmiotów, od innych osób czy też z istniejących źródeł administracyjnych. Rzeczony uczciwy statystyk, przewidując stosowne potrzeby, zabiega o zgodę na wtórne i pośrednie wykorzystanie danych (w tym powiązań z innymi

zbiorami danych, gdy uzyskał pozwolenie od badanych podmiotów) oraz otrzymuje właściwe zezwolenie na dokonanie recenzji i niezależne powtórzenie analiz.

Można zatem zauważyć, że amerykańskie rozwiązania etyczne kładą istotny nacisk na imienną odpowiedzialność statystyka za pozyskane i opracowane dane oraz na potwierdzenie wartości uzyskanych wyników w „obrocie" naukowym.

Stosowne zasady etyczne w statystyce promuje również Organizacja Narodów Zjednoczonych. Formalnie w obecnym kształcie zostały one przyjęte uchwałą 68 sesji Zgromadzenia Ogólnego tejże organizacji z dnia 29 stycznia 2014 r. jako *Fundamentalne zasady statystyki publicznej* (United Nations General Assembly, 2014). Zasad owych jest 10. Określają one bardzo ramowo podstawy efektywnego i rzetelnego funkcjonowania instytucji statystyki publicznej. Zasada 6 określona w tym dokumencie stanowi, że dane indywidualne zgromadzone przez instytucje statystyczne w celu opracowania statystycznego muszą być ściśle poufne i wykorzystywane wyłącznie do celów statystycznych – bez względu na to, czy dotyczą osób fizycznych, czy też prawnych.

Opierając się na tych ogólnych zasadach (jednak odnosząc się do ich wcześniejszej wobec powyższej, ale – w odniesieniu do rozpatrywanych zagadnień – w zasadzie identycznej wersji) Komisja Gospodarcza Organizacji Narodów Zjednoczonych dla Europy opracowała reguły i wskazówki w zakresie zarządzania poufnością statystyczną i dostępem do mikrodanych (UNECE, 2007). Opracowanie jest próbą osiągnięcia większej jednolitości rozwiązań stosowanych w poszczególnych krajach dotyczących udostępniania mikrodanych. Ma to na celu efektywniejszy dostęp do nich przez środowisko naukowe dla potrzeb zasługujących na uznanie oraz – poprzez omówienie konkretnych przykładów praktycznych stosowanych na świecie i stosowne wsparcie – umożliwienie wprowadzenia odpowiednich ulepszeń w tym zakresie. W przywołanym opracowaniu UNECE (2007) można zatem znaleźć rekomendacje odnośnie do rozwiązań prawnych dotyczących ochrony tajemnicy statystycznej i zasad ujawniania danych (w tym mikrodanych), omówienie najważniejszych form udostępniania mikrodanych (w tym różnic pomiędzy plikami do użytku publicznego i ograniczonego do specjalnie uprawnionych osób czy instytucji), a także wskazanie najważniejszych aspektów zarządzania ryzykiem ujawniania mikrodanych w relacji ze środowiskiem naukowym oraz zarządzania poufnością danych i metadanymi w tym kontekście. Istotnej wartości rzeczonemu opracowaniu dodają przykłady praktycznych rozwiązań w rozpatrywanym zakresie stosowanych w Australii, Brazylii, Danii, Finlandii, Holandii, Kanadzie, Nowej Zelandii, Słowenii, Szwecji, Stanach Zjednoczonych i we Włoszech. Pragmatyka każdego kraju w omawianej dziedzinie została przeanalizowana przy użyciu schematu SWOT (ang. *strengths, weaknesses, opportunities, threats*), czyli poprzez wskazanie mocnych i słabych stron oraz szans na pomyślny rozwój tudzież zagrożeń występujących w rozpatrywanych systemach.

Ważne kwestie dotyczące kontroli ujawniania danych można znaleźć także w innym podobnym podręczniku Komisji, poświęconym aspektom poufności w integracji danych (UNECE, 2009). Dokument koncentruje się na ośmiu zasadach dotyczących:

- przeznaczenia integracji danych wyłącznie do celów statystycznych i naukowych,
- podejmowania przez krajowe instytucje statystyczne działań w zakresie integracji danych jedynie zgodnie z zakresem ich kompetencji oraz prerogatyw,
- osiągnięcia wystarczającej przewagi korzyści płynących z integracji danych nad ograniczeniami wynikającymi z ochrony prywatności i poufności informacji oraz nad ryzykiem związanym z integralnością systemu statystyki publicznej,
- niepodejmowania działań związanych z integracją danych, jeśli zobowiązania wobec respondentów wykluczyłyby taką możliwość,
- wymogu używania zintegrowanych danych wyłącznie do zaakceptowanych celów statystycznych lub badawczych oraz konieczności ponownego uzyskania stosownych uprawnień w wypadku wprowadzenia istotnych zmian w bieżących czynnościach,
- ograniczenia liczby jednostek i zmiennych zawartych w połączonym zbiorze do poziomu nieprzekraczającego zakresu wymaganego danym projektem,
- prowadzenia integracji danych w otwarty i przejrzysty sposób,
- ograniczenia dostępu do zintegrowanych danych jednostkowych zwykle wyłącznie do upoważnionych pracowników instytucji statystycznej; ewentualny dostęp do tychże mikrodanych osób z zewnątrz musi być oparty na jasnych podstawach prawnych oraz być spójny z zasadami wykorzystania danych w statystyce publicznej; każda taka osoba powinna zapewnić prawnie i logistycznie umocowane gwarancje, że dane te zostaną użyte zgodnie z zadeklarowanym zakresem oraz że żadne osoby postronne nie będą mieć do nich dostępu.

Opracowanie opisuje też przykładowe ramy procedury uzyskania dostępu do zintegrowanych mikrodanych przez odpowiedni podmiot gospodarczy.

Spośród nowszych opracowań tej organizacji wart przytoczenia jest dokument pt. *Ogólne prawo o statystyce publicznej dla Europy Wschodniej, Kaukazu i Azji Środkowej* (ang. *Generic law on official statistics for Eastern Europe, Caucasus and Central Asia* (UNECE, 2016). Znajdują się w nim zalecenia i wytyczne dotyczące dobrych praktyk w zakresie ogólnego prawa o statystyce publicznej. Z punktu widzenia niniejszego opracowania najistotniejszy jest niewątpliwie rozdział siódmy zatytułowany Poufność statystyczna. Wskazano tam, jakie dane statystyczne mają charakter poufny: są to dane jednostkowe bądź zagregowane o indywidualnych osobach fizycznych lub prawnych, w których możliwa jest bezpośrednia

bądź pośrednia identyfikacja jednostki. Należy podkreślić, że w rozdziale tym pojawia się m.in. nawiązanie do tworzenia agregatów z wykorzystaniem co najmniej trzech jednostek (zapis ten jest podobny, lecz nie tożsamy z obecnie funkcjonującym w Polsce zapisem wynikającym z ustawy o statystyce publicznej). Nawiązano w nim do kwestii udostępniania danych dotyczących władz szczebla lokalnego oraz centralnego. Przytoczono także cele, w których władze oraz organizacje międzynarodowe nie mogą wykorzystywać danych jednostkowych. Są to wszelkie dochodzenia, nadzory, postępowania sądowe, administracyjne procesy decyzyjne lub inne podobne postępowania w sprawach dotyczących osób fizycznych lub prawnych. Podkreślono także, że gestor danych wytwarzający oficjalne statystyki powinien chronić dane jednostkowe, agregaty oraz statystyki o charakterze poufnym przed ich upowszechnieniem, jak również powinien podjąć niezbędne środki regulacyjne, administracyjne, techniczne i prawne, aby zapobiec nieautoryzowanemu dostępowi do tych danych. Dodatkowo jedynie w czasie potrzebnym do realizacji celów statystycznych może on przechowywać dane jednostkowe zawierające identyfikatory i korzystać z nich. Po tym czasie oryginalne zbiory danych należy zniszczyć lub usunąć. Poruszono tam również aspekty związane z udostępnianiem poufnych danych w celach naukowych.

Kolejnym interesującym opracowaniem, które w pewnych elementach odnosi się do ochrony poufności danych, jest dokument zatytułowany *Wskazania na temat modernizacji ustawodawstwa statystycznego* (ang. *Guidance on modernizing statistical legislation*) (UNECE, 2018a). Wytyczne w nim przedstawione zostały zatwierdzone przez dyrektorów urzędów statystycznych z ponad 60 krajów na sesji plenarnej Konferencji Statystyków Europejskich w 2018 r. Dokument ten jest zgodny z podstawowymi zasadami statystyki publicznej wypracowanymi przez ONZ i zawiera wytyczne dla krajów w zakresie opracowania przepisów statystycznych niezbędnych do wsparcia modernizacji systemów statystycznych i osiągnięcia pełnej wartości statystyki publicznej. W porównaniu z opracowaniem pt. *Ogólne prawo o statystyce publicznej dla Europy Wschodniej, Kaukazu i Azji Środkowej*, zaprezentowane w tym dokumencie wskazania nie zawierają szczegółowych przepisów dotyczących poufności agregatów. Zaproponowano tu nowy zapis dotyczący ochrony poufnych danych, zgodny z europejskimi regulacjami statystycznymi, który stanowi, że instytucje statystyki publicznej chronią poufne dane w taki sposób, aby osoby trzecie, wykorzystując wszelkie dostępne środki, nie były w stanie zidentyfikować, czy to bezpośrednio, czy pośrednio, tożsamości osób fizycznych lub prawnych. Ponadto obecne systemy umożliwiają rozpowszechnianie statystyk nawet wówczas, gdy tożsamość osób fizycznych lub prawnych daje się zidentyfikować tylko pod warunkiem, że osoba fizyczna lub prawna jednoznacznie wyraziła zgodę na ujawnienie danych. Podkreślono również, że usługi dostępu naukowców do danych jednostkowych nie są dobrze rozwinięte. W związku z tym problemem w wytycznych zapisano, że zapewnienie dostępu

do danych jednostkowych dla naukowców jest ważne z punktu widzenia możliwości zdobycia nowych informacji o zmianach zachodzących w społeczeństwie i lepszego zrozumienia rozwoju gospodarczego, społecznego i środowiskowego. Ze względu na obowiązek zachowania poufności taki dostęp do mikrodanych dla celów badawczych powinien być ściśle regulowany. Nakreślono również ogólne warunki i zasady dostępu do tego typu danych, które powinny zostać uwzględnione w uregulowaniach prawnych przez poszczególne krajowe urzędy statystyczne. W dokumencie zamieszczono ponadto wiele innych zaleceń dotyczących prywatności i poufności danych.

Ostatnim wartym omówienia dokumentem autorstwa UNECE odnoszącym się w pewnych elementach do zagadnienia ochrony poufności danych jest opracowanie z 2018 r. zatytułowane *Wytyczne na temat wykorzystania rejestrów i danych administracyjnych w spisach ludności* (ang. *Guidelines on the use of registers and administrative data for population and housing censuses*) (UNECE, 2018b), które dotyczy wykorzystania danych administracyjnych oraz pochodzących z rejestrów w spisach powszechnych. W opisie podstaw prawnych można znaleźć zapis związany z otrzymywanymi danymi administracyjnymi, zgodnie z którym krajowe urzędy statystyczne są prawnie zobowiązane do ochrony poufności takich danych. Przewiduje się również możliwość zaostrzenia odpowiednich przepisów poprzez wprowadzenie ogólnego zakazu udostępniania innym gestorom danych przechowywanych w rejestrach statystycznych krajowych urzędów statystycznych. Jeżeli chodzi o warte uwagi informacje w odniesieniu do jakości wyników analiz, w podrozdziale, który został poświęcony jakości i poufności danych, można znaleźć stwierdzenie, że zapewnienie jakości i ochrona poufności danych osobowych są w pewnym sensie celami sprzecznymi, choć stanowią dwa istotne wymiary wyników spisu powszechnego. Jak podkreślono, należy mieć świadomość, że rzeczą trudną (a często wręcz niemożliwą) jest publikowanie dokładnych danych (szczególnie w wypadku małych obszarów bądź dziedzin), przy jednoczesnym przyjęciu odpowiedniego poziomu kontroli ujawniania informacji w celu ochrony poufności. Warto zauważyć, że przy ustalaniu tych poziomów należy wziąć pod uwagę inne publikacje, szczególnie te oparte na tych samych danych administracyjnych co spis powszechny. W kontekście raportów jakości przygotowywanych po każdej rundzie spisów powszechnych podkreślono konieczność sprawdzenia, czy zastosowanie zabiegów mających na celu ochronę poufności danych doprowadziło do powstania różnic w danych, jak również w ich skali.

Warte podkreślenia jest również to, że kontrola ujawniania danych statystycznych jest jednym z elementów, które były brane pod uwagę przez UNECE w kontekście organizowanych spotkań grup eksperckich. Z dwuletnią częstotliwością Europejska Komisja Gospodarcza ONZ organizuje spotkania przedstawicieli państw członkowskich poświęcone temu zagadnieniu, podczas których przedstawiane są podejścia, narzędzia czy nowe rozwiązania dla procesu ochrony różnej

postaci danych statystycznych. Materiały ze spotkań (czyli przedstawiane i omawiane prezentacje lub pliki tekstowe – najczęściej w wersji anglojęzycznej) można znaleźć na stronie internetowej UNECE (2023). Do tego nurtu należała między innymi międzynarodowa konferencja 2021 Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality, która odbyła się w dniach 1–3 grudnia 2021 r. w Poznaniu. W jej organizację włączyli się m.in.: Katedra Statystyki Uniwersytetu Ekonomicznego w Poznaniu, Główny Urząd Statystyczny oraz Urząd Statystyczny w Poznaniu. Opracowania wygłoszonych wówczas wystąpień są dostępne na stronie UNECE (2021)[9]. Na bazie stosownego referatu z tej konferencji powstał też artykuł Młodaka i in. (2022).

Wkład w promowanie rozwiązań z obszaru SDC wniósł również Bank Światowy, pod którego egidą opracowano podręczniki pt. *Statistical disclosure control for microdata: A theory guid*e (autorstwa dwóch pracowników Banku Światowego – Thijsa Benschopa oraz Matthew Welcha) (Benschop i Welch, 2019) oraz *Statistical Disclosure Control: A practice guide* (tych samych autorów, do których dołączyła inna ekspertka Banku – Cathrine Machingauta) (Benschop i in., 2022). Są one warte przywołania ze względu na to, że w pierwszym z nich przedstawiono – oprócz fundamentalnych zagadnień z zakresu kontroli ujawniania mikrodanych (w tym najpopularniejszych metod ochrony tajemnicy statystycznej czy charakterystyki tego procesu „krok po kroku") – także liczną egzemplifikację zagadnień teoretycznych, której źródłem są krajowe urzędy statystyczne. Z kolei drugi z wyżej wskazanych podręczników zawiera wskazówki dotyczące korzystania z programu `sdcMicro`, który ma zaimplementowane narzędzia odnoszące się do kontroli poufności w zbiorach mikrodanych.

Ponadto spośród dokumentów autorstwa Banku Światowego odnoszących się do szeroko rozumianej kontroli ujawniania danych statystycznych warto przywołać dokument z 2018 r. zatytułowany *Odpowiedzialne zarządzanie danymi osobowymi: polityka danych osobowych Grupy Banku Światowego* (ang. *Managing personal data responsibly: The World Bank Group personal data policy*) (The World Bank, 2018). Przedstawiono w nim propozycje Zarządu dla zatwierdzenia pierwszej Grupy Banku Światowego zajmującej się polityką prywatności danych osobowych. Ze względu na powszechne naruszenia prywatności, które niosą poważne konsekwencje zarówno dla osób fizycznych oraz prawnych, jak również dla organizacji i ogólnie dla całego społeczeństwa, ochrona prywatności stała się koniecznością. Z tego względu Bank Światowy utworzył specjalną grupę zajmującą się tą kwestią. W przywołanym dokumencie można znaleźć informacje na temat przyczyn konieczności ochrony danych osobowych, wzmianki o badaniach, które zostały wykonane w tym kontekście (np. na temat wpływu rozporządzenia

[9] Obszerne sprawozdanie z tej konferencji opublikowano również w czasopiśmie *Wiadomości Statystyczne. The Polish Statistician* (Dehnel i in., 2022).

o ochronie danych osobowych – RODO), zasady prywatności, które mają być dostosowane do globalnych standardów ochrony danych osobowych, czy też szczegóły techniczne (m.in. koszty oraz zasoby kadrowe Grupy Banku Światowego do spraw ochrony prywatności).

Organizacja ta ma również własną politykę dostępu do informacji. Materiały z podstawowymi informacjami o polityce dostępu do informacji, zapytaniach, odwołaniach, aktualnościach i notatkach, jak również raporty roczne oraz raporty z badań nad dostępem do informacji są zamieszczone na stronie internetowej Banku Światowego (The World Bank, b.d.a). Przygotowana jest również strona internetowa dotycząca udzielania dostępu do pozostających w dyspozycji tej instytucji innych zbiorów danych (The World Bank, b.d.b).

Kolejną instytucją podnoszącą w swoich opracowaniach zagadnienia z zakresu ochrony poufności danych jest OECD. Przykładem jest rekomendacja pt. *Zalecenie Rady dotyczące dostępu do danych badawczych z funduszy publicznych* (ang. *Recommendation of the Council concerning access to research data from public funding*) (OECD, 2021). Zalecenie to, na wniosek złożony przez Komitet do spraw Polityki Naukowej i Technologicznej, zostało przyjęte przez Radę OECD pod koniec 2006 r. Omawiany dokument ma trzy zasadnicze cele. Pierwszym z nich jest pomoc rządom, organizacjom wspierającym i finansującym badania, instytucjom badawczym, a także badaczom w radzeniu sobie z barierami i wyzwaniami w zakresie poprawy międzynarodowego udostępniania i udzielania dostępu do danych badawczych z publicznych funduszy. Ma on także promować udzielanie dostępu do danych wśród badaczy, instytucji badawczych i krajowych agencji badawczych, a jednocześnie uwzględniać krajowe regulacje prawne, polityki badawcze, jak też krajowe struktury organizacyjne. Nakierowany jest również na poprawę wydajności i skuteczności globalnego systemu nauki. Należy dodać, że nie wiąże się to ani z uciążliwymi obowiązkami oraz regulacjami, ani z nowymi kosztami.

Kolejnym wartym wspomnienia opracowaniem autorstwa OECD jest dokument-rekomendacja pt. *Zalecenie Rady w sprawie lepszego dostępu i bardziej efektywnego wykorzystania informacji sektora publicznego* (ang. *Recommendation of the Council for enhanced access and more effective use of public sector information*) (OECD, 2008). Zalecenie dotyczące lepszego dostępu i bardziej efektywnego użycia informacji pochodzących z sektora prywatnego zostało przyjęte, na wniosek Komitetu do spraw Polityki Informacyjnej, Komputerowej i Komunikacyjnej (noszącego obecnie nazwę Komitetu ds. Polityki Gospodarki Cyfrowej), przez Radę OECD w 2008 r. Kraje członkowskie ustanowiły liczne przepisy prawne, strategie polityczne i praktyki związane z dostępem i ponownym wykorzystaniem informacji pochodzących z sektora publicznego – zarówno na poziomie krajowym, jak i międzynarodowym. Celem *Zalecenia* był wpływ na globalną wymianę i użycie informacji pozostających do dyspozycji tego sektora, a także dostarczenie szerszych i bardziej kompatybilnych ram dla rozwoju i wdrażania

podejść oraz wytycznych mających ułatwić dostęp i ponowne wykorzystanie takich informacji.

Następnym dokumentem OECD, w którym przedstawiono rekomendację przyjętą pod koniec 2015 r. przez Radę OECD, jest *Zalecenie Rady OECD w sprawie dobrej praktyki statystycznej* (ang. *Recommendation of the OECD Council on good statistical practice*) (OECD, 2015). W prace nad nim zaangażowany był Komitet OECD do spraw Statystyki i Polityki Statystycznej. Należy podkreślić, że jest to pierwsza rekomendacja związana w sposób bezpośredni ze statystyką. Odzwierciedla ona to, że jakość statystyk ma fundamentalne znaczenie dla jakości prac analitycznych prowadzonych przez OECD, a także dla jakości statystycznych publikacji i baz danych tworzonych przez tę organizację. Dokument ma również na celu uzupełnienie istniejących już i przyjętych międzynarodowych wytycznych, w tym: fundamentalnych *Zasad dla oficjalnych statystyk ONZ* oraz *Europejskiego kodeksu praktyk statystycznych*, a także *Kodeksu dobrych praktyk statystycznych dla Ameryki Łacińskiej oraz Karaibów*. Dokument obejmuje 12 rekomendacji. Czwarta rekomendacja bezpośrednio jest związana z procesem kontroli ujawniania danych statystycznych. W jej myśl zapewnione muszą zostać: ochrona poufności wszelkich dostawców danych, a także prawne zagwarantowanie poufności informacji o indywidualnych jednostkach i ich wykorzystanie wyłącznie w celach statystycznych. Z kolei dziewiąta rekomendacja pośrednio jest związana z zagadnieniem ochrony tajemnicy statystycznej. Rekomenduje się w niej bowiem zapewnienie przystępnego dla użytkowników rozwiązania w zakresie dostępu do danych oraz ich rozpowszechniania, aby statystyki były prezentowane w przejrzystej i zrozumiałej postaci, publikowane w odpowiedni i wygodny sposób – w formie otwartych danych, były łatwe do odnalezienia, a także dostępne wraz z metadanymi je opisującymi i odpowiednimi wytycznymi. Oczekuje się również reagowania na poważniejsze błędy popełniane przez użytkowników takich zasobów. Także inne rekomendacje Rady OECD ujęte w tym dokumencie można powiązać w mniejszym lub większym stopniu z procesem ochrony tajemnicy statystycznej.

Ostatni scharakteryzowany tutaj dokument OECD to *Poprawa dostępu do danych i ich udostępniania* (ang. *Enhancing access to and sharing of data*), w skrócie EASD (OECD, 2019). Jego autorami są Christian Reimsbach-Kounatze i Elettra Ronchi, przy współudziale Suguru Iwaya. Jest to raport z 2019 r., którego celem było przeanalizowanie możliwości poprawy dostępu do danych oraz ich udostępniania w kontekście rosnącego znaczenia sztucznej inteligencji i Internetu Rzeczy. W raporcie omówiono, w jaki sposób EASD może zmaksymalizować społeczną oraz ekonomiczną wartość ponownego wykorzystywania danych, a także jak można zminimalizować związane z tym ryzyko i sprostać innym wyzwaniom. W szczególności przeanalizowano w nim podejścia mające na celu ustanowienie ram zarządzania danymi w zakresie dostępu do nich oraz ich udostępniania – co

jest bardzo ważnym aspektem z punktu widzenia kontroli ujawniania danych. W tym względzie szczególną uwagę zwrócono na takie kwestie jak: typologia danych, kluczowe mechanizmy dostępu do danych czy główne typy podmiotów i ich role. W raporcie przedstawiono ponadto dostępne dowody bezpośrednich i pośrednich korzyści ekonomicznych oraz społecznych płynących z dostępu do danych tudzież ich udostępniania. Oprócz tego w raporcie zwrócono uwagę na główne wyzwania stojące przed decydentami oraz przedstawiono aktualne trendy w publicznych strategiach politycznych mających na celu sprostanie tym wyzwaniom, wraz z przykładami podejść EASD oraz inicjatyw politycznych w krajach OECD i gospodarkach partnerskich. Raport ten stanowi część projektu dotyczącego EASD, który jest tworzony w ramach trzech komitetów partnerskich: Komitetu ds. Polityki Gospodarki Cyfrowej (CDEP), Komitetu Polityki Naukowej i Technologicznej oraz Komitetu Zarządzania Publicznego, stanowiąc przy tym uzupełnienie innych raportów opracowanych w ramach tych komitetów. Z punktu widzenia kontroli ujawniania danych statystycznych najbardziej istotnymi rozdziałami w tym raporcie są: rozdział drugi, w którym opisano dostęp do danych oraz ich udostępnianie, a także rozdział czwarty, gdzie informuje się o ryzyku, a także o wyzwaniach związanych z dostępem do danych i ich udostępnianiem.

Dla instytucji statystycznych w krajach Unii Europejskiej szczególną wartość posiada *Europejski kodeks praktyk statystycznych* (Eurostat, 2011). Kodeks obejmuje 15 zasad. Poufności danych poświęcona jest zasada 5 głosząca, że należy bezwzględnie zapewnić poufność dostawców danych (gospodarstw domowych, przedsiębiorstw, organów administracji i innych respondentów), poufność informacji przez nich przekazywanych oraz wykorzystywanie tych informacji wyłącznie do celów statystycznych. W tym celu przepisy prawa powinny ową poufność gwarantować, a pracownicy statystyki, podejmując w niej zatrudnienie, podpisać stosowne prawne zobowiązanie do zachowania poufności. Wszelkie umyślne naruszenia poufności danych statystycznych muszą podlegać karze. Kodeks zapewnia pracownikom prawo do otrzymania instrukcji i wytycznych dotyczących ochrony poufności danych statystycznych w procesie ich tworzenia i rozpowszechniania. Informacje o polityce zapewnienia poufności powinny być podawane do wiadomości publicznej. Kodeks zakłada także, że obowiązują przepisy regulujące kwestie fizycznego i technicznego bezpieczeństwa oraz integralności statystycznych baz danych tudzież protokoły, które regulują dostęp do mikrodanych statystycznych dla użytkowników zewnętrznych wykorzystujących je do celów związanych z badaniami naukowymi.

Do tych postulatów nawiązują również ustalenia zawarte w niektórych innych regułach kodeksu. W zasadzie 14 (spójność i porównywalność) zapisano, że możliwe jest łączenie i zbiorcze wykorzystanie powiązanych danych pochodzących z różnych źródeł. To powoduje konieczność zapewnienia odpowiednich mechanizmów kontroli ujawniania danych, które nie zawsze bywają proste. Z kolei za-

sada 15 (dostępność i przejrzystość) orzeka między innymi, że można zezwolić na dostęp do mikrodanych dla celów badawczych (dostęp ten powinien jednak być uregulowany ścisłymi zasadami lub protokołami – punkt 15.4) oraz że należy poinformować użytkowników danych o metodologii stosowanej w procesach statystycznych, w tym również o wykorzystaniu danych administracyjnych (punkt 15.6). Ten ostatni zapis obejmuje ponadto zaopatrzenie publikacji danych w opis zastosowanych metod kontroli ich ujawniania, a także wpływu, jaki owa kontrola miała na jakość informacji wynikowych (co wiąże się też z kolejnym punktem 15.7, w którym zapisano prawo użytkowników danych do uzyskania informacji o jakości wyników badań statystycznych w odniesieniu do kryteriów jakości statystyk europejskich).

Jak można zauważyć, wszystkie opracowania, wytyczne i rekomendacje opierają się na tym samym kanonie zasad ochrony poufności danych (nieujawnianie danych indywidualnych, zminimalizowanie ryzyka ich odtworzenia z mikrodanych zanonimizowanych lub zbiorczych oraz wykorzystywanie danych jednostkowych wyłącznie do celów statystycznych lub naukowych). Występuje jednak między nimi pewne zróżnicowanie rozłożenia określonych akcentów w zakresie poszczególnych kwestii, co wynika głównie ze specyfiki osób lub instytucji, do których są adresowane.

2

Pomiar i ocena ryzyka ujawnienia informacji poufnych

2.1. Mikrodane

W podrozdziale 1.1 wyjaśniono już terminy *ujawnienie* (tożsamości, atrybutu oraz dedukcyjne) oraz *ryzyko ujawnienia*. Z kolei w podrozdziale 1.2 opisano, czym są mikrodane, jakie mogą być ich źródła, sklasyfikowano zmienne w nich występujące oraz wskazano inne zasoby, które z nimi są udostępniane. Wszystkie te pojęcia i koncepcje okazują się istotne w kontekście zagadnienia ryzyka ujawnienia informacji poufnych.

Dla przypomnienia, ryzyko ujawnienia to prawdopodobieństwo rozpoznania jednostki statystycznej (czyli poznania jej tożsamości) lub uzyskania nowych, wcześniej nieznanych informacji na jej temat (a zatem poznania jej atrybutów) na podstawie danych udostępnionych przez krajowy urząd statystyczny lub innego gestora danych wynikowych.

Ryzyko obydwu wspomnianych w definicji typów ujawnienia jest w procesie kontroli ujawniania mikrodanych redukowane. Sposób tej redukcji zależy od tego, na jakich zmiennych są stosowane metody ochrony poufności. Jeżeli metody maskowania niezakłóceniowego lub zakłóceniowego (por. rozdz. 3) zostaną użyte na quasi-identyfikatorach, to minimalizowane jest ryzyko ujawnienia tożsamości jednostki statystycznej, a jednocześnie – pośrednio – również jej atrybutu. Ujawnienie tożsamości zawsze prowadzi bowiem do ujawnienia atrybutu. Jeżeli rekord w zbiorze mikrodanych zostanie przyporządkowany do jednostki statystycznej, to osoba, która naruszyła poufność, pozna wszystkie objęte badaniem statystycznym charakterystyki dla tejże jednostki (w drugą stronę taka zależność nie zachodzi: ujawnienie atrybutu nie prowadzi do ujawnienia tożsamości). Jeżeli jednak stosowne metody zastosuje się do zmiennych wrażliwych, to redukowane jest ryzyko ujawnienia atrybutu. Do ujawnienia wartości cechy może dojść nawet wtedy, gdy osoba naruszająca poufność danych wynikowych nie przyporządkuje jednostki statystycznej do konkretnego rekordu w mikrodanych. Wystarczy bowiem, by grupa jednostek statystycznych o wspólnych wartościach przyjmowanych przez zmienne kluczowe współdzieliła również tę samą wartość innej wrażliwej zmiennej.

Przed przystąpieniem do pomiaru lub oceny ryzyka ujawnienia informacji poufnych w zbiorze danych jednostkowych konieczne jest wytypowanie zestawu zmiennych, które mogą służyć użytkownikowi zewnętrznemu do naruszenia poufności. Na etapie planowania procesu kontroli ujawniania mikrodanych przeprowadza się więc wstępną klasyfikację zmiennych pod względem pełnionych przez nie funkcji w tymże procesie oraz ich nośności informacji wrażliwych (por. podrozdz. 1.2). Następnie, w kolejnym kroku tego samego etapu, formułuje się jeden lub kilka tzw. **scenariuszy ujawnienia**. Ich celem jest wyselekcjonowanie zmiennych w zbiorze mikrodanych, który ma zostać udostępniony (w różnej postaci lub formie – por. podrozdz. 6.2 i 6.3) użytkownikowi zewnętrznemu do realizacji celów naukowo-badawczych, ostatecznego zestawu quasi-identyfikatorów oraz zmiennych wrażliwych. W scenariuszu takim opisuje się, do jakich innych zbiorów danych jednostkowych – czy to z ogólnodostępnych źródeł, czy to z takich o ograniczonej dostępności – ten użytkownik może mieć dostęp, a także, w jaki sposób mógłby z nich skorzystać do naruszenia poufności. Zakłada się, że te zewnętrzne bazy danych będą stanowić wykaz jednostek statystycznych wraz z ich identyfikatorami (bezpośrednio określającymi ich tożsamość) oraz zestawem quasi-identyfikatorów (w skład którego wejdą zmienne, które także są objęte badaniem statystycznym – i to na ich podstawie dojdzie do pośredniej identyfikacji jednostki statystycznej). Można sformułować tylko jeden taki scenariusz ujawnienia, ale równie dobrze może być ich więcej. Jeżeli jest ich wiele, zaleca się ograniczenie ich liczby do kilku, tych najbardziej prawdopodobnych do zaistnienia.

Założenia tzw. **najgorszego możliwego scenariusza**, które często przyjmuje się w procesie kontroli ujawniania mikrodanych, szczegółowo omówili m.in. Hundepool i in. (2012). Jak podkreślają ci autorzy, rodzaj identyfikacji jednostki statystycznej w udostępnionych zasobach jest zależny od tego, komu takiego dostępu udzielono oraz od formy tego dostępu. Jeżeli mikrodane kierowane są do osób ze środowiska naukowego, a dostęp do tych zasobów jest ograniczony, to wysoce prawdopodobne wydaje się wyłącznie niezamierzone ujawnienie, w przypadku zaś gdy zasoby te mają być powszechnie dostępne, bardziej prawdopodobne staje się raczej celowe złamanie poufności (zwłaszcza gdy zestaw rozpatrywanych quasi-identyfikatorów jest liczniejszy).

W związku z powyższym w literaturze przedmiotu wyróżnia się dwa rodzaje identyfikacji jednostek statystycznych. Pierwszym jest **identyfikacja spontaniczna**. Mówimy o niej wówczas, gdy użytkownik zewnętrzny ma szczegółową, osobistą wiedzę o jednej lub kilku jednostkach objętych badaniem i – wypadkowo lub celowo – rozpoznaje je w dostępnych dlań zasobach. Drugim rodzajem jest **identyfikacja poprzez dopasowanie/łączenie rekordów**. W takim wypadku osoba z zewnątrz ma dostęp do wykazu jednostek statystycznych (zawierającego identyfikatory) z innego źródła i próbuje dopasować te informacje, używając quasi-identyfikatorów – tych, które są dostępne zarówno w jej bazie danych,

jak i w udostępnionych przez gestora danych zasobach – dla identyfikacji przebadanych w badaniu statystycznym jednostek.

Wypracowane zostały różne koncepcje w zakresie zdefiniowania ryzyka ujawnienia informacji poufnych, a także miary i reguły służące jego ocenie. Opierają się one na rodzaju quasi-identyfikatorów, którymi potencjalnie ma się posłużyć osoba dążąca do naruszenia poufności przy pośredniej identyfikacji jednostek statystycznych, o których informacje (które podlegać będą udostępnieniu) zebrano np. w badaniu statystycznym. Koncepcje te, zgodnie z Benschopem i in. (2022), to:

- **koncepcja unikatowości kombinacji wartości quasi-identyfikatorów** (kluczy) do identyfikacji jednostek zagrożonych – dla zmiennych jakościowych; są to zazwyczaj miary *a priori*, których wartości wyznacza się jeszcze przed zastosowaniem wybranych metod maskowania (niezakłóceniowego lub zakłóceniowego) na etapie wdrożenia procesu kontroli ujawniania mikrodanych,
- **koncepcja unikatowości wartości w sąsiedztwie oryginalnych wartości** – dla zmiennych ilościowych; są to miary *a posteriori*, a więc takie, których wartości oblicza się poprzez porównanie mikrodanych przed i po zastosowaniu metod kontroli ujawniania mikrodanych.

Pierwsza wymieniona koncepcja najczęściej znajduje zastosowanie w wypadku mikrodanych z badań społecznych, druga z kolei – w wypadku tych z badań przedsiębiorstw. Jest to związane z typem charakterystyk jednostek statystycznych zbieranych w takich badaniach. W pierwszych przeważają zmienne jakościowe, w drugich zaś – ilościowe. W związku z tym wytypowane quasi-identyfikatory będą najczęściej takiego właśnie odpowiednio typu. Nie wyklucza się jednak możliwości, by w zbiorze danych jednostkowych znajdowały się zarówno jakościowe, jak i ilościowe zmienne kluczowe.

Konsekwencją przyjęcia powyższych koncepcji jest to, że gdy w zbiorze danych jednostkowych zostały wyróżnione wyłącznie quasi-identyfikatory typu jakościowego, wówczas możliwe jest wyznaczenie wartości miar ryzyka ujawnienia informacji poufnych dla oryginalnych mikrodanych. Na ich podstawie można m.in. podjąć decyzję, czy konieczna jest ochrona poufności (gdy ryzyko ujawnienia okaże się zbyt wysokie, a wybrane przez gestora miary osiągają wartości zdecydowanie wyższe od przyjętych dla każdej z nich progów wyznaczających akceptowalny poziom ryzyka ujawnienia), jak również traktować je jako punkt odniesienia przy porównywaniu różnych rozwiązań w zakresie SDC[10].

W wypadku wytypowania wyłącznie quasi-identyfikatorów typu ilościowego, nie da się wyznaczyć wartości miar ryzyka ujawnienia informacji poufnych w za-

[10] Pisząc o różnych rozwiązaniach SDC, autorzy mają na myśli wybraną metodę lub kombinację kilku metod kontroli ujawniania mikrodanych, które różnią się zarówno doborem metod, jak również ich parametryzacją.

sobach oryginalnej postaci przed przystąpieniem do stosowania metod ochrony poufności. Ich wartości uzyskane *ex post* mogą natomiast posłużyć do wyboru oraz parametryzacji jednej lub kilku metod dla zmiennych ilościowych (stosuje się je w celach porównania różnych rozwiązań SDC).

Podejście do oceny ryzyka ujawnienia informacji poufnych w udostępnianych mikrodanych komplikuje się, gdy występują w nich quasi-identyfikatory zarówno typu jakościowego, jak i ilościowego. W takiej sytuacji można przeprowadzić ocenę tego ryzyka z użyciem właściwych miar oddzielnie na podstawie zmiennych jakościowych i oddzielnie na podstawie zmiennych ilościowych. Niestety, wiele podejść do oceny ryzyka ujawnienia opisanych w literaturze międzynarodowej, które są dostępne w narzędziach informatycznych przeznaczonych do kontroli ujawniania mikrodanych i stosowane w praktyce, nie umożliwia łącznego pomiaru ani oszacowania ryzyka w charakteryzowanym tu przypadku. W książce Hundepoola i in. (2012) zasugerowano, by w takiej sytuacji dokonać podziału zbiorowości według wariantów quasi-identyfikatorów typu jakościowego, a następnie te podzbiorowości potraktować tak, jakby były to zbiorowości z samymi quasi-identyfikatorami typu ilościowego. Zdaniem autorów niniejszego opracowania przeprowadzanie odrębnych procesów kontroli ujawniania mikrodanych na każdym zbiorze z osobna nie jest do końca trafnym podejściem, zważywszy na to, że dane jednostkowe i tak zostaną udostępnione użytkownikom zewnętrznym w całości. Zatem taka ich łączna struktura powinna być wzięta pod uwagę przy ochronie poufności. Ponadto każdy podzbiór oryginalnych mikrodanych może się charakteryzować innym poziomem ryzyka ujawnienia, co może nie w pełni oddawać jego poziom dla zbioru danych jednostkowych jako całości.

Przy ocenie poufności z wykorzystaniem jakościowych quasi-identyfikatorów, dla niektórych reguł bądź metod ma znaczenie to, w jakim badaniu – reprezentacyjnym czy pełnym – pozyskano dane jednostkowe. Wagi z losowania są bowiem w odpowiedni sposób wykorzystywane przy wyznaczaniu wartości przyjętych miar ryzyka ujawnienia informacji poufnych. Można jednak przyjąć w uproszczeniu, że w zbiorze danych jednostkowych z badania pełnego waga uogólniająca wynosi jeden dla każdej obserwacji. Podobne założenie przyjmuje się, jeśli w badaniu mieszanym pewna subpopulacja wyszczególniona w ramach badanej populacji generalnej w całości podlega temu badaniu, a z pozostałej części owej populacji generalnej losowana jest próba losowa. W takim wypadku obserwacjom z subpopulacji przypisuje się mnożnik uogólniający równy jeden.

W literaturze obcojęzycznej z zakresu kontroli ujawniania mikrodanych przywołuje się przede wszystkim miary oceny ryzyka ujawnienia informacji poufnych bazujące na częstościach kluczy (czyli kombinacji wartości quasi-identyfikatorów) w próbie lub w populacji – dla zmiennych typu jakościowego – bądź oparte na porównaniu/łączeniu oryginalnej wartości z tą powstałą po zakłóceniu wskutek ochrony poufności – dla zmiennych typu ilościowego.

Ze względu na możliwości użytkownika w zakresie posiadania alternatywnych źródeł danych z dziedziny, której dotyczą udostępniane informacje, można wyróżnić także (Młodak i in., 2022):

- **ryzyko wewnętrzne** – gdy potencjalna identyfikacja jednostki może nastąpić tylko na podstawie udostępnionych mu danych,
- **ryzyko zewnętrzne** – gdy użytkownik ma dostęp także do alternatywnych źródeł danych, które może skutecznie powiązać z udostępnionym mu zasobem; ocena tego ryzyka jest znacznie utrudniona ze względu na występujący zazwyczaj brak lub niedostatek informacji o dostępie użytkownika do określonych baz danych; pewną wskazówką w tym względzie może być jego miejsce pracy (na przykład jeśli użytkownik pracuje w urzędzie pracy, to jest prawdopodobne, że ma dostęp do bazy danych o bezrobotnych, którą można powiązać z mikrodanymi z Badania Aktywności Ekonomicznej Ludności).

Ocenę ryzyka ujawnienia można przeprowadzić dla każdego z rekordów z osobna i wykorzystać ją do wybiórczej ochrony danych (dla obserwacji uznanych za zagrożone), ale można też na podstawie cech owych rekordów przedstawić ogólną definicję ryzyka ujawnienia informacji poufnych dla udostępnianego zbioru danych jednostkowych jako całości. W związku z tym w kontroli ujawniania mikrodanych wyróżnia się następujące poziomy oceny ryzyka ujawnienia informacji poufnych:

- **indywidualny** – dla pojedynczego rekordu w mikrodanych,
- **globalny** – dla zbioru danych jednostkowych jako całości,
- **związany z hierarchiczną strukturą danych** – dla pojedynczego rekordu lub dla całego zbioru oceniany jest wpływ występowania hierarchicznej struktury jednostek statystycznych w mikrodanych na ryzyko ujawnienia informacji poufnych na odpowiednim poziomie rzeczonej hierarchii.

W wypadku ryzyka ujawnienia mierzonego lub szacowanego na poziomie indywidualnym można powiedzieć, że im rzadsza jest kombinacja wartości przyjmowanych przez jakościowe quasi-identyfikatory w próbie, tym większe jest ryzyko ujawnienia tożsamości dla określonej jednostki statystycznej, której rekord ten odpowiada. Prawdopodobieństwo tego, że osoba dążąca do naruszenia poufności poprawnie dopasuje taki rekord do zewnętrznej bazy danych (zawierającej identyfikatory oraz quasi-identyfikatory dla poszczególnych jednostek statystycznych i pozostającej do jego dyspozycji) lub go z nią połączy, jest bowiem wówczas większe niż prawdopodobieństwo identyfikacji w sytuacji, gdyby wielu respondentów współdzieliło wartości zmiennych kluczowych. W literaturze traktującej o kontroli ujawniania mikrodanych mówi się, że rekord o unikatowym kluczu, występującym tylko jeden raz w całej próbie, jest rekordem *sample unique*.

W ustalaniu ryzyka ujawnienia informacji poufnych można uwzględnić również częstości wystąpień poszczególnych kluczy w populacji. W wypadku badań

reprezentacyjnych częstości te nie są znane i zazwyczaj szacuje się je z wykorzystaniem wag z losowania na podstawie próby losowej. Wyjątek stanowią badania pełne, w których dla uproszczenia przyjmuje się, że częstości kluczy w próbie i w populacji generalnej są takie same.

Ze względu na to, że w próbie losowej każda jednostka odznaczająca się określonym kluczem reprezentuje taką samą liczbę jednostek o tym kluczu w populacji (czyli ma taką samą częstość klucza w populacji), ryzyko indywidualne oznaczające prawdopodobieństwo identyfikacji można obliczyć jako odwrotność tej częstości. Do szacowania ryzyka indywidualnego wykorzystuje się m.in. model Benedettiego-Franconi (Templ, 2017) – jest on dostępny we wszystkich narzędziach informatycznych przeznaczonych do ochrony poufności w zbiorach danych jednostkowych.

Ryzyko ujawnienia na poziomie rekordu będzie przyjmowało taką samą wartość dla każdego rekordu współdzielącego określony klucz, a to ze względu na takie same częstości tychże kluczy w próbie i w populacji. Można przyjąć, że wyraża ono prawdopodobieństwo ujawnienia jednostki statystycznej lub – inaczej – prawdopodobieństwo poprawnego dopasowania do rekordu odpowiadającego tej jednostce losowo wybranego rekordu z zewnętrznej bazy danych po wartościach jakościowych quasi-identyfikatorów.

W szczególnym przypadku o rekordzie powiemy, że jest *population unique*, jeżeli częstość jego klucza w populacji wynosi jeden. Każdy rekord o takiej etykiecie jest jednocześnie rekordem *sample unique*. Musi bowiem odznaczać się unikatowym z punktu widzenia próby kluczem. Ponadto odpowiadająca mu waga uogólniająca musi być równa jeden, by częstość jego klucza w populacji generalnej również wynosiła jeden.

Przejdźmy do ryzyka ujawnienia informacji poufnych mierzonego na poziomie całego zbioru danych jednostkowych (na tzw. poziomie globalnym). Można je wyznaczyć poprzez odpowiednią agregację ryzyka indywidualnego wyznaczonego dla każdego rekordu w rozpatrywanych mikrodanych. Przykłady miar przytoczyli Benschop i in. (2022). Są nimi:

- **średnia miara ryzyka indywidualnego** – wartość tę wystarczy przemnożyć przez wielkość próby losowej, by się dowiedzieć, jaką szacunkowo część wszystkich rekordów uda się zidentyfikować osobie dążącej do naruszenia poufności,
- **liczba jednostek z ryzykiem indywidualnym większym niż ustalony (w sposób bezwzględny lub względny) próg** – przydatność tej miary polega na tym, że nawet gdy ryzyko globalne jest niewielkie, w mikrodanych nadal mogą występować jednostki statystyczne o wysokim ryzyku ujawnienia mierzonym na poziomie indywidualnym.

Pomiar ryzyka ujawnienia komplikuje się, gdy w mikrodanych występuje hierarchiczny układ jednostek statystycznych. Dzieje się tak wówczas, gdy podrzęd-

ne jednostki statystyczne (np. osoby) są grupowane w nadrzędne jednostki statystyczne wyższego rzędu (np. w gospodarstwa domowe). Jeżeli osoba, której udało się naruszyć poufność, dokonała poprawnej identyfikacji choćby jednej osoby należącej do określonego gospodarstwa domowego, to skutkiem wiedzy o tej przynależności może być identyfikacja pozostałych osób wchodzących w skład tegoż gospodarstwa. Ryzyko hierarchiczne wyznacza się zarówno na poziomie indywidualnym, jak i globalnym. Nigdy nie przyjmuje ono niższych wartości niż ryzyko ujawnienia wyznaczone na tych samych poziomach bez uwzględnienia takiej struktury danych. Co więcej, wszystkie podrzędne jednostki statystyczne, które przynależą do jednej nadrzędnej jednostki statystycznej, współdzielą wartość ryzyka hierarchicznego na poziomie indywidualnym.

Alternatywnym podejściem do oceny ryzyka ujawnienia, które również wykorzystuje informacje o częstościach kluczy (ale – w wypadku badań reprezentacyjnych – tylko w próbie losowej), jest zasada ***k-anonimowości*** zaproponowana jako swoisty kompromis pomiędzy stratą informacji a ryzykiem ujawnienia informacji poufnych. W myśl tej zasady sprawdza się, czy dla przyjętego progu $k > 1$ każda możliwa kombinacja wartości zmiennych kluczowych (najczęściej typu jakościowego) jest współdzielona przez co najmniej k rekordów. Jeżeli tak jest, to mówimy, że zbiór danych jednostkowych spełnia zasadę k-anonimowości. Za miarę ryzyka globalnego przyjmuje się tutaj np. liczbę lub odsetek rekordów, które naruszają tę zasadę dla przyjętego progu k. Jeżeli spełnienie zasady k-anonimowości (dla przyjętego progu k) zostanie uznane za wystarczającą ochronę, można się skupić na zapewnieniu mikrodanym największej możliwej użyteczności przy jedynym warunku: spełnieniu tej zasady. W celu spełnienia zasady k-anonimowości jej autorzy zalecają stosować połączenie przekodowania (globalnego) oraz lokalnego ukrywania danych (zob. podrozdz. 3.1) dla jakościowych quasi-identyfikatorów. Niestety, często zasada ta nie jest gwarantem zapewnienia mikrodanym wystarczającej ochrony. Osiągnięcie jej spełnienia przez wszystkie rekordy w mikrodanych przekłada się bowiem na redukcję ryzyka ujawnienia tożsamości, nadal jednak – o czym wspomniano na samym początku tego podrozdziału – może zajść ujawnienie atrybutu (Hundepool i in., 2012).

Jeżeli w zbiorze danych jednostkowych występują zmienne wrażliwe, to w celu niedopuszczenia do ujawnienia atrybutu, należy łączyć zasadę k-anonimowości z zasadą ***l-różnorodności***, która stanowi remedium na słabe strony zasady k-anonimowości. Zgodnie z definicją, jest ona spełniona, gdy dla liczby naturalnej $l > 1$ zbiór danych jednostkowych zawiera przynajmniej l różnych wrażliwych wartości każdej zmiennej wrażliwej takich, że są one najczęściej przyjmowanymi wartościami, a ich częstości są identyczne lub prawie identyczne. Wybór progu l zależy od liczby kategorii możliwych do przyjęcia przez każdą zmienną wrażliwą. Oznacza to, że dla każdej zmiennej wrażliwej z osobna ustala się wartość progu l, a następnie sprawdza się spełnienie zasady l-różnorodności przez wszystkie obserwa-

cje w mikrodanych. Ponadto zasadę tę można stosować jedynie do jakościowych zmiennych wrażliwych, które jednocześnie nie są rozpatrywane jako quasi-identyfikatory. W literaturze przedmiotu wyróżnia się różne warianty tej metody, które np. znajdą zastosowanie w przypadkach wielowymiarowych (Templ, 2017).

Podejście do oceny ryzyka ujawnienia informacji poufnych, gdy w mikrodanych wyróżniono quasi-identyfikatory typu ilościowego, różni się od podejścia stosowanego dla zmiennych jakościowych. Ze względu na przyjętą koncepcję unikatowości, wartości miar ryzyka ujawnienia wyznacza się przede wszystkim na poziomie całego zbioru danych jednostkowych (ryzyko globalne), a nie na poziomie rekordów. Może być ono wyrażone poprzez miarę o wartościach bezwzględnych (miara przedziałowa) lub względnych (na podstawie odpowiednio przeprowadzonego łączenia rekordów) (Templ i in., 2015; Templ, 2017; Benschop i in., 2022).

- **Miara przedziałowa** opiera się na założeniu, że wartości zakłócone nie powinny być zbyt bliskie wartości oryginalnych. W związku z tym buduje się wokół każdej zakłóconej wartości zmiennej ilościowej (a więc dla każdego rekordu z osobna) pewien przedział. Długość tego przedziału może być zależna na przykład od odchylenia standardowego zmiennej ilościowej oraz od parametru skalowania. Następnie sprawdza się, czy oryginalna wartość tej zmiennej dla określonej obserwacji zawiera się w tym przedziale. Jeżeli tak jest, to jednostka statystyczna jest uznawana za zagrożoną identyfikacją, gdyż zakłócona wartość zmiennej jest zbyt podobna do tej oryginalnej i w związku z tym wymaga większego zakłócenia. Jeżeli natomiast tak nie jest, to rekord dla jednostki statystycznej należy uznać za bezpieczny. Podejście takie może się okazać niewystarczające, gdy rozkład zmiennej ilościowej charakteryzuje się występowaniem wartości odstających. Wówczas, nawet po znacznym zakłóceniu wartości zmiennej ilościowej, wartości w ogonach rozkładu nadal mogą takimi pozostać, a jednostki statystyczne, którym one odpowiadają, pozostaną łatwe do identyfikacji.
- **Łączenie rekordów** – ocenia się liczbę poprawnych powiązań przy łączeniu obserwacji po wartościach zmiennej ilościowej w mikrodanych z badania statystycznego w dwóch postaciach: zakłóconej metodą ochrony poufności oraz oryginalnej. Połączenie następuje pomiędzy najbliższymi rekordami. Miara nie daje informacji o ryzyku ujawnienia w zasobach oryginalnej postaci, a jedynie służy do porównań i wyboru metody zakłóceniowej (jeżeli rozważanych jest kilka takich metod) oraz sposobu jej parametryzacji. W podejściu tym wykorzystuje się różne miary dystansu. Metody łączenia rekordów opisali szerzej m.in. Domingo-Ferrer i in. (2001), Domingo-Ferrer i Torra (2004), Hundepool i in. (2012) czy Templ i in. (2014).

Bardziej formalne zapisy wybranych miar, reguł i metod przedstawiono w dalszej części opracowania. Osobom chcącym zgłębić zaprezentowane powyżej zagad-

nienie autorzy polecają m.in podręczniki Hundepoola i in. (2010; 2012), w których bardziej szczegółowo omówiono założenia najgorszego możliwego scenariusza, wyprowadzono wzory na ryzyko ujawnienia na poziomie indywidualnym i globalnym, przedstawiono koncepcje podejść do oceny poufności w zależności od typu quasi-identyfikatorów, jak również zasygnalizowano newralgiczne aspekty związane z wyznaczaniem ryzyka ujawnienia. Z kolei w książce Templa (2017), oprócz ugruntowania teoretycznego, w przystępny sposób zaprezentowano praktyczne wykorzystanie podejść z użyciem pakietu `sdcMicro` w środowisku R.

2.2. Dane tabelaryczne

W niniejszym podrozdziale omówiono kwestię ryzyka ujawnienia informacji poufnych towarzyszącego udostępnianiu wynikowych lub publikacyjnych tablic statystycznych, które zostały naliczone przez krajowy urząd statystyczny lub innego gestora danych (np. do publikacji wyników z przeprowadzonego badania statystycznego lub przygotowanych specjalnie na zamówienie użytkownika zewnętrznego) na podstawie poufnego zbioru danych jednostkowych. Pojęcia związane z tablicami statystycznymi zostały już omówione w podrozdziale 1.2. Podejścia stosowane przy ochronie poufności danych tabelarycznych różnią się od siebie w zależności od postaci danych wynikowych.

Ogólnie rzecz ujmując, można powiedzieć, że tablice częstości są szczególnym przypadkiem tablic wielkości – z tą tylko różnicą, że wartość badanej cechy jest równa jeden, jeśli jednostka statystyczna należy do przekroju, który reprezentuje określona komórka tablicy, a zero w przeciwnym razie (Duncan i in., 2011). Ponadto w tablicach częstości agregaty mają zawsze wartość dodatnią lub zerową, gdy żaden respondent w populacji nie należy do wyróżnionego przekroju (w badaniach pełnych) lub – dodatkowa możliwość – nie został wylosowany do próby losowej (w badaniach reprezentacyjnych). Jest to jednak zero znaczące i wartość ta nie podlega obowiązkowi ochrony tajemnicy statystycznej. W wypadku tablic wielkości wartości agregatów mogą być – w zależności od wybranej zmiennej ilościowej, na podstawie której obliczono wartości sumaryczne – zarówno dodatnie, jak i ujemne, a wartość zero nie musi koniecznie oznaczać, że żadna jednostka statystyczna nie należy do rozpatrywanego przekroju. Ogólnie, w tablicach statystycznych występowanie małych liczebności w przekrojach jest zwykle związane z wysokim stopniem szczegółowości takiej tablicy, jak również z nierównym rozkładem jednostek pomiędzy kategorie zmiennej jakościowej, na podstawie której zostały utworzone owe przekroje. W skrajnym przypadku może to prowadzić do tego, że komórka tablicy będzie reprezentować wartość cechy tylko dla jednej jednostki.

Problem poufności w tablicach częstości należy wiązać przede wszystkim z występowaniem komórek o bardzo małych liczebnościach, w szczególności zaś

wynoszących jeden. Dla tablic wielkości jest on dodatkowo związany z przeważającym udziałem (dominacją) wartości zmiennej dla jednej jednostki statystycznej w ogólnej wartości agregatu.

Dostosowując wprowadzoną w podrozdziale 1.1 definicję ujawnienia do omawianej tu postaci danych wynikowych, możemy wyróżnić następujące jej przypadki:

- **Ujawnienie jednostki** – w tym kontekście identyfikację należy rozumieć szerzej, nie tylko jako rozpoznanie jednostki statystycznej jako członka grupy w tablicy, ale również rozpoznanie jej przez nią samą (tzw. **samoidentyfikacja**). Staje się ono realne, jeśli wartość komórki w tablicy częstości wynosi jeden. Jak twierdzą Hundepool i in. (2012), wiele urzędów statystycznych nie postrzega samoidentyfikacji jako naruszenia poufności. Jednak ktoś, kto sam rozpoznaje siebie jako członka grupy w tablicy statystycznej, może się obawiać, że inni też go rozpoznają. Identyfikacja lub samoidentyfikacja jednostki statystycznej może prowadzić do ujawnienia atrybutu dla zidentyfikowanej jednostki lub grupy.
- **Ujawnienie atrybutu (dla jednostki)** – dotyczy sytuacji, kiedy w pierwszej kolejności na podstawie niewielkich liczebności w komórkach tablicy częstości dochodzi do identyfikacji jednostki statystycznej, a następnie – z wykorzystaniem innych tablic wynikowych lub publikacyjnych naliczonych na podstawie tego samego źródła danych – ujawnione zostają dodatkowe informacje poufne o zidentyfikowanej jednostce.
- **Ujawnienie atrybutu (dla grupy)** – zachodzi wówczas, gdy użytkownik zewnętrzny dążący do naruszenia prywatności uzyskuje dodatkowe informacje o zidentyfikowanej grupie lub o tym, że zidentyfikowana grupa jakiegoś atrybutu nie posiada. W literaturze przedmiotu można znaleźć stwierdzenie, że jest to bardzo często zaniedbywany element ochrony poufności (Hundepool i in., 2012). Niebezpieczeństwo ujawnienia danych wrażliwych występuje wówczas, gdy wszystkie lub prawie wszystkie jednostki statystyczne należą do jednego przekroju tablicy.
- **Ujawnienie przez łączenie** – do identyfikacji dochodzi nie na podstawie pojedynczej tablicy statystycznej, lecz wskutek wykorzystania relacji łączącej dwie lub więcej takich tablic. Analiza wspólnych przekrojów występujących w różnych tablicach (tzw. tablicach łączonych) stwarza ryzyko uzyskania informacji identyfikującej jednostkę, a następnie dodatkowych informacji poufnych na jej temat.
- **Postrzeganie ryzyka naruszenia poufności** – oprócz obiektywnego ryzyka naruszenia poufności, na którego ograniczenie starają się wpływać krajowe urzędy statystyczne, istnieje ryzyko subiektywne. Użytkownik może postrzegać publikację danych tabelarycznych jako potencjalnie ryzykowną. Może to wynikać po pierwsze z charakteru danych postrzeganych jako wraż-

liwe, po drugie zaś – z braku wiedzy na temat stosowanych metod ochrony. Duncan i in. (2011) wskazali, że przykładowymi konsekwencjami ryzyka subiektywnego mogą być np. wzrost skali odmów odpowiedzi w badaniach statystycznych, jak również dodatkowa motywacja dla potencjalnych osób planujących naruszenie poufności do podjęcia takich prób. Przyczyn rzeczonego ryzyka natomiast doszukują się m.in. w małych wartościach komórek w tablicach, które bywają uważane za obarczone ryzykiem, bez względu na to, czy takimi rzeczywiście są, jak również w znajomości podstawowych kwestii demograficzno-społecznych, która może sugerować użytkownikom możliwość występowania nietypowych jednostek w mikrodanych.

Tablice wielkości najczęściej naliczane są na podstawie poufnych zasobów pozyskanych w toku badań statystyki przedsiębiorstw. Należy pamiętać, że liczba jednostek w populacjach przedsiębiorstw jest zwykle mniejsza niż w populacjach osób, a także, że wiele cech w populacji podmiotów gospodarczych charakteryzuje się silną asymetrią rozkładu. Ryzyko naruszenia poufności przede wszystkim może być związane z sytuacją, gdy do danego obszaru lub domeny:

- należy tylko jeden podmiot gospodarczy (samoidentyfikacja, identyfikacja i ujawnienie atrybutu przez inną jednostkę),
- należą dwa podmioty gospodarcze (identyfikacja i ujawnienie atrybutu przez konkurenta, który zna swój udział w wartości agregatu – jest to tzw. **ujawnienie dokładne**),
- należą co najmniej dwa podmioty gospodarcze, przy czym jeden z nich ma przeważający udział w wartości agregatu, przy niewielkim udziale pozostałych jednostek (identyfikacja i ujawnienie atrybutu w pewnym przedziale przez konkurenta – jest to tzw. **ujawnienie przybliżone**).

Komórka w tablicy statystycznej, która niesie ze sobą ryzyko ujawnienia informacji poufnych – a więc może prowadzić do naruszenia poufności danych – jest nazywana **komórką wrażliwą** lub **komórką z ryzykiem pierwotnym**. Mianem tym określa się komórki tablicy o małych częstościach (w wypadku tablic częstości) lub dodatkowo o wysokiej dominacji pojedynczych jednostek (w wypadku tablic wielkości). Ochrona poufności polega na uniemożliwieniu nie tylko poznania dokładnej wartości cechy, ale również poznania tej wartości w pewnym przedziale.

Żeby wytypować komórki obarczone ryzykiem pierwotnym, należy skorzystać z tzw. **reguł wrażliwości**. Poniżej wymieniono reguły najczęściej stosowane (Hundepool i in., 2012):

- **reguła minimalnej liczby respondentów reprezentowanych w komórce tablicy**: komórka jest wrażliwa, jeśli liczba respondentów reprezentowana w komórce okazuje się mniejsza niż n (zwykle za n przyjmuje się liczbę 3; w uzasadnionych przypadkach (Hundepool i in., 2012) można przyjąć liczbę większą niż 3); reguła ma zastosowanie zarówno dla tablic częstości, jak i tablic wielkości,

– **reguła (*n*, *k*)-dominacji**: komórka jest obarczona ryzykiem pierwotnym, jeśli *n* największych podmiotów gospodarczych w komórce tablicy reprezentuje więcej niż *k*% całkowitej wartości cechy ilościowej w komórce; zakłada się, że wartości cech są nieujemne; minimalna liczebność komórki wynosi $100n/k$; reguła ma zastosowanie wyłącznie w wypadku tablic wielkości,
– **reguła koncentracji *p*%**: komórka jest obarczona ryzykiem pierwotnym, jeśli całkowita wartość komórki po odjęciu wartości dla dwóch respondentów o kolejno największym udziale w komórce jest mniejsza niż *p*% wartości dla jednostki o największym udziale; zakłada się nieujemność wartości cech jednostek; minimalna wartość jednostek składających się na wartość komórki wynosi 3; reguła ma zastosowanie wyłącznie w wypadku tablic wielkości,
– **reguła *p*/*q***: komórka jest zagrożona ryzykiem pierwotnym, jeśli respondent należący do niej (i znający do *q*% pozostałych w niej udziałów) może oszacować dane wielkości dla innego respondenta z precyzją *p*% prawdziwej wartości; reguła ma zastosowanie wyłącznie w wypadku tablic wielkości.

Reguły (*n*, *k*)-dominacji, koncentracji *p*% oraz *p*/*q* należą do liniowych miar wrażliwości związanych z koncentracją jednostek. Należy wyraźnie podkreślić, że wartości parametrów dla reguł użytych w poszczególnych przypadkach powinny być traktowane jako poufne i nie wolno ich udostępniać użytkownikom zewnętrznym (o ile nie wynika to np. wprost z przepisów ustawy). Przy ocenie ryzyka pierwotnego część krajowych urzędów statystycznych stosuje regułę minimalnej liczby reprezentowanej przez komórkę tablicy wraz z regułą dominacji (1, *k*), jednak Hundepool i in. (2012) na przykładzie wykazali nieadekwatność takiego podejścia i stwierdzili, że lepszym rozwiązaniem może się okazać zastosowanie zarówno reguły (2, *k*)-dominacji, jak i reguły *p*%.

Po ustaleniu komórek z ryzykiem pierwotnym należy zapewnić, że wartości tych komórek nie zostaną ujawnione. Zakłada się przy tym, że wartości brzegowe (czyli sumy wartości w wierszach i kolumnach) są częścią tablicy statystycznej, a na podstawie relacji liniowych między cechami oraz operacji arytmetycznych na owych wartościach może dojść do odtworzenia wartości komórek wrażliwych. Wyeliminowanie takiej możliwości powoduje konieczność zastosowania odpowiednich technik, zwanych *ochroną wtórną* lub *uzupełniającą*. **Ryzykiem wtórnym** nazwiemy ryzyko ujawnienia informacji wrażliwej po wyeliminowaniu ryzyka pierwotnego, na przykład poprzez wykorzystanie związków i zależności pomiędzy poszczególnymi informacjami w tablicy (w tym wartości brzegowych) bądź w innych pozostających do dyspozycji intruza źródłach danych. Równoważnie możemy nazwać tym określeniem prawdopodobieństwo prawidłowego dopasowania danych do konkretnej jednostki. **Komórki z ryzykiem wtórnym** zaś to komórki tablic, w wypadku których występuje wtórne ryzyko ujawnienia informacji wrażliwych.

Bardziej sformalizowane zapisy reguł stosowanych do wyznaczania komórek wrażliwych, a także opisy algorytmów służących do detekcji komórek z ryzykiem wtórnym, zostały zaprezentowane w dalszej części monografii (por. rozdział 3).

2.3. Wyniki analiz

Poufne dane jednostkowe pozyskiwane w toku badań statystycznych realizowanych przez krajowe urzędy statystyczne mogą być udostępniane użytkownikom zewnętrznym w celach naukowo-badawczych na różne sposoby, które szerzej omówiono w rozdziale 6. Na potrzeby niniejszego podrozdziału należy jednak wspomnieć, że jedno z rozwiązań zakłada utworzenie przez gestora danych ściśle kontrolowanego środowiska pracy dla odbiorcy, które zostanie odizolowane od świata zewnętrznego i na którym będzie zapewniona możliwość pracy stacjonarnej (w siedzibie gestora danych lub w innym miejscu przez niego wskazanym) lub zdalnej. Krajowe urzędy statystyczne muszą zachować balans pomiędzy umożliwieniem użytkownikom danych wynikowych swobodnego prowadzenia prac naukowo-badawczych a dbałością o zapewnienie poufności. Praca w kontrolowanym środowisku gwarantuje dostęp do zbiorów danych jednostkowych (lub – w razie takiej konieczności – do danych w zagregowanej postaci) podlegających ochronie prawnej i etycznej. Użytkownicy realizują założone zamierzenia badawcze, przeprowadzając zaprojektowane przez siebie analizy. W tym celu na ogół przetwarzają dane z badań statystycznych w oryginalnej postaci (na której pracują również pracownicy służb statystyki publicznej) na różne (często nowe) sposoby, a wyniki tych analiz mogą w zasadzie przybrać dowolną postać. Efekty prac użytkowników zewnętrznych (czyli wyniki przeprowadzonych analiz statystycznych i ekonometrycznych) najczęściej są przeznaczane do udostępnienia poza kontrolowanym środowiskiem gestora danych, np. w formie artykułu zamieszczonego w czasopiśmie naukowym lub referatu wygłoszonego podczas konferencji naukowej. W związku z tym odbiorcy informacji statystycznych nie mogą mieć możliwości samowolnego pobrania tych wyników, ponieważ wszystkie zasoby pozostają pod kontrolą gestora danych i stanowią jego własność. Jeśli zechcą oni opublikować te rezultaty, muszą zwrócić się z prośbą o ich udostępnienie. I tu właśnie zaczyna się rola kontroli ujawniania wyników analiz (ang. *output checking*). Zapewnia ona poufność danych wynikowych w zagregowanej postaci i polega na weryfikacji wszystkich wyników analiz, które zostały opracowane w kontrolowanym przez gestora danych środowisku i które mają zostać udostępnione poza nim. Weryfikacja takich danych wynikowych jest więc zdecydowanie trudniejszym i odmiennym zadaniem w porównaniu z zapewnianiem ochrony poufności dostawców danych przy opracowywaniu oficjalnych statystyk, analiz i publikacji, gdzie dane wynikowe mają dobrze określoną i ugruntowaną formę.

Trudno jest bowiem opracować zestaw reguł, które obejmowałyby każdą możliwą postać danych wynikowych.

W celu wprowadzenia porządku w procesie kontroli ujawniania wyników analiz dane wynikowe opracowane przez użytkownika zewnętrznego na podstawie poufnych zbiorów danych jednostkowych dzieli się na skończoną liczbę klas. Kryterium podziału jest tutaj ich postać, nie zaś zakres informacyjny. Każdej z tych klas przyporządkowana jest etykieta *bezpieczne* lub *niebezpieczne*. Jeżeli jakiegoś wyniku analizy nie da się przyporządkować do żadnej z utworzonych wcześniej klas, to domyślną jego etykietą jest ta druga. Etykiety te mają jednak jedynie charakter informacyjny – zarówno dla użytkownika danych, jak i osoby oddelegowanej przez gestora danych do zapewnienia poufności udostępnianym informacjom statystycznym. To, że jakieś wyniki analiz otrzymają etykietę *niebezpieczne*, nie oznacza, że nie zostaną one udostępnione użytkownikowi zewnętrznemu. I podobnie, wyniki analiz sklasyfikowane jako *bezpieczne* niekoniecznie zostaną udostępnione poza chronione i ściśle kontrolowane środowisko. Dane wynikowe, którym przyporządkowano etykietę *bezpieczne*, to te, co do których uznaje się, że użytkownik danych otrzyma je w takiej samej postaci, w jakiej je opracował lub – co najwyżej – po naniesieniu nań niewielkich zmian. Wyjątki od tej reguły, w których pomimo oznaczenia wyniku analizy etykietą *bezpieczny* nie zostanie on udostępniony, powinny być ściśle określone i nie mogą być one liczne. Z kolei za *niebezpieczne* uznaje się takie dane wynikowe, co do których prawdopodobieństwo konieczności naniesienia znacznych zmian w procesie kontroli ujawniania wyników analiz jest wysokie – czyli takie, które z wysokim prawdopodobieństwem nie będą mogły zostać udostępnione użytkownikowi w opracowanej przez niego postaci. W zależności od przyjętego podejścia do tego procesu, w jego toku konieczne może być udowodnienie osobie odpowiedzialnej za ochronę poufności, że kontekst i zawartość danych wynikowych nie ujawnia informacji poufnych. Na ogół w wypadku wyników analiz opatrzonych etykietą *bezpieczne* to osoba przeprowadzająca proces SDC musi zebrać argumenty i wskazać przyczyny, z powodu których nie mogą być one udostępnione. Z kolei w wypadku danych z etykietą *niebezpieczne* to użytkownik danych, który je opracował, musi wykazać, że ich poufność nie jest zagrożona i że mogą być one opublikowane. Należy jednak podkreślić, że zawsze to w gestii i kompetencjach osoby sprawdzającej wyniki analiz leży podjęcie ostatecznej decyzji co do tego, które dane wynikowe mogą zostać udostępnione użytkownikowi zewnętrznemu, a które nie.

Jak już wspomniano, wpływ na podział danych wynikowych, które mają zostać poddane procesowi kontroli ujawniania wyników analiz, na te z etykietą *bezpieczne* oraz te z etykietą *niebezpieczne* ma jedynie ich struktura, a nie zakres informacji w zbiorze danych jednostkowych, na podstawie których przeprowadzono analizę. Wszystkie wyniki analiz, które zostały opracowane w ściśle kontrolowanym środowisku z wykorzystaniem poufnych mikrodanych, niezależnie

od ich etykiety, powinny podlegać sprawdzeniu, bez wyjątku. Etykieta *bezpieczne* informuje jedynie, że żadne wartości z bezpiecznych rezultatów nie będą usunięte. Ponadto, co również już zasygnalizowano, kontrola ujawniania wyników analiz każdorazowo zależy od kontekstu. Przed rozpoczęciem prac naukowo-badawczych przez użytkownika zewnętrznego nie da się określić, czy ich rezultaty będą mogły być udostępnione, czy też nie. Podział na *bezpieczne* i *niebezpieczne* informuje jedynie o prawdopodobieństwie udostępnienia opracowanych wyników analiz, jak również dostarcza wytycznych, co należy zrobić, aby dane wynikowe uznane za *niebezpieczne* ostatecznie stały się całkowicie *bezpieczne* i by można je udostępnić poza ściśle kontrolowane środowisko.

Krajowe urzędy statystyczne mogą podjąć liczne działania, by na koniec prac użytkownika zewnętrznego na poufnych zbiorach danych jednostkowych nie okazało się, że żadne opracowane przez niego wyniki analiz nie mogą zostać mu przekazane ze względu na naruszenie poufności. Do działań tego rodzaju należy między innymi wymóg określenia zakresu prowadzonych prac naukowo-badawczych oraz przewidywanych form ich zakończenia już we wniosku o udzielenie dostępu do zasobów. Choć struktura i zakres danych wynikowych mogą być z początku trudne do przewidzenia, jak również mogą się zmieniać w toku prac, to jednak w pewnych specyficznych sytuacjach, na podstawie wieloletniego doświadczenia, gestor danych może być w stanie podjąć decyzję o odrzuceniu wniosku o dostęp oraz może udzielić stosownej odpowiedzi zwrotnej. Ponadto proces SDC może być przeprowadzany wielokrotnie, bezpośrednio po przekazaniu przez użytkownika danych informacji, że naliczył on wyniki analiz i że mogą one zostać sprawdzone pod kątem poufności. Rozwiązanie takie zmniejsza prawdopodobieństwo zaistnienia sytuacji, w której na koniec projektu naukowo-badawczego okazałoby się, że żadne wyniki analiz opracowane w ściśle kontrolowanym przez gestora danych środowisku na podstawie poufnych mikrodanych, nie mogą zostać udostępnione ich odbiorcy. Tym samym doprowadziłoby to do marnotrawstwa jego czasu, pieniędzy oraz nakładu pracy.

Więcej informacji o klasyfikowaniu wyników analiz na *bezpieczne* i *niebezpieczne* można znaleźć m.in. w pracach: Hundepool i in. (2012), Bond i in. (2015), Brandt i in. (2010), Höninger i in. (2010).

Najważniejszym celem procesu kontroli ujawniania wyników analiz jest to, aby żadne informacje o jednostce statystycznej podlegającej ochronie w myśl obowiązku zachowania tajemnicy statystycznej nie zostały ujawnione przy udostępnianiu wszelkiej postaci wyników analiz, a w szczególności, by żaden respondent nie został zidentyfikowany. Osoba nieuprawniona nie może również uzyskać dostępu do poufnych informacji o grupach, które wcześniej były jej nieznane i dla niej niedostępne. Czynniki brane pod uwagę przy przeprowadzaniu tego procesu, które dodatkowo mogą zaważyć na decyzji odnośnie do udostępnienia lub nie danych wynikowych, obejmują sytuacje, w których identyfikacja wymaga sporych

nakładów czasu i wysiłku, dodatkowej wiedzy (której posiadania – lub łatwego uzyskania – nie oczekuje się od większości użytkowników), a także możliwości technologicznych pozostających w dyspozycji użytkownika.

Podczas sprawdzania danych wynikowych można spotkać błędy dwojakiego rodzaju:

- **błędy poufności**, polegające na zaakceptowaniu do publikacji bądź udostępnienia wyników analiz, które jednak naruszają reguły poufności,
- **błędy nieefektywności**, zachodzące wówczas, gdy wstrzymuje się publikację lub udostępnienie danych wynikowych, które nie naruszają reguł poufności.

Sposób radzenia sobie z oboma rodzajami błędów polega w jednym podejściu na minimalizowaniu obu, a w drugim – na minimalizowaniu błędu poufności przy akceptacji nieefektywności.

Więcej informacji na temat tych błędów można znaleźć np. w pracach: Hundepool i in. (2012), Bond i in. (2015), Brandt i in. (2010), Höninger i in. (2010) oraz Ritchie i Elliot (2015).

Jak już wspomniano, opracowanie odpowiednich reguł kontroli ujawniania wyników analiz do użycia przez gestora danych w wypadku udzielania dostępu użytkownikom zewnętrznym do poufnych zasobów w ściśle kontrolowanym przez niego środowisku jest trudne. Wynika to z potrzeby pozostawiania maksymalnej możliwej swobody przy prowadzeniu analiz na danych jednostkowych oraz z nieznanej postaci (pod względem rozmiaru czy ilości) danych wynikowych.

W zakresie oceny ryzyka ujawnienia informacji poufnych dla wyników analiz wypracowane zostały dwa modelowe podejścia:

- model oparty na zasadach,
- reguła kciuka.

Zastosowanie **modelu opartego na zasadach** (ang. *principles-based model*) pozwala na zminimalizowanie błędów obu wymienionych wyżej rodzajów. Drugie podejście z kolei, tzw. **zasada kciuka** (ang. *rule-of-thumb model*), zapobiega błędom poufności, lecz błędy nieefektywności są w nim akceptowane.

Model oparty na zasadach wymaga dobrej współpracy pomiędzy użytkownikami zewnętrznymi a osobami oddelegowanymi przez gestora danych do rozstrzygania kwestii zapewniania ochrony poufności. Ze względu na potrzebę zapobiegania błędom nieefektywności ściśle sprecyzowane reguły nie są tutaj wskazane, nie uwzględniają bowiem nigdy pełnej złożoności danych wynikowych. W podejściu tym wyniki analiz są najpierw rozpatrywane w całym kontekście, a dopiero później jest podejmowana decyzja co do ich bezpieczeństwa. Osoba przeprowadzająca proces kontroli ujawniania wyników analiz powinna zweryfikować nie tylko same dane wynikowe, ale także ich powiązanie z innymi, wcześniej udostępnionymi rezultatami. Jest bowiem prawdopodobne, że dane wynikowe, które z osobna nie

ujawniają informacji poufnych, w kombinacji z innymi mogą pozwolić na uzyskanie takich informacji o objętej badaniem jednostce statystycznej bądź grupie jednostek. Dodatkowo osoby prowadzące prace naukowo-badawcze często zwracają się z prośbą o sprawdzenie kilku wyników jednocześnie (np. zestawu powiązanych tablic) – w tym wypadku również występuje podobne ryzyko. Konieczne zatem wydaje się odbycie szkoleń w zakresie kontroli ujawniania wyników analiz, zarówno przez badaczy, jak również przez osoby reprezentujące krajowy urząd statystyczny czy innego gestora, aby zasady rządzące przebiegiem tego procesu były dla wszystkich klarowne.

Zaletą podejścia opartego na zasadach jest pozostawienie badaczowi maksymalnej elastyczności przy pracy, co przełoży się niewątpliwie pozytywnie na kompleksowość wykorzystania przezeń dostępnych dla niego danych. Wadą tej opcji natomiast jest to, że wymaga ona przeprowadzenia szkoleń dla pracowników gestora danych oraz użytkowników zewnętrznych, którzy muszą poświęcić na ten cel swój czas i wysiłek. Omawiany model przenosi również odpowiedzialność za weryfikację danych wynikowych z osób, które opracowują reguły (w podejściu opartym na regułach), na osoby sprawdzające poszczególne dane wynikowe. Nie ma tu ścisłych standardów do stosowania, a osoba weryfikująca wyniki analiz musi się kierować posiadanym doświadczeniem i zrozumieniem podstawowych zasad kontroli ujawniania danych.

Zasada kciuka opiera się na ścisłych regułach, które w mniej lub bardziej zautomatyzowany sposób mogą zostać zastosowane, nawet bez obszernej wiedzy i doświadczenia z zakresu kontroli ujawniania wyników analiz. W przeciwieństwie do tego podejścia, pierwszy model dostarcza jedynie wskazówek do sprawdzania danych wynikowych, wymaga bardziej szczegółowego spojrzenia na nie, nie zaś zastosowania prostych reguł. Omawiane podejścia mogą być odrębnie stosowane, można je również z sukcesem połączyć i wdrożyć do praktycznego użycia jednocześnie. Ideą w takim wypadku jest, by najpierw zastosować zasadę kciuka – wdrożyć jej ściśle określone reguły i sprawdzić poprawność ich wykonania – a następnie, po zidentyfikowaniu wyników uznanych za niebezpieczne z perspektywy tychże reguł, przeprowadzić ich bardziej kompleksowe sprawdzenie z zastosowaniem modelu opartego na zasadach.

Jak już wspomniano, zasada kciuka zapobiega błędom poufności, akceptując niektóre błędy nieefektywności i przyjmując je za pewne. Podejście to opiera się na restrykcyjnych, sztywnych regułach. Prawdopodobieństwo tego, że wyniki analiz, które spełniają te reguły, doprowadzą do ujawnienia informacji poufnych, jest bardzo niskie. Niezmienny charakter reguł sprawia, że możliwa jest ich automatyzacja w większym lub mniejszym stopniu. Dzięki niej posiadanie zaawansowanej wiedzy z obszaru kontroli ujawniania wyników analiz (zarówno po stronie użytkowników, jak również pracowników krajowego urzędu statystycznego czy innego gestora danych) nie jest wymagane, aczkolwiek staje się wskazane.

Z drugiej jednak strony, ze względu na sztywność reguł i nieuwzględnienie w nich pełnego kontekstu danych wynikowych, zasada kciuka nie gwarantuje, że pomimo spełnienia przez owe dane stosownych reguł, nie zostaną ujawnione informacje poufne. Wśród kluczowych zalet przemawiających jednak za wdrożeniem tego podejścia do wykorzystania w praktyce należy również wskazać większą praktyczność oraz mniejszą czasochłonność w porównaniu z modelem opartym na zasadach. W literaturze przedmiotu wskazuje się, że podejście to jest polecane w niezbyt skomplikowanych, prostych przypadkach, gdy dane wynikowe mają ograniczoną szczegółowość. Zalecane jest ono również dla niedoświadczonych gestorów danych, którzy dopiero wprowadzają udostępnianie poufnych zbiorów danych jednostkowych w ściśle kontrolowanym środowisku i mają w związku z tym zbyt słabe doświadczenie w zakresie modelu opartego na zasadach. Zasada kciuka zapewnia wówczas maksymalne bezpieczeństwo, pozwala na zdobycie doświadczenia i w miarę upływu czasu na bardziej złożone spojrzenie na problem. Stosowanie ściśle określonych reguł jest również zalecane, gdy proces kontroli ujawniania wyników analiz ma być przeprowadzany w sposób częściowo lub w pełni automatyczny. Dla zredukowania stopnia trudności realizowanego procesu podejście to można także łączyć z modelem opartym na zasadach. W połączonym podejściu zasada kciuka może być punktem wyjścia przy sprawdzaniu dowolnych wyników. Pozwala ona wskazać ich krytyczne (niebezpieczne) elementy, które nie spełniają określonych reguł i dla których następnie – w celu podjęcia ostatecznej decyzji o bezpieczeństwie udostępnienia danych wynikowych – może być stosowany model oparty na zasadach.

Więcej informacji o opisanych powyżej podejściach do procesu SDC dla wyników analiz przedstawili m.in: Hundepool i in. (2012), Bond i in. (2015), Brandt i in. (2010), Höninger i in. (2010) oraz Ritchie i Elliot (2015).

Podsumowując charakterystykę kontroli ujawniania wyników analiz w kontekście ochrony poufności udostępnianych danych wynikowych, należy podkreślić, że przebiega ona inaczej niż w procesach wykorzystujących mikrodane bądź dane tabelaryczne. W procesach kontroli ujawniania mikrodanych oraz kontroli ujawniania tablic statystycznych wypracowano bowiem liczne reguły, miary i metody pozwalające na dokonanie pomiaru lub szacunku ryzyka ujawnienia informacji poufnych, które dodatkowo mają stosowne interpretacje ułatwiające proces decyzyjny. Dla wyników analiz tego ryzyka się nie mierzy, a przynajmniej nie w takim samym kontekście. Tutaj jedynie klasyfikuje się je jako bezpieczne lub nie, a następnie poddaje się procesowi kontroli ujawniania wyników analiz, który dostarcza odpowiedzi na pytanie, czy mogą one zostać udostępnione użytkownikowi zewnętrznemu, czy wymagane jest dokonanie w nich modyfikacji, czy też w części lub w całości nie mogą one zostać przekazane. Spośród trzech filarów kontroli ujawniania danych statystycznych, którymi są mikrodane, tablice statystyczne oraz wyniki analiz, to właśnie te ostatnie wydają się najmniej rozpozna-

nym obszarem, z najskromniejszym wachlarzem dostępnych metod i technik, jak również z brakiem odpowiednich narzędzi informatycznych (obecnie nie ma żadnego powszechnie dostępnego programu zawierającego stosowne funkcjonalności i pozwalającego na przeprowadzenie tego procesu) oraz nielicznymi tekstami naukowymi poświęconymi temu zagadnieniu. Częstym problemem jest również brak dostępu do wykwalifikowanej kadry pracowniczej oraz odpowiedniego środowiska pracy. Kontrola ujawniania wyników analiz pozostaje więc nadal sporym wyzwaniem dla mniej doświadczonych krajowych urzędów statystycznych, a także w zakresie harmonizacji i uspójniania podejść stosowanych na arenie międzynarodowej.

3

Metody i techniki kontroli ujawniania danych wynikowych

W tej części zaprezentowano najistotniejsze metody kontroli ujawniania danych w odniesieniu do ochrony mikrodanych. Znalazły się tutaj zarówno sposoby proste, jak również podejścia bardziej złożone, oparte na zaawansowanych algorytmach i narzędziach matematycznych. Oprócz tego omówiono dokładnie, ilustrując to odpowiednimi przykładami, poszczególne narzędzia ochrony mikrodanych i danych tabelarycznych przed ujawnieniem. Warto podkreślić, że specyfika metod SDC dla każdego z tych typów jest odrębna. Natomiast w obrębie każdej z rzeczonych form danych dokonano stosownych porównań podstawowych właściwości rozpatrywanych metod. Następnie omówiono sposoby ochrony informacji publikowanych w opracowaniach statystycznych – zwłaszcza w formie statystyk opisowych, rezultatów różnorodnych analiz oraz wykresów i kartogramów. Na zakończenie zarysowano istotę, metody i dylematy związane z generowaniem danych syntetycznych. Trzeba pamiętać, że nie istnieje uniwersalna hierarchia zaprezentowanych tutaj metod pod względem ryzyka ujawnienia czy straty informacji – ich przydatność zależy od konkretnych danych i przyjętych parametrów.

3.1. Metody niezakłóceniowe

Pierwszym i oczywistym sposobem ochrony poufności mikrodanych jest ich **anonimizacja**. Oznacza to usunięcie ze zbioru mikrodanych zmiennych – identyfikatorów oraz tych spośród quasi-identyfikatorów, które w danym układzie stały się identyfikatorami w całej lub w znacznej części rekordów rozpatrywanego zbioru. Przykład 3.1 ilustruje, na czym owa czynność polega.

Przykład 3.1. Załóżmy, że z pewnego badania pozyskano bazę danych ukazaną w tablicy 3.1. Jednoznacznymi identyfikatorami są: numer PESEL oraz – rozpatrywane łącznie – imię i nazwisko respondenta, dlatego te informacje należy usunąć. Stosowne kolumny zostały zaznaczone kolorem.

Tabl. 3.1. Przykład anonimizowanej bazy danych

PESEL	Imię	Nazwisko	Wiek	Z1	Z2	Z3	Z4	Z5
37121106358	Jan	Kowalski	73	1	6		2	2
37092211337	Wojciech	Ziębiński	73	1	6		2	3
53010822762	Anna	Nowak	57	1	1	2	2	4
61043044523	Weronika	Malinowska	49	1	9		1	2
91110561191	Jacek	Kwiatkowski	19	1	9		2	1
74060922015	Jerzy	Zieliński	36	3	1	2	2	1
81032751846	Grażyna	Jarzębska	29	1	3		2	4
55021230511	Andrzej	Kowalewski	55	1	1	2	2	4
95070402349	Marlena	Witkowska	15	1	6		2	1
94081125737	Jakub	Wolski	16	2	1	2	2	1
64113088123	Mirosława	Janowska	47	1	3		2	2

Źródło: Dane fikcyjne.

Z pojęciem anonimizacji ściśle wiąże się **pseudonimizacja**. Jest to przetworzenie danych osobowych w taki sposób, by nie można ich było już przypisać konkretnej osobie, której dane dotyczą, bez użycia dodatkowych informacji, pod warunkiem że takie dodatkowe informacje są przechowywane osobno i są objęte środkami technicznymi i organizacyjnymi uniemożliwiającymi ich przypisanie zidentyfikowanej lub możliwej do zidentyfikowania osobie fizycznej. W odróżnieniu od anonimizacji, pseudonimizacja jest procesem odwracalnym: zawsze można użyć dodatkowych informacji z osobnego pliku do odtworzenia danych oryginalnych. Dlatego też te dodatkowe informacje powinny być szczególnie chronione przed dostępem do nich osób niepowołanych.

Pozostałe sposoby, które opisano poniżej, mają zastosowanie do baz, z których już usunięto identyfikatory lub quasi-identyfikatory umożliwiające w powiązaniu ze sobą jednoznaczną identyfikację jednostek. Niemniej nadal pozostaje kwestia ryzyka identyfikacji jednostki na podstawie pozostałych w zbiorze informacji. Trzeba zatem je zredukować do minimum.

Pierwsze omówione tu sposoby są zaliczane do **maskowania niezakłóceniowego** (ang. *non-perturbative masking*). Prowadzą one bowiem do tego, że wrażliwe dane stają się – w różny sposób – niewidoczne dla zewnętrznego użytkownika. Tak więc w finalnym udostępnianym zbiorze określona informacja jednostkowa albo figuruje w dokładnej postaci, albo jej nie ma wcale.

Prostą technicznie metodą takiej redukcji jest **podpróbkowanie** (ang. *subsampling*). Oznacza to nic innego jak udostępnianie pewnej próbki rekordów spośród figurujących w bazie danych zgromadzonych w trakcie badania statystycznego. Próbka ta może być dobrana w sposób losowy albo nielosowy. Usta-

lając metodę doboru, należy uwzględnić potrzebę zredukowania zasobu przede wszystkim tych informacji, które stwarzają największe ryzyko identyfikacji jednostek. Na przykład, jeśli w danej gminie mieszka tylko jedna osoba w wieku od 45–50 lat z wyższym wykształceniem chemicznym, to w losowaniu podpróbki (np. przy użyciu schematu losowania prostego bez zwracania) z dużym prawdopodobieństwem ten rekord zostanie pominięty. Jeśli zaś do tego losowania wprowadzimy ograniczenie wiekowe (np. ustalimy, że pod uwagę bierzemy tylko osoby w wieku do 44 lat) lub co do poziomu albo kierunku wykształcenia (np. uwzględniamy jedynie osoby z wykształceniem wyższym humanistycznym lub ekonomicznym), to zostanie pominięty na pewno. Podpróbkowanie redukuje wiedzę o odpowiedziach z wyjściowej próby badania. Z drugiej jednak strony prowadzi faktycznie do usunięcia wszystkich wartości w pewnym podzbiorze rekordów. Może mieć to istotny wpływ na jakość prowadzonych przez użytkownika tych danych obliczeń, np. estymacji określonych wielkości na podstawie tych danych. Im mniejsza próbka, tym precyzja szacunków gorsza (nawet jeśli zastosować optymalizację wag dla wybranych do próbki rekordów). Metoda podpróbkowania jest efektywna w przypadku danych jakościowych (tzn. wyrażonych na skali nominalnej lub porządkowej). Nieco gorzej sprawdza się ona dla danych ilościowych (skala różnicowa bądź ilorazowa). Tutaj zawsze istnieje ryzyko, że pewne zmienne tego rodzaju mogą się znajdować w niezależnym zbiorze administracyjnym będącym w posiadaniu osoby nieuprawnionej do dostępu do danych jednostkowych statystyki publicznej. Jest wtedy wielce prawdopodobne, że jeśli wartość danej zmiennej w obu zbiorach będzie identyczna, to dotyczy to tej samej jednostki. W przypadku gdy w owym uzyskanym przez rzeczoną osobę niezależnie zbiorze znajdą się dane identyfikacyjne, osoba taka otrzyma bez trudu wszystkie dane – także chronione – dla owej jednostki ze zbioru stworzonego przez statystykę. Tak więc użycie podpróbkowania dla danych ilościowych ma sens tylko wówczas, gdy wyklucza lub poważnie utrudnia taką możliwość (na przykład poprzez równości przybliżone).

Innym podejściem jest **przekodowanie** (ang. *recoding*) określonych wrażliwych zmiennych. Jeśli zmienne mają charakter jakościowy, polega ono na połączeniu kilku kategorii w jedną – bardziej zgrubną i o większej liczbie należących do niej jednostek, która pozwala ukryć informację wrażliwą. Dla zmiennej ilościowej przekodowanie następuje na drodze zastąpienia owej zmiennej przez jej odpowiednik w postaci jakościowej. Przykłady 3.2 oraz 3.3 ilustrują tę technikę.

Przykład 3.2. Załóżmy, że na pewnym poziomie agregacji występuje pojedynczy rekord, który opisuje osobę w wieku 62 lat zatrudnioną w swym głównym miejscu pracy w niepełnym wymiarze czasu pracy. Przypuśćmy – jak czyniono to na przykład w Narodowym Spisie Powszechnym Ludności i Mieszkań w 2011 r. – że status zatrudnienia w głównym miejscu pracy był kodowany następująco:

1 – pracownik najemny pełnozatrudniony,
2 – pracownik najemny niepełnozatrudniony,
3 – pracodawca,
4 – pracujący na własny rachunek (niezatrudniający pracowników),
5 – pomagający członek rodziny,
9 – nieustalony.

Można wówczas połączyć kategorie 1 i 2 w jedną: pracownik najemny i owemu wrażliwemu rekordowi (a także pozostałym, u których zmienna ta oryginalnie przyjmuje kategorie 1 lub 2) przypisać tę nową. Wtedy nowe kodowanie byłoby takie:

1 – pracownik najemny,
2 – pracodawca,
3 – pracujący na własny rachunek (niezatrudniający pracowników),
4 – pomagający członek rodziny,
9 – nieustalony.

Przykład 3.3. Przypuśćmy, że w bazie danych znajduje się zmienna „odległość w kilometrach od miejsca zamieszkania do głównego miejsca pracy" oraz że występują pojedyncze rekordy, które na podstawie jednorazowo występujących wartości tej zmiennej (np. 7 km lub 33 km), a także innych charakterystyk, da się łatwo zidentyfikować. Można wówczas zamienić tę zmienną na zmienną jakościową, ukazującą przynależność obserwacji zmiennej wyjściowej do określonego przedziału jej wartości, na przykład:

1 – do 5 km,
2 – od 6 do 10 km,
3 – od 11 do 15 km,
4 – od 16 do 20 km,
5 – od 21 do 25 km,
6 – od 26 do 30 km,
7 – od 31 do 35 km,
8 – od 36 do 40 km,
9 – od 41 do 45 km,
10 – od 46 do 50 km,
11 – 51 i więcej km.

Ma to sens oczywiście tylko wtedy, gdy do każdej z tych kategorii należy wystarczająco duża liczba rekordów. W przeciwnym razie trzeba zastosować szersze przedziały (a zatem mniejszą liczbę kategorii).

Specyficzną wariacją przekodowania jest **przekodowanie górne i dolne** (ang. *top and bottom coding*). Stosuje się je do zmiennych, których wartości mogą być uporządkowane (a zatem wyrażone są na skali porządkowej, różnicowej lub ilorazowej). Chodzi tutaj o to, że nowe, zgrubne kategorie są tworzone tylko dla wartości najwyższych i najniższych.

W wypadku danych z przykładu 3.3 oznacza to, że jeżeli w bazie występują pojedyncze rekordy z wielkościami 7 km czy 48 km, to pierwszą kategorią byłoby np. „do 10 km", a ostatnią „46 i więcej km" (z odpowiednimi etykietami). Pozostałe przedziały wartości zostają zachowane.

Lokalne ukrywanie danych (ang. *local suppression*) polega na usuwaniu pewnych wartości niektórych zmiennych dla konkretnych jednostek w celu uniknięcia identyfikacji jakiejkolwiek jednostki, o której dane zgromadzono w zbiorze mikrodanych. Efektem lokalnego ukrywania jest zwiększenie liczby rekordów, dla których kombinacja określonych wartości pewnych innych zmiennych, uznanych za kluczowe, jest taka sama. Dzieje się tak dlatego, że brak danych – w domyślnym podejściu – traktowany jest jako dowolna możliwa do przyjęcia przez określoną zmienną wartość. Metoda ta ma zastosowanie przede wszystkim do danych jakościowych. W wypadku zmiennych ilościowych (czyli o wartościach wyrażonych na skali różnicowej lub ilorazowej) każda wartość może być bowiem z dużym prawdopodobieństwem – w powiązaniu z innymi kluczowymi zmiennymi – unikatowa. Stąd nie ma sensu ukrywać wartości takiej zmiennej (można natomiast ją np. przekodować), warto się zaś skupić na ukrywaniu wartości zmiennych jakościowych. To, które wartości należy ukryć, zależy od tego, czy kombinacja pozostałych nie zidentyfikuje danej jednostki, ale także od ważności danej cechy dla badacza. Każdorazowo należy jednak dążyć do tego, aby liczba ukrytych wartości była jak najmniejsza. Ilustruje to przykład 3.4.

Przykład 3.4. Załóżmy, że z pewnego badania pozyskano bazę danych ukazaną w tablicy 3.2. Jest to baza odpersonalizowana, w której zastosowano następujące kodowanie (oparte częściowo na metodologii Narodowego Spisu Powszechnego Ludności i Mieszkań w 2011 r.):

- płeć: M – mężczyzna, K – kobieta,
- stan cywilny prawny: 1 – kawaler/panna, 2 – żonaty/zamężna, 3 – wdowiec/wdowa, 4 – rozwiedziony/rozwiedziona, 9 – nieustalony,
- wykształcenie: 1 – wyższe ze stopniem naukowym co najmniej doktora, 2 – wyższe z tytułem magistra, lekarza lub równorzędnym, 3 – wyższe z tytułem inżyniera, licencjata, dyplomowanego ekonomisty, 4 – dyplom ukończenia kolegium, 5 – policealne z maturą, pomaturalne, 6 – policealne bez matury, 7 – średnie zawodowe z maturą, 8 – średnie zawodowe bez matury, 9 – średnie ogólnokształcące z maturą, 10 – średnie ogólnokształcące bez matury, 11 – zasadnicze zawodowe, 12 – gimnazjalne, 13 – podstawowe, 14 – podstawowe nieukończone i bez wykształcenia, 99 – nieustalone,
- status na rynku pracy: 1 – pracujący, 2 – bezrobotny, 3 – bierny zawodowo, 9 – nieustalony.

Wyróżnione rekordy to te, w których występują unikatowe kombinacje kategorii badanych zmiennych. Niektóre z tych wartości muszą zatem być ukryte. Ponieważ zmienna płeć ma tylko dwie opcje, a – co jeszcze ważniejsze – jest bardzo

Tabl. 3.2. Przykład ukrywania wrażliwych danych

ID	Płeć	Stan cywilny prawny	Wykształcenie	Status na rynku pracy
1	M	1	6	3
2	M	1	6	3
3	K	3	5	2
4	K	1	11	1
5	K	1	9	1
6	M	3	1	2
7	K	1	3	2
8	M	1	6	3
9	K	1	6	2
10	M	2	4	2
11	K	1	3	1
12	K	1	9	1
13	M	2	4	2
14	M	2	7	1
15	K	3	5	2
16	K	3	5	2
17	M	3	1	2
18	M	3	1	3
19	K	1	9	3
20	K	1	9	2
21	M	1	2	1
22	M	2	7	1
23	K	1	6	2
24	M	2	4	2
25	M	2	12	1
26	M	2	7	1

Objaśnienia: ID – sztuczny identyfikator osoby. Wyróżniono rekordy z danymi zagrożonymi identyfikacją. Czerwonym kolorem zaznaczono informacje proponowane do usunięcia.

Źródło: Dane fikcyjne.

istotna, zatem tutaj niewskazane byłoby ukrywanie jakichkolwiek informacji. Popatrzmy zatem na pozostałe zmienne. W wypadku rekordu o ID = 4 tylko poziom wykształcenia odnośnej osoby jest unikatowy (nikt inny takiego poziomu nie ma). Natomiast kobiet będących pannami pracującymi jest w bazie więcej (ponad 3). Tym samym to właśnie wykształcenie należy ukryć. Z podobnych względów ukry-

wamy też status na rynku pracy dla rekordu o ID = 12. Ukrycie stanu cywilnego prawnego niewiele pomoże, gdyż istnieje jeszcze tylko jeden rekord (ID = 5), dla którego wykształcenie = 9 (średnie ogólnokształcące z maturą), a zatem – np. na podstawie liczby kobiet panien ogółem o takim wykształceniu (uzyskanej z publikacji zbiorczej czy innego źródła) – użytkownik mógłby łatwo odtworzyć stan cywilny owej osoby. Stąd należy usunąć raczej wykształcenie lub status na rynku pracy. Ponieważ kobiet panien o zmiennej wykształcenie = 9 jest więcej niż kobiet panien pracujących, ryzyko owego odtworzenia danej wrażliwej w pierwszym przypadku jest dużo mniejsze niż w drugim. Stąd należy usunąć status na rynku pracy (o ile wykształcenie nie jest dla badacza ważniejsze). Te oraz podobnie uzyskane propozycje usunięć zaznaczono w tablicy 3.2 kolorem czerwonym.

Jak widać, w wypadku lokalnego ukrywania danych badacz ma pewną swobodę odnośnie do wyboru zmiennych czy wartości, które mają być ukryte, aby liczba koniecznych ukryć była jak najmniejsza. Oczywiście ramy owej swobody są wyznaczone przez ważność zmiennych (czasami ustaloną subiektywnie) oraz przez ryzyko identyfikacji jednostki (to kryterium jest przede wszystkim obiektywne).

Obecnie przyjrzymy się rodzajom i sposobom stosowania metod niezakłóceniowych dla tablic statystycznych. Na początek jednak kilka uwag ogólnych. W wypadku tablic częstości ochronę poufności można zastosować na wcześniejszym etapie niż naliczenie tablicy. Z drugiej zaś strony – metody ochrony poufności mogą zostać użyte także wówczas, gdy tablica została już naliczona. Z tego punktu widzenia metoda, która została zastosowana do danych jednostkowych, może być również uznana za sposób ochrony danych zawartych w tablicy. Pod tym względem wśród metod ochrony poufności rozróżnia się metody *pretablicowe* i *posttablicowe*. Według innego kryterium metody ochrony poufności dzieli się na takie, które modyfikują dane w celu ochrony poufności, oraz takie, które nie zmieniają danych, a tylko np. ukrywają wybrane wartości w tablicach. Biorąc to pod uwagę, Hundepool i in. (2012) wprowadzili podział metod ochrony poufności, w tym wypadku dla tablic częstości, na:

- Ukrywanie komórek – metoda polega na niepublikowaniu wartości komórki z ryzykiem ujawnienia oraz zastąpieniu jej umownym znakiem, np. X (szczegóły związane z tą metodą omówiono w podrozdziale 3.4.2). Zastosowanie ukrywania komórek składa się z dwóch etapów. W pierwszym wyznaczone zostają komórki, które ze względu na małą liczność mogą stanowić ryzyko naruszenia poufności, określane jako komórki z ryzykiem pierwotnym. Ze względu na reguły addytywności występujące zazwyczaj w tablicy, nie będzie to ochrona wystarczająca. W związku z tym konieczne jest ukrycie dodatkowych komórek, określane jako ukrycie wtórne. W opublikowanej tablicy nie ma rozróżnienia, na którym etapie poszczególne komórki zostały ukryte.
- Metody, które stosuje się przed zaprojektowaniem i naliczeniem tablicy – są to metody stosowane do mikrodanych. Polegają na wprowadzeniu w nich

określonych zmian. Jako zalety tych metod wymienia się to, że proces ochrony jest realizowany jednokrotnie, co ułatwia „produkcję" tablic. Z drugiej strony jako ich wadę wskazuje się wpływ dokonanej korekty na rozkład cech w zbiorze danych (tzn. po zastosowaniu ochrony rozkład cechy zmienia się w stosunku do rozkładu danych pierwotnych). Ponadto rozwiązanie to często nie pozwala przekazać badaczom wystarczająco czytelnych kryteriów dających tymże użytkownikom możność dokonania w swoich analizach stosownych korekt wynikających ze zmian wprowadzonych w danych.

- Metody, które dotyczą korekty kształtu tablicy – ogólnie, restrukturyzacja w celu ochrony poufności polega na zmniejszeniu stopnia szczegółowości tablicy przed jej publikacją. Restrukturyzacja ma na celu redukcję liczby komórek z ryzykiem pierwotnym. Metoda ta nie wyklucza zastosowania równocześnie innych metod – zarówno na etapie wcześniejszym (tzn. na etapie danych jednostkowych), jak również na etapie późniejszym.
- Metody, które wprowadzają korektę w wartościach komórek tablicy – polegają na publikowaniu wartości zmienionej w miejsce rzeczywistej. Zmianie może podlegać część komórek lub nawet wszystkie, w zależności od zastosowanej metody. W angielskiej terminologii określa się to jako zakłócanie danych (ang. *perturbation*) – w odróżnieniu od modyfikacji tablicy polegającej tylko na ukryciu wybranych wartości i zastąpieniu ich umownym symbolem.

Metody zachowania poufności dla tablic dzieli się na określone klasy – zarówno ze względu na etap procesu statystycznego, na którym się je stosuje, jak też ze względu na sposób manipulacji danymi.

Ze względu na moment zastosowania ochrony poufności wyróżnia się następujące rodzaje metod SDC:

- **pretablicowe** – w tej klasie metod ochronę poufności stosuje się jeszcze przed naliczeniem tablic; w związku z tym metody te są niezależne od konkretnego naliczenia danej tablicy; w tym znaczeniu każdą z metod stosowanych do mikrodanych można uznać za pretablicową ochronę danych,
- **posttablicowe** – metody, które służą do ochrony poufności po naliczeniu tablicy; oprócz ochrony zastosowanie tych metod ma na celu również ocenę ryzyka naruszenia poufności oraz maksymalizację użyteczności danych.

Inną klasyfikację metod można przedstawić jako podział na:

- metody, które zmieniają dane zawarte w tablicy,
- metody, które nie zmieniają danych zawartych w tablicy, ale wprowadzają ograniczenia na publikację pewnych widniejących w niej informacji, zwykle poprzez ukrycie określonych wartości w komórce.

Obecnie omówimy te spośród powyższych metod, które należą do klasy podejść niezakłóceniowych dla tablic statystycznych, w części 3.2 zaś – podejścia zakłóceniowe.

Restrukturyzacja tablicy (ang. *table re-design*) jest rekomendowana jako proste w użyciu narzędzie, które po pierwsze – chroni ryzykowne komórki, a po drugie – zachowuje ich rzeczywiste wartości (Hundepool i in., 2012). Wartości komórek, które niosą ze sobą ryzyko naruszenia poufności, są łączone z innymi w celu zmniejszenia stopnia ich szczegółowości. Duncan i in. (2011) zwrócili uwagę, że restrukturyzacja – czyli ponowne konstruowanie struktury tablicy – jest szczególnym przypadkiem całego procesu projektowania publikacji wyników w formie tablic. W tym jednak wypadku mamy zazwyczaj do czynienia z sytuacją, że intencją instytucji statystycznej lub innego gestora było uzyskanie określonego poziomu szczegółowości, ale ze względu na zachowanie poufności poziom ów musiał ulec zmianie (np. na bardziej zgrubny), co z punktu widzenia użyteczności danych ma istotne znaczenie. Strata informacji jest zresztą wymieniana jako poważna wada tej metody. Do zalet rozpatrywanego podejścia należy natomiast łatwość jego wdrożenia, zrozumiałość dla użytkownika oraz zachowanie reguł addytywności.

Restrukturyzacja polega na rozbiciu tablicy na tablice o niższych wymiarach. W tym względzie jej szczególnym przypadkiem jest przekodowanie. Załóżmy na przykład, że w kolumnach tablicy mamy płeć (kobiety, mężczyźni) × stopień niepełnosprawności (brak, lekki, umiarkowany, ciężki), a w wierszach wiek w latach. Przekodowaniem w tym wypadku może być użycie grup wieku (np. 5-letnich: 0–4 lat, 5–9 lat itd.) zamiast liczby przeżytych lat, bez zmiany pozostałych elementów tablicy (czyli kolumnami będą kolejno: kobiety pełnosprawne, kobiety niepełnosprawne w stopniu lekkim, kobiety niepełnosprawne w stopniu umiarkowanym, kobiety niepełnosprawne w stopniu ciężkim, mężczyźni pełnosprawni, mężczyźni niepełnosprawni w stopniu lekkim, mężczyźni niepełnosprawni w stopniu umiarkowanym oraz mężczyźni niepełnosprawni w stopniu ciężkim). Z kolei restrukturyzacją będzie przebudowanie tablicy w ten sposób, by w kolumnach oddzielnie, a nie krzyżowo, znalazły się kategorie płci i stopnia niepełnosprawności. Oznacza to na przykład, że w pierwszej kolumnie znajdą się kobiety, w drugiej – mężczyźni, w trzeciej – osoby pełnosprawne, w czwartej – niepełnosprawni w stopniu lekkim, w piątej – niepełnosprawni w stopniu umiarkowanym, w szóstej zaś – niepełnosprawni w stopniu ciężkim.

Poniżej zaprezentowano przykładową restrukturyzację tablicy. Tablica 3.3 została zrestrukturyzowana przez zmniejszenie liczby kategorii, w wyniku czego uzyskano tablicę 3.4. Nie zapewnia to jednak wystarczającej ochrony poufności z powodu równoczesnej publikacji tablicy 3.5, która pozostaje w relacji z tablicą 3.3.

W zakresie tablic wielkości natomiast, ponieważ użytkownicy mają bardzo różne oczekiwania w stosunku do udostępnianych wyników badań, krajowe urzędy statystyczne prowadzą liczne klasyfikacje, według których publikowane są rzeczone wyniki. Klasyfikacje owe tworzą hierarchie, a tablice łączą zmienne klasyfikacyjne według różnych przekrojów. Jeżeli dwie tablice przedstawiają dane,

Tabl. 3.3. Liczba biernych zawodowo według płci na obszarze X

Wyszczególnienie	Mężczyźni	Kobiety	Razem
Obszar 1	2	9	11
Obszar 2	38	0	38
Obszar 3	25	14	39
Razem	65	23	88

Źródło: Obliczenia własne, dane fikcyjne.

Tabl. 3.4. Liczba biernych zawodowo według płci na obszarze X (tablica po restrukturyzacji)

Wyszczególnienie	Mężczyźni	Kobiety	Razem
Obszar 1+2	40	9	49
Obszar 3	25	14	39
Razem	65	23	88

Źródło: Obliczenia własne, dane fikcyjne.

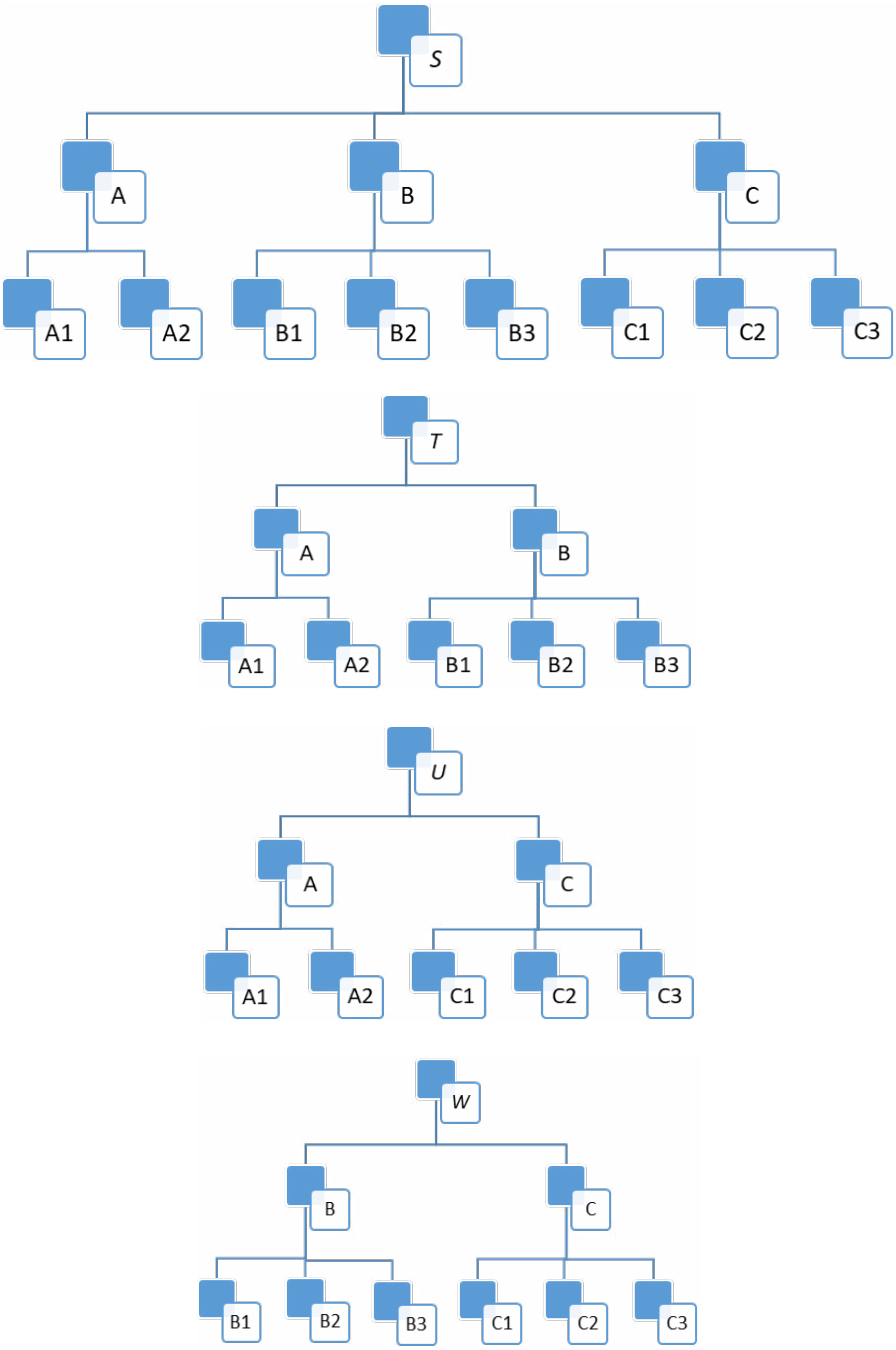
Tabl. 3.5. Liczba biernych zawodowo na obszarze X

Obszar 1	Obszar 2	Obszar 3	Razem
11	38	39	88

Źródło: Obliczenia własne, dane fikcyjne.

które da się połączyć według przynajmniej jednej wspólnej zmiennej występującej w obu tablicach, można powiedzieć, że mamy do czynienia z tablicami łączonymi. Przykład ukazany na rysunku 3.1 oraz w tablicach 3.6–3.8 opiera się na publikacji de Wolfa (2007), który wprowadził hierarchię w sposób formalny.

Niech publikacje dotyczą na przykład wartości sprzedaży w regionie C i podregionach C1, C2 i C3. Innym wymiarem dla publikacji niech będzie dział klasyfikacji działalności B, składający się z grup B1, B2 i B3. Za ostatni wymiar można wówczas przyjąć klasę wielkości A dzielącą zbiorowość na podklasy A1 i A2. Dane są wówczas publikowane w ramach trzech zmiennych hierarchicznych – *T*: A×B, *U*: A×C, oraz *W*: B×C. Zmienna klasyfikacyjna *S* definiuje hierarchię bazową dla zmiennych *T*, *U*, *W*. Jednocześnie zmienne te są podhierarchiami zmiennej *S*. Tablice 3.6, 3.7. i 3.8 są tablicami łączącymi (czyli dającymi w połączeniu hierarchię bazową), które pokrywa zmienna *S*. Dla ochrony komórek wrażliwych niezbędne jest wzięcie pod uwagę bazowej zmiennej hierarchicznej, gdyż ustalenie i ukrycie lub zniekształcenie komórek wrażliwych jedynie dla każdej z tablic z osobna może nadal prowadzić do naruszenia poufności poprzez wykorzystanie relacji zależności pomiędzy tablicami z uwzględnieniem łączącej je hierarchii *S*. Stosowny przykład wskazuje wspomniany artykuł (de Wolf, 2007).



Rys. 3.1. Schematy hierarchii tablic

Źródło: Opracowano na podstawie: (de Wolf, 2007).

Tabl. 3.6. Tablica zmiennych A i B

T	B1	B2	B3	Ogółem
A1	…			
A2		…		
Ogółem			…	

Źródło: Opracowano na podstawie: (de Wolf, 2007).

Tabl. 3.7. Tablica zmiennych A i C

U	C1	C2	C3	Ogółem
A1	…			
A2		…		
Ogółem			…	

Źródło: Opracowano na podstawie: (de Wolf, 2007).

Tabl. 3.8. Tablica zmiennych B i C

W	C1	C2	C3	Ogółem
B1	…			
B2		…		
B3			…	
Ogółem				…

Źródło: Opracowano na podstawie: (de Wolf, 2007).

Metodą często stosowaną w celu ochrony tablic zawierających w komórkach dane ilościowe jest metoda nazwana **przekodowywaniem globalnym** (ang. *global recoding*), co można też rozumieć jako restrukturyzację tablicy. Polega ona na redukcji szczegółowości wymiaru tablicy, czyli zmniejszeniu przed opublikowaniem liczby kategorii zmiennej klasyfikacyjnej. Prowadzi to nierzadko do dużej utraty użyteczności publikowanych danych. Z tego powodu podejście to jest często stosowane wraz z innymi metodami ochrony. Czasami wymogi publikacji nie pozwalają na zmianę liczby kategorii zmiennej klasyfikacyjnej, co z kolei uniemożliwia zastosowanie tej metody jako jedynej metody ochrony.

Ukrywanie komórek tablicy jest najpopularniejszą metodą stosowaną w celu ochrony danych w tablicach wyników badań statystyki gospodarczej. Wartości wszystkich komórek wskazanych jako wrażliwe zostają zastąpione ustalonym symbolem, np. X. Proces ukrywania polega w pierwszej kolejności na wytypowaniu komórek z ryzykiem pierwotnym. Po tym etapie następuje wtórne wyznaczenie komórek, które również zostaną ukryte. Docelowo po opublikowaniu tablicy nie ma możliwości rozróżnienia, z jakiego powodu komórka została ukryta, tzn.

czy stało się tak ze względu na ryzyko pierwotne, czy też wtórne. Podczas wytypowania komórek z ryzykiem wtórnym – które to wytypowanie jest działaniem uzupełniającym – pojawia się problem optymalnego wyznaczenia komórek do ukrycia. Najlepszym rozwiązaniem jest tu wyznaczenie zbioru komórek przeznaczonych do ukrycia, dla którego strata informacji jest najmniejsza. Fischetti i Salazar-Gonzalez (2003) wskazali, że z matematycznego punktu widzenia znalezienie jednoznacznego, optymalnego rozwiązania dla wszystkich przypadków w efektywnym czasie jest bardzo mało prawdopodobne. Poszukiwanie rozwiązań koncentruje się więc przede wszystkim na podejściu heurystycznym – i to głównie na tablicach dwu- i trójwymiarowych. Innym problemem jest występowanie w tablicy komórek określanych jako *singletony* i *wielokomórkowe ryzyka*. Singleton to komórka, której wartość jest reprezentowana tylko przez jednego respondenta. Z kolei sytuacja, w której respondent ma udział w wartości więcej niż jednej komórki tablicy na tym samym poziomie agregacji, jest nazywana ryzykiem wielokomórkowym. Uwzględnienie tego ryzyka może prowadzić do zbytniego poziomu utraty informacji.

Główne reguły dla wyznaczania poufnych komórek z ryzykiem pierwotnym w tablicy to (por. podrozdz. 2.2):

- minimalna liczba jednostek, które składają się daną agregację w komórce (najczęściej przyjmowaną wartością jest 3),
- reguła dominacji (n, k),
- reguła p%.

W podręczniku Hundepoola i in. (2012) można znaleźć wyrażenie pewnych zależności pomiędzy regułami (n, k) i p% w ten sposób, że np. dla (2, k) można przyjąć wartość $p = 100(100 - k)/k$. Wtedy na ogół zbiór komórek, które według tak wyrażonej reguły p% staną się poufne, będzie podzbiorem zbiorowości komórek poufnych wedle reguły (2, k). Z drugiej zaś strony zbiór ten okaże się znacznie bardziej liczny niż zbiór komórek w tej tablicy uznanych za poufne wedle reguły (1, k). Będzie więc $U_{(1,\,k)} \subseteq U_p \subseteq U_{(2,\,k)}$, gdzie $U_{(1,\,k)}$ to zbiór komórek poufnych pierwotnie zakwalifikowanych w danej tablicy według reguły (1, k), U_p – zbiór komórek poufnych wedle reguły p%, $U_{(2,\,k)}$ zaś – zbiór komórek poufnych zgodnie z regułą (2, k). Na przykład, gdyby przyjąć za k liczbę 75, wówczas wartość p według wspomnianej zależności wynosiłaby 33. Należy też podkreślić, że w przypadku stosowania reguły p%, aby komórka mogła zostać uznana za bezpieczną, minimalna liczba jednostek, które będą składały się na daną agregację komórki, wynosi 3.

Idea **przedziału poufności** (ang. *confidentiality interval*) bazuje na tym, że ze względu na potencjalną liniową zależność pomiędzy komórkami, które nie są ukryte, a ukrytymi zawsze istnieje możliwość wyznaczenia zakresu możliwych wartości komórki ukrytej w pewnym przedziale, tzn. wyznaczenia górnego i dolnego progu przedziału, w którym znajduje się faktyczna wartość owej komórki.

Dotyczy to najczęściej sytuacji, gdy tablica zawiera wartości nieujemne. Właściwa ochrona poufności danych w tablicy poprzez ukrycie znajdujących się w niej wartości dla części komórek polega w tym wypadku na uniemożliwieniu poznania wartości komórki wrażliwej z dokładnością, której granice są zawarte w przedziale określonym przez reguły poufności. Granice przedziału poufności są wyznaczane dla komórek w ramach ukrycia pierwotnego. W tablicy 3.9 zaprezentowano górne wartości przedziału dla reguł koncentracji. Dolną granicę przedziału określa się z reguły zgodnie z zasadami symetrii.

Tabl. 3.9. Górne granice przedziałów poufności według najważniejszych reguł SDC

Reguła	Górna granica przedziału
$(1, k)$	$\left(\frac{100}{k}\right)x_1 - X$
(n, k)	$\left(\frac{100}{k}\right)(x_1 + x_2 + \ldots + x_n) - X$
$p\%$	$\left(\frac{p}{100}\right)x_1 - (X - x_1 - x_2)$

Źródło: Hundepool i in. (2012).

Tablica 3.10 ilustruje tablicę, w której zastosowano ukrywanie komórek. Następnie z pomocą występujących w tejże tablicy zależności wyznaczono przedział, w którym musi się znajdować wartość ukrytej komórki.

Tabl. 3.10. Przykład tablicy z ukrytymi komórkami

Wyszczególnienie	A	B	C	Razem
I	X_{11}	8	X_{13}	30
II	X_{21}	40	X_{23}	50
III	17	16	30	63
Razem	30	64	49	143

Źródło: Opracowano na podstawie (Hundepool i in., 2012).

Z liniowej zależności pomiędzy komórkami w tablicy oraz z tego, że X_{11}, X_{21}, X_{13}, $X_{23} \geq 0$, otrzymujemy:

$$X_{11} + X_{13} = 22,$$
$$X_{21} + X_{23} = 10,$$
$$X_{11} + X_{21} = 13,$$
$$X_{13} + X_{23} = 19.$$

Dla komórki X_{11} górna granica przedziału wartości wynosi zatem 13, a dolna 3, tzn. $X_{11}^{max} = 13$ oraz $X_{11}^{min} = 3$ (gdyż $X_{21} \leq 10$). Wykorzystując metodykę programowania liniowego, można wyznaczyć granice górną i dolną przedziałów dla każdej ukrytej komórki w tablicy. De Waal i Coutinho (2020) zaproponowali algorytm, który uwzględnia agregację wszystkich zależności (także tych pośrednio wynikających z równości podstawowych) i zarazem wydajniejszy obliczeniowo od klasycznego, a to dzięki zastąpieniu jednego dużego problemu kilkoma mniejszymi, łatwiej rozwiązywalnymi.

Daalmans i de Waal (2010) przedstawili ogólne sformułowanie problemu wtórnego ukrywania komórek. Wtórne ukrywanie zalicza się do klasy szczególnie trudnych zagadnień *NP*[11], co w praktyce oznacza, że nie można się spodziewać znalezienia algorytmu, który wyznaczy jednoznaczne rozwiązanie w efektywnym czasie dla wszystkich przypadków zbiorów wejściowych. W celu dokonania ukrycia komórek w pierwszym kroku ustala się, które komórki w tablicy są wrażliwe. Realizacja tej procedury bazuje na uprzednio zaakceptowanych kryteriach. Często stosowanym w tym kontekście kryterium, w uzupełnieniu do wcześniej wymienianych, jest reguła (p, q), $p < q$, nazywana regułą *a priori – a posteriori*. Jest ona rozszerzeniem reguły p% (por. podrozdz. 2.2). Przyjmuje się, że przed opublikowaniem tablicy każdy respondent może oszacować udział każdego innego respondenta w każdej komórce tablicy w zakresie q%. Komórka jest uznawana za wrażliwą, jeśli udział respondenta dostawcy w wartości komórki może być oszacowany po opublikowaniu tablicy przez innego respondenta dostawcę do tej komórki w zakresie p%. Reguła (p, q) redukuje się do reguły p%, gdy przyjmiemy $q = 100$. Wartości komórek uznanych za wrażliwe na podstawie tej reguły nie są publikowane. Drugi krok w procedurze zachowania poufności to wyznaczenie dodatkowych komórek do ukrycia, które to ukrycie nie pozwoli na wyznaczenie wartości w komórkach wrażliwych w ustalonym wcześniej za pomocą wybranej reguły zakresie. Im szerszy zakres zostanie przyjęty, tym ukryte komórki będą bardziej „bezpieczne". Wiązać się to jednak będzie z większą stratą informacji. Natomiast od strony obliczeniowej wyznaczenie dodatkowych komórek do ukrycia polega na znalezieniu takiego rozwiązania, które ograniczy do minimum stratę informacji.

Duncan i in. (2011) zwrócili uwagę, że problem ten był w ostatnich latach szeroko eksplorowany poprzez odwołanie się do teorii grafów, programowania matematycznego oraz sztucznej inteligencji. Punktem wyjścia jest tu określenie dwóch kryteriów rozwiązania: maksymalnej ochrony i minimalizacji straty informacji. Problem sformułowany zostaje następnie jako jednowymiarowy, gdzie przy przyjętych z góry ograniczeniach dla pierwszego z tych kryteriów poszuku-

[11] Problem NP (ang. *nondeterministic polynomial*, niedeterministycznie wielomianowy) to problem decyzyjny, dla którego rozwiązanie można zweryfikować w czasie wielomianowym.

je się optimum dla drugiego. Ograniczeniami owymi będą granice przedziałów dla ochrony komórek pierwotnych oraz poszukiwanie optymalnego rozwiązania z punktu widzenia minimum dla sumy wartości komórek wybranych do ukrycia wtórnego. Problem ten jest zagadnieniem programowania matematycznego, w którym warunki ograniczające mają postać liniową. Jeśli dodatkowo narzuci się ograniczenie, że wartości w komórkach są całkowite, zagadnienie sprowadza się do liniowego programowania całkowitoliczbowego (ang. *integer linear programming* – ILP). Ponieważ wraz ze zwiększaniem się rozmiaru tablic złożoność obliczeniowa gwałtownie rośnie, rozwiązuje się tu na ogół problem programowania liniowego (ang. *linear programming* – LP) bez ograniczenia do liczb całkowitych. Fischetti i Salazar-Gonzalez (2003) zaproponowali w tym kontekście metodę ukrywania częściowego, w której zamiast totalnego ukrywania komórki publikowany byłby przyjęty pewien górny i dolny zakres przedziału, w którym mieści się jej wartość.

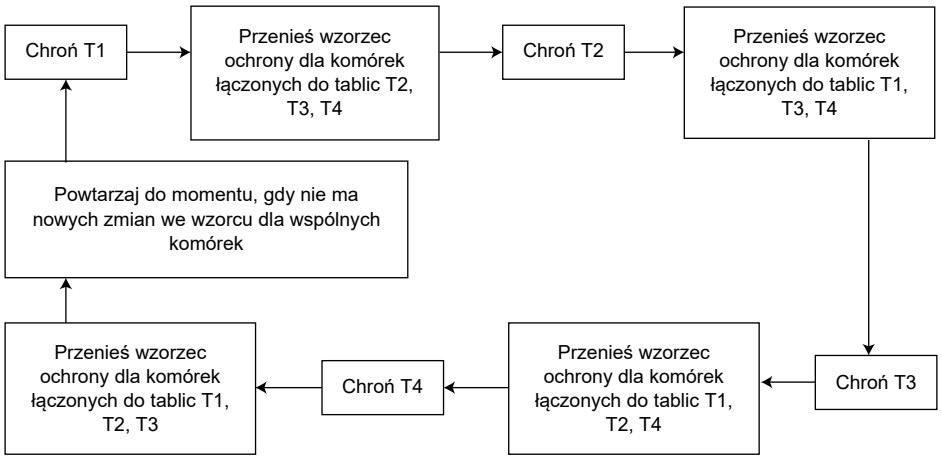
Złożoność problemu poszukiwania wtórnego ukrywania szybko rośnie wraz z wielkością tablic. Dodatkowo tablice zawierające hierarchie są tablicami złożonymi, które mogą być rozpatrywane jako rodzaj tablic łączonych. Dla złożonych i „dużych" tablic stosuje się rozwiązania heurystyczne. Są one oparte na programowaniu liniowym, a ich celem jest zaproponowanie algorytmów, które znajdą rozwiązanie bliskie optymalnego w efektywnym czasie. Do takich rozwiązań należą:

- **Podejście modułowe** (HiTaS) – tablicę hierarchiczną dzieli się na tablice bez hierarchii i poszukuje optimum ukrycia dla każdej wydzielonej tablicy z osobna. Łącząc wyniki w odpowiedni sposób, uzyskuje się rozwiązanie dla całej tablicy, które nie musi być optymalne. Celem tego podejścia jest szybsze znalezienie rozwiązania, które może jednak prowadzić do większej straty informacji.
- **Podejście hiperkostki** (ang. *hypercube*) – podobnie jak dla podejścia modułowego, tablica zostaje podzielona na proste podtablice, a następnie poszukuje się rozwiązania dla każdej z podtablic w sposób iteracyjny. W odróżnieniu od podejścia modułowego, dla każdej z podtablic znalezione rozwiązanie nie musi być rozwiązaniem optymalnym, co z kolei zwykle prowadzi do bardziej restrykcyjnego ukrycia komórek, niż byłoby to konieczne.
- **Przepływy sieciowe** (ang. *Network*) – możliwe do zastosowania w tablicach dwuwymiarowych z co najwyżej jedną zmienną hierarchiczną. Sieci są metodą bardzo efektywną, ale niedającą gwarancji znalezienia optymalnego rozwiązania. Finalnie jednak często otrzymuje się tu rezultat bliski optymalnego. Sieć reprezentuje relacje agregacji według kolumn i wierszy w tablicy dwuwymiarowej. Zapewnia relację addytywności względem wartości brzegowych i wartości globalnej.

Idea ukrywania komórek dla tablic hierarchicznych polega na wydzieleniu stosownych podtablic i traktowaniu ich jak osobne tablice. Trzeba jednak pamiętać

o dokonanych przypadkach ukrycia komórek w innych podtablicach. W związku z tym procedura ukrywania jest stosowana kilkukrotnie do każdej podtablicy. Pomimo to okazuje się ona efektywniejsza, niż gdyby była zastosowana do tablicy bez podziału na podtablice. Dodatkowo w metodzie modułowej ważna jest kolejność wyboru podtablic przeznaczonych do ukrywania.

W wypadku tablic łączonych właściwa kolejność owego wyboru jest trudniejsza do określenia i nieoczywista. Spośród zaproponowanych rozwiązań tego problemu można wymienić rozszerzone podejście modułowe. Polega ono na wyznaczeniu dla zbioru tablic łączonych najpierw najmniejszej tablicy, która zawiera wszystkie podtablice wymagające ochrony. Następnie do tych tablic stosowane jest klasyczne podejście modułowe. Powodem, dla którego wskazuje się słabości ukrywania komórek i podkreśla konieczność stosowania innych metod w tym kontekście, są zmiany technologiczne i rosnące możliwości generowania tablic na życzenie użytkownika. Stosowanie ukrywania komórek dla tablic łączonych wymaga bowiem jednoczesnego wykonania procedury ochrony dla wszystkich tablic, które zostają połączone. Jeżeli przynajmniej jedna z tablic została opublikowana wcześniej, wyznaczenie optymalnego rozwiązania może nie być możliwe. Trzeba bowiem zapewnić, żeby ukryte były wspólne komórki występujące we wszystkich tablicach. Dodatkowo problem pogłębia się wraz z większą liczbą tablic. Na rysunku 3.2 przedstawiono schemat ukrywania dla czterech tablic. Podejście to, nazywane tradycyjnym, jest podejściem iteracyjnym i zostało opisane np. w artykule Giessing (2009). Jednak przy większej liczbie tablic jest to zbyt uciążliwe.



Rys. 3.2. Schematyczny opis algorytmu ukrywania komórek w tablicach oznaczonych jako T1, T2, T3, T4, jeżeli uwzględni się, że są to tablice łączone (podejście tradycyjne)

Źródło: Opracowano własne.

3.2. Metody zakłóceniowe

Innym typem metod kontroli ujawniania mikrodanych jest **maskowanie zakłóceniowe** (ang. *perturbative masking*). Polega ono na zakłócaniu samych wrażliwych wartości zmiennych w celu uniemożliwienia dokładnego ich odtworzenia przez nieuprawnionego użytkownika przy jednoczesnej minimalizacji strat informacyjnych. Ogólnie rzecz biorąc, maskowanie zakłóceniowe można przedstawić jako transformację wyjściowego zbioru danych **X** na zbiór danych **Z** postaci:

$$\mathbf{Z} = \mathbf{AXB} + \mathbf{C}.$$

Macierze **X** i **Z** są tutaj rozumiane jako macierze o rozmiarze $n \times m$, **A** jest macierzą transformacji rekordów (o rozmiarze $n \times n$), **B** – macierzą transformacji zmiennych (o rozmiarze $m \times m$), **C** – macierzą przesunięć lub szumu (o rozmiarze $n \times m$).

Najprostszym rodzajem zakłóceń jest **dodawanie szumu** (ang. *noise addition*). Polega ono na dodawaniu do oryginalnych danych wrażliwych specjalnie zdefiniowanych zakłóceń w celu zniekształcenia owych informacji uniemożliwiającego odtworzenie ich faktycznej postaci, jednak przy minimalizacji negatywnych efektów dla jakości odpowiednich wielkości zagregowanych dla populacji (a najlepiej zachowaniu ich bez zmian). Szum taki modeluje się często za pomocą odpowiednich zmiennych losowych. Warto zaznaczyć, że określenie „dodawanie" nie ma tutaj ścisłej konotacji matematycznej: szum można nakładać w rozmaity sposób przy użyciu różnorodnych operacji. Poniżej zaprezentowano najpowszechniejsze techniki z tej rodziny.

Szum addytywny (zwany także *białym szumem*) polega na dodaniu do wartości zmiennej odpowiedniego zakłócenia:

$$z = x + \varepsilon,$$

gdzie x jest oryginalną daną, z – daną zniekształconą, ε – zmienną losową o rozkładzie normalnym, $\varepsilon \sim N(0, \sigma^2)$. Zatem oryginalna zmienna X_j, której wartości opisuje wektor $x_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$ będzie przekształcona do postaci

$$z_{ij} = x_{ij} + \varepsilon_{ij}, \tag{3.1}$$

przy czym ε_{ij} jest odpowiednim zakłóceniem uzyskanym z rozkładu normalnego $\varepsilon_{ij} \sim N(0, \sigma_j^2)$. Jeśli w szumie nie występuje autokorelacja, tzn. jeżeli $\mathrm{Cov}(\boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_k) = 0$ ($\boldsymbol{\varepsilon}_l = (\varepsilon_{1l}, \varepsilon_{2l}, \ldots, \varepsilon_{nl})$), dla każdych $j, k = 1, 2, \ldots, m$, $j \neq k$, to ten szum addytywny nazywamy *nieskorelowanym* (ang. *uncorrelated additive noise*). Istotną cechą

tego rodzaju zakłóceń jest to, że kowariancja między zmiennymi zostaje zachowana, tzn. $\mathrm{Cov}(Z_j, Z_k) = \mathrm{Cov}(X_j, X_k)$ dla każdych $j, k = 1, 2, \ldots, m, j \neq k$. Opcja ta nie zachowuje jednak wariancji i korelacji oryginalnych zmiennych (w sensie Pearsona). Jeśli chcemy to osiągnąć, to konieczne jest zastosowanie *skorelowanego* szumu addytywnego (ang. *correlated additive noise*), czyli przyjęcie założenia, że $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1^T \boldsymbol{\varepsilon}_2^T \ldots \boldsymbol{\varepsilon}_m^T] \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$, gdzie $\boldsymbol{\Sigma}_\varepsilon = a\boldsymbol{\Sigma}_X$, dla pewnej dodatniej liczby rzeczywistej a, przy czym $\boldsymbol{\Sigma}_X$ oznacza oryginalną macierz kowariancji zmiennych $X_1, X_2, \ldots, X_m$ znajdujących się bazie. Szerzej na ten temat piszą np. Domingo-Ferrer i in. (2004).

Szumu można też użyć tak, aby próbkowa macierz kowariancji zmiennych zniekształconych była nieobciążonym estymatorem macierzy kowariancji zmiennych oryginalnych. Może mieć to istotne znaczenie dla utrzymania odpowiedniej jakości estymacji, choćby dla małych domen. Stosuje się zatem **szum z transformacją liniową**. Postępowanie jest wówczas dwuetapowe. W pierwszym kroku nakłada się szum według wzoru (3.1). Następnie zmienne Z_j są transformowane do postaci (Domingo-Ferrer i in., 2004):

$$G_j = cZ_j + d_j, \qquad j = 1, 2, \ldots, m.$$

Parametry c i d_j są wyznaczone tak, aby $E(G_j) = E(X_j)$ oraz $\mathrm{Var}(G_j) = \mathrm{Var}(X_j)$ dla każdego $j = 1, 2, \ldots, m$, co daje $d_j = (1 - c)E(X_j)$, $j = 1, 2, \ldots, m$.

Istnieją też możliwości wykorzystania w konstrukcji szumu **transformacji nieliniowych**. Kwestiami tymi zajmował się m.in. Sullivan (1989). Warto przytoczyć tylko za Domingo-Ferrerem i in. (2004), że podejście takie opiera się na następującej procedurze składającej się z sześciu kroków:

1. Wyznaczamy dystrybuantę empiryczną dla każdej zmiennej wyjściowej.
2. Wygładzamy dystrybuantę empiryczną.
3. Konwertujemy wygładzoną dystrybuantę empiryczną na zmienną losową o rozkładzie jednostajnym; zmienną tę przekształcamy następnie na zmienną losową o standaryzowanym rozkładzie normalnym.
4. Dodajemy szum do zmiennej losowej o standaryzowanym rozkładzie normalnym.
5. Transformujemy powrotnie otrzymaną zmienną na wartości dystrybuanty.
6. Dokonujemy transformacji powrotnej do oryginalnej skali pomiarowej.

Zaletą tego podejścia jest zachowanie rozkładów jednowymiarowych. Transformacja powrotna do oryginalnej dystrybuanty jest dokonywana z wykorzystaniem specjalnego algorytmu iteracyjnego opartego na aproksymacji odpowiednich wielkości przy użyciu średnich wartości oryginalnej funkcji lub wartości funkcji dla środków przedziałów, do których „zaszumione" dane należą. Procedury takie wymagają jednak specjalistycznej wiedzy i bywają czasochłonne. Z drugiej strony, jeśli na przykład parametr c w wypadku transformacji liniowej nie zosta-

nie utajniony, to osoba nieupoważniona może – korzystając z takiego algorytmu transformacji powrotnej – odtworzyć wiedzę o szumie i jego rozkładzie, co prowadzi do ujawnienia chronionych informacji. Wykazali to właśnie Domingo-Ferrer i in. (2004).

Innym rodzajem szumu jest **szum multyplikatywny** (ang. *multiplicative noise*)[12]. Jego nakładanie polega na mnożeniu wartości wyjściowych zmiennych przez liczby losowe wygenerowane z rozkładu – zazwyczaj normalnego – o średniej jeden i niskiej wariancji:

$$z = x\eta,$$

gdzie η to zmienna losowa o rozkładzie normalnym, $\eta \sim N(1, \sigma^2)$, przy czym σ^2 jest małe. Tak więc oryginalna zmienna X_j będzie przekształcona do postaci:

$$z_{ij} = x_{ij}\, \eta_{ij},$$

przy czym η_{ij} jest odpowiednim zakłóceniem uzyskanym z rozkładu normalnego $\eta_{ij} \sim N(1, \sigma^2)$.

Szum multyplikatywny może też występować **w wersji logarytmicznej** (ang. *logarithmic multiplicative noise*). Zmienne wyjściowe poddajemy wówczas transformacji logarytmicznej:

$$Y_j = \ln X_j.$$

Do danych po transformacji dodajemy następnie szum w postaci liczb losowych generowanych z odpowiedniego rozkładu, np. standaryzowanego rozkładu normalnego:

$$Z_j = Y_j + \varepsilon_j.$$

Finalnie „zaszumione" dane otrzymujemy, biorąc potęgę o podstawie e i wykładniku Z_j, czyli obliczając e^{Z_j}. „Zaszumienie" to w istocie ma postać multyplikatywną, gdyż można je przedstawić jako $Z_j = X_j \cdot \xi_j$, przy czym ξ_j jest zmienną losową. Jeśli ε_j ma rozkład normalny, to ξ_j ma rozkład logarytmiczno-normalny, $j = 1, 2, \ldots, m$.

Należy podkreślić, że nakładanie szumu może być stosowane tylko do danych wyrażonych na skali różnicowej lub ilorazowej. Jedynie bowiem w tym przypad-

[12] W tym kontekście w języku polskim określenia „multyplikatywny" i „multiplikatywny" są poprawne (NFJP, b.d.; SJP, b.d.; PWN, 1968, s. 407), choć forma „multyplikatywny" uznawana jest dzisiaj za nieco przestarzałą. Jednak ze względu na poszanowanie tradycji językowej używamy właśnie tego ostatniego przymiotnika.

ku dopuszczalne jest dodawanie. Mnożenie zaś jest możliwe wyłącznie dla danych wyrażonych na skali ilorazowej. Oznacza to, że szum multyplikatywny ma sens jedynie w odniesieniu do tego rodzaju informacji statystycznych. Powinny to też być dane ilościowe. Pewna „furtka" w zakresie nakładania szumu na dane jakościowe otwiera się wszakże w wypadku oparcia „zaszumienia" danych na transformacjach nieliniowych i podanym wyżej związanym z nimi algorytmie. Dzięki posługiwaniu się tam *de facto* rozkładami empirycznymi zamiast zwykłymi danymi można w takim wypadku do pewnego stopnia nałożyć szum i na dane tego rodzaju. Ciekawą dyskusję na temat efektywności stosowania różnego rodzaju szumu przeprowadził m.in. Mivule (2012).

Przykład 3.5 ilustruje efekty nakładania szumu.

Przykład 3.5. Załóżmy, że w pewnej bazie zgromadzono mikrodane na temat wynagrodzeń miesięcznych wypłaconych pracownikom firm określonej branży działającym na danym obszarze (tabl. 3.11) w danym miesiącu. Są to wielkości wyrażone w złotych, na skali ilorazowej, z dwoma miejscami po przecinku (grosze). Zastosowano następujące metody nakładania szumu:

- szum addytywny: zakłócenia wygenerowane z rozkładu normalnego $N(10, 4)$,
- szum addytywny z transformacją liniową:

$$0{,}8 \cdot FAKTYCZNE + 0{,}8 \cdot \varepsilon + 0{,}2 \cdot \overline{FAKTYCZNE},$$

  gdzie ε to zakłócenie dla danego rekordu wygenerowane ze standaryzowanego rozkładu normalnego ($N(0, 1)$), a $\overline{FAKTYCZNE}$ – średnia wielkości faktycznych,
- szum multyplikatywny: zakłócenia wygenerowane z rozkładu jednostajnego na w przedziale $[0{,}5, 1{,}5]$,
- szum multyplikatywny w wersji logarytmicznej – zakłócenia wygenerowane z rozkładu normalnego $N(0, 2)$.

Parametry rozkładów, z których generowano zakłócenia, oraz parametry przekształceń dobrano tak, aby zakłócenia mieściły się w pewnych rozsądnych granicach odchyleń od wielkości prawdziwej. Wychodząc zaś z założenia, że dąży się do jak najmniejszego zakłócenia wyniku zagregowanego, w tablicy 3.11. podano także sumy wypłaconych wynagrodzeń.

Najlepszy pod względem odchylenia sum „zaszumionych" danych od sumy danych faktycznych jest oczywiście wariant szumu addytywnego z transformacją liniową – wynika to z jego definicji. Relatywnie niezbyt duże odchylenia tego rodzaju cechują także pozostałe warianty szumu. Należy jednak pamiętać o tym, że precyzja ta zależy od przyjętych parametrów szumowania. Na przykład, gdyby w wypadku szumu multyplikatywnego przyjąć inny zakres rozkładu jednostajnego, albo gdyby w wariancie logarytmicznym wariancję rozkładu normalnego ustalić na jeden, to różnice sum byłyby o wiele większe.

Tabl. 3.11. Przykład nakładania szumu

ID	FAKTYCZNE	SZUM_ADD	SZUM_ADD_L	SZUM_MUL	SZUM_MUL_LG
1	2852,34	2863,40	2977,01	3907,99	2998,57
2	3927,55	3941,85	3837,82	5720,17	3905,51
3	2258,33	2271,60	2501,59	2675,50	1500,89
4	2594,17	2601,96	2769,95	1635,04	2682,93
5	3263,22	3279,38	3305,58	4871,22	2633,73
6	2965,84	2970,90	3067,90	2162,44	3243,32
7	3552,11	3561,54	3535,88	3495,46	3894,26
8	3147,53	3161,70	3213,75	4107,24	3600,04
9	2074,88	2085,14	2354,60	2662,60	1132,27
10	3475,12	3490,02	3473,68	1960,26	4207,18
11	4021,44	4030,85	3910,24	4582,91	3770,02
12	2384,65	2395,81	2602,83	2954,78	2569,40
13	5213,51	5218,91	4866,09	7640,35	7390,28
14	4100,89	4108,96	3976,18	3542,23	4430,25
15	2003,77	2011,96	2296,85	2359,29	1852,06
16	2551,28	2562,38	2736,52	2083,51	2032,44
17	3571,19	3580,29	3551,19	4005,72	4021,18
18	3128,73	3139,60	3196,69	2015,42	2663,83
19	4215,39	4223,71	4066,58	4611,57	3170,60
20	4521,76	4532,75	4312,80	6472,40	4196,99
21	4410,82	4424,45	4221,71	3642,11	2725,61
22	6017,94	6033,88	5508,64	4190,97	6518,99
23	4056,83	4073,48	3940,13	5217,29	3972,35
24	2204,59	2207,86	2457,47	3230,16	1620,81
25	4315,27	4322,23	4147,48	4068,01	5065,52
SUMA	**86829,15**	**87094,61**	**86829,17**	**93814,64**	**85799,03**

Objaśnienia: ID – identyfikator osoby, FAKTYCZNE – dane faktycznie zebrane, SZUM_ADD – dane po nałożeniu szumu addytywnego, SZUM_ADD_L – dane po nałożeniu szumu addytywnego z transformacją liniową, SZUM_MUL – dane po nałożeniu szumu multyplikatywnego, SZUM_MUL_LG – dane po nałożeniu szumu multiplikatywnego w wersji logarytmicznej.

Źródło: Opracowano z wykorzystaniem możliwości środowiska SAS Enterprise Guide 4.3. Dane fikcyjne.

Ze względu na wspomniane wyżej własności poszczególnych metod nakładania szumu oraz możliwy zakres ich parametrów, jak również formę danych, w praktyce konieczne jest efektywne dobranie metody wstawiania szumu. Można tego dokonać na przykład symulacyjnie: ustalić kilka wariantów, które w najbardziej prawdopodobny sposób nadawałyby się do wykorzystania w danych okolicznościach, a następnie, losując wielokrotnie i niezależnie próby o tej samej liczebności z posiadanej bazy, weryfikować odchylenie wartości sumarycznych zakłóconych zmiennych od odpowiednich wartości faktycznych. Uśrednione wyniki takich pomiarów dla wszystkich wylosowanych prób uwidocznią nam, która metoda jest w danej w sytuacji najlepsza.

Interesującym podejściem zakłócającym stosowanym do danych ilościowych jest także **mikroagregacja** (ang. *microaggregation*). Obejmuje ona w istocie pewną rodzinę narzędzi zapewniających ochronę poufności danych w ujęciu makro poprzez odpowiednie działania na poziomie mikro. U podstaw stosowania mikroagregacji leżą fundamentalne reguły publikacyjne, dopuszczające publikowanie zbiorów mikrodanych, gdy zawierają one co najmniej k rekordów, a żaden z nich nie dominuje pod danym względem (tzn. jego udział w danej wielkości ogółem dla grupy nie jest większy niż p%). Parametry naturalne k i p ($0 < p < 100$) są zazwyczaj arbitralnie ustalone. Na przykład w polskiej statystyce publicznej obowiązują (co nawet – jak wskazano w części 1.5.1 – zapisano w ustawie o statystyce publicznej) progi $k = 3$ i $p = 75$.

Bezpośrednie zastosowanie tych reguł prowadzi do zastąpienia przed publikacją wartości indywidualnych odpowiednimi wartościami sumarycznymi wyznaczonymi dla niewielkich poziomów agregacji (stąd właśnie nazwa *mikroagregacja*). W najogólniejszym ujęciu mikroagregacja polega na tym, że n rekordów zawartych w bazie mikrodanych jest łączonych w g grup, z których każda liczy co najmniej k elementów (n, g i k są liczbami naturalnymi, g, $k < n$). Grupy te są wyznaczane z uwzględnieniem kryteriów możliwie najlepszej homogeniczności (czyli wewnętrznej jednorodności) oraz heterogeniczności (jak największych różnic między grupami). Następnie dla każdej z tych grup wyznacza się średnią wartość określonej zmiennej i zastępuje się nią oryginalne wartości. Tak zmodyfikowane dane mogą zostać opublikowane.

Wyróżnia się kilka wariantów mikroagregacji. Biorąc pod uwagę kwestię liczebności grup, możemy wskazać opcje, gdy liczebność ta jest ustalona arbitralnie oraz gdy zostaje ona ukształtowana endogenicznie w wyniku odpowiedniego algorytmu grupującego.

Załóżmy zatem najpierw, że liczba elementów w każdej grupie jest ustalona jako k. Najprostszy mechanizm podziału zbioru rekordów na grupy polega wówczas na posortowaniu wektorów danych obrazujących rekordy w kolejności rosnącej lub malejącej według ustalonego kryterium. Jeśli mamy do czynienia z jedną zmienną grupującą, to sortowanie będzie się odbywać – rzecz jasna – według jej

wartości. W wypadku większej liczby takich zmiennych można się oprzeć w tym kontekście np. na malejącym porządku leksykograficznym i określonej ważności poszczególnych zmiennych: najpierw porządkujemy wektor według wartości najważniejszej zmiennej, a gdy jej wartości dla dwu lub więcej wektorów są jednakowe, ustawiamy je według wartości zmiennej drugiej co do ważności i tak dalej[13]. Następnie dzielimy ten ciąg wartości kolejno na k-elementowe podzbiory. Jeśli n nie jest wielokrotnością k, to ostatnią grupę tworzymy z ostatnich $n - (c + 1)$ k rekordów, gdzie c jest liczbą naturalną taką, że $ck < n < (c + 1)k$. W tej grupie znajdzie się zatem więcej niż k elementów. Dla każdej grupy obliczane są średnie wartości zmiennych z danymi wrażliwymi – i tymi średnimi zastępuje się oryginalne wartości. Podejście jednowymiarowe w tym ujęciu może być stosowane przede wszystkim wówczas, gdy wszystkie zmienne są wysoko ze sobą skorelowane. W przeciwnym razie lepiej jest przeprowadzać odrębne grupowania dla każdej zmiennej z osobna.

Liczba elementów należących do każdej grupy może też – jako się rzekło – być ustalana przez odpowiedni algorytm grupujący, z zachowaniem wszakże warunku, aby żadna z tych grup nie liczyła mniej niż k elementów. Przydatne okazują się tutaj metody ograniczonej analizy skupień. Na przykład Mateo-Sanz i Domingo-Ferrer (1998) proponują tzw. metodę k-Warda, opartą na algorytmie 1 (i przy założeniu, że $n \geq 2k$):

Algorytm 1. Metoda k-Warda

1. Tworzymy grupę z k pierwszych (najmniejszych według przyjętego porządku) rekordów oraz drugą grupę z k ostatnich (największych) rekordów. Wyznaczamy dla nich średnie arytmetyczne ich elementów, które będą je reprezentowały w kolejnych krokach.
2. Za pomocą metody Warda grupujemy rekordy w skupienia zawierające nie mniej niż k elementów, przy czym na żadnym etapie aglomeracji nie łączymy dwóch skupień, z których każde liczy co najmniej k elementów.
3. Dla każdej grupy z finalnie uzyskanego podziału zawierającej $2k$ lub więcej elementów stosujemy kroki 1 i 2. Algorytm kończy się wówczas, gdy uzyskamy skupienia, których liczebności są większe lub równe k, ale jednocześnie mniejsze niż $2k$.

Otrzymany podział jest najlepszy, jako że nie istnieje podział, który byłby lepszy – w tym sensie, że żaden inny podział na skupienia co najmniej k-elementowe nie składa się z grup, z których każda byłaby zawarta w jakiejś grupie podziału otrzymanego w wyniku zastosowania algorytmu 1.

[13] Alternatywnie można w tym kontekście użyć pierwszej – tzn. o największym udziale w zasobie wspólnej zmienności zmiennych – głównej składowej (uzyskanej w wyniku analizy głównych składowych (Jolliffe; 2002; Panek i Zwierzchowski, 2013) bądź też miernika kompleksowego (Młodak, 2006).

Przypomnijmy, że metoda Warda to narzędzie hierarchicznego aglomeracyjnego grupowania obiektów (Ward, 1963). Opiera się ona na algorytmie, którego przebieg jest następujący (Młodak, 2006). Dla $u = 1$ traktujemy obiekty jako skupienia jednoelementowe. Na każdym poziomie $u = 2, 3, \ldots$ łączymy dwa obiekty rzędu $u - 1$ o najmniejszej znormalizowanej odległości euklidesowej między środkami ciężkości skupień, aż do wyczerpania zbioru obiektów. Młodak (2006) podaje też pozycyjny wariant tej metody. Zasady grupowania są identyczne, zmienia się tylko formuła odległości międzyskupieniowej, w której wykorzystywana jest mediana.

Jednakże technika sortowania oparta na podejściu liniowym jest nieco toporna, pomija kontekst wzajemnych powiązań między zjawiskami, a sortowanie danych indywidualnych zwiększa ryzyko ich ujawnienia. Dlatego też Mateo-Sanz i Domingo-Ferrer (1998) zaproponowali pewną rodzinę metod mikroagregacji opartą na tworzeniu grup k-elementowych bez sprowadzania danych wielowymiarowych do jednego wymiaru. Zamiast tego stosuje się odległość wielowymiarową, traktując rekord (lub jego część) jako wektor w przestrzeni wielowymiarowej. Algorytm 2 uwidacznia tę koncepcję.

Algorytm 2. Wielowymiarowa mikroagregacja

1. Tworzymy jedną grupę z k „pierwszych" wektorów danych i inną grupę z k „ostatnich" wektorów danych.
2. Jeżeli poza dwiema grupami utworzonymi w kroku 1 pozostaje co najmniej $2k$ wektorów danych, powtarzamy krok 1, biorąc jako zbiór danych poprzednio rozważane rekordy z wyłączeniem tych zaklasyfikowanych już uprzednio do określonych grup; w przeciwnym razie przechodzimy do kroku 3.
3. Jeżeli liczba wyjściowych wektorów nienależących do żadnej grupy utworzonych w kroku 1 wynosi od k do $2k - 1$, to tworzymy nową grupę z wszystkich tych elementów i kończymy algorytm.
4. Jeżeli poza dwiema grupami utworzonymi w kroku 1 pozostaje mniej niż k wektorów, dodajemy takie wektory do najbliższej z grup utworzonych w kroku 1.

Jak już wspomniano, wektory „pierwsze" i „ostatnie" są wyznaczane na podstawie odległości wielowymiarowej. Rozpoczyna się to od wskazania dwóch wektorów ekstremalnych, czyli takich, że ich odległość jest największa spośród odległości między wszystkimi parami wektorów. Następnie dla każdego z owych ekstremalnych wektorów bierzemy $k - 1$ wektorów najbliższych niego. W ten sposób otrzymujemy grupy wektorów „pierwszych" i „ostatnich". Kryterium wykorzystywane do wyodrębnienia tych grup zwie się **kryterium największej odległości** (ang. *maximum-distance* (*MD*) *criterion*). Wynik tego grupowania zależy

w pewnej mierze od tego, który z dwóch ekstremalnie odległych wektorów uznany zostanie za „pierwszy", a który za „ostatni" (brak tu jednoznacznego kryterium). Jednak różnice pomiędzy tymi opcjami w kształcie finalnego grupowania uzyskanego z zastosowaniem algorytmu 2 nie są zazwyczaj nazbyt znaczące. Podejście to ma charakter heurystyczny, z ustaloną liczebnością grup. Jeśli zaś chcemy zastosować zmienną liczebność grup w ujęciu wielowymiarowym, stosujemy algorytm 1, przy czym w kroku 1 kwalifikację wektorów do „pierwszych" i „ostatnich" przeprowadzamy nie według porządku, a według metody maksymalnej odległości. Inne ciekawe podejścia w tym zakresie zaprezentowali i porównali m.in. Solanas i Martinez-Balleste (2006).

Podejściu heurystycznemu przeciwstawia się czasem podejście optymalizacyjne. Finalny podział na skupienia otrzymuje się tutaj, minimalizując względem owego podziału na skupienia C_1, C_2, …, C_g sumę kwadratów błędów wewnątrzgrupowych:

$$SSE = \sum_{j=1}^{g} \sum_{i \in C_j} (x_{ij} - \overline{x}_j)^2,$$

gdzie x_{ij} to obserwacja zmiennej grupującej dla obiektu i-tego w skupieniu C_j, $\overline{x}_j$ – średnia wartości tejże zmiennej dla obiektów skupienia C_j, tzn. $\overline{x}_j = \sum_{i \in C_j} x_{ij} / n_j$, przy czym n_j oznacza liczebność skupienia C_j, $j = 1, 2, \ldots, g$. Przykład tej metody podali Hansen i Mukherjee (2003). Oczywiście, optymalizując podział na skupienia, należy pamiętać o respektowaniu granic liczebności skupień (co najmniej k i nie więcej niż $2k - 1$).

W ostatnich latach próbuje się stosować mikroagregację również do danych jakościowych (wyrażonych na skali nominalnej lub porządkowej). Grupowanie odbywa się na zasadzie algorytmu k-mód (ang. *k-modes algorithm*). Opiera się on na podejściu k-średnich. Oznacza to, że przyjmuje się wstępny podział danej zbiorowości na ustaloną liczbę skupień, po czym w kolejnych krokach przesuwa się obiekty pomiędzy skupieniami, tak aby każdy obiekt należał do najbliższego skupienia. Algorytm kończy się, gdy między skupieniami nie można zrobić już żadnego przesunięcia. Specyfika tego podejścia w odniesieniu do danych jakościowych polega na tym, że nie stosuje się średniej – a zatem środków ciężkości skupień (na podstawie odległości obrazujących dystans obiektów od odpowiednich skupień). W wypadku skali porządkowej używa się bowiem mediany, natomiast jeśli dane są wyrażone na skali nominalnej – reguły większości. Stosowane są także specyficzne formuły odległości (Sharma i Gaud, 2015). Tym samym wrażliwe dane zastępuje się odpowiednio medianą lub dominantą dla odpowiedniej grupy.

Przykład 3.6 ukazuje efekty przeprowadzonej mikroagregacji.

Przykład 3.6. Załóżmy, że w wyniku pewnego badania zgromadzono dane o 25 pracujących osobach. Dane te dotyczą: miesięcznego wynagrodzenia brutto, stażu pracy oraz odległości od miejsca zamieszkania do miejsca pracy. Są to dane bardzo wrażliwe, a zatem konieczna jest mikroagregacja. Dokonano jej za pomocą algorytmu 2, z tym, że w kroku 1 tworzenie grupy „pierwszych" i „ostatnich" rekordów rozpoczyna się od wyznaczenia średniego rekordu (tzn. sztucznego rekordu składającego się ze średnich arytmetycznych badanych zmiennych), a następnie wyznaczenia rekordu x_a najbardziej od owego „średniego wektora" odległego. Następnie znajduje się rekord x_b, najbardziej odległy od x_a. Po $k-1$ rekordów najbliższych x_a oraz x_b wraz z nimi odpowiednio tworzy skrajne skupienia.

Krok 2 wykonujemy wówczas, gdy poza skupieniami utworzonymi w kroku 1 pozostaje co najmniej $3k$ rekordów, krok 3 – jeżeli liczba niesklasyfikowanych rekordów wynosi od $2k$ do $3k-1$ (wtedy rekordy takie dzielimy na dwie grupy: zawierającą $k-1$ rekordów najbliższych rekordu najbardziej odległego od średniego w tym wypadku oraz składającą się z pozostałych rekordów), krok 4 zaś – gdy bez przynależności grupowej pozostaje mniej niż $2k$ rekordów. Takie podejście[14] zastosowano w programie μ-Argus, który tutaj wykorzystano (szerzej opisano go w części 5.2). Tablica 3.12 ukazuje oryginalną postać danych oraz danych po mikroagregacji. Minimalną liczebność grupy ustalono na 3 rekordy.

Tabl. 3.12. Dane poddane mikroagregacji i jej efekty

ID	WYNAGR_O	STAZ_O	ODL_O	WYNAGR_M	STAZ_M	ODL_M
1	2852,34	5	2	2278,72	6	6
2	3927,55	7	5	4044,80	10	3
3	2258,33	4	8	2278,72	6	6
4	2594,17	8	1	3077,64	13	3
5	3263,22	10	4	4044,80	10	3
6	2965,84	11	3	3077,64	13	3
7	3552,11	3	7	3077,64	13	3
8	3147,53	19	2	3077,64	13	3
9	2074,88	9	9	2278,72	6	6
10	3475,12	6	12	3207,76	13	12
11	4021,44	21	3	4756,66	22	4
12	2384,65	15	10	3207,76	13	12

[14] Jest ono nazywane *algorytmem* MDAV (wielowymiarowej mikroagregacji opartej na maksymalnej odległości od wektora średnich, ang. *multivariate microaggregation based on maximum distance to average vector*).

ID	WYNAGR_O	STAZ_O	ODL_O	WYNAGR_M	STAZ_M	ODL_M
13	5213,51	17	2	4756,66	22	4
14	4100,89	8	3	4044,80	10	3
15	2003,77	4	5	2278,72	6	6
16	2551,28	12	15	3207,76	13	12
17	3571,19	16	11	3207,76	13	12
18	3128,73	23	1	3077,64	13	3
19	4215,39	28	8	4756,66	22	4
20	4521,76	13	3	4044,80	10	3
21	4410,82	10	1	4044,80	10	3
22	6017,94	25	6	4756,66	22	4
23	4056,83	17	10	3207,76	13	12
24	2204,59	8	4	2278,72	6	6
25	4315,27	18	2	4756,66	22	4

Objaśnienia: WYNAGR – wynagrodzenie miesięczne brutto w złotych, STAZ – staż pracy w latach, ODL – odległość od miejsca zamieszkania do miejsca pracy w kilometrach (xxxx_O – wartości oryginalne zmiennej xxxx przed mikroagregacją, xxxx_M – wartości zmiennej xxxx po mikroagregacji).

Źródło: Opracowano z wykorzystaniem programu μ-Argus, wersja 5.1. Dane fikcyjne.

Warto nadmienić, że istnieje też wersja mikroagregacji umożliwiająca jej przeprowadzenie również dla zmiennych jakościowych. Wykorzystuje się tutaj tzw. odległość Gowera, którą można wyznaczać dla różnych skal pomiarowych. Zakłócenia mogą być w tym wypadku dobierane losowo z prawdopodobieństwami odpowiadającymi występowaniu określonych kategorii wśród najbliższych – w myśl odległości Gowera – sąsiadów danego rekordu. Alternatywnie za zaburzoną wartość można też tutaj przyjąć poziom zmiennej kategorialnej z największą liczbą wystąpień albo losowy, jeśli owo maksimum wystąpień jest osiągalne dla więcej niż jednej kategorii. Pietrzak i in. (2022) pokazali, jak metodę tę – w powiązaniu z celowaną wymianą rekordów, postrandomizacją (PRAM) i nakładaniem szumu – można zastosować do mikrodanych z badania wypadków przy pracy. Koncepcję odległości Gowera (z pewnymi modyfikacjami) wykorzystamy też w rozdziale 4 do konstrukcji miar straty informacji.

Kolejny sposób zaliczany do kategorii maskowania zakłóceniowego to **wymiana danych** (ang. *data swapping*). Polega na przekształceniu bazy danych poprzez zamianę między rekordami objętych ochroną wartości danej zmiennej. Czyni się to w taki sposób, aby zachować odpowiednie wielkości agregatowe lub częstości dla poziomów wyższego rzędu. Historię rozwoju tego podejścia przedstawili m.in. Hundepool i in. (2006). Nadmieńmy, że wymiana danych może zostać zastosowa-

na zarówno do informacji statystycznych wyrażonych w formie jakościowej, jak też do zmiennych ilościowych (Reiss i in., 1982; Reiss, 1984).

Najbardziej znanym wariantem wymiany danych jest **wymiana rang** (ang. *rank swapping*). Można ją stosować do danych wyrażonych zarówno na skali porządkowej, jak i na skalach silniejszych. Przebiega ona następująco. Załóżmy, że ochronie poddajemy zmienną X. Wartości tej zmiennej porządkujemy zatem w kolejności rosnącej. Następnie każda zrangowana w ten sposób wartość zmiennej X jest zamieniana z inną wartością losowo wybraną spośród tych wartości, których rangi zawierają się w pewnym ograniczonym przedziale – na przykład spośród tych, których rangi nie różnią się od rangi danej wartości więcej niż o p% całkowitej liczby rekordów, gdzie $p \in (0, 100)$ jest ustalonym parametrem. Pewna zaleta tego algorytmu polega na tym, że może on być stosowany niezależnie dla każdej zmiennej z osobna. Również skala nałożonych zakłóceń jest tu mniejsza. Do ułomności owego podejścia można wszakże zaliczyć przede wszystkim pomijanie kwestii bezwzględnych różnic między wartościami danej zmiennej (co nie zawsze znajduje odzwierciedlenie w rangach) oraz nieuwzględnianie powiązań między zmiennymi. Przykład 3.7 egzemplifikuje zastosowanie wymiany rangowej. Warto nadmienić, że w niektórych programach komputerowych (np. w pakiecie `sdcMicro` środowiska R) dostępny jest szczególny wariant tejże metody, w którym istnieje możliwość przeprowadzenia grupowania ustalonego odsetka najmniejszych lub najmniejszych wartości przed wymianą. Wpływa to istotnie na ochronę wartości w ogonach rozkładu, w tym wartości odstających.

Przykład 3.7. Załóżmy, że posługujemy się danymi uwidocznionymi w tablicy 3.12, z tym że uzupełniono je dodatkowo o zmienną jakościową – stan cywilny prawny (według kategorii wskazanych w przykładzie 3.4). Wyniki wymiany rangowej ukazuje tablica 3.13. Parametr p ustalono tutaj na 20, co oznacza, że wymianę prowadzimy w obrębie wartości, których rangi nie są odległe bardziej niż o 5. Wymianę przeprowadzono, opierając się na zmiennych ilościowych (czyli WYNAGR_O, STAZ_O i ODL_O).

Tabl. 3.13. Rezultaty wymiany rangowej

ID	STANC_O	WYNAGR_O	STAZ_O	ODL_O	STANC_R	WYNAGR_R	STAZ_R	ODL_R
1	2	2852,34	5	2	2	2258,33	8	3
2	2	3927,55	7	5	2	3263,22	3	3
3	1	2258,33	4	8	1	2852,34	8	10
4	2	2594,17	8	1	2	2074,88	4	2

ID	STANC_O	WYNAGR_O	STAZ_O	ODL_O	STANC_R	WYNAGR_R	STAZ_R	ODL_R
5	3	3263,22	10	4	4	3927,55	11	6
6	1	2965,84	11	3	1	3552,11	10	4
7	2	3552,11	3	7	1	2965,84	7	10
8	1	3147,53	19	2	1	3475,12	17	1
9	1	2074,88	9	9	1	2594,17	13	12
10	1	3475,12	6	12	1	3147,53	4	9
11	2	4021,44	21	3	2	4410,82	16	5
12	4	2384,65	15	10	4	2204,59	10	8
13	2	5213,51	17	2	2	4215,39	19	1
14	3	4100,89	8	3	4	4315,27	12	5
15	4	2003,77	4	5	2	2551,28	6	3
16	4	2551,28	12	15	3	2003,77	8	15
17	1	3571,19	16	11	1	3128,73	21	8
18	2	3128,73	23	1	4	3571,19	17	2
19	1	4215,39	28	8	2	5213,51	28	11
20	3	4521,76	13	3	3	4056,83	9	2
21	2	4410,82	10	1	2	4021,44	15	2
22	2	6017,94	25	6	2	6017,94	18	4
23	2	4056,83	17	10	2	4521,76	23	7
24	4	2204,59	8	4	3	2384,65	5	3
25	3	4315,27	18	2	3	4100,89	25	1

Objaśnienia: STANC – stan cywilny prawny (klasyfikacja jak w przykładzie 3.4, xxxx_O – wartości oryginalne zmiennej xxxx przed wymianą rangową, xxxx_R – wartości zmiennej xxxx po wymianie rangowej. Pozostałe zmienne jak w tablicy 3.12.

Źródło: Opracowano z wykorzystaniem programu μ-Argus, wersja 5.1. Wykorzystano też dane z tablicy 3.12. Dane fikcyjne.

Widać zatem, że wymiana rangowa jest skuteczna, jednak bardzo sporadycznie może się zdarzyć, że wielkości oryginalne pozostaną niezmienione (tutaj tak zdarzyło się w przypadku osoby o ID = 22).

Wymianę danych można łączyć z innym rodzajami zakłócania. Na przykład Calviño (2017) pokazał, że w kontroli ujawniania danych da się efektywnie wykorzystywać główne składowe, dokonując ich zamiany, a następnie nakładając szum.

Szczególny przypadek wymiany danych stanowi **celowana wymiana rekordów** (ang. *targeted record swapping* – TRS). Stosowanie TRS jest zalecane przez Eurostat. W pierwszym rzędzie rekomendacja ta dotyczy mikrodanych ze spisów powszechnych, które stanowią podstawę naliczania tablic spisowych. Główne cechy charakterystyczne TRS są następujące:

- w TRS wymienia się predefiniowane grupy rekordów; na przykład, jeśli rekordy dotyczą osób, a ich grupy to gospodarstwa domowe, to wymiana następuje nie pomiędzy poszczególnymi osobami, lecz pomiędzy całymi gospodarstwami domowymi; grupami mogą być także na przykład klasy wielkościowe podmiotów gospodarczych czy specjalne strefy ekonomiczne (gdy rekordy dotyczą podmiotów gospodarczych),
- dla każdego poziomu hierarchii geograficznej wyznaczane są grupy rekordów o najwyższym ryzyku ujawnienia informacji poufnych,
- ryzyko to jest obliczane na podstawie kombinacji wartości zmiennych kluczowych (tzw. zmiennych ryzyka) na każdym poziomie hierarchii geograficznej dla każdego rekordu, z zastosowaniem zasady *k*-anonimowości; następnie ustala się ryzyko ujawnienia dla grupy rekordów jako maksymalną indywidualną wartość ryzyka ujawnienia dla rekordów ją tworzących (również na każdym szczeblu hierarchii),
- wymianie podlegają dane z zakresu zmiennych definiujących poziomy geograficzne/administracyjne (tzw. zmienne hierarchiczne) – wymiana owa jest dokonywana między grupami rekordów najbliższymi według odległości wyznaczanej na podstawie wartości określonych zmiennych (tzw. zmienne podobieństwa); zamiast zmiennych geograficznych mogą być też przyjęte inne zmienne definiujące dokładnie określoną hierarchię klasyfikacyjną (np. szczeble Polskiej Klasyfikacji Działalności dla podmiotów gospodarczych),
- dla każdej grupy rekordów zagrożonej ryzykiem ujawnienia i dobranej do próby (prawdopodobieństwo wylosowania do próby jest ustalane na podstawie ryzyka ujawnienia) tworzona jest lista grup rekordów o identycznych wartościach pewnych charakterystyk (tzw. zmiennych podobieństwa), spośród których losowana jest jedna grupa rekordów – tzw. dawca,
- wymiana danych jest dokonywana w obrębie poszczególnych poziomów wskazanej wyżej hierarchii (np. geograficznej) na tym jej poziomie, na którym grupa rekordów zostanie uznana za zagrożoną identyfikacją; dawca jest dla niej losowany na tym samym poziomie hierarchii, lecz z innego obszaru/domeny,
- wymieniane są dane o grupach rekordów, dla których nie jest spełniona zasada *k*-anonimowości (choć planowane jest rozszerzenie metody TRS w ten sposób, by umożliwić ocenę ryzyka ujawnienia z użyciem innego podejścia),
- na najniższym poziomie hierarchicznym wymiana może być dokonana także między pewną dodatkową liczbą grup jednostek, dobraną w taki sposób,

aby ogółem wymienić dane między możliwie najmniejszą liczbą grup przy zachowaniu ustalonego odsetka minimalnej liczby wymian,
- dodatkowo, oprócz wymiany wartości zmiennych hierarchicznych pomiędzy zagrożoną grupą rekordów i jej dawcą, wymianie podlegać mogą również wartości innych zmiennych, które zadeklarowane zostaną jako tzw. zmienne przenoszone,
- zaletą wymiany całych grup rekordów jest m.in. zachowanie rozkładów brzegowych (dla jednostek geograficznych wysokiego poziomu) oraz zachowanie struktury grupy rekordów (nie powstaną niemożliwe do zaistnienia grupy rekordów); zaletą wskazania zmiennych podobieństwa jest zachowanie rozkładów brzegowych dla tych zmiennych oraz minimalizacja obciążenia zakłóconych danych; zaletą wskazania zmiennych przenoszonych jest zachowanie spójności we wszystkich zmiennych przestrzennych, wykluczona zostanie możliwość ich „wymieszania" (ich wartości również będą przenoszone).

Kolejne podejście zakłóceniowe polega na **zaokrąglaniu** (ang. *rounding*) wrażliwych wielkości. Oznacza to, że oryginalne wartości zastępuje się ich wersjami zaokrąglonymi. Dla danej zmiennej wartość zaokrąglenia jest wybrana zazwyczaj ze zbioru punktów zaokrągleń definiującego zestaw zaokrągleń. Najpopularniejszym rozwiązaniem w tym zakresie jest przyjęcie za punkty zaokrągleń $p_i = b \cdot i$, dla $i = 1, 2, \ldots, l$, gdzie liczba naturalna b jest podstawą zaokrąglania, a l to maksymalny zakres zaokrągleń. W ten sposób, jeśli zmienna X przyjmuje wartości nieujemne, to zbiory „przyciągania" zaokrągleń (ang. *sets of attraction*) – tzn. przedziały faktycznych wartości zmiennej X, które można zaokrąglić do wielkości p_i, $i = 1, 2, \ldots, l$ – są odpowiednio postaci: $[0, p_1 + (b/2)]$, $[p_i - (b/2), p_i + (b/2)]$, $i = 2, 3, \ldots, l - 1$, $[p_l - b/2, x_{\max}]$, gdzie $x_{\max}$ to maksymalna wartość zmiennej X. Tym samym każda oryginalna wartość zmiennej X zostaje zastąpiona stosownym zaokrągleniem odpowiadającym przedziałowi przyciągania, do którego ona należy. Na przykład, jeśli $b = 10$, a $[x_{\max}/10] = 200$, to $l = 200$ oraz $p_i = i \cdot 10$, $i = 1, 2, \ldots, 200$[15]. Wobec tego np. dla wielkości 1875,24 przedział przyciągania to $[188 \cdot 10 - (10/2), 188 \cdot 10 + (10/2)] = [1880 - 5, 1880 + 5] = [1875, 1885]$. Tym samym liczbę tę można zaokrąglić do 1880. Zaokrąglenie może mieć też formę kontrolowaną. W tym wypadku dana wielkość zostaje zaokrąglona do jednej z dwóch najbliższych wielokrotności podstawy b z odpowiednimi prawdopodobieństwami, tak aby oczekiwany błąd zaokrąglenia wynosił zero. Na przykład, jeśli reszta z dzielenia liczby b przez 10 wynosi k, to zaokrąglenie ma postać $10 \cdot [b/10]$ z prawdopodobieństwem $1 - (k/10)$ oraz $10 \cdot ([b/10] + 1)$ z prawdopodobieństwem $k/10$, $k \in (0, 10)$. Jeśli zatem chcemy – tak jak w poprzedniej egzemplifikacji – zaokrą-

[15] Zapis $[a]$ oznacza część całkowitą liczby rzeczywistej a, czyli największą liczbę całkowitą nie większą od a.

glić liczbę 1875,24 do wielokrotności 10, to zastąpimy ją liczbą 1870 z prawdopodobieństwem 1 – 0,524 = 0,476 lub liczbą 1880 z prawdopodobieństwem 0,524.

Warto nadmienić, że w wielowymiarowym zbiorze danych zaokrąglenie przeprowadza się na ogół jednowymiarowo, dla każdej zmiennej z osobna. Możliwe jest wszakże także zaokrąglanie wielowymiarowe – na przykład z wykorzystaniem diagramu Woronoja opartego na metryce euklidesowej lub inne (zob. np. Willenborg i de Waal (2001)).

Do ochrony mikrodanych stosuje się też czasem **próbkowanie wtórne** (ang. *resampling*). Polega ono na tym, że z obserwacji danej zmiennej X losujemy niezależnie p próbek $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_p$, z których każda ma rozmiar n (przypomnijmy, że n to liczba rekordów). Stąd $\boldsymbol{s}_l = (x_{li_1}, x_{li_2}, \ldots, x_{li_n})$, $i_k \in \{1, 2, \ldots, n\}$, $k = 1, 2, \ldots, n$, $l = 1, 2, \ldots, p$. Następnie wartości należące do każdej z tych próbek sortujemy w ten sam sposób (np. w kolejności niemalejącej) i otrzymujemy posortowane próbki $\boldsymbol{s}_{(l)} = (x_{li_{(1)}}, x_{li_{(2)}}, \ldots, x_{li_{(n)}})$, $l = 1, 2, \ldots, p$. Obserwacje zmiennej X zastępujemy finalnie średnimi arytmetycznymi wartości odpowiedniej rangi z tych próbek. Oznacza to, że w miejsce wartości x_k wstawiamy $\overline{x}_{pk} = \sum_{l=1}^{p}(x_{li_{(k)}} / p)$, $k = 1, 2, \ldots, n$. Oczywiście podejście to można stosować tylko dla danych wyrażonych na skali różnicowej bądź ilorazowej. Zgodnie z bootstrapowym centralnym twierdzeniem granicznym (Bickel i Freedman, 1981), dla dostatecznie dużych n odchylenia wielkości dla populacji uzyskane ze skorygowanych danych od odpowiednich wielkości oryginalnych są nieznaczne.

Wśród szerokiego wachlarza narzędzi kontroli ujawniania mikrodanych jakościowych poczesne miejsce zajmuje **metoda postrandomizacyjna** (ang. *the post–randomization method* – PRAM). Jak jej nazwa wskazuje, jest to metoda probabilistyczna, generująca określone zakłócenia. Najogólniej rzecz ujmując, wartości zmiennych jakościowych dla pewnych rekordów zostają tutaj zamienione na inne z wykorzystaniem specyficznego mechanizmu probabilistycznego, a konkretnie – macierzy przejść Markowa. Dzięki takiemu podejściu PRAM łączy w sobie dodawanie szumu, ukrywanie danych oraz przekodowywanie.

Formalnie metodę tę można opisać następująco. Załóżmy, że oryginalnie zebrano dane o zmiennej X. Oznaczmy przez X^* tę samą zmienną, ale z przekształconymi, zakłóconymi wartościami. Przyjmijmy też, że zmienne X i X^* są zmiennymi jakościowymi, obejmującymi k kategorii (gdzie k to liczba naturalna). Określmy macierz przejścia między kategoriami j i l jako $p_{jl} \overset{\text{def}}{=} P\{X^* = l \mid X = j\}$. Zatem $p_{jl} \in [0, 1]$ jest prawdopodobieństwem tego, że oryginalna kategoria j zostanie zamieniona na kategorię l. Prawdopodobieństwa p_{jl}, $j, l = 1, 2, \ldots, k$, tworzą macierz o rozmiarze $k \times k$, która jest właśnie macierzą przejść Markowa. Tym samym każda kategoria j zmiennej X zostanie zastąpiona kategorią l wylosowaną ze zbioru $\{1, 2, \ldots, k\}$ z prawdopodobieństwem p_{jl}. Mamy wobec tego $\sum_{l=1}^{k} p_{jl} = 1$ dla

każdego $j = 1, 2, \ldots, k$. Oczywiście p_{jj} oznacza prawdopodobieństwo zdarzenia polegającego na tym, że kategoria j pozostanie bez zmian, $j = 1, 2, \ldots, k$.

Efektywność metody PRAM, wyrażona ryzykiem ujawienia danych wrażliwych i stratą informacji, zależy w dużym stopniu od wyboru macierzy przejścia Markowa. Można też ograniczyć zakres możliwych zmian przez wskazanie pewnej liczby najbliższych kategorii, w obrębie których możliwe jest dokonanie zmian. Jak wspomniano, PRAM można stosować wyłącznie do danych jakościowych. Więcej na temat PRAM pisali np. de Wolf i in. (1999) oraz Hundepool i in. (2012). Woo i Slavković (2014) badali możliwości uzyskania nieobciążonych oszacowań parametrów modeli regresyjnych w sytuacji zastosowania PRAM przy użyciu odpowiedniej postaci algorytmu największej wiarygodności. Z kolei Nayak i Adeshiyan (2016) rozważali niezmienny wariant metody PRAM, pozwalający na uzyskanie nieobciążonych oszacowań określonych wielkości kosztem zwiększenia wariancji.

Przykład 3.8 uwidacznia, jak można zastosować PRAM w praktyce.

Przykład 3.8. Zastosujemy metodę PRAM do danych jakościowych wskazanych w przykładzie 3.4. (tabl. 3.2). Ustalono następujące prawdopodobieństwa pozostawienia danej kategorii bez zmian:

- płeć: M – 0,8, K – 0,8,
- stan cywilny prawny: 1 – 0,5, 2 – 0,8, 3 – 0,5 (kategorie 4 i 9 nie wystąpiły),
- wykształcenie: 1 – 0,8, 2 – 0,7, 3 – 0,6, 4 – 0,6, 5 – 0,6, 6 – 0,6, 7 – 0,7, 9 – 0,8, 11 – 0,8, 12 – 0,5 (kategorie: 8, 10, 13, 14 i 99 nie wystąpiły), tutaj też ograniczono możliwe zmiany do trzech najbliższych kategorii,
- status na rynku pracy: 1 – 0,7, 2 – 0,7, 3 – 0,8 (kategoria 9 nie wystąpiła).

Tabl. 3.14. Przykład zastosowania metody PRAM

ID	PŁEĆ_O	STAN CYWILNY PRAWNY_O	WYKSZTAŁCENIE_O	STATUS NA RYNKU PRACY_O	PŁEĆ_P	STAN CYWILNY PRAWNY_P	WYKSZTAŁCENIE_P	STATUS NA RYNKU PRACY_P
1	M	1	6	3	M	2	7	3
2	M	1	6	3	K	3	6	3
3	K	3	5	2	K	2	5	2
4	K	1	11	1	K	2	11	1
5	K	1	9	1	K	1	9	1
6	M	3	1	2	M	3	1	3

ID	PŁEĆ_O	STAN CYWILNY PRAWNY_O	WYKSZTAŁCENIE_O	STATUS NA RYNKU PRACY_O	PŁEĆ_P	STAN CYWILNY PRAWNY_P	WYKSZTAŁCENIE_P	STATUS NA RYNKU PRACY_P
7	K	1	3	2	K	3	6	3
8	M	1	6	3	M	1	6	3
9	K	1	6	2	M	3	6	2
10	M	2	4	2	K	2	1	2
11	K	1	3	1	K	1	3	1
12	K	1	9	1	M	1	9	3
13	M	2	4	2	K	1	7	2
14	M	2	7	1	M	3	7	1
15	K	3	5	2	K	1	9	2
16	K	3	5	2	K	1	3	3
17	M	3	1	2	M	2	5	3
18	M	3	1	3	M	3	1	2
19	K	1	9	3	M	1	5	1
20	K	1	9	2	K	2	7	2
21	M	1	2	1	M	2	2	1
22	M	2	7	1	M	2	6	2
23	K	1	6	2	M	2	5	1
24	M	2	4	2	M	2	2	1
25	M	2	12	1	M	2	12	1
26	M	2	7	1	M	2	7	1
27	M	3	1	2	M	3	1	3

Objaśnienia: xxxx_O – wartości oryginalne zmiennej xxxx przed zastosowaniem metody PRAM, xxxx_P – wartości zmiennej xxxx po zastosowaniu metody PRAM. Pozostałe zmienna jak w tablicy 3.12.

Źródło: Opracowano wykorzystaniem programu μ-Argus, wersja 5.1. Wykorzystano też dane z tabl. 3.2. Dane są fikcyjne.

Ze względu na założenia stosowanego oprogramowania, prawdopodobieństwa zmiany można było ustalać tylko dla tych kategorii zmiennych, które faktycznie wystąpiły w bazie. Z uzyskanych wyników można wysnuć wniosek, że w wypadku zbiorów danych o niskiej liczebności ryzyko pozostawienia danych bez zmiany

informacji wrażliwej jest dość spore (chodzi tutaj np. o rekordy nr 4 i 25, które są jedynymi, gdzie poziom wykształcenia wynosi 11 i 12. W takim wypadku warto rozszerzyć zakres dopuszczalnych zamian.

Na zakończenie naszej prezentacji warto wspomnieć jeszcze o metodzie **MASSC**, będącej połączeniem czterech kroków: mikroaglomeracji, podstawiania, podpróbkowania i kalibracji (ang. *microagglomeration, substitution, subsampling and calibration*). Pozwala to na równoczesną kontrolę ryzyka ujawnienia i straty informacji z powodu zastosowania odpowiednich mechanizmów SDC. Metoda ta pojawiła się w pracy Singha i in. (2004).

Mikroaglomeracja polega tutaj na dokonaniu podziału zbioru rekordów na skupienia o podobnym ryzyku ujawnienia danych wrażliwych. W tym celu wykorzystuje się zmienne kluczowe odpowiednie z punktu widzenia tego celu, tzn. quasi-identyfikatory. Następnie dokonuje się probabilistycznego podstawienia w celu zakłócenia oryginalnych danych (np. poprzez generowanie szumu czy macierzy przejść Markowa – podobnie jak w podejściu PRAM). W dalszej kolejności za pomocą probabilistycznego losowania podpróbek ukrywa się pewne zmienne bądź nawet całe rekordy (ukrywanie odbywa się zatem z określonym prawdopodobieństwem). Finalnie wreszcie dokonuje się stosownej kalibracji wag użytych w losowaniu zasadniczej próby do badania w celu zachowania odpowiedniej jakości oszacowań dla zmiennych wyjściowych, których precyzja ma istotne znaczenie dla użytkowników danych.

Na podkreślenie zasługuje tutaj kwantyfikacja ryzyka ujawniania informacji wrażliwych i jego zastosowania w generowaniu zakłóceń maskujących. Jednakże – jak zauważają Hundepool i in. (2006) – ułomnością podejścia MASSC jest nadmierne uproszczenie rzeczywistości wskutek ograniczenia analizy ujawniania tylko do jego komponentów wynikających z powiązań kluczowych zmiennych ze źródłami zewnętrznymi. Ponieważ zmienne kluczowe są na ogół jakościowe, wyznaczając ryzyko ujawniania danych wrażliwych, ignoruje się tutaj to, że pewne ilościowe zmienne wynikowe także mogą zostać użyte do odtworzenia jednostkowych danych wrażliwych. Może mieć to znaczenie na przykład w wypadku danych gospodarczych. Tak więc metoda MASSC staje się najefektywniejsza przede wszystkim wtedy, gdy zmienne ilościowe nie występują wcale.

Oprócz wskazanych wyżej pozycji literaturowych ciekawy przegląd narzędzi SDC dla mikrodanych można znaleźć książce Templa i in. (2014). Szerokiego omówienia tych zagadnień wraz z przykładami praktycznych algorytmów napisanych w środowisku R i ich empirycznych realizacji dokonali również Benschop i in. (2022). Z kolei Domingo-Ferrer i in. (2001) przeprowadzili interesujące eksperymentalne porównanie efektywności metod SDC dla mikrodanych z punktu widzenia straty informacji i ryzyka ujawnienia danych wrażliwych. Okazało się, że najlepsze wyniki daje wymiana rang (jednak jest ryzykowna w wypadku baz

on-line z możliwością powtarzania zapytań, gdyż wtedy ryzyko natrafienia na rekord z identycznymi danymi jak rekord chroniony staje się znaczne), a w drugiej kolejności – mikroagregacja.

Obecnie przyjrzymy się, które metody zakłóceniowe można stosować w ochronie danych tabelarycznych.

W wypadku tablic, a szczególnie tablic częstości, może zachodzić potrzeba opublikowania wszystkich wartości zawartych w tego rodzaju tablicy. Ze względu na reguły poufności nie jest to jednak możliwe. Rozwiązaniem tego problemu jest publikowanie wartości zmienionych w specjalny, metodyczny sposób. Duncan i in. (2011) zauważyli, że nie jest to dezinformacja, gdyż odbiorca otrzymuje informację o tym, że ze względu na poufność musiała nastąpić niewielka korekta. Użyteczność zmienionych danych będzie jednak zależała od charakteru danych oraz od użytej metody. **Zaokrąglanie** polega na dodaniu niewielkiej – ale akceptowalnej – liczby do rzeczywistych danych. Zapewnia ono ochronę komórkom o niewielkiej wartości i wartościom komórek równych zero. Hundepool i in. (2012) zwrócili uwagę, że zaokrąglanie wartości komórek jest uważane za efektywną metodę ochrony poufności, szczególnie w sytuacji, w której wiele tablic powstaje na podstawie tego samego źródła danych jednostkowych. Problemem pozostaje sytuacja, w której użytkownik staje się kreatorem tablicy, tzn. ma dużą dowolność w jej tworzeniu, a tablica jest tworzona dynamicznie. Może to prowadzić do sytuacji, w której niektóre komórki zostaną zmienione w znaczący sposób. Co więcej, może to mieć wpływ na zwiększenie siły związku pomiędzy zmiennymi oraz częstościowej wariancji komórek. Do tej kategorii ochrony danych należą takie metody jak: zaokrąglanie standardowe, zaokrąglanie losowe oraz zaokrąglanie kontrolowane.

W wypadku **zaokrąglania standardowego** (ang. *conventional rounding*) każda komórka jest zaokrąglona do najbliższej wielokrotności przyjętej podstawy. Na ogół za podstawę zaokrąglania przyjmuje się 3 lub 5. Do zalet tej metody zalicza się to, że wartości brzegowe tablicy są zaokrąglane niezależnie od pozostałych komórek, a zatem zachowana jest spójność wartości globalnych oraz to, że różne tablice, które reprezentują to samo źródło, będą reprezentować takie same wartości dla tych samych rekordów. Wadą rozpatrywanego podejścia jest natomiast niezachowywanie addytywności względem wartości brzegowych. Tym samym zaokrąglanie standardowe cechuje mniejsza efektywność w stosunku do innych metod zaokrąglania. Przykład zaokrąglenia przy podstawie 5 danych zawartych w tablicy 3.15 przedstawia tablica 3.16.

W **zaokrąglaniu losowym** (ang. *random rounding*), podobnie jak przy standardowym zaokrąglaniu, zmiana wartości w komórkach następuje drogą zaokrąglania przy ustalonej podstawie, ale dokonywanego w sposób losowy. Każda komórka jest zaokrąglana niezależnie od innych, lecz w ten sposób, że większe prawdopodobieństwo jest związane z tym, że wartość tej komórki będzie zmieniona na wartość bliższą wielokrotności podstawy. Schematy przypisanych praw-

Tabl. 3.15. Liczba zamieszkałych według płci

Wyszczególnienie	Mężczyźni	Kobiety	Razem
Obszar 1	1	21	22
Obszar 2	18	0	18
Obszar 3	12	2	14
Razem	31	23	54

Źródło: Obliczenia własne na podstawie Hundepool i in. (2012).

Tabl. 3.16. Liczba zamieszkałych według płci (tablica z zaokrągleniami)

Wyszczególnienie	Mężczyźni	Kobiety	Razem
Obszar 1	0	20	20
Obszar 2	20	0	20
Obszar 3	10	0	15
Razem	30	25	55

Źródło: Obliczenia własne na podstawie Hundepool i in. (2012).

dopodobieństw mogą być różne, ale nie powinny być obciążone, to znaczy, że wartość oczekiwana różnicy między komórką zmienioną a źródłową wynosi zero. Na przykład, jeśli reszta z dzielenia danej liczby naturalnej b przez 5 wynosi k, to zaokrąglenie ma postać $5 \cdot [b/5]$ z prawdopodobieństwem $1 - (k/5)$ oraz $5 \cdot ([b/5] + 1)$ z prawdopodobieństwem $k/5$, $k = 0, 1, 2, 3, 4$ (przy czym $[a]$ to część całkowita liczby rzeczywistej a). Jeśli zatem chcemy zaokrąglić liczbę 247 do wielokrotności 5, to efektem tego działania będzie 245 z prawdopodobieństwem $1 - (2/5) = 3/5$ lub liczba 250 z prawdopodobieństwem 2/5. Egzemplifikację schematu losowania oraz zaokrąglania dla przykładowych wartości w tablicy przy podstawie 3 przedstawia tablica 3.17.

Tabl. 3.17. Przykład zaokrąglania losowego

Wartość źródłowa	Zaokrąglona wartość wraz z prawdopodobieństwem jej uzyskania (w nawiasie)
0	0 (1)
1	0 (2/3) lub 3 (1/3)
2	3 (2/3) lub 0 (1/3)
3	3 (1)
4	3 (2/3) lub 6 (1/3)
5	6 (2/3) lub 3 (1/3)
6	6 (1)

Źródło: Hundepool i in. (2012).

Do wad tej metody należy zaliczyć to, że nie zachowuje ona addytywności. Ponadto występuje tutaj możliwość dużej straty informacji, gdyż zmianie podlegają wszystkie komórki w tablicy. Jej zaletę stanowi natomiast łatwość implementacji.

Zaokrąglanie kontrolowane (ang. *controlled rounding*) jest wskazywane przez Hundepoola i in. (2012) jako najbardziej efektywna metoda zakłóceniowa dla tablic częstości. Polega ona na wykorzystaniu programowania liniowego w celu zmiany wartości komórek o bardzo niewielką wartość wraz z zachowaniem addytywności. Możliwe jest tu przyjęcie ustalonego progu ochrony dla komórek chronionych. Komórki są wówczas zaokrąglane tak, aby zmieniona wartość różniła się od rzeczywistych wartości co najmniej o założoną wartość progową, wyrażoną np. w procentach. Problemem w tym wypadku pozostaje trudność implementacji, gdy rozmiary tablicy rosną. Złożoność obliczeniowa rozpatrywanej metody szybko wtedy rośnie. Dodatkową niedogodnością jest wówczas zachowanie poufności w wypadku tablic łączonych. W celu poprawy możliwości implementacji poszukuje się metod, które będą swojego rodzaju kompromisem pomiędzy zwiększeniem efektywności w stosunku do standardowego lub losowego zaokrąglania a częściowym zachowaniem addytywności.

Innym sposobem częściowej, niewielkiej korekty wartości komórek jest tzw. **barnardyzacja** (ang. *barnardisation*). Polega ona na dodawaniu do / odejmowaniu od wszystkich wartości komórek liczby 1 ze stosunkowo małym prawdopodobieństwem. Zerowe wartości komórek nie są zmieniane. Wartości brzegowe koryguje się odpowiednio w celu zachowania addytywności, a liczba zmienionych komórek jest nieznana użytkownikom tablicy. Wadą tej metody jest trudność implementacji oraz zachowania spójności między tablicami, które pozostają między sobą we wzajemnej relacji, tzn. między tablicami łączonymi. Duncan i Roehrig (2007) zaproponowali sposób ochrony poufności (nazwany **korektą cykliczną** – ang. *cyclic perturbation*), który zachowuje addytywność wartości brzegowych. Dodatkowo użytkownicy zyskują częściową wiedzę na temat zastosowanej metody. Ochronę komórek przeprowadza się w sposób cykliczny. Polega ona na tym, że wartości brzegowe nie są zmieniane, a wartości komórek wewnętrznych zostają zmodyfikowane w sposób losowy z zachowaniem addytywności. Pewien ustalony ciąg zmian, polegający na tym, że wartości komórek mogą zostać zwiększone lub

Tabl. 3.18. Przykład schematu korekty cyklicznej

Wyszczególnienie	Cykl 1			Cykl 2			Cykl 3		
	A	B	C	A	B	C	A	B	C
I	+	–		–		+		+	–
II		+	–	+	–		–		+
III	–		+		+	–	+	–	

Źródło: Duncan i in. (2011).

zmniejszone o wartość jeden, tworzy cykl. Tablica podlegająca ochronie przechodzi przez ustalony szereg takich cykli. Tablica 3.18 przedstawia trzy przykładowe cykle dla tablicy dwuwymiarowej 3 × 3, w której zwiększenie o jeden symbolizuje znak „+", a zmniejszenie o jeden znak „–". Duncan i in. (2011) podkreślają, że istotne jest, aby w praktyce zadbać o to, by każda komórka miała szansę być poddana zmianie w takiej samej liczbie cykli. W tablicy 3.18 każda komórka ma szansę zostać zmodyfikowana w dwóch cyklach. Dla zapewnienia efektywności metody kluczowa jest kolejność cykli. Trzem sytuacjom: A, B i C przypisuje się prawdopodobieństwa α, β oraz $\gamma = 1 - \alpha - \beta$. W zależności od tego, jaka sytuacja zostanie wylosowana, podejmowane są specyficzne dla niej działania. Dla A tam, gdzie wystąpi symbol „+", wartość komórki zostanie zwiększona, a tam, gdzie wystąpi symbol „–", zostanie zmniejszona. W wypadku opcji B operacje będą przeprowadzone w sposób odwrotny, a dla C – wartości komórek pozostaną niezmienione. W celu zapewnienia poufności prawdopodobieństwa α i β mogą być dobrane na wystarczająco wysokim poziomie.

Przyjrzymy się obecnie różnym sposobom wykorzystania metod zakłóceniowych w spisach powszechnych. Spisy powszechne są jednym z przykładów istotnych wyzwań dla ochrony poufności ze względu na dużą liczbę tablic generowanych na podstawie jednego źródła oraz znaczny stopień ich szczegółowości. Hundepool i in. (2012) opisali różne podejścia stosowane dla ochrony poufności w tym zakresie na przykładzie praktyki trzech urzędów statystycznych: brytyjskiego urzędu statystycznego – ONS, australijskiego biura statystycznego – ABS (Australian Bureau of Statistics) oraz nowozelandzkiego statystycznego urzędu – SNZ (Statistics New Zealand).

Urząd brytyjski stosował w spisie powszechnym w 2001 r. równolegle dwie metody ochrony poufności: metodę używaną do mikrodanych – zamianę rekordów (ang. *record swapping*), a dodatkowo na poziomie tablic – korektę komórek o małych wartościach. Ponieważ w Szkocji nie zastosowano ochrony dla tablic, a tylko na poziomie mikrodanych, porównania danych wynikowych okazały się trudne, co spotkało się z krytyką. Drugim powodem owej krytyki były niespójności występujące pomiędzy tablicami. Z tego powodu w 2011 r. zastosowano tylko jedną metodę ochrony na poziomie mikrodanych – ponownie zamianę rekordów, przy następujących kryteriach: tablice powinny zachować addytywność, parametryzacja powinna być łatwa do wyjaśnienia oraz być używana w podobnej formie w poprzednich spisach. Gospodarstwa domowe zostały ocenione według ryzyka ujawnienia informacji wrażliwych oraz nadano im stosowne rangi według stopnia tegoż ryzyka. Obowiązkiem prawnym było wprowadzenie do publikacji odpowiedniej niepewności, chociaż dokładnego znaczenia pojęcia „odpowiedniej niepewności" nie określono.

Australijskie biuro statystyczne w spisie powszechnym przeprowadzonym w 2006 r. również zastosowało nową metodę dla ochrony poufności. Dla każde-

go rekordu został przydzielony specyficzny numer losowy, a dla każdej komórki naliczanej tablicy – na podstawie kluczy dla jednostek wchodzących w skład owej komórki – tworzony był odpowiedni klucz. Klucz ów stanowił podstawę oceny tego, jaki stopień ochrony jest wymagany dla tej komórki. Komórka miała taką samą ocenę – niezależnie od tablicy, w której się znajdowała – jeśli spełniała te same kryteria. Gwarantowało to spójność oraz elastyczność przy tworzeniu tablic według kryteriów użytkownika.

Z kolei nowozelandzki urząd statystyczny stosował w 2006 r. ukrywanie komórek w tablicach według następujących kryteriów:

- ukrycie wartości dla obszarów, dla których przeciętna liczba mieszkańców wynosi nie więcej niż 100,
- tworzenie tablic o największej szczegółowości oraz zawierających dane o dochodach,
- mała przeciętna wartość komórki w tablicy (zdefiniowana jako iloraz liczby ludności obszaru przez liczbę komórek w tablicy).

Dodatkowo wszystkie wartości komórek w tablicach zostały zaokrąglone przy podstawie 3. W późniejszych latach SNZ upowszechniał tablice bardziej szczegółowe. Użytkownicy tablic publikowanych przez SNZ akceptują losowe zaokrąglanie i nie wnoszą krytycznych uwag odnośnie do braku addytywności. Dodatkowo SNZ udostępnia bardziej szczegółowe dane spisowe za opłatą licencyjną.

Metody stosowane w powyższych trzech urzędach różnią się, ale proces dojścia do wyboru optymalnego rozwiązania w tym zakresie przebiega podobnie. Oznacza to, że bazuje się na poprzednio wypracowanych metodach oraz uwzględnia się zachodzące zmiany technologiczne w połączeniu z możliwością implementacji stosownych narzędzi, a także postrzeganie przez użytkowników jakości publikowanych danych po zastosowaniu ochrony ich poufności.

Inną metodą jest **kontrolowane dopasowanie tablicy** (ang. *controlled tabular adjustment* – CTA). Polega ono na wyznaczeniu nowej tablicy z zachowaniem addytywności wedle tablicy oryginalnej w taki sposób, że wartości komórek z ryzykiem pierwotnym zostają zmienione na inne wartości, odpowiednio odległe od stosownych danych faktycznych. W praktyce przyjmuje się, że odległości między wartościami zamiennymi a faktycznymi powinny należeć do pewnego ustalonego przedziału tolerancji. Rzeczona metoda wraz z różnymi jej wariantami opiera się – tak jak metody stosowane do ukrywania komórek – na programowaniu liniowym. Różnica między nią a innymi rozwiązaniami podobnego rodzaju tkwi w odmiennym sformułowaniu warunków brzegowych dla optymalizacji. Wartości komórek w docelowej tablicy są zmienione na bezpieczne. W celu zachowania addytywności zmodyfikowane muszą być również niektóre komórki oryginalnie bezpieczne. Chodzi o to, aby ingerencja w dane była możliwie najmniejsza. Stratę informacji mierzy się odległością pomiędzy wartościami komórek pierwotnych a wartościami komórek po zmianie.

Z kolei **metoda kluczy komórkowych** (ang. *cell-key method* – CKM) jest podejściem posttablicowym, polegającym na dodawaniu do każdej komórki tablicy szumu losowanego określonym mechanizmem z ustalonego rozkładu. Procedura ta składa się z kilku kroków. Najpierw każdemu rekordowi w zbiorze mikrodanych stanowiących podstawę konstrukcji tablicy przyporządkowuje się **klucz rekordu** (ang. *record key*) losowany z rozkładu jednostajnego $U(0, 1)$. W wyniku tego podczas konstruowania tablic nie tylko zlicza się rekordy należące do danej komórki (czyli spełniające wymogi określone jej definicją), ale odpowiednio dodaje się także klucze rekordów. Część ułamkowa sumy kluczy rekordów dla danej komórki jest **kluczem komórki** (ang. *cell-key*), uważanym za zmienną losową o rozkładzie jednostajnym $U(0, 1)$. Następnie na podstawie predefiniowanych prawdopodobieństw przejścia (tzw. *p*-tablica – ang. *p-table*) wyznacza się szum jako funkcję kluczy komórek i wartości komórek, po czym dodaje się go do odpowiednich danych (Dove i in., 2018). Sama *p*-tablica jest zaś tak definiowana, aby wartość oczekiwana przejść wynosiła zero oraz by różne wartości komórek cechowały się różnym poziomem zakłóceń, np. niskie wartości miały ów poziom wyższy. O tworzeniu *p*-tablicy szczegółowo pisał m.in. Giessing (2016). Dzięki swej konstrukcji metoda CKM zapewnia spójność: dla tak samo zdefiniowanych komórek zarówno wartości owych komórek, jak i ich klucze będą takie same, a więc i szum będzie identyczny. Stosowanie szumu losowego w CKM nie gwarantuje addytywności, gdyż wtedy szum byłby nakładany odrębnie na każdą komórkę, bez uwzględniania addytywności. Europejskie Centrum Doskonalenia SDC (ang. European Centre of Excellence on SDC – CoE on SDC), którego celem jest harmonizacja podejść w zakresie kontroli ujawniania danych stosowanych w różnych krajach UE, uznaje jednak, że spójność okazuje się tutaj ważniejsza.

Metody celowanej wymiany rekordów i kluczy komórkowych stały się podstawą studiów i analiz prowadzonych w różnych krajach Europejskiego Systemu Statystycznego (ESS) mających na celu zbadanie możliwości jak najefektywniejszego wdrożenia tychże narzędzi z uwzględnieniem specyfiki statystyki w poszczególnych państwach.

3.3. Porównanie rozpatrywanych metod kontroli ujawniania dla mikrodanych i tablic statystycznych

W tablicy 3.19 zebrano w celu porównawczym najważniejsze własności rozpatrywanych tu narzędzi SDC dla ochrony mikrodanych. Widać, że metody niezakłóceniowe najczęściej stosuje się do zmiennych jakościowych, rzadko zachodzi tutaj potrzeba stosowania zaawansowanych narzędzi probabilistycznych. Odwrotnie jest w wypadku metod zakłóceniowych – z wyjątkiem PRAM mogą być

Tabl. 3.19. Syntetyczne porównanie własności metod SDC dla mikrodanych

Metoda	Zastosowanie do zmiennych		Wykorzystanie narzędzi probabilistycznych
	jakościowych	ilościowych	
Maskowanie niezakłóceniowe			
Podpróbkowanie	+	+	+/–[a)]
Przekodowanie	+	+	–
Lokalne ukrywanie danych	+	–	–
Maskowanie zakłóceniowe			
Dodawanie szumu	–	+	+
Mikroagregacja	–/+ [b)]	+	–
Wymiana rang	–/+[c)]	+	+
Celowana wymiana rekordów	+	–	+
Zaokrąglanie	–	+	–/+[d)]
Próbkowanie wtórne	–	+	+
PRAM	+	–	+
MASSC	–	+	+

a) Próbkę można dobierać losowo lub nielosowo. b) Dotyczy mikroagregacji z wykorzystaniem odległości Gowera. c) Metody nie stosuje się do zmiennych wyrażonych na skali nominalnej. d) W zależności od opcji.

Źródło: Opracowanie własne.

Tabl. 3.20. Sumaryczne zestawienie własności metod SDC dla tablic (w ujęciu posttablicowym)

Metoda	Ingerencja w dane źródłowe	Redukcja szczegółowości tablicy	Wykorzystanie narzędzi probabilistycznych
Tablice częstości			
Ukrywanie komórek	–	–	–
Restrukturyzacja tablicy	–	+	–
Zaokrąglanie standardowe	+	–	–
Zaokrąglanie losowe	+	–	+
Zaokrąglanie kontrolowane	+	–	+
Barnardyzacja i korekta cykliczna	+	–	+
Metoda kluczy komórkowych	+	–	+
Tablice wielkości			
Przekodowywanie globalne	–	+	–
Ukrywanie komórek	–	–	–
Przedział poufności	–	+	–
Zaokrąglanie standardowe	+	–	–
Zaokrąglanie losowe	+	–	+
Zaokrąglanie kontrolowane	+	–	+
Kontrolowane dopasowanie tablic CTA	+	–	+
Metoda kluczy komórkowych	+	–	+

Źródło: Opracowanie własne.

one stosowane głównie do zmiennych ilościowych. W znacznej mierze bazują też na możliwościach, jakie daje teoria prawdopodobieństwa.

W tablicy 3.20 syntetycznie porównuje się przedstawione we wcześniejszych częściach metody kontroli ujawniania danych dla tablic. Widać, że część metod SDC może być stosowana do obu rodzajów tablic. Chodzi tu przede wszystkim o ukrywanie komórek lub zaokrąglanie ich wartości. Warto ponadto zauważyć, że restrukturyzacja (dla tablic częstości) i przekodowywanie globalne (dla tablic wielkości) w istocie mogą być tożsame znaczeniowo, gdy prowadzą do tego samego, czyli np. zastąpienia szczegółowszych poziomów agregacji bardziej zgrubnymi.

Jak już zasygnalizowano, z metod ochrony poufności przeznaczonych dla mikrodanych można również skorzystać w procesie kontroli ujawniania tablic statystycznych. W takim podejściu stosuje się je na zbiorach danych jednostkowych, na podstawie których mają zostać naliczone tablice statystyczne, jeszcze przed przystąpieniem do tej czynności. Następnie, po zastosowaniu wybranych metod SDC, można przystąpić do agregowania jednostkowych danych statystycznych. W zależności od przyjętego rozwiązania podejście to może być jedynym źródłem ochronnym (w takim wypadku uznajemy, że agregatom tym zapewniono już wystarczającą ochronę poufności). Równie dobrze jednak można je połączyć z podejściem posttablicowym i po naliczeniu tablic statystycznych lub agregatów prezentowanych w innym układzie zastosować dodatkowo metodę ochrony poufności przeznaczoną dla danych tabelarycznych. Należy podkreślić, że jeżeli podejścia pre- i posttablicowe stosowane są w kombinacji, to można podejść do ich parametryzacji mniej restrykcyjnie. W przeciwnym razie każda metoda powinna mieć zadeklarowane wartości przyjmowanych przez nie parametrów w sposób bardziej obostrzony.

Zdaniem autorów część metod stosowanych w procesie kontroli ujawniania mikrodanych wydaje się właściwsza do stosowania w podejściu pretablicowym. Na przykład, gdyby w celu ochrony poufności na poziomie mikrodanych użyto lokalnego ukrywania danych, wówczas w zbiorze takim pojawiłyby się braki danych. Przy naliczaniu tablicy statystycznej trzeba by potraktować je jako odrębną kategorię zmiennej jakościowej lub zniwelować ich wpływ na jakość i bezpieczeństwo danych wynikowych poprzez przeprowadzenie imputacji tychże braków danych bądź odpowiednie skalibrowanie wag uogólniających w badaniach reprezentacyjnych. Pierwsza wspomniana tutaj metoda radzenia sobie z brakami danych, tzn. imputacja, może skutkować wzrostem ryzyka ujawnienia informacji poufnych (nie mamy pewności, czy zaimputowana wartość nie jest tą pierwotną, prawdziwą dla określonej jednostki statystycznej). Gdy zaś skorzystamy z metody globalnego przekodowania, naliczenie wybranej tablicy statystycznej może się okazać niemożliwe w sytuacji, gdyby przekodowano zmienną grupującą i nie przyjmowałaby ona już tak szczegółowych kategorii, które wyznaczają przekroje w tejże tablicy.

Podsumowując, można stwierdzić, że zastosowanie metod niezakłóceniowych w podejściu pretablicowym może się wiązać z licznymi konsekwencjami, które są w stanie utrudnić lub nawet uniemożliwić naliczenie tablic statystycznych. W związku z tym bardziej odpowiednie wydaje się tutaj skorzystanie z metod zakłóceniowych.

3.4. Ochrona wyników analiz

W podrozdziale 1.2.4 wprowadzono pojęcie wyników analiz jako typu danych wynikowych, które mogą mieć poufny charakter. Z tego względu przed ich udostępnieniem, w celu zapewnienia ochrony poufności, należy je poddać procesowi kontroli ujawniania. Z kolei w podrozdziale 2.3 zaprezentowano ogólną koncepcję zagadnienia ryzyka ujawniania informacji poufnych w przypadku takiej postaci danych wynikowych – w szczególności zaś omówiono ich klasyfikację pod względem struktury i czynność przyporządkowywania im etykiet. Przedstawiono dwa podejścia w procesie SDC stosowane z osobna lub w kombinacji, tj. podejście oparte na zasadach oraz zasadę kciuka. W niniejszej części zaprezentowano podział podstawowych typów wyników analiz na klasy wraz z ich etykietami. Podano również ogólną postać zasady kciuka. Następnie przytoczono zasady lub wskazówki (w zależności od podejścia) dla kilku przykładowych postaci danych wynikowych. Wskazano również literaturę, w której zainteresowany czytelnik będzie mógł znaleźć zasady bądź wskazówki w zakresie SDC dla innej postaci wyników analiz niż zaprezentowane poniżej. Na koniec omówiono podejście do weryfikacji ilustracji graficznych pod kątem ochrony poufności, wraz z jego autorską egzemplifikacją.

Na podstawie tekstów naukowych z zakresu SDC, pochodzących z publikacji z wielu krajów, wśród bezpiecznych lub niebezpiecznych klas danych wynikowych wyróżnić można następujące wyniki analiz:

- **Klasy wyników analiz z etykietą *niebezpieczne*:**
  - Dane tabelaryczne:
    - tablice częstości
    - tablice wielkości
  - Statystyki opisowe:
    - wartość minimalna
    - wartość maksymalna
    - kwartyle (w tym mediana)
    - kwintyle
    - decyle
    - percentyle
    - średnia
    - wskaźniki

  - Rezultaty analiz:
    - reszty z modelu regresji
  - Ilustracje graficzne.
- **Klasy wyników analiz z etykietą *bezpieczne*:**
  - Statystyki opisowe:
    - dominanta
    - wariancja
    - odchylenie standardowe
    - skośność
    - kurtoza
    - wskaźniki koncentracji
  - Rezultaty analiz:
    - współczynniki modelu regresji liniowej
    - współczynniki modelu regresji nieliniowej
    - statystyki podsumowujące i testowe
    - współczynniki korelacji
    - analiza czynnikowa
    - analiza korespondencji.

Jak już wspomniano, w literaturze przedmiotu wypracowano tzw. **ogólną wersję zasady kciuka**, której zapisy znajdą zastosowanie w SDC niezależnie od struktury danych wynikowych (Hundepool i in., 2012; Bond i in. (2015), Brandt i in., 2010; Höninger i in., 2010). Odniesienia do punktów tej zasady można również znaleźć w opisach reguł lub wskazówek w wypadku dwóch podejść stosowanych w procesie SDC dla poszczególnych klas wyników analiz. Ogólna postać zasady kciuka obejmuje – spisane w sposób jasny, prosty, ale też rygorystyczny – cztery poniższe reguły.

- Pierwsza odnosi się do liczby jednostek: w każdej komórce (punkcie danych itp.) we wszystkich zestawieniach tabelarycznych oraz w innej postaci zagregowanych danych wynikowych powinno się znajdować co najmniej 10 jednostek (w wypadku badań reprezentacyjnych bez przeważania) – jest to tzw. **zasada progu** (ang. *threshold rule*).
- Druga reguła mówi o **10 stopniach swobody**: wszystkie dane wynikowe uzyskane w drodze zastosowania dowolnego modelu powinny mieć co najmniej właśnie tyle stopni swobody (liczba stopni swobody to liczba obserwacji pomniejszona o liczbę parametrów oraz o inne ograniczenia wynikające z modelu), a także do konstrukcji tego modelu powinno być wykorzystanych co najmniej 10 jednostek.
- Trzecia reguła dotyczy **ujawniania informacji o grupach**: aby zapobiec ujawnieniu grupy, w zestawieniach tabelarycznych i innych podobnych wartość żadnego agregatu (np. w komórce) nie powinna przekraczać 90% odpowiedniej sumy brzegowej (np. wierszowej lub kolumnowej w tabelach).

Ujawnienie miałoby również miejsce, gdyby pewne zmienne wyznaczające przekroje tablicy statystycznej definiowały grupę jednostek, a inne zmienne ujawniłyby informację o charakterze poufnym lub wrażliwym dla każdego członka grupy. Nawet w przypadku niezidentyfikowania żadnej jednostki statystycznej poufność informacji zostaje wówczas naruszona, ponieważ informacja jest prawdziwa w odniesieniu do każdej jednostki należącej do grupy, a tę jako całość daje się zidentyfikować.

– Ostatnia reguła jest powiązana z **przewagą**: w zestawieniach tabelarycznych lub podobnych udział jednostki o największym udziale w wartości w komórce nie powinien stanowić więcej niż 50% owej wartości. W praktyce w wielu krajach reguła ta nie jest jednak przestrzegana, ponieważ wymaga dostępu do dodatkowych informacji (np. danych o wartościach cechy dla każdej jednostki statystycznej, na podstawie których naliczono każdy z agregatów, czy też danych o wartościach cech tylko dla tych respondentów, którzy odznaczają się największym udziałem w danym agregacie, lecz dla wszystkich komórek zestawienia tabelarycznego). Wiązałoby się to bowiem z dodatkowym nakładem pracy oraz ryzykiem ujawnienia informacji naruszającej poufność. Ponadto nie każde narzędzie informatyczne umożliwia zastosowanie takiego rozwiązania. Z tego powodu rzeczoną zasadę stosuje się przede wszystkim przy danych z obszaru statystyki gospodarczej, zmiennych sklasyfikowanych jako wrażliwe oraz gdy rozkład zmiennych charakteryzuje się silną skośnością i występowaniem jednostek odstających.

Szczegółowy opis reguł dla zasady kciuka oraz wskazówek dla podejścia opartego na zasadach można znaleźć w następujących publikacjach: Hundepool i in. (2012), Bond i in. (2015), Brandt i in. (2010), Höninger i in. (2010), a także Hochfellner i in. (2012). W ostatniej wymienionej pozycji przedstawione zostały wybrane reguły stosowane w kontroli ujawniania wyników analiz, które są wykorzystywane w praktyce przy stacjonarnym udostępnianiu poufnych danych jednostkowych przez Niemiecką Federalną Agencję Pracy w Instytucie Badań nad Zatrudnieniem. Poniżej w celu egzemplifikacji obu scharakteryzowanych wcześniej podejść przytoczono zalecenia w zakresie kilku podstawowych postaci wyników analiz: wśród statystyk opisowych wybrano wartość minimalną i maksymalną, kwartyle, dominantę i średnią, wśród rezultatów analiz zaś – współczynniki modelu regresji oraz reszty z modelu.

Wartość minimalna, wartość maksymalna

– **Zasada kciuka**: nie przewiduje się możliwości udostępniania informacji o minimalnej lub maksymalnej wartości cechy, ponieważ przeważnie odnoszą się one tylko do jednej jednostki statystycznej.
– **Podejście oparte na zasadach**: w uzasadnionych przypadkach może się okazać możliwe udostępnienie informacji o najniższej bądź najwyższej wartości

badanej cechy – gdy zostaną one uznane za bezpieczne. Mianowicie, jeżeli wartość minimalna lub maksymalna jest przyjmowana dla większej liczby jednostek statystycznych (co najmniej 10), to można ją uznać za bezpieczną. Istotną rolę odgrywają więc w tym wypadku dobra dokumentacja oraz fachowe informacje dostarczone do osób odpowiedzialnych za proces SDC. Warto wszakże podkreślić, że nie wolno publikować tychże wartości, gdy powszechnie znany jest ranking jednostek według badanej cechy lub gdyby wynik analizy prowadził do ujawnienia bardzo wąskiej grupy jednostek statystycznych o określonej cesze. Niezwykle ważne jest przyjrzenie się zwłaszcza wartości maksymalnej badanej cechy, ponieważ respondent ją posiadający zazwyczaj bywa powszechnie znany. Wskazane jest tu zastosowanie zasady progu, przewagi oraz ujawniania grupy z ogólnej zasady kciuka.

Kwartyle (dolny, mediana i górny)

- **Zasada kciuka**: nie zaleca się udostępniania dokładnej wartości kwartyli, gdyż często reprezentują one wartość cechy dla pojedynczego respondenta. Wartości kwartyli wyznacza się dla uporządkowanych danych jednostkowych, a liczebności przedziałów zdeterminowanych przez te kwartyle traktuje się podobnie jak komórki w tablicy wielkości – gdy pozycja jednostki w rankingu jednostek statystycznych jest znana, wówczas zachodzi ryzyko ujawnienia informacji poufnych o tej jednostce. Jak wysokie jest to ryzyko, zależy od liczby jednostek należących do określonego przedziału oraz od wyboru przedziału,
- **Podejście oparte na zasadach**: konieczne jest ustalenie, czy respondent mógłby zostać zidentyfikowany na podstawie opublikowanych wartości kwartyli. Jeżeli znany jest ranking jednostek, kwartyle nie powinny być publikowane. Jeżeli zróżnicowanie wartości cechy wokół wybranego kwartyla jest małe lub zerowe, zachodzi ryzyko ujawnienia grupy jednostek statystycznych. W sytuacji dużego rozproszenia wartości cechy wokół kwartyla istnieje zaś ryzyko ujawnienia informacji o pojedynczym respondencie. Zaleca się tutaj użycie tych samych reguł, które znajdują zastosowanie dla tablic wielkości.

Dominanta

- **Zasada kciuka**: wartość dominanty powinna być współdzielona przez co najmniej 10 jednostek statystycznych (bez ich przeważania), a rozkład jednostek powinien być taki, żeby wartość najczęściej występująca była obserwowana dla nie więcej niż 90% ogólnej liczby jednostek w wierszu lub w kolumnie tablicy (ryzyko ujawnienia grupy lub pojedynczych respondentów, gdy wszystkie jednostki osiągają tę samą wartość badanej cechy). Zaleca się tutaj użycie tych samych reguł, które znajdują zastosowanie dla tablic wielkości.
- **Podejście oparte na zasadach**: dopuszcza się wyjątek, zgodnie z którym wartość dominanty mogłaby zostać udostępniona (nawet gdyby pozwalała

ona na zidentyfikowanie jednostki) – mianowicie może to wystąpić w sytuacji dysponowania zgodą takiej jednostki na opublikowanie rzeczonej informacji statystycznej. Zaleca się tutaj również m.in. stosować kryterium związane z przewagą (z ogólnej zasady kciuka), a także użycie tych samych reguł, które znajdują zastosowanie dla tablic wielkości.

Średnia

- **Zasada kciuka**: nakazuje się w tym wypadku sprawdzenie spełnienia pierwszej i czwartej reguły ogólnej zasady kciuka. Wartość średnia powinna być obliczona na podstawie co najmniej dziesięciu nieprzeważonych obserwacji, a wartość cechy dla jednostki statystycznej z największym udziałem w wartości takiego agregatu nie powinna przekraczać 50%. Zaleca się tu użycie tych samych reguł, które znajdują zastosowanie dla tablic wielkości.
- **Podejście oparte na zasadach**: poddaje się pod rozwagę, że wraz ze wzrostem stopnia złożoności wskaźnika ryzyko ujawnienia informacji poufnych maleje, a tym samym rośnie prawdopodobieństwo ich udostępnienia. Gdyby osoby przeprowadzające proces SDC dysponowały zgodą od respondenta na udostępnienie takich informacji, mogłyby one zostać opublikowane.

Współczynniki regresji liniowej

- **Zasada kciuka**: jeżeli co najmniej jeden spośród współczynników (np. wyraz wolny) nie zostanie udostępniony, to pozostałe mogą być bezpiecznie udostępnione.
- **Podejście oparte na zasadach:** zakłada, że kompletny model ekonometryczny można opublikować, jeżeli ma on co najmniej 10 stopni swobody, nie bazuje wyłącznie na zmiennych jakościowych, a także nie dotyczy tylko jednej jednostki (wykluczona jest np. prezentacja szeregu czasowego dla jednej jednostki statystycznej). Gdyby w modelu znalazły się tylko zmienne niezależne typu binarnego, wówczas współczynniki regresji odpowiadałyby wartościom średnim dla agregatów – i tak powinny być traktowane. W wypadku modeli nasyconych (czyli modeli o liczbie parametrów równej liczbie obserwacji) dane wynikowe teoretycznie mogłyby zostać wykorzystane do odtworzenia danych zagregowanych. Jeżeli taki model został oszacowany, to współczynniki owego modelu są ściśle związane z agregatami.

Reszty

- **Zasada kciuka**: nie pozwala się na udostępnianie wartości reszt, a także ich ilustracji graficznych, ponieważ reszty są mikrodanymi o charakterze poufnym.
- **Podejście oparte na zasadach**: również nie pozwala się na publikację reszt. W wypadku ich wykresów podejście to poleca udostępnienie jedynie ich tekstowej analizy, bez ilustracji graficznej. Wyjątkiem od tej zasady są stan-

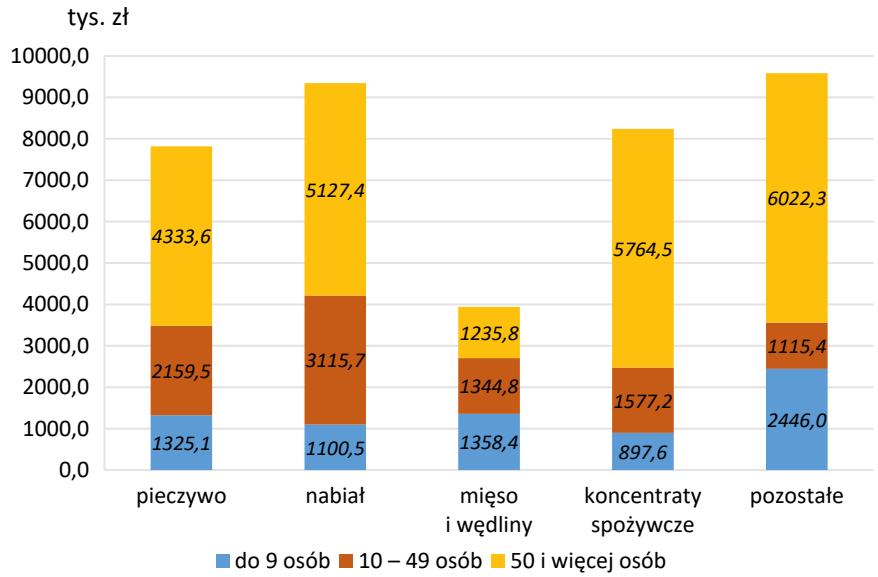
daryzowane reszty dla agregatów (a nie danych jednostkowych). W takim wypadku reszty dla komórek mogą być udostępnione, jeżeli każda z tych komórek obejmuje co najmniej 10 jednostek.

Ilustracje graficzne (wykresy, kartogramy, piktogramy) są – jak powszechnie wiadomo – nieodłącznym elementem profesjonalnej prezentacji wynikowych informacji statystycznych. Dlatego też kontrola ujawniania danych powinna być stosowana także do nich.

Ilustracje graficzne powstają w zasadzie na podstawie finalnych mikrodanych czy tablic wynikowych, tak więc kontroli SDC powinny być w pierwszym rzędzie poddane owe źródła. Jednak nie zwalnia to z konieczności weryfikacji także ilustracji graficznych pod tym kątem, gdyż mogą one powstać na przykład na podstawie tablic nieprzeznaczonych do publikacji bądź też pewne wrażliwe kombinacje wartości zmiennych zostaną uwidocznione dopiero na wykresie czy kartogramie.

Dlatego kontrola ujawniania danych na obiektach graficznych powinna przebiegać w czterech głównych kierunkach. Pierwszym z nich jest sprawdzenie, **czy nie staje się realna identyfikacja indywidualnego respondenta**. Rysunek 3.3 ukazuje przykład wykresu z taką możliwością.

Jeśli w powiecie, którego dotyczy ta ilustracja, działa na przykład tylko jeden producent koncentratów spożywczych zatrudniający 50 i więcej pracowników, to jego konkurenci bez trudu odczytają ze wskazanego wykresu informacje o wiel-

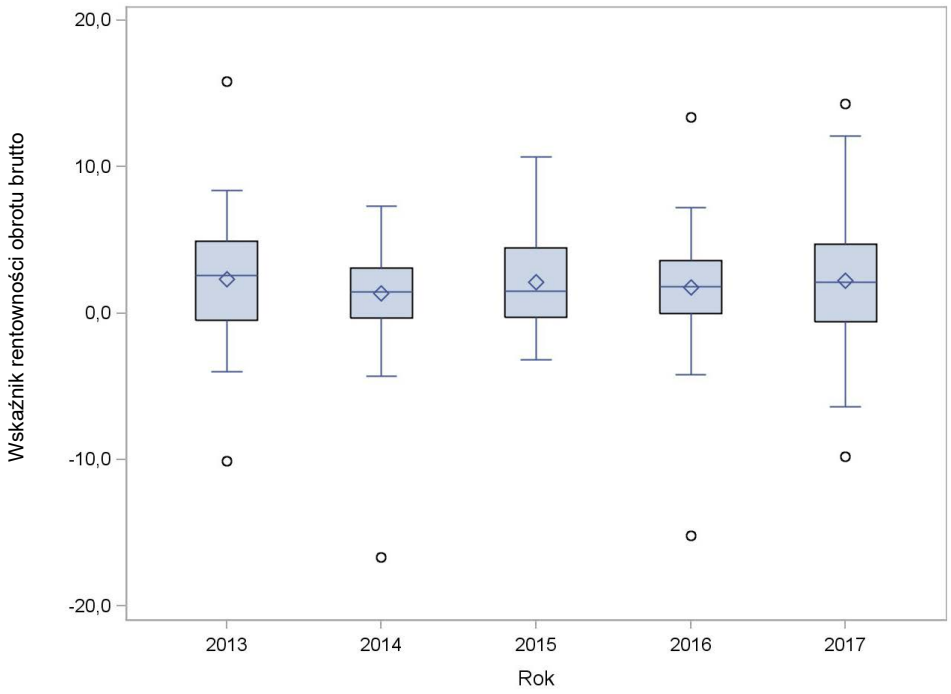


Rys. 3.3. Przychody netto ze sprzedaży produktów w przedsiębiorstwach w pewnym powiecie według liczby zatrudnionych w 2017 r. (w tys. zł)

Źródło: Dane fikcyjne.

kości jego produkcji. Podobnie, jeśli np. mamy tutaj dwóch producentów nabiału zatrudniających od 10 do 49 osób, to każdy z nich też bez większego wysiłku odczyta stąd dane konkurenta.

Kolejna kwestia to **występowanie obserwacji odstających** (ang. *outliers*). W analizie rozkładów empirycznych rozmaitych zjawisk i ich parametrów, teoretycznie rzecz ujmując, do ujawniania wrażliwych danych dojść nie powinno. Rozpatrywane są bowiem tutaj pewne informacje zbiorcze. Jednak czasami występują pojedyncze obserwacje odstające, których obecność na histogramach czy innych wykresach jest wyraźnie zaznaczona. A to już często pozwala na identyfikację jednostek, dla których takie wielkości są przyjmowane. Rysunek 3.4 ukazuje, jak może to wyglądać na wykresie pudełkowym.



Rys. 3.4. Wskaźnik rentowności obrotu brutto w przedsiębiorstwach produkujących odzież w pewnym podregionie w latach 2013–2017 (w %)

Źródło: Opracowano z wykorzystaniem programu SAS Enterprise Guide 4.3. Dane fikcyjne.

Warto przypomnieć, że brzegi pudełka oznaczają w tym wypadku kwartyle: górny i dolny, linia środkowa pudełka obrazuje medianę, obrócony zaś kwadrat – średnią arytmetyczną. Tym samym pudełko obejmuje około połowy obserwacji. Linie pionowe (zwane „wąsami") oznaczają zakres zmienności obserwacji, tutaj

określany przez 1,5 odległości międzykwartylowej od górnego i dolnego brzegu pudełka. Obserwacje znajdujące się poza pudełkiem i wąsami (oznaczone kółkami) uważa się za odstające.

Tego typu wykres stwarza ryzyko identyfikacji jednostek, dla których występują owe odstające obserwacje. Ryzyko takie występuje, jeśli użytkownik danych na podstawie innych informacji wie, kto w danej branży ma największe kłopoty finansowe lub jest w najlepszej kondycji ekonomicznej, lub też jeśli inne tablice w tej samej publikacji pozwalają mu taką wiedzę – choćby częściową – uzyskać. Willenborg i de Waal (1996) zauważyli jednak, że jeśli jednostka jest pod pewnym względem jedyną w populacji, ale nikt inny nie jest tego świadomy, to dane jej dotyczące nie zostaną odtworzone. Z drugiej strony, jeżeli jednostka nie jest unikatem, ale w populacji znajduje się inna, dla której wartości pewnych zmiennych kluczowych są identyczne, to dane owej odstającej jednostki mogą zostać dostrzeżone przez tę drugą. Oczywiście statystyk nie ma wpływu na wiedzę posiadaną przez użytkownika, a pochodzącą z niezależnych źródeł. Jednak należy w pierwszym rzędzie zadbać o to, aby wyeliminować lub zredukować do absolutnego minimum możliwość identyfikacji wrażliwych danych w obrębie publikacji lub udostępnianych zasobów informacji statystycznych. Na przykład, jeżeli inna tablica podaje kwartyle (bądź inne kwantyle) rozkładu zjawiska wskazanego na rysunku 3.4 w przekroju według wielkości firmy, to zestawiając te informacje, łatwo ocenić, ile osób zatrudnia firma, dla której obserwacja rozpatrywanego zjawiska jest odstająca (im bardziej szczegółowe klasowe przedziały liczby pracujących, tym precyzja szacunku większa). Tak więc kontrola ujawniania danych między poszczególnymi elementami przekazu statystycznego musi być dość silnie skorelowana.

Z tymi problemami wiąże się też **bezpieczeństwo danych podstawowych**. Chodzi mianowicie o to, czy dane uwidocznione na wykresie lub kartogramie zostały poddane odpowiedniej kontroli ujawniania. W istocie oczekuje się tutaj zastosowania metod SDC analogicznych do przypadku danych tabelarycznych, tzn. przeprojektowywania (poprzez redukcję rozmiaru lub przekodowywania) albo też zaokrąglania, z zastosowaniem reguły progu, p% czy (n, k)-dominacji. Ponadto, jeśli na ilustracji ukazane są wyłącznie statystyki opisowe lub indeksy, to można sprawdzić, czy zostały one wyznaczone na podstawie wystarczająco dużej liczby jednostek oraz czy w danej agregacji jedna z tych jednostek nie ma udziału większego niż 50%. Natomiast nie zaleca się stosowania w tym wypadku ukrywania danych, gdyż czyni to je wówczas trudnymi – a nierzadko wręcz niemożliwymi – do prezentacji na wykresie. Chodzi zatem tutaj o bezpieczeństwo końcowego produktu, nie zaś danych źródłowych. Na przykład z tego punktu widzenia dane ukazane na rysunkach 3.3. i 3.4 mogą być same w sobie bezpieczne, jeśli taka kontrola zostanie dokonana, mimo że ich pierwotne źródło do bezpiecznych należeć nie musi.

I tutaj dochodzimy właśnie do ostatniego warunku bezpieczeństwa danych uwidocznionych na ilustracji, którym jest **eliminacja danych „zagnieżdżonych" w ilustracji**. To, co widać na rysunku, może być bowiem przekazem bezpiecznym. Jednakże bardzo często przekaz taki powstaje z wykorzystaniem mikrodanych, na podstawie których program komputerowy dokonuje stosownych obliczeń i ukazuje ich wyniki na wykresie lub kartogramie. Mikrodane te zaś mogą być wrażliwe. Najprostszym przykładem tego rodzaju sytuacji jest arkusz Excela z danymi będący integralną częścią wykresu sporządzonego w pakiecie Office. Tutaj mogą znajdować się nie tylko dane finalnie użyte do opracowania wykresu, ale także informacje chronione, które posłużyły do wyznaczenia stosownych statystyk sumarycznych. Podobne zagnieżdżenia mogą występować także w programach służących do sporządzania wykresów on-line, w internecie, na podstawie stosownych mikrodanych (np. w publicznie dostępnych bazach i bankach danych). Tak więc należy uczynić wszystko, aby dostęp do tych danych jednostkowych był niemożliwy. W wypadku publikacji oznacza to, że – biorąc pod uwagę także oczekiwaną jakość opracowania – należy osadzać ilustracje w formie grafiki wektorowej (formaty CGM, WMF, EMF, EPS itp.) lub rastrowej o wysokiej rozdzielczości, co najmniej 300 dpi (JPEG, TIFF, BMP, PNG itp.). Natomiast narzędzia konstruujące wykresy on-line powinny blokować dostęp do mikrodanych źródłowych, a same wykresy zapisywać także wyłącznie w wyżej wskazanych formatach.

Podsumowując zagadnienie kontroli ujawniania wyników analiz pod względem metod służących ochronie poufności udostępnianych danych wynikowych opracowanych na podstawie poufnych mikrodanych w kontrolowanym przez gestora danych środowisku, należy podkreślić wyraźną różnicę pomiędzy nimi a metodami stosowanymi w procesach SDC przeznaczonych dla mikrodanych bądź tablic statystycznych. Opisane we wcześniejszych podrozdziałach tego rozdziału reguły i metody, które mają zastosowanie dla danych jednostkowych i tabelarycznych (por. podrozdziały 3.1–3.2), oprócz ukrycia poufnej informacji, dopuszczają bowiem inne rozwiązania, polegające na odpowiednim zmniejszeniu szczegółowości danych (lub też na ich zniekształceniu) przed udzieleniem do nich dostępu użytkownikom zewnętrznym. W wypadku wyników analiz sprawdza się je pod kątem ich bezpieczeństwa, a proces SDC daje jasną odpowiedź: dane wynikowe niestwarzające zagrożenia ujawnienia informacji poufnej mogą zostać przekazane użytkownikowi zewnętrznemu, a te, które mogą stanowić zagrożenie – nie. W razie stwierdzenia potencjalnego zagrożenia dopuszcza się dwie możliwości – usunięcie tych elementów wyników analiz, które stwarzają ryzyko dla ochrony poufności, i udostępnienie reszty z nich albo zwrócenie się z prośbą do użytkownika zewnętrznego, by ponownie przeprowadził swe prace (o ile oczywiście wyrazi taką wolę), wprowadzając doń stosowne zmiany, by efekty tych prac nie stanowiły dłużej zagrożenia.

3.5. Dane syntetyczne

Dane syntetyczne to dane generowane przy użyciu mechanizmów stochastycznych na podstawie oryginalnej struktury i definicji zmiennych, zachowujące wartość analityczną zbioru wyjściowego (czyli związki między zmiennymi i kształty ich rozkładów), a jednocześnie cechujące się maksymalnie wysokim poziomem ochrony przed identyfikacją jednostki i ujawnieniem informacji wrażliwych. Tworzone są one na drodze generowania lub modelowania, których źródła metodologiczne wywodzą się z metod edycji i imputacji. Jest to na ogół proces dość złożony, ale bardzo efektywny. Dzięki rozwojowi odpowiednich narzędzi informatycznych zmniejsza się też czaso- i nakładochłonność. Dane syntetyczne dostarczają użytkownikom anonimowych informacji, których nie można powiązać z konkretnymi jednostkami.

Zastosowanie danych syntetycznych umożliwia znaczne zwiększenie użyteczności zbiorów danych przy jednoczesnym zachowaniu poufności informacji. Ma to istotne znaczenie w sytuacji udostępniania mikrodanych dotyczących konkretnych osób lub podmiotów gospodarczych. Dane syntetyczne nie muszą już być tak pieczołowicie chronione przed wszelkim nieuprawnionym dostępem jak dane oryginalne. Ich efektywna konstrukcja umożliwia też naliczanie tablic i innych informacji zagregowanych w taki sposób, że wielkości w nich zawarte będą odpowiadać wielkościom, jakie można by uzyskać, używając danych oryginalnych. Niemniej udostępnianie danych syntetycznych także odbywa się na określonych warunkach, przede wszystkim w celach naukowych.

Wyróżnia się kilka zasadniczych rodzajów zbiorów z danymi syntetycznymi. Są to:

- **Zbiory nieme** (ang. *dummy files*). Są to zestawy danych, w których na ogół nie jest zachowana wartość analityczna oryginału. Główny nacisk jest tutaj kładziony na utrzymanie struktury i reguł logicznych obowiązujących w danych oryginalnych. Zbiory nieme stosuje się najczęściej do testowania różnego rodzaju algorytmów i procesów oraz gdy oczekiwania użytkownika ograniczają się jedynie do korzystania ze strukturalnie podobnych zasobów. Ponieważ te pliki nie mają wartości analitycznej, ryzyko ujawnienia informacji wrażliwej tutaj w praktyce nie występuje.
- **Zbiory w pełni syntetyczne** (ang. *fully synthetic files*). Wszystkie zmienne są tu syntetyzowane dla zachowania odpowiedniej wartości analitycznej zgromadzonych informacji w porównaniu z oryginalnym zbiorem danych. Wynikiem ich tworzenia jest zachowanie rozkładów jednowymiarowych lub jednego albo więcej rozkładów wielowymiarowych czy łącznych występujących w danych wyjściowych. Dzięki temu utrzymuje się wartości wszystkich lub określonych oryginalnych wartości brzegowych bądź statystyk opi-

sowych tych rozkładów. Zbiory te mają takie samo zastosowanie do użytku publicznego, ale większą wartość niż tradycyjne mikrodane.

– **Zbiory częściowo syntetyczne** (ang. *partially synthetic files*). W plikach tego rodzaju syntetyzuje się tylko niektóre zmienne, najczęściej te będące nośnikami informacji wrażliwych w największym stopniu. Inne zmienne pozostawia się w oryginalnym kształcie.

Istnieją różne metody konstruowania danych syntetycznych. Bardzo popularnym podejściem w tym zakresie jest modelowanie sekwencyjne, a konkretnie – **metoda w pełni warunkowej specyfikacji** (ang. *the fully conditional specification* – FCS). Została ona zaproponowana wprawdzie raczej z myślą o imputacji brakujących danych, ale tworzenie danych syntetycznych można postrzegać w istocie jako szczególny przypadek imputacji masowej. Stąd FCS może być zastosowana także do realizacji tego celu. Ogólnie rzecz ujmując, generowanie liczb losowych z – zazwyczaj nieznanego – rozkładu *k*-wymiarowego jest zastępowane w tym wypadku przez łatwiejsze *k*-krotne losowanie z rozkładów jednowymiarowych (gdzie *k* jest liczbą naturalną). Każda zmienna jest tutaj syntetyzowana osobno, za pomocą odpowiedniego modelu regresyjnego. Mówiąc bardziej precyzyjnie, istotą FCS jest dekompozycja wielowymiarowego rozkładu łącznego na szereg jednowymiarowych rozkładów warunkowych:

$$f_{X_1, X_2, \ldots, X_k} = f_{X_1} \cdot f_{X_2|X_1} \cdot \ldots \cdot f_{X_k|X_1, X_2, \ldots, X_{k-1}}. \tag{3.2}$$

Innymi słowy, zamiast próbować od razu wyjaśnić wszystkie relacje między zmiennymi znajdującymi się w zbiorze danych, syntezator postępuje krok po kroku, modelując i generując jedną zmienną naraz, warunkowo względem poprzednich. Kolejność syntetyzacji zmiennych pozostaje kwestią indywidualnej decyzji w konkretnym przypadku, opartej na odpowiedniej wiedzy merytorycznej, logicznej (np. synteza statusu na rynku pracy i poziomu wykształcenia przed syntezą miesięcznych dochodów) i doświadczeniu badacza. Wobec powyższego FCS jest przeprowadzana w dwóch krokach: najpierw oryginalny zestaw danych jest używany do oszacowania każdego z rozkładów warunkowych przedstawionych po prawej stronie równości (3.2), po czym wartości syntetyczne dla danej zmiennej zostają wygenerowane za pomocą oszacowanego modelu dla tej zmiennej, w którym jako dane wejściowe wykorzystuje się wartości syntetyczne już wytworzone dla poprzednich zmiennych. Modelem tym jest najczęściej model regresji logistycznej (a konkretnie – logitowy) (Drechsler, 2011). Jednak dobór modelu powinien zależeć od specyfiki danej zmiennej objaśnianej oraz od własności zmiennych wcześniejszych. Zatem w zależności od potrzeb można tutaj stosować modele parametryczne, nieparametryczne lub mieszane.

Zaletą FCS jest zatem jej łatwość metodologiczna oraz zachowanie wszystkich relacji między zmiennymi, jak również powiązanie z odpowiednią edycją danych.

Do ułomności FCS zaliczyć natomiast należy występowanie obserwacji odstających w wypadku rozkładów asymetrycznych, co może być ryzykowne z punktu widzenia możliwości identyfikacji jednostki. Przy dużej liczbie zmiennych znaczna jest też czasochłonność rzeczonego procesu.

Jeżeli rozkład pewnych zmiennych jest nieregularny, można zastosować też podejście uczenia maszynowego **drzewa klasyfikacji i regresji** (ang. *classification and regression tree* – CART). Zasadniczo model CART dzieli zbiór potencjalnych regresorów na względnie jednorodne z predykcyjnego punktu widzenia podzbiory. Szereg takich podziałów można efektywnie przedstawić za pomocą struktury drzewiastej, której liście odpowiadają owym podzbiorom. Na przykład, jeśli X_1 to płeć, a X_2 – poziom ukończonego wykształcenia, to jednostki, dla których X_1 = *„kobieta"* umieszczane są na liściu L_1 (bez względu na wartość zmiennej X_2), te, dla których X_1 = *„mężczyzna"* i $X_2 \leq$ *„średnie"* – na liściu L_2, te zaś, dla których X_1 = *„mężczyzna"* oraz $X_2 >$ *„średnie"* – na liściu L_3. Za pomocą metody CART można w szczególności uchwycić nieliniowe relacje między zmiennymi, które mogły nie zostać odpowiednio uwzględnione w modelowaniu parametrycznym. Więcej informacji o tej metodzie podają np. Drechsler i Reiter (2011) czy Reiter (2005). Warto wspomnieć, że w wypadku zmiennych ilościowych zamiast drzewa klasyfikacyjnego można równie efektywnie stosować analizę skupień.

Innym rodzajem algorytmu konstrukcji danych syntetycznych jest metoda **statystycznego zaciemniania zachowującego informacje** (ang. *information preserving statistical obfuscation* – IPSO). Generuje ona sztuczne wartości danych syntetycznych, zachowując jednocześnie wartości określonych statystyk i esencję wniosków statystycznych. Zakłada się tutaj, że zbiór danych składa się z dwóch podzbiorów zmiennych: macierzy zmiennych poufnych $\mathbf{Y}$ i macierzy zmiennych niebędących poufnymi $\mathbf{X}$. Zmienne zawarte w macierzy $\mathbf{Y}$ uznaje się za zależne, a zmienne z macierzy $\mathbf{X}$ – za niezależne. Klasycznym celem IPSO jest syntetyzacja subbazy $\mathbf{Y}$ (w wyniku której powstaje zbiór $\mathbf{Y}'$). Możemy wtedy udostępniać albo tylko w pełni syntetyczny zbiór $\mathbf{Y}'$, albo częściowo syntetyczne połączenie $\mathbf{Y}'$ i $\mathbf{X}$. Kluczową kwestią jest tutaj to, aby w modelu regresji wielorakiej $\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$ syntetyczne wartości $\mathbf{Y}'$ dawały takie same (lub bardzo zbliżone) oszacowania parametrów oraz błędów standardowych (i macierzy kowariancji), jak gdyby zastosowano oryginalny zbiór $\mathbf{Y}$. Najpierw zatem konstruuje się model regresyjny oparty na danych oryginalnych, po czym na jego podstawie wyznacza się teoretyczne wartości $\hat{\mathbf{Y}}$. Następnie do $\hat{\mathbf{Y}}$ dodawany jest szum o rozkładzie normalnym, w wyniku czego otrzymujemy syntetyczne wartości $\mathbf{Y}' = \hat{\mathbf{Y}} + \boldsymbol{\Theta}$. Jednak użytkownik będzie badał model $\mathbf{Y}' = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$, w którym oszacowania parametrów $\boldsymbol{\beta}$ (czyli $\hat{\boldsymbol{\beta}}$) i macierzy kowariancji składnika losowego $\hat{\boldsymbol{\Sigma}}$ z dużym prawdopodobieństwem będą się różniły od tych, które uzyskano by na podstawie danych oryginalnych. Dlatego też w odpowiednich algorytmach uwzględnia się często stosowne kroki

korekcyjne (np. dodawanie do $\hat{\mathbf{Y}}$ szumu skorelowanego – zachowującego oryginalną macierz korelacji), które znoszą te rozbieżności (Hundepool i in., 2012).

Metoda IPSO umożliwia zatem zachowanie wartości określonych parametrów i statystyk z danych oryginalnych. Cechuje ją też łatwość wdrażania do rozwiązań bardziej złożonych. Opiera się ona jednak na założeniu normalności rozkładu zmiennych, które często w praktyce nie jest spełnione.

Symulacja danych jest dość powszechnie stosowana do generowania sztucznych danych w celu przeprowadzenia analiz empirycznych, takich jak testowanie hipotezy lub szacowanie konkretnych statystyk, albo weryfikacji efektywności danego modelu, estymatora czy algorytmu analitycznego. Istotą symulacji jest duża liczba powtarzających się losowych procesów próbkowania mających na celu uzyskanie określonych wartości liczbowych i wyników (np. szacowanie gęstości, bootstrapping itp.). Procesy symulacyjne mogą być jednak wykorzystywane także do tworzenia sztucznych danych jako danych syntetycznych. Na przykład z *k*-wymiarowego rozkładu normalnego moglibyśmy – przy użyciu odpowiedniego generatora liczb losowych – wygenerować *k* wektorów X_1, X_2, …, X_k o rozmiarze *n* (gdzie *k* i *n* są liczbami naturalnymi) i otrzymać w ten sposób syntetyczne dane złożone z *n* syntetycznych jednostek i *k* syntetycznych zmiennych. Kwestią arbitralnej decyzji pozostaje tutaj ustalenie wektora wartości oczekiwanych i macierzy kowariancji takiego rozkładu. Jeżeli parametry te są całkowicie oderwane od rzeczywistości, to ryzyko ujawnienia nie istnieje. Jak już wspomniano, takie nieme pliki mogą być przydatne w celach testowych. Jednak można również tu wykorzystać stosowne wielkości wskazanych wyżej parametrów wyznaczone na podstawie danych oryginalnych oraz zbadać (za pomocą odpowiednich testów), czy można rozkład *k* określonych zmiennych uznać za wielowymiarowo normalny. Jeśli faktycznie wyjściowy rozkład taki jest, to można drogą symulacji wygenerować dane syntetyczne, w dużym stopniu odzwierciedlające informacje faktycznie zebrane. Nawet jeżeli ograniczymy się tylko do identyczności parametrów, to otrzymany symulacyjnie zbiór też może być użyteczny. Oczywiście ta droga dotyczy zmiennych ilościowych. W wypadku zmiennych jakościowych odpowiednie wartości losuje się z oszacowanego rozkładu wielomianowego lub tworzy na podstawie CART.

Tak więc procesy symulacyjne są łatwe do zrozumienia i wdrożenia, mogą przy tym prowadzić do otrzymania całkowicie bezpiecznych danych, umożliwiają też znaczne zachowanie oryginalnej wartości analityczne zgromadzonych informacji. Jednak często nie pozwalają na zaspokojenie złożonych potrzeb analitycznych.

Kolejną metodą jest **głębokie uczenie** (ang. *deep learning*). Należy do klasy metod uczenia maszynowego, a zatem opiera się na narzędziach sztucznej inteligencji. Pozwala efektywnie określać modele predykcyjne dla bardzo dużych zbiorów danych. Najpopularniejsze podejście w tym zakresie polega na wykorzystaniu generatywnej sieci kontradyktoryjnej (ang. *generative adversarial network* – GAN).

Model próbuje poznać podstawową strukturę oryginalnych danych, generując nowe dane (a dokładniej – nowe próbki) z tego samego rozkładu statystycznego co dane oryginalne. Główną ideą GAN jest posiadanie dwóch konkurencyjnych modeli sieci neuronowych: generatora (pobierającego szum lub wartości losowe jako dane wejściowe i generującego próbki) oraz dyskryminatora (próbującego odróżnić dane z generatora od danych uczących). Innymi słowy, dyskryminator pobiera dane rzeczywiste i oblicza prawdopodobieństwo oryginalności, które jest porównywane z arbitralnie ustaloną wartością progową. Jeżeli rzeczone prawdopodobieństwo jest niższe od tego progu, to dane uważa się za wygenerowane, w przeciwnym razie – za rzeczywiste. Proces uczenia jest procesem iteracyjnym, w którym dwie sieci toczą ciągłą grę, podczas której generator uczy się wytwarzać coraz bardziej realistyczne próbki, a dyskryminator uczy się coraz lepiej odróżniać wygenerowane dane od danych rzeczywistych. Taka nauka wzajemnym kosztem ma pozwolić osiągnąć równowagę w czasie. Generatywna sieć kontradyktoryjna jest procesem dość złożonym, wymagającym wiedzy na temat sieci neuronowych, nie zawsze też przejrzystym. Metoda okazuje się też czasochłonna i wymaga dużych zasobów obliczeniowych, jest także wrażliwa na przerwy i awarie. Trudne bywa też tutaj modelowanie zmiennych jakościowych.

Z kolei **metoda pseudowiarygodności** (ang. *pseudo likelihood*) polega na generowaniu danych syntetycznych dla populacji z uwzględnieniem wag wynikających z losowania próby w odpowiednich modelach (Kim i in., 2021). Po oszacowaniu gęstości rozkładów określonych zmiennych dla skończonej populacji syntezator może generować w pełni syntetyczne populacje poprzez wielokrotne losowanie z owych rozkładów stosownych wartości. Można tu stosować algorytm Monte Carlo łańcucha Markowa (ang. *Markov chain Monte Carlo* – MCMC) z odpowiednimi rozkładami warunkowymi. Podejście to uwzględnia specyfikę losowania próby i jego wpływ na jakość finalnych rezultatów. Poza tym dostarczenie użytkownikowi danych dla populacji jest dla niego zazwyczaj wygodniejsze. Problemem może być w tym wypadku jedynie wybór do MCMC rozkładu *a priori*.

Ryzyko ujawnienia informacji wrażliwych może być oceniane z wykorzystaniem różnorodnych reguł i miar opisanych w tej książce, w zależności od charakteru zmiennych. Istnieją także pewne mierniki użyteczności danych syntetycznych skonstruowane specjalnie do oceny wartości informacyjnej takich danych (Snoke i in., 2018). Dodatkowo warto tutaj poświęcić nieco więcej uwagi tzw. **różnicowaniu prywatności** (ang. *differential privacy* – DP). Od niedawna odgrywa ona bowiem w ochronie danych wrażliwych (zwłaszcza statystycznych) niepoślednią rolę. Jest ona właściwie ideą wspierającą matematyczne ramy kontroli ujawniania, swoistą alternatywę dla tradycyjnych ram kontroli ujawniania. Dwa główne typy DP to: jednoparametrowy DP_{ε} i słabszy dwuparametrowy $DP_{\varepsilon-\delta}$. Najpowszechniejszy jest wariant DP_{ε}. Parametr ε to arbitralnie określany stopień

ochrony przed identyfikacją jednostki, jaki oferuje dana metoda zgodna z tymże wariantem, czyli górny limit informacji, który może zostać w określony sposób bezpiecznie ujawniony. Najczęściej jest to związane w wielkością nakładanych zakłóceń. Koncepcję DP można zobrazować następująco. Załóżmy, że ktoś powiedział, że zarobki sąsiada są dwukrotnie większe od średniej ogólnej w jego powiecie. Jeśli zatem zajrzeć do odpowiednich danych statystycznych, to łatwo można wyciągnąć wniosek co do faktycznych dochodów owego sąsiada. Warto przy tym zauważyć, że ów sąsiad wcale nie musiał przekazywać komukolwiek informacji o swoich zarobkach, a zatem nikt nie może zapobiec ujawnieniu tej informacji. Wobec tego przyjęta metoda SDC i udostępniania nie musi być „odpowiedzialna" za ujawnienie informacji wrażliwej. Potrzebna jest natomiast ocena jej roli w tym zakresie poprzez porównanie tego, co jej zastosowanie może ujawnić o jednostce, gdy informacje na jej temat są obecne w udostępnianym zbiorze danych, z tym, co może się stać na jej temat jawne, gdy w bazie ona nie występuje lub nie podała do niej stosownych informacji. Oczywiście, im mniejszy wkład jednostki w dany zbiór, tym ryzyko ujawnienia informacji wrażliwej drogą DP jest mniejsze.

Na przykład, zaokrąglanie arbitralne lub losowe do odpowiedniej wielokrotności ustalonego mnożnika nie zapewnia DP (nie da się bowiem zagwarantować, że w innym wariancie udostępnianego zbioru zaokrąglona wartość dla tej samej wartości oryginalnej nie będzie inna, co może prowadzić do odkrycia informacji wrażliwej), ale zaokrąglanie danej liczby do najbliższej wielokrotności owego mnożnika – już tak. Metoda zapewniająca DP musi mieć zatem składnik losowy, ale dobrany w odpowiedni sposób – na ogół tak, aby zakres możliwych zakłóceń mieścił się w pewnych granicach określonych poziomem ochrony ε. Autorzy wielu prac z tego zakresu proponują dodanie w tym celu do danej oryginalnej szumu wygenerowanego z rozkładu Laplace'a. Parametr prywatności ε jest tu bezpośrednio powiązany z wariancją bazowego rozkładu Laplace'a. Oprócz tego – w przeciwieństwie do innych rodzajów nakładania szumu – można bezpiecznie ujawnić wariancję zastosowanego rozkładu Laplace'a (lub równoważnie ε). Problemem może być w tym wypadku interpretacja uzyskanych wyników, wymagająca większego doświadczenia. Szerzej – w tym bardziej formalnie – o idei DP pisali m.in. Dwork i Smith (2009).

Tworzenie danych syntetycznych z zastosowaniem różnicowania prywatności może przebiegać na przykład przy zastosowaniu metody „histogramu". Składa się ona z następujących kroków:

- pogrupowania wszystkich zmiennych w odpowiednie kategorie,
- utworzenia tabeli kontyngencji obejmującej wszystkie kombinacje (histogramu),
- dodania szumu z rozkładu Laplace'a do każdej komórki tabeli,
- zaokrąglenia wyników do liczb całkowitych i odtworzenia oryginalnych danych.

Metoda ta spełnia założenie DP, ponieważ dana jednostka może wystąpić tylko w jednej komórce tej tabeli. W wypadku większej liczby zmiennych należy się jednak liczyć z tym, że otrzymana tablica będzie zawierała wiele komórek zerowych. Dlatego alternatywnie stosuje się podejście oparte na statystykach dostatecznych (tzn. zapewniających redukcję danych bez utraty informacji o estymowanej wielkości). Na przykład można skonstruować model syntetyzacyjny na podstawie zestawu danych brzegowych (czyli np. sum określonych wielkości dla jednostek na przyjętych poziomach agregacji) z szumem nałożonym na nie w ten sposób, aby spełniona była reguła DP. Należy to czynić odrębnie dla każdej oddzielnej konstrukcji modelu. Jeżeli jednak model taki zostanie skonstruowany dla dwóch zbiorów z różnych edycji danego badania, to oba utworzone tą drogą zbiory syntetyczne mogą być swobodnie udostępniane.

Do tworzenia danych syntetycznych spełniających wymóg DP można także używać algorytmów iteracyjnych, opartych np. na sieciach GAN (Jordon i in., 2019). Wartość ε jest wówczas sumą odpowiednich wartości progowych z każdego kroku. Liczba kroków musi być wszakże ustalona z góry, co pozwala zachować właściwość DP.

Metody oparte na DP nie muszą zapewniać wystarczającej ochrony informacji wrażliwych. Wszystko zależy od właściwego zaprojektowania tworzenia danych syntetycznych i procesu SDC w danym przypadku. Trudności w tym zakresie mogą wynikać np. z dużej liczby zmiennych, małej liczby rekordów lub znacznej liczby możliwych wartości zmiennych w oryginalnym zbiorze. Dodatkowo, jeśli parametr prywatności ε jest nadmiernie duży, a zastosowana metoda przetwarza dane w minimalnym stopniu, to poziom zapewnionej ochrony może się pogorszyć do tego stopnia, że stanie się ona bezsensowna.

Więcej informacji na temat różnorodnych metod generowania danych syntetycznych podają np. Gjaltema i in. (2022), Drechsler (2011) czy El Emam i in. (2020). W rozdziale 5 zasygnalizowano możliwości przeprowadzenia tejże konstrukcji przy użyciu określonych narzędzi informatycznych.

4

Strata informacji

Immanentną cechą kontroli ujawniania danych jest wprowadzanie niepewności odnośnie do prawdziwej wartości udostępnianych zmiennych. Niepewność owa w naturalny sposób wiąże się z ukryciem lub zmianą wartości w rekordzie zbioru mikrodanych czy w komórce tablicy. W konsekwencji na skutek zastosowania SDC powstaje określona strata informacji źródłowej, która może wpłynąć na jakość udostępnianych danych oraz obliczeń i szacunków dokonywanych przez ich użytkownika. Stąd użytkownik ów powinien razem z danymi otrzymywać wiedzę na temat oczekiwanej wielkości owej straty spowodowanej przez SDC. Jak zaznaczono we wprowadzeniu, minimalizacja owej straty jest drugim – obok minimalizacji ryzyka identyfikacji jednostki i ujawniania informacji wrażliwych – kryterium optymalizacji SDC. Oznacza to udostępnianie mikrodanych oraz publikowanie tablic i analiz w możliwie największym niezmienionym zakresie. Konieczny staje się więc efektywny pomiar wielkości straty informacji.

Jak już zasygnalizowano w poprzednim rozdziale, w odniesieniu do procesu kontroli ujawniania wyników analiz na jego wyjściu otrzymujemy informację dwojakiego rodzaju. Albo dane wynikowe nie stwarzają zagrożenia ujawnienia informacji poufnej i mogą zostać przekazane użytkownikowi zewnętrznemu, albo mogą stanowić potencjalne zagrożenie – wówczas nie mogą zostać przekazane. W wypadku stwierdzenia potencjalnego zagrożenia dopuszczalne jest usunięcie tych elementów wyników analiz, które stwarzają ryzyko dla ochrony poufności, i udostępnienie reszty z nich albo zwrócenie się z prośbą do analityka, by ponownie przeprowadził swe prace, tak by ich efekty nie stanowiły dłużej zagrożenia. Co jednak jest najważniejsze, analityk ma dostęp do pełnego zestawu mikrodanych, pozbawionych jedynie identyfikatorów, i to na ich podstawie przeprowadza badanie empiryczne z użyciem odpowiednich metod. Użyteczność takich zasobów nie jest w żaden sposób ograniczona. Proces SDC jest przeprowadzany dopiero na samym końcu, kiedy dane wynikowe mają zostać przekazane użytkownikowi zewnętrznemu, który pracował na danych jednostkowych w chronionych warunkach i pod pełną kontrolą gestora danych. Ponadto analityk ma też świadomość co do zawartości informacyjnej wszelkiej maści wyników analiz przed ich uzyskaniem, nie zachodzi więc dodatkowa konieczność szacowa-

nia/pomiaru ich użyteczności na etapie, gdy są one przekazywane poza chronione środowisko pracy.

W związku z przedstawionymi tu przyczynami w niniejszym rozdziale omówiono miary straty informacji dla danych jednostkowych oraz dla danych zagregowanych do postaci tablic statystycznych. To właśnie takiej postaci dane wynikowe, poddane procesowi SDC przed ich opublikowaniem bądź udostępnieniem przez gestora danych, będą dla analityków podstawą do przeprowadzenia badań, a – wskutek zapewnienia ochrony tajemnicy statystycznej – ich użyteczność zostanie zredukowana względem zestawów danych w oryginalnej postaci. Z tej przyczyny nie każda analiza będzie mogła zostać na nich przeprowadzona i właściwie dobrane miary straty informacji powinny być dla użytkownika wskazówką, w jakim zakresie udostępnione mu zasoby może wykorzystać.

W rozdziale tym omówiono kolejno: koncepcję i najważniejsze grupy miar straty informacji, ich konstrukcję oraz istotność dla różnych rodzajów przeznaczenia udostępnianych danych, w tym oceny jakości estymacji.

4.1. Istota straty informacji[16]

Pojęcie straty informacji na skutek zastosowania metod ochrony poufności jest pojęciem subiektywnym. Istotne w tym kontekście staje się to, w jakim celu i do jakich analiz będą stosowane opublikowane tablice wynikowe. Z punktu widzenia decyzji, które muszą zostać podjęte w celu zachowania poufności oraz konsekwencji ich wyboru, strata informacji może być oceniana w różny sposób. Shlomo i Young (2006) wskazują na trzy główne grupy oceny rzeczonej straty, w zależności od statystycznego aspektu rozpatrywanych danych. Czynią to co prawda z myślą o tablicach częstości, ale klasyfikację tę bez trudu można zastosować również do innego rodzaju informacji statystycznych (mikrodanych, tablic wielkości, prezentacji wynikowych itp.). Grupy miar oceny straty informacji można zatem podzielić na miary: zakłócenia rozkładu, wpływu na wariancję szacunków i wpływu na siłę związku.

- Miary zakłócenia rozkładu są oparte na metrykach odległości pomiędzy rzeczywistymi a zmienionymi wartościami zmiennych. Jeśli na przykład jednostką podstawową jest dana jednostka geograficzna, to dla każdej jednostki z osobna mierzona jest odległość pomiędzy wartościami oryginalnymi a zmodyfikowanymi, po czym oblicza się średnią tychże odległości.
- W miarach wpływu na wariancję szacunków bierze się pod uwagę różnice pomiędzy wariancjami dla przeciętnych wartości określonych podzbiorów lub całego zbioru (w wypadku tablic – kolumn, wierszy lub całej tablicy)

[16] Podrozdziały 4.1–4.4 stanowią zmodyfikowaną i ulepszoną wersję punktów 2–5 artykułu Młodaka (2020).

przed procesem SDC i po jego przeprowadzeniu. Innym sposobem w tym zakresie jest jednoczynnikowa analiza wariancji ANOVA dla wybranej zmiennej zależnej względem wybranych niezależnych zmiennych jakościowych. Miarą straty jest wówczas porównanie, jak zmieniają się komponenty współczynnika determinacji R^2 (poprzez podział na wariancję wewnątrzgrupową i międzygrupową) dla zbioru źródłowego opartego / tablicy źródłowej opartej na danych rzeczywistych oraz dla zbioru zmienionego / tablicy zmienionej poprzez zastosowanie danej metody ochrony poufności. Zastosowanie owo może spowodować utratę homogeniczności grup, tzn. wariancja międzygrupowa może maleć, a wewnątrzgrupowa – rosnąć. Możliwa jest też sytuacja odwrotna, czyli wzrost wariancji międzygrupowej przy spadku wariancji wewnątrzgrupowej.

– Miary wpływu na siłę związku bazują na analizie oddziaływania SDC na kierunek i siłę związku między określonymi zjawiskami w porównaniu z tymi powiązaniami dla oryginalnych zmiennych. Do oceny skali odmienności w tym względzie można wykorzystać współczynniki korelacji, często jednak wykonuje się także test niezależności pomiędzy odpowiednimi wymiarami w stosownych przekrojach. Oznacza to badanie niezależności przy użyciu pewnej tablicy kontyngencji. Możliwe są także inne podejścia w tym zakresie.

Można zauważyć, że mierniki zakłócenia rozkładu są użyteczne, gdy korzystającego z udostępnionych danych interesują rozkłady określonych zjawisk (np. w czasie i przestrzeni). Wpływ SDC na wariancję szacunków ma istotne znaczenie w wypadku estymacji określonych wielkości dla populacji. Natomiast obserwowanie zmiany siły związków na skutek zastosowania SDC może być szczególnie ważne podczas analizy współzależności zjawisk. Warto wszakże zauważyć, że niektóre rodzaje badań mogą wymagać mierników straty idących w więcej niż jednym kierunku (np. podczas estymacji dokonywanej z wykorzystaniem estymatorów regresyjnych pożądany będzie pomiar straty informacji zarówno w kierunku wpływu na wariancję szacunków, jak i na siłę związku między określonymi zjawiskami). W kolejnych podrozdziałach opisano szczegółowo konkretne metody oceny straty informacji należące do tych grup.

4.2. Miary zakłócenia rozkładu

Jak wspomniano, pomiar straty informacji w kontekście zakłócenia rozkładu oryginalnych zmiennych wskutek zastosowania SDC opiera się na unormowanych różnicach między odpowiednimi wartościami w zbiorze danych oryginalnych oraz w zbiorze danych zniekształconych. Należy przy tym uwzględnić skalę

pomiarową, na jakiej mierzone są poszczególne obserwacje, oraz dopuszczalność wykonywania na nich odpowiednich działań arytmetycznych.

Niech zatem zbiory danych liczą n obserwacji i m zmiennych (gdzie n i m to liczby naturalne). Oznaczmy przez x_{ij} wartość obserwacji zmiennej X_j dla jednostki i, a przez x_{ij}^* – odpowiednią wartość w zbiorze powstałym w wyniku zastosowania metod SDC, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Ogólna postać miary straty może być zatem następująca:

$$\lambda = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} d(x_{ij}, x_{ij}^*)}{mn}, \tag{4.1}$$

gdzie $d(\cdot,\cdot)$ jest miarą odległości spełniającą klasyczne warunki zwrotności, symetrii i nierówności trójkąta, przyjmującą wartości należące do przedziału [0, 1]. Wartości wskaźnika λ także należą wtedy do tego przedziału, przy czym im większa wartość, tym dokuczliwsza strata informacji. Sytuacja, gdy $\lambda = 0$ – tzn. w zbiorze nie ma żadnych zmian – jest w SDC oczywiście tylko teoretyczna.

Formuła (4.1) daje też możliwość wyznaczenia straty informacji dla poszczególnych zmiennych. Dla zmiennej X_j strata ta wyniesie:

$$\lambda_j = \frac{\sum_{i=1}^{n} d(x_{ij}, x_{ij}^*)}{n}, \quad j = 1, 2, \ldots, m.$$

Definicja miary d zależy od skali pomiarowej, na której wyraża się wartości zmiennej X_j. Jeśli obserwacje tej zmiennej są mierzone na skali nominalnej, to odległość owa wynosi:

$$d(x_{ij}, x_{ij}^*) = \begin{cases} 0, \text{ gdy } x_{ij} = x_{ij}^*, \\ 1, \text{ gdy } x_{ij} \neq x_{ij}^*. \end{cases} \tag{4.2}$$

Jeżeli wartości zmiennej X_j są mierzone na skali porządkowej, to miarą ową jest relacja

$$d(x_{ij}, x_{ij}^*) = \frac{r(x_{ij}, x_{ij}^*)}{k_j - 1}, \tag{4.3}$$

gdzie $r(x_{ij}, x_{ij}^*)$ oznacza liczbę kategorii zmiennej X_j, o którą różnią się wartości x_{ij} i x_{ij}^*, a k_j – liczbę kategorii, które może przyjmować zmienna X_j.

Dla zmiennej ilościowej odległość ta może być np. znormalizowaną wartością bezwzględną lub znormalizowanym kwadratem różnicy między wartością ze zbioru oryginalnego a odpowiednią wartością ze zbioru ukształtowanego w wyniku SDC, czyli

$$d(x_{ij}, x_{ij}^{*}) = \frac{\left|x_{ij} - x_{ij}^{*}\right|}{\max\limits_{k=1,2,\ldots,n}\left|x_{kj} - x_{kj}^{*}\right|} \tag{4.4}$$

lub

$$d(x_{ij}, x_{ij}^{*}) = \frac{(x_{ij} - x_{ij}^{*})^2}{\max\limits_{k=1,2,\ldots,n}(x_{kj} - x_{kj}^{*})^2}, \tag{4.5}$$

$i = 1, 2, \ldots, n, j = 1, 2, \ldots, m$.

Młodak (2020) zauważył, że miary (4.4) i (4.5) nie są funkcjami rosnącymi ze względu na poszczególne cząstkowe straty informacji. Innymi słowy, jeśli na przykład dla pewnego $i \in \{1, 2, 3, \ldots, n\}$ wartość $\left|x_{ij} - x_{ij}^{*}\right|$ zwiększy się, a wszystkie wartości $\left|x_{hj} - x_{hj}^{*}\right|$ dla $h \neq i$ pozostają takie same, to wartość miernika powinna wzrosnąć. Tymczasem w wypadku wskazanych formuł tak nie będzie. Jeśli bowiem dla tegoż i wskazana bezwzględna różnica (lub kwadrat różnicy, odpowiednio) między wartością oryginalną a wartością po SDC osiągnie maksimum, to cząstkowa strata informacji dla i pozostanie bez zmian – wyniesie jeden, a dla pozostałych jednostek okaże się mniejsza. W rezultacie otrzymamy mniejszą wartość miernika, podczas gdy strata informacji tak naprawdę się zwiększyła. Dlatego Młodak zaproponował jeszcze inną miarę, pozwalającą na zniwelowanie tego problemu, mianowicie:

$$d(x_{ij}, x_{ij}^{*}) = \frac{2}{\pi}\operatorname{arctg}\left|x_{ij} - x_{ij}^{*}\right|, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m. \tag{4.6}$$

Z kolei w innej pracy Młodaka (2019) dyskutowano także możliwości wykorzystania do oceny straty informacji mierników kompleksowych.

Stosowanie tych miar pociąga za sobą pewne problemy. Jednym z narzędzi SDC dla zmiennych jakościowych jest bowiem przekodowywanie. Wówczas liczby kategorii takiej przekodowywanej zmiennej w zbiorze oryginalnym i w zbiorze powstałym w wyniku SDC będą różne. Zatem w pierwszym rzędzie należy zadbać o to, aby numery kategorii pozostawionych bez zmian w obu wariantach były identyczne. Na przykład, jeśli przed przekodowaniem zmienna X_j liczyła $k_j = 8$ kategorii oznaczonych jako 1, 2, 3, 4, 5, 6, 7, 8, a w wyniku przekodowania połączono kategorie 2 i 3 oraz 6 i 7, to nowe kategorie powinny mieć odpowiednio numery 1, 2, 4, 5, 6, 8. Wtedy podejście (4.3) stosuje się i w tej sytuacji.

Odmienny problem pojawia się w wypadku ukrywania. W zbiorze poddanym SDC powstają bowiem luki w danych. Przyjmijmy w takiej sytuacji, że dane o zmiennej X_j są wyrażone na skali nominalnej. Wówczas jeżeli obserwacja x_{ij}^{*} w równości (4.2) jest ukryta, to przyjmujemy $d(x_{ij}, x_{ij}^{*}) = 1$. Jeśli obserwacje zmiennej X_j wyrażają się na skali porządkowej, to dla celów obliczeniowych we

wzorze (4.3) ukrytej wartości $x_{ij}^{\star}$ przyporządkowujemy arbitralnie $x_{ij}^{\star} := 1$, gdy x_{ij} jest bliższe k_j, lub $x_{ij}^{\star} := k_j$, gdy x_{ij} jest bliższe 1. Jeżeli zmienna X_j ma charakter ilościowy, to w obliczeniach podstawiamy $x_{ij}^{\star} := \max\limits_{k=1,2,\ldots,n} x_{kj}$, gdy $x_{kj} \leq \operatorname{med}\limits_{k=1,2,\ldots,n} x_{kj}$, oraz $x_{ij}^{*} := \min\limits_{k=1,2,\ldots,n} x_{kj}$, gdy $x_{kj} > \operatorname{med}\limits_{k=1,2,\ldots,n} x_{kj}$, $j = 1, 2, \ldots, m$. Pozwala to na uzyskanie wyraźnego obrazu powstałych różnic.

Miara postaci (4.1) jest uogólnieniem i normalizacją miar proponowanych przez Domingo-Ferrera i in. (2001). Oprócz tego – dla zmiennych ilościowych – autorzy ci zasugerowali także miary (niekoniecznie ograniczone od góry) oparte na różnicy między średnimi:

$$\lambda = \frac{1}{m}\sum_{j=1}^{m}\left|\overline{x}_j - \overline{x}_j^{\star}\right| \tag{4.7}$$

lub

$$\lambda = \frac{1}{m}\sum_{j=1}^{m}(\overline{x}_j - \overline{x}_j^{\star})^2 \tag{4.8}$$

lub

$$\lambda = \frac{1}{m}\sum_{j=1}^{m}\frac{\left|\overline{x}_j - \overline{x}_j^{\star}\right|}{\left|\overline{x}_j\right|}, \tag{4.9}$$

gdzie $\overline{x}_j = \sum_{i=1}^{n} x_{ij} / n$, a $\overline{x}_j^{\star} = \sum_{i=1}^{n} x_{ij}^{\star} / n$ to średnie arytmetyczne wartości zmiennej X_j odpowiednio przed i po zastosowaniu SDC, $j = 1, 2, \ldots, m$.

Warto też wspomnieć, że w wypadku stosowania metody PRAM (do zmiennych jakościowych, rzecz jasna) stratę informacji można obliczać także za pomocą wskaźnika entropii postaci (Domingo-Ferrer i in., 2001; Młodak, 2020):

$$\mathrm{EBIL}_j = \sum_{i=1}^{n}\mathcal{H}(X_j|X_j^{\star} = q), \tag{4.10}$$

gdzie

$$\mathcal{H}(X_j|X_j^{\star} = q) = -\sum_{r=1}^{k_j} P(X_j = r|X_j^{\star} = q)\log P(X_j = r|X_j^{\star} = q),$$

k_j jest liczbą kategorii zmiennej X_j, a $\mathbf{P}_j = \left[P(X_j = r|X_j^{\star} = q)\right]$, $r, q = 1, 2, \ldots, k_j$, to macierz prawdopodobieństw przejść Markowa w PRAM dla tejże zmiennej. Skrót EBIL pochodzi od angielskiego określenia *entropy-based information loss measure* – miara straty informacji oparta na entropii. Im wyższa wartość entropii, tym strata informacji jest większa.

Opisane podejścia mają mocne i słabe strony. Do pierwszych należy przede wszystkim pełne odzwierciedlenie odmienności między danymi oryginalnymi a udostępnianymi po wykonaniu SDC, zwłaszcza przez miarę postaci (4.1). Dzięki bezwzględnym różnicom między poszczególnymi obserwacjami każde ukrycie bądź zniekształcenie informacji staje się istotnym składnikiem straty łącznej. Jest to bardzo użyteczne do oceny skali wprowadzonych ingerencji oraz ich wpływu na rozkłady poszczególnych zmiennych. W szacowaniu określonych wielkości dla populacji oraz – dla danych wyrażonych na skali ilorazowej – w wyznaczaniu wskaźników strata w tym ujęciu będzie jednak najczęściej przeszacowana. Chodzi mianowicie o to, że jednostkowe różnice w procesie sumowania mogą się w określonym stopniu wzajemnie niwelować, wskutek czego faktyczne odchylenie szacunków od ich wielkości, które mogłyby zostać uzyskane, gdyby SDC nie zastosowano, będą niezbyt duże. Podobnie rzecz ma się dla wskaźników: zbliżone straty jednostkowe dla obu zmiennych będących podstawą wyznaczania wskaźnika powodują, że wartości owego wskaźnika będą bardzo bliskie wartości, które można by otrzymać przy użyciu danych oryginalnych. Dobrze widać to choćby na przykładzie miar wyrażonych wzorami (4.7)–(4.9), które w pewnym stopniu ten problem zmniejszają (mimo istotnych odchyleń indywidualnych średnie mogą się różnić nieznacznie), jednak nie dają zbyt dużo informacji o wpływie straty na jakość estymacji.

Innym problemem w stosowaniu tych miar jest kwestia braków danych. Wiele metod SDC (np. w ramach mikroagregacji) może wypełniać takie luki w mikrodanych informacjami przez siebie generowanymi, co w istocie jest pewną formą imputacji. Pominięcie tego rodzaju sytuacji w ogólnym rozrachunku straty może się okazać zbyt kosztowne. Najlepszym rozwiązaniem byłoby zatem przyjęcie takich „oryginalnych" wartości owych cech, aby odpowiednie jednostkowe straty $d(x_{ij}, x_{ij}^{\star})$ były możliwie największe (minimalizuje to ryzyko utraty istotnej wiedzy w tym zakresie). Zależeć one będą od skali pomiarowej, na której są wyrażone dane dotyczące rozpatrywanej zmiennej (Młodak, 2020):

- dla danych wyrażonych na skali nominalnej: „oryginalna" kategoria x_{ij} powinna być inna niż otrzymana w SDC ($x_{ij}^{\star}$),
- dla danych wyrażonych na skali porządkowej: „oryginalna" kategoria x_{ij} powinna być odległa od nałożonej w SDC ($x_{ij}^{\star}$) o maksymalną możliwą liczbę kategorii w myśl danej kategoryzacji,
- dla danych wyrażonych na skalach: różnicowej lub ilorazowej: jeśli wartość otrzymana z SDC ($x_{ij}^{\star}$) jest bliższa maksymalnej wartości zmiennej dla danych wejściowych, to za „oryginalną" wielkość x_{ij} przyjmujemy minimum faktycznych zgromadzonych wartości zmiennej ($\min_{k \in Z_j} x_{kj}$), w przeciwnym razie – maksimum ($\max_{k \in Z_j} x_{kj}$), gdzie $Z_j \subseteq \{1, 2, \ldots, n\}$ – jest zbiorem tych jednostek, dla których otrzymano dane z zakresu zmiennej X_j.

Podobnie postępujemy, gdy znamy dane oryginalne, a w SDC dla mikrodanych zastosowano metody niezakłóceniowe, wskutek czego wrażliwe dane zostały ukryte. Wtedy musimy przyjąć pewne „robocze" wartości odpowiednich danych po zastosowaniu SDC, które umożliwią otrzymanie możliwie największych indywidualnych strat $d(x_{ij}, x_{ij}^*)$. Oznacza to, że (Młodak, 2020):

- dla danych wyrażonych na skali nominalnej: „robocza" kategoria x_{ij}^* w danych po zastosowaniu SDC powinna być inna niż oryginalna x_{ij},
- dla danych wyrażonych na skali porządkowej: „robocza" kategoria x_{ij}^* w danych po zastosowaniu SDC powinna być odległa od oryginalnej x_{ij} o maksymalną możliwą liczbę kategorii w myśl danej kategoryzacji,
- dla danych wyrażonych na skalach różnicowej lub ilorazowej: jeśli wartość oryginalna jest bliższa maksymalnej wartości zmiennej dla danych wejściowych, to za „roboczą" wielkość x_{ij}^* po SDC przyjmujemy minimum faktycznych zgromadzonych wartości zmiennej ($\min_{k \in Z_j} x_{kj}$), w przeciwnym razie – maksimum ($\max_{k \in Z_j} x_{kj}$).

Gdy z kolei stosujemy SDC dla tablic, wówczas, stosując metody oparte na ukrywaniu komórek, do porównań trzeba zasymulować tablicę, w której dla brakujących komórek ich wartości zostaną zaimputowane. W wypadku ukrywania komórek strata informacji zostanie wyrażona jako suma kosztów poniesionych wskutek wtórnego ukrycia komórek. Problemem pozostaje to, czy waga każdej komórki w tablicy jest taka sama, czy też komórki o większej wartości mają wagę większą. W praktyce ukrycie zbyt dużej liczby komórek o wysokich wartościach może znacznie obniżyć użyteczność publikowanych danych.

W zależności od preferencji i potrzeb użytkowników zagadnienie straty informacji może być różnie wyrażone. W ten sposób da się wpłynąć na działanie algorytmu wyboru komórek do wtórnego ukrycia. Hundepool i in. (2012) wskazali najbardziej popularne kryteria brane pod uwagę przy formułowaniu funkcji kosztu dla ukrywania komórek:

- jednakowa waga dla wszystkich komórek – celem jest minimalizacja liczby wtórnie ukrytych komórek,
- liczba jednostek w agregacie, który komórka reprezentuje – prowadzi do poszukiwania możliwości ukrycia tylko takich komórek, które łącznie będą reprezentować jak najmniejszą liczbę jednostek,
- wartość komórki – optymalnym rozwiązaniem będzie pozostawienie w publikacji jak największej liczby komórek o największych wartościach.

W sytuacji występowania silnej asymetrii danych preferowanie zachowania komórek o największej wartości może prowadzić do zbyt dużej nierównowagi dla funkcji kosztu. W tym wypadku zalecana jest transformacja funkcji kosztu. Jednym z możliwych i często stosowanych podejść w tym zakresie jest transformacja potęgowa (Box i Cox, 1964):

$$y = \begin{cases} x^{\lambda} & \text{dla} \quad \lambda \neq 0, \\ \log(x) & \text{dla} \quad \lambda = 0. \end{cases}$$

W ochronie tablicy, w której wartości komórek są modyfikowane (metodami zakłóceniowymi), proponuje się miary straty oparte na odległości między wartościami komórek zmienionymi a pierwotnymi. Dla układu: tablica pierwotna – tablica zmieniona strata informacji będzie mierzona sumą odległości (wyznaczanych według formuły (4.4) lub (4.5) bądź podobnej) pomiędzy wartościami komórek zmienionych i pierwotnych. Problem braków danych ma wtedy mniejsze znaczenie – podczas wyznaczania wartości w komórkach można je pominąć lub wcześniej dokonać imputacji. Jedyny kłopot może występować wówczas, gdy nie będzie żadnych danych dotyczących kategorii wyznaczonej daną komórką. Wtedy trzeba albo zrezygnować z danej konstrukcji tablicy (np. poprzez połączenie pewnych jej kategorii w inną, bardziej zgrubną), albo też miarę straty oprzeć na pomiarze straty na poziomie mikrodanych, odpowiadających tejże komórce, dokonanym w opisany wcześniej sposób.

4.3. Miary wpływu na wariancję szacunków

Druga grupa miar straty informacji bazuje na wpływie zmian dokonywanych w wyniku zastosowania metod kontroli ujawniania danych na zmienność rozpatrywanych wielkości statystycznych. Wpływ ten można ocenić, stosując na przykład jednoczynnikową analizę wariancji ANOVA. Ma ona jednak dość ograniczone możliwości zastosowania, albowiem opiera się na podziale jednostek na grupy w myśl określonej klasyfikacji i na badaniu zmienności w tych grupach. W SDC podziału na grupy dokonuje się przede wszystkim, opierając się na predefiniowanych klasach lub podobieństwie rekordów pod względem zmiennych kluczowych.

Przykład takiego podejścia zaprezentowali Mateo-Sanz i Domingo-Ferrer (1998). Zakładamy zgodnie z nim, że zbiór n jednostek został podzielony na p (gdzie p jest liczbą naturalną, $p < n$) grup $G_1, G_2, \ldots, G_p$, z których każda liczy n_l jednostek, $l = 1, 2, \ldots, p$, $\sum_{l=1}^{p} n_l = n$. Każda jednostka i opisana jest zatem przez wektor $\boldsymbol{x}_i = (x_{i1}, x_{i1}, \ldots, x_{im})$, $i = 1, 2, \ldots, n$. Zgodnie z regułą minimalnej liczby respondentów przyjmuje się też, że $n_l \geq k$, $l = 1, 2, \ldots, p$, gdzie liczba naturalna $k < n$ oznacza arbitralnie ustaloną minimalną liczbę jednostek, które mogą należeć do każdej grupy.

Wtedy wewnątrzgrupowa suma kwadratów *SSE* (ang. s*um of squared errors of all observations vs respective means* – suma kwadratów błędów dla wszystkich obserwacji względem ich odpowiednich średnich grupowych) dana jest wzorem

$$SSE = \sum_{l=1}^{p} \sum_{i \in \{1,2,\ldots,n\},\, i \in G_l} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_l)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_l)^{\mathrm{T}},$$

gdzie $\overline{\boldsymbol{x}}_l = \sum_{i \in \{1,2,\ldots,n\},\, i \in G_l} \boldsymbol{x}_i / n_l$ to wektor średnich arytmetycznych badanych zmiennych w grupy G_l, $l = 1, 2, \ldots, p$.

Międzygrupowa suma kwadratów *SSA* (ang. *sum of squared errors of all treatment means vs grand mean* – suma kwadratów błędów dla wszystkich średnich grupowych w odniesieniu do średniej globalnej) to:

$$SSA = \sum_{l=1}^{p} n_l (\overline{\boldsymbol{x}}_l - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_l - \overline{\boldsymbol{x}})^{\mathrm{T}},$$

gdzie $\overline{\boldsymbol{x}} = \sum_{i=1}^{n} \boldsymbol{x}_i / n$ oznacza wektor średnich arytmetycznych badanych zmiennych ogółem.

Suma kwadratów ogółem (ang. *total sum of squares* – TSS) ma zatem postać:

$$TSS = SSA + SSE = \sum_{l=1}^{p} \sum_{i \in \{1,2,\ldots,n\},\, i \in G_l} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\mathrm{T}}.$$

Miarą straty informacji jest wówczas udział wewnątrzgrupowej sumy kwadratów *SSE* w sumie kwadratów ogółem, czyli:

$$\lambda = \frac{SSE}{TSS} = 1 - \frac{SSA}{TSS}. \tag{4.11}$$

Miara dana wzorem (4.11) przyjmuje wartości z przedziału [0, 1]. Im większa wartość, tym strata informacji znaczniejsza, gdyż zastąpienie faktycznych wartości dla jednostek należących do danej grupy przez stosowną średnią arytmetyczną zmniejsza zróżnicowanie wewnątrzgrupowe. Efektywne zastosowanie SDC zmierza do minimalizacji tej redukcji. Miara ta może być stosowana tylko do zmiennych wyrażonych na skali różnicowej lub ilorazowej.

Oceny wpływu zastosowanych metod SDC na zmienność rozpatrywanych informacji można też dokonywać, porównując wariancję/kowariancję badanych zmiennych przed i po dokonaniu kontroli ujawniania danych. Domingo-Ferrer i in. (2001) podali następujące przykłady takich miar:

– miary oparte na macierzy kowariancji zmiennych $X_1, X_2, \ldots, X_m$ lub tylko na ich elementach diagonalnych:
  - $\lambda = 2 \sum_{l=1}^{m} \sum_{j:1 \le j \le l} \frac{\left| v_{jl} - v_{jl}^{*} \right|}{m(m+1)}$,

- $\lambda = 2\sum_{l=1}^{m} \sum_{j:1 \le j \le l} \frac{(v_{jl} - v_{jl}^{*})^2}{m(m+1)}$,
- $\lambda = 2\sum_{l=1}^{m} \sum_{j:1 \le j \le l} \frac{\left|v_{jl} - v_{jl}^{*}\right|}{\left|v_{jl}\right| \cdot m(m+1)}$,
- $\lambda = \sum_{j=1}^{m} \frac{\left|v_{jj} - v_{jj}^{*}\right|}{m}$,
- $\lambda = \sum_{j=1}^{m} \frac{(v_{jj} - v_{jj}^{*})^2}{m}$,
- $\lambda = \sum_{j=1}^{m} \frac{\left|v_{jj} - v_{jj}^{*}\right|}{\left|v_{jj}\right| \cdot m}$,

gdzie v_{jl} i v_{jl}^{*} to kowariancje zmiennych X_j i X_l przed i po zastosowaniu SDC, $j, l = 1, 2, \ldots, m$,

– miary oparte na macierzy korelacji zmiennych $X_1, X_2, \ldots, X_m$:

- $\lambda = 2\sum_{l=1}^{m} \sum_{j:1 \le j \le l} \frac{\left|\rho_{jl} - \rho_{jl}^{*}\right|}{m(m+1)}$,
- $\lambda = 2\sum_{l=1}^{m} \sum_{j:1 \le j \le l} \frac{(\rho_{jl} - \rho_{jl}^{*})^2}{m(m+1)}$,
- $\lambda = 2\sum_{l=1}^{m} \sum_{j:1 \le j \le l} \frac{\left|\rho_{jl} - \rho_{jl}^{*}\right|}{\left|\rho_{jl}\right| \cdot m(m+1)}$,

gdzie ρ_{jl} i ρ_{jl}^{*} to współczynniki korelacji zmiennych X_j i X_l przed i po zastosowaniu SDC, $j, l = 1, 2, \ldots, m$.

Warto zauważyć, że miary oparte na macierzy korelacji zmiennych mogą być użyte także do danych wyrażonych na skali porządkowej – wówczas należy oprzeć je na współczynniku korelacji τ-Kendalla. Ponadto współczynniki bazujące na kowariancji i korelacji odnoszą się w znacznej mierze także do związków między określonymi zjawiskami. Mogą być zatem zaliczone również do miar wpływu na siłę związku.

4.4. Miary wpływu na siłę związku

Dla praktycznej efektywności zastosowania kontroli ujawniania danych bardzo istotne jest zachowanie kierunków i siły związków między badanymi zjawiskami, które odzwierciedlają zebrane dane, a przynajmniej możliwie najmniejszy

ubytek w tym zakresie. Tylko wtedy bowiem utrzymany jest sens posługiwania się takimi danymi i szansa na adekwatność wniosków sformułowanych na podstawie analizy udostępnionych danych w stosunku do realiów. Stratę informacji w tym aspekcie można oceniać rozmaicie. Najbardziej naturalnym narzędziem służącym do realizacji tego celu wydaje się współczynnik korelacji. Można się tutaj posłużyć współczynnikiem korelacji liniowej Pearsona, który jednak ma zastosowanie tylko do danych wyrażonych na skali różnicowej lub ilorazowej, w dodatku w razie występowania zależności innego typu niż liniowa może nie być dostatecznie użyteczny. Dlatego warto w tym kontekście rozważyć także alternatywne użycie współczynnika korelacji τ-Kendalla.

Dokładny kształt miar straty informacji opartych na współczynniku korelacji może być taki sam jak przedstawiono w poprzednim podrozdziale. Oznacza to, że rzeczone miary mogą być funkcją wartości bezwzględnych lub kwadratów różnic między współczynnikami korelacji odpowiednich zmiennych przed i po przeprowadzeniu SDC. Inną opcją w tym zakresie może być analiza macierzy odwrotnych do macierzy korelacji: wejściowej $\mathbf{R} = [\rho_{jl}]$ i po dokonaniu SDC $\mathbf{R}^{\star} = [\rho^{\star}_{jl}]$. Owe macierze odwrotne to odpowiednio $\mathbf{R}^{-1} = [\rho^{(-1)}_{jl}]$ i $(\mathbf{R}^{\star})^{-1} = [\rho^{\star(-1)}_{jl}]$, $j, l = 1, 2, \ldots, m$. Diagonalne elementy każdej z takich macierzy $\rho^{(-1)}_{jj}$ i $\rho^{\star(-1)}_{jj}$, $j = 1, 2, \ldots, m$, odpowiednio, należą do przedziału $[1, \infty)$ i wskazują siłę związku informacyjnego odpowiedniej zmiennej z pozostałymi, z uwzględnieniem także powiązań nieuchwytnych formalnie. Tym samym suma wartości bezwzględnych różnic między owymi elementami może być dobrym miernikiem straty informacji:

$$\lambda = \sum_{j=1}^{m} \left| \rho^{(-1)}_{jj} - \rho^{\star(-1)}_{jj} \right|.$$

Można też oczywiście rozważać taki miernik w postaci znormalizowanej, opartej na sferze jednostkowej i odległości euklidesowej:

$$\lambda = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{m} \left(\frac{\rho^{(-1)}_{jj}}{\sqrt{\sum_{l=1}^{m} (\rho^{(-1)}_{ll})^2}} - \frac{\rho^{\star(-1)}_{jj}}{\sqrt{\sum_{l=1}^{m} (\rho^{\star(-1)}_{ll})^2}} \right)^2} \in [0, 1]. \tag{4.12}$$

Miernik ten zaproponował Młodak (2020). Jednak w pracy tej występuje mnożnik $\frac{1}{2}$, a nie $\frac{1}{\sqrt{2}}$. Modyfikację tę wprowadzono nieco później. Wynikło to stąd, że formuła (4.12) została oparta na odległości euklidesowej między dwoma punktami należącymi do kuli jednostkowej w przestrzeni $\mathbb{R}^m$ (tzn. kuli o środku w początku układu współrzędnych – $(0, 0, \ldots, 0)$ – i promieniu 1). Wówczas mak-

symalna możliwa odległość między tymi punktami wynosi 2. Stąd właśnie, aby otrzymać wartości znormalizowane na przedziale [0, 1], zastosowano mnożnik $\frac{1}{2}$. Jednakże elementy diagonalne odwróconej macierzy korelacji są zawsze nie mniejsze niż 1. Tym samym tak naprawdę oblicza się tutaj odległości punktów leżących w hiperćwiartce rzeczonej kuli obejmującej jedynie punkty o wszystkich współrzędnych nieujemnych. Maksymalna odległość między punktami w tym wypadku jest równa $\sqrt{2}$. Stąd w pierwotnej wersji indeks miał tendencję do zaniżania straty informacji. Dlatego też zdecydowano się zamienić mnożnik na $\frac{1}{\sqrt{2}}$.

W każdym z tych przypadków wyższa wartość miernika świadczy o wyższej stracie informacyjnej dotyczącej związków między obserwowanymi zjawiskami. Ilustracją zastosowania tego miernika jest przykład 4.1, w którym wykorzystano podobną egzemplifikację jak w pracy Młodaka (2020), jednak tutaj użyto skorygowanego miernika danego wzorem (4.12).

Przykład 4.1. Rozpatrzmy dane prezentowane w przykładzie 3.6 (miesięczne wynagrodzenie brutto, staż pracy i odległość od miejsca zamieszkania do miejsca pracy – są to zmienne wyrażone na skali ilorazowej) oraz rezultaty zastosowanej w stosunku do nich metody kontroli ujawniania danych, jaką była wymiana rang (tabl. 3.13). Wówczas macierze korelacji τ-Kendalla dla zmiennych przed i po tejże kontroli mają odpowiednio postaci uwidocznione w tablicy 4.1. Tablica 4.2. obrazuje odpowiednie macierze do nich odwrotne.

Tym samym wartość współczynnika straty informacji w postaci „surowej" wynosi 0,3549, a w postaci znormalizowanej 0,0450. Oznacza to, że zastosowanie wymiany rangowej spowodowało stratę 4,5% informacji o związkach pomiędzy badanymi zmiennymi. Ubytek ten można uznać za niewielki.

Tabl. 4.1. Macierze korelacji τ-Kendalla przed i po zastosowaniu wymiany rang

	WYNAGR	STAZ	ODL
Przed wymianą rang			
WYNAGR	1,0000	0,3435	–0,1481
STAZ	0,3435	1,0000	–0,1079
ODL	–0,1481	–0,1079	1,0000
Po wymianie rang			
WYNAGR	1,0000	0,4849	–0,1550
STAZ	0,4849	1,0000	–0,0661
ODL	–0,1550	–0,0661	1,0000

Źródło: Opracowano z wykorzystaniem danych z tablicy 3.13 (użytych też w pracy Młodaka (2020)) oraz programu SAS Enterprise Guide 4.3.

Tabl. 4.2. Macierze odwrotne do macierzy korelacji τ-Kendalla przed i po zastosowaniu wymiany rang

	WYNAGR	STAZ	ODL
Przed wymianą rang			
WYNAGR	1,1500	−0,3810	0,1292
STAZ	−0,3810	1,1380	0,0663
ODL	0,1292	0,0663	1,0263
Po wymianie rang			
WYNAGR	1,3338	−0,6358	0,1647
STAZ	−0,6358	1,3075	−0,0121
ODL	0,1647	−0,0121	1,0247

Źródło: Opracowano z wykorzystaniem danych z tablicy 4.1.

Młodak i in. (2022) wykorzystali ten miernik – podobnie jak miarę zakłócenia rozkładu daną wzorem (4.1) z odległościami określonymi wzorami (4.2), (4.3) i (4.6) – do oceny straty informacji na skutek zastosowania kontroli ujawniania danych w badaniu wypadków przy pracy.

Inne podejście, szczególnie przydatne w wypadku zmiennych jakościowych, polega na konstruowaniu tablic kontyngencji dotyczących porównywanych zmiennych. Na podstawie tych tablic wykonuje się test niezależności pomiędzy odpowiednimi zmiennymi. Test niezależności dla tablicy dwuwymiarowej tego rodzaju jest oparty na współczynniku zgodności chi-kwadrat pomiędzy wartościami obserwowanymi a teoretycznymi. Alternatywnie można tutaj skorzystać z testu ilorazu wiarygodności chi-kwadrat lub testu Mantela-Haenszlea. Miarę związku ocenia współczynnik V Cramera, ewentualnie współczynnik φ czy współczynnik kontyngencji Pearsona. Ocena straty opiera się na procentowej względnej różnicy pomiędzy wartościami danego współczynnika obliczonymi dla tablicy źródłowej oraz dla tablicy opracowanej na podstawie danych, dla których zastosowano ochronę poufności. Empirycznie tego rodzaju postępowanie pokazano w podrozdziale 5.5 (przykład 5.5). Dla tablic wielowymiarowych zależności warunkowe oraz wartości teoretyczne można wyrazić za pomocą modeli log-liniowych.

Kolejne możliwości w tym kierunku rodzi porównywanie parametrów strukturalnych (tzn. parametrów funkcji regresji) oraz parametrów struktury stochastycznej (czyli cech rozkładu czynnika losowego) wraz z ocenami i ich jakości w odpowiednich modelach ekonometrycznych. Wówczas można uzyskać obraz potencjalnego zniekształcenia informacyjnego wnioskowania o współzależności zjawisk powstałego na skutek zastosowania SDC.

4.5. Strata informacji w estymacji

Zastosowanie metod niezakłóceniowych lub zakłóceniowych w celu zapewnienia odpowiedniego poziomu poufności danych jednostkowych, czyli wyeliminowania lub absolutnego zminimalizowania ryzyka ujawnienia, skutkuje nie tylko poniesieniem straty informacji, ale również odciska swoje piętno na agregacji danych (w wypadku badań pełnych) lub na estymacji wybranych parametrów populacji oraz na jej jakości (w wypadku badań reprezentacyjnych). Fakt ten trzeba mieć na względzie, gdy dane zabezpieczone wspomnianymi metodami są udostępniane dla celów naukowych. Zarówno dla gestora, jak również dla użytkownika danych ważne jest, aby pomimo zastosowania na mikrodanych metod niezakłóceniowych lub zakłóceniowych możliwe było uzyskanie agregatów lub ocen parametrów populacji (ogółem lub w bardziej szczegółowych przekrojach) tożsamych z tymi, które mogłyby zostać otrzymane na podstawie oryginalnych danych jednostkowych bądź niewiele różniących się od nich. Użytkownik danych chce bowiem jak najlepiej poznać badane zjawisko. Z drugiej zaś strony gestor danych, udostępniający określone informacje, będzie kojarzony z każdymi wynikami opracowywanymi i publikowanymi przez osoby ze środowiska naukowego.

Jak wspomniano wcześniej, skutkiem zastosowania metod niezakłóceniowych do danych jednostkowych może być zmniejszenie szczegółowości lub ukrycie części informacji przed ich użytkownikiem. Zmniejszenie szczegółowości danych – polegające na przykład na zastąpieniu dokładnych wartości zmiennej przedziałami, w których ona się mieści, przejściu na słabszą skalę pomiarową czy na łączeniu kategorii zmiennej w bardziej zgrubne – przełoży się niewątpliwie na zmniejszenie szczegółowości możliwych do uzyskania agregatów, jak również na zakres estymacji punktowej bądź przedziałowej parametrów populacji. Udostępnianie jedynie części obserwacji z badania pełnego będzie wymagać utworzenia specjalnych wag pozwalających na uogólnienie wyników na całą populację generalną. Gdy udostępniana jest podpróba obserwacji z próby objętej badaniem reprezentacyjnym, niezbędne staje się odpowiednie skalibrowanie wag z losowania, tak aby uzyskiwane szacunki odtwarzały informacje o całej populacji generalnej. Jeżeli w wyniku zastosowania metod niezakłóceniowych część wartości zmiennej zostanie zastąpionych brakami danych, konieczne okaże się odpowiednie ich zaimputowanie lub skalibrowanie wag z losowania. Wybór metody imputacji będzie skutkować rozbieżnościami w uzyskiwanych agregatach lub szacunkach. Jeżeli z kolei braki danych wystąpią w wielu zmiennych, a przyjętym rozwiązaniem niwelacji ich negatywnego znaczenia będzie kalibracja wag, to dla każdej z nich niezbędne stanie się utworzenie oddzielnej wagi, aby możliwe było odtworzenie informacji o populacji generalnej.

O ile użytkownik danych nie musi uwzględniać zmniejszonej szczegółowości otrzymanych informacji w wykorzystywanych przez siebie metodach staty-

stycznych, o tyle konieczność zastosowania odpowiednich technik imputacji lub kalibracji może się okazać dla niego problematyczna. Dotyczy to zwłaszcza niedoświadczonych użytkowników, nieposiadających zaawansowanej wiedzy i umiejętności z zakresu statystyki.

Metody zakłóceniowe nie powodują zmniejszenia szczegółowości bądź ukrycia wartości zmiennej dla (części) obserwacji. Wartości zmiennej – dla niektórych lub wszystkich rekordów – zostają jednak w odpowiedni sposób zniekształcone. Im większe będzie to zniekształcenie, czyli im większa będzie różnica pomiędzy wartościami zakłóconymi a oryginalnymi, tym większa będzie rozbieżność pomiędzy agregatami/szacunkami uzyskiwanymi na podstawie zniekształconych informacji a tymi ustalanymi na podstawie oryginalnych danych jednostkowych. Należy podkreślić, że metody zakłóceniowe mają w założeniu pozwalać na uzyskanie agregatów bądź szacunków parametrów populacji niewiele różniących się od bazujących na danych oryginalnych. Ta bliskość nie musi jednak być zachowana w bardziej szczegółowych przekrojach – a to właśnie na coraz bardziej szczegółowe informacje statystyczne występuje zapotrzebowanie. Na przykład, w zależności od wybranej metody dodawania szumu, zachowywane są wybrane parametry charakteryzujące zakłócaną zmienną – w przypadku szumu dodawanego addytywnie, nieskorelowanego, zachowywana jest wartość średnia oraz kowariancja, niezachowywane natomiast są wariancja oraz współczynniki korelacji (Hundepool i in., 2012; Templ i in., 2015).

W klasycznej postaci metody wymiany rang wszelkie statystyki opisowe charakteryzujące rozkład zmiennej dla populacji nie są zniekształcane – ponieważ metoda ta jedynie przetasowuje wartości zmiennej pomiędzy rekordami (Hundepool i in., 2012). Zmiany (w wartościach minimalnej i maksymalnej) można zaobserwować, gdy metoda jest aplikowana na danych z wykorzystaniem pakietu `sdcMicro` środowiska R, w którym możliwe jest dokonanie grupowania ustalonego procenta najniższych i najwyższych wartości zmiennej (i zastąpienia ich wartością średnią) przed rangowaniem. W wypadku metody PRAM, polegającej na zakłócaniu zmiennych jakościowych, możliwe jest wybranie wariantu *invariant* – wtedy zachowywane są jednowymiarowe rozkłady zniekształcanych zmiennych (tzn. liczebności jednostek według kategorii zmiennej przed i po zakłóceniu będą identyczne), jednak rozkłady wielowymiarowe (liczebności w przekrojach dwóch lub większej liczby zmiennych) nie są zachowywane (Hundepool i in., 2012; Benschop i in. 2022). Prawidłowość ta w wypadku badań częściowych dotyczy jednakże jedynie rozkładów jednowymiarowych w próbie – ze względu na to, że każdy rekord może mieć przypisaną inną wagę z losowania, po uogólnieniu wyników na populację generalną nawet rozkłady jednowymiarowe mogą być zniekształcone. Dodatkowo należy podkreślić, że stosowanie metod zakłóceniowych może doprowadzić do powstania nielogicznych wartości zakłóconych zmiennych, jak również do niezachowania związków pomiędzy zmiennymi. Z tego powodu po

zakłóceniu mikrodanych, a przed ich udostępnieniem, niezbędna wydaje się ich weryfikacja – sprawdzenie, czy wartości każdej zmiennej z osobna, jak również w kombinacji z wartościami pozostałych zmiennych, są prawidłowe.

W procesie agregowania danych lub przeprowadzania estymacji konieczne jest uwzględnienie zakłócenia danych. Nie jest to jednak zagadnienie dokładnie opisane w literaturze. Na przykład, jak sugerują (Hundepool i in., 2012), gdy zmienne jakościowe zostały zakłócone metodą PRAM, można sięgnąć do technik mających zapewnić korektę, zbliżoną do zastosowanej w przypadku błędnej klasyfikacji lub przypadków losowych odpowiedzi. Podobnie dostęp do parametrów dodanego szumu i uwzględnienie ich w analizie może się przełożyć na jakość otrzymanych szacunków.

Bez względu na zastosowaną metodę kontroli ujawniania mikrodanych konieczne jest każdorazowe dokładne zbadanie jej wpływu na jakość estymacji – na precyzję estymatora, jego obciążenie oraz dokładność. Na przykład wpływ metod ograniczających ryzyko ujawnienia – m.in. dodawania szumu oraz wymiany rang – na całkowity błąd estymacji opisali Biemer i in. (2017).

5

Przegląd wybranych narzędzi informatycznych

We współczesnych realiach gromadzenia, przetwarzania i analizy danych żadne tego typu działania nie mogą się obejść bez odpowiednich narzędzi informatycznych. Podobnie rzecz się ma z kontrolą ujawniania danych. Chociaż rozwój tej dziedziny statystyki dopiero nabiera tempa, to już funkcjonuje kilka interesujących rozwiązań programistycznych umożliwiających efektywne przeprowadzanie takiej kontroli na zbiorach informacji liczbowych lub symbolicznych posiadanych w formie elektronicznej. Rozdział wieńczy ukazanie najpowszechniej stosowanych narzędzi informatycznych służących do sprawnego i efektywnego przeprowadzenia kontroli ujawniania danych, a zatem programów τ-Argus i μ-Argus, a także narzędzi środowiska R i innych. Wszystkie ukazane treści zostały zilustrowane stosownymi przykładami.

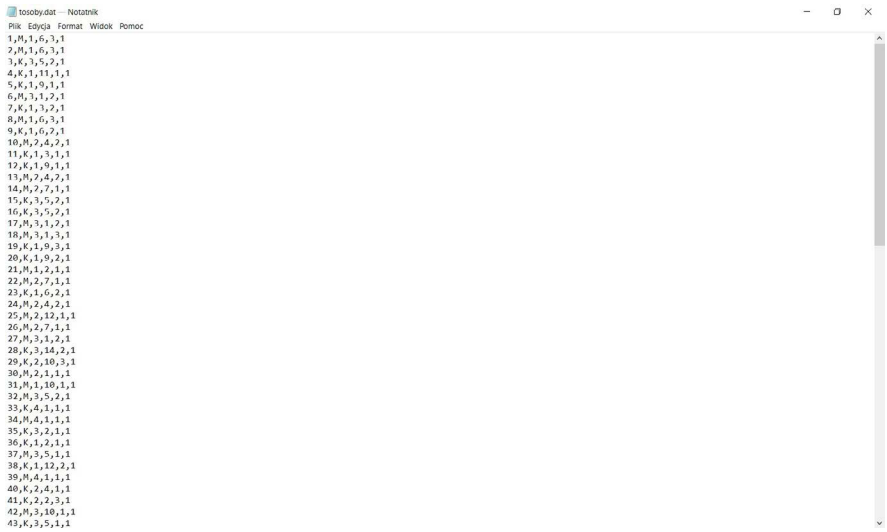
W tej części opracowania omówiono najważniejsze programy służące do przeprowadzenia procesu SDC. Prezentację rozpoczniemy ukazaniem najbardziej chyba znanych i nieodpłatnie udostępnionych programów τ-Argus i μ-Argus. Zostały one opracowane w języku Java przez Centraal Bureau voor de Statistiek, Statistics Netherlands w wyniku kilku projektów europejskich. Istnieją jednak ich obszerniejsze wersje „paczkowe" (ang. *bundle*), w których kompilator Java jest już wbudowany, a zatem wersje te nie wymagają dodatkowej instalacji tego narzędzia. W dalszej kolejności wskazano najnowsze możliwości, jakie zarysowały się w środowisku R w tym zakresie (między innymi `sdcTable` i `sdcMicro`). Na koniec wreszcie zasygnalizowano (obecnie raczej bardzo skromne) możliwości wykorzystania potencjału innych programów w kierunku SDC.

5.1. Program τ-Argus

Jest to program służący do sprawnego przeprowadzania kontroli danych tabelarycznych. Szczegóły dotyczące jego stosowania są zawarte w dołączonym do plików programu podręczniku użytkownika (de Wolf i in., 2014). Tutaj zatem wskażemy tylko najistotniejsze aspekty związane z jego użytkowaniem.

Na wstępie należy wybrać rodzaj solvera (czyli narzędzia do rozwiązywania równań). Czyni się to, wybierając z menu Help pozycję Options. Do wyboru są tam narzędzia Xpress, CPlex i Free solver. Jeśli nie posiadamy licencji na dwa pierwsze, można wybrać trzecie rozwiązanie. Równocześnie warto ustawić ograniczenia czasowe dla kontroli ujawniania danych w pojedynczej tablicy (Max. time per suitable (modular)). Następnie można przystąpić do wczytywania danych.

W praktyce najczęściej posługujemy się zbiorami mikrodanych powstałymi przy użyciu narzędzi wykorzystywanych w rozmaitych badaniach statystycznych. Każde z nich daje możliwość konwersji zapisanych zbiorów na bardziej uniwersalne formaty. W wypadku rozpatrywanego tu programu należy wyeksportować dane do formatu CSV (z danymi rozdzielonymi przecinkami)[17], wstawiając uprzednio do zbioru dodatkową zmienną składającą się z samych jedynek (co umożliwia sumowanie, niezbędne do tworzenia tablic kontyngencji różnego wymiaru). Następnie z takiego pliku trzeba usunąć pierwszy wiersz z nazwami zmiennych. Jeśli wśród danych znajdują się liczby z miejscami dziesiętnymi (np. 0,345), to przecinek trzeba zmienić na kropkę (w Europie Zachodniej stosuje się taki właśnie separator dziesiętny). Na zakończenie zmieniamy rozszerzenie tak przygotowanego pliku na DAT. Finalnie przygotowany plik ma więc np. postać pokazaną na rysunku 5.1.

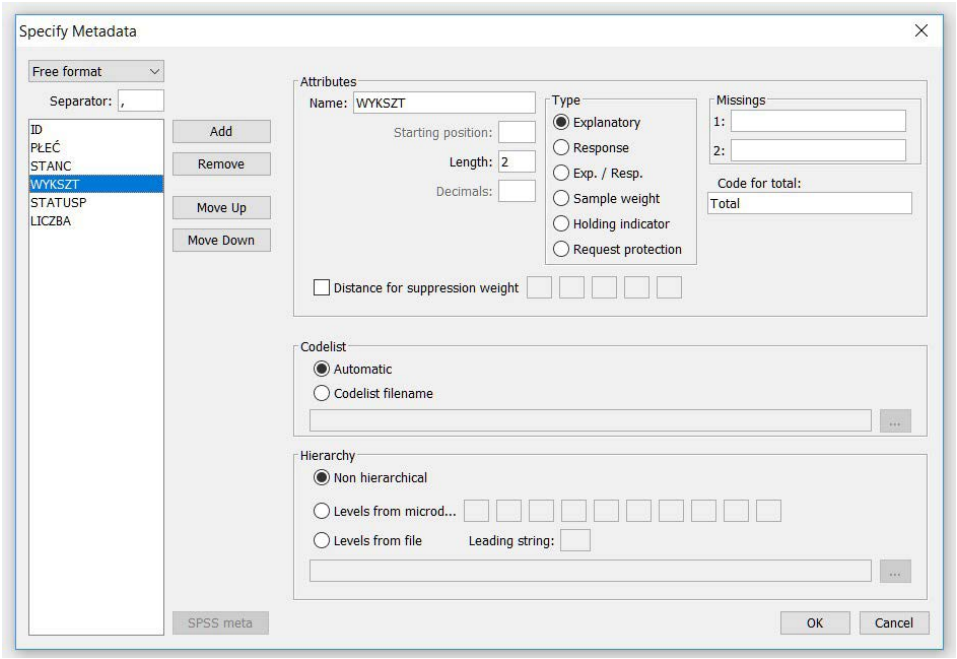


Rys. 5.1. Przykład właściwie przygotowanego pliku z danymi do wykorzystania w programie τ-Argus

Źródło: Dane fikcyjne.

[17] Czasami (np. w Office 2013) zdarza się, że mimo wszystko program generuje plik CSV z danymi rozdzielanymi średnikami. Wtedy średniki należy zamienić na przecinki.

Następnie z menu File programu τ-Argus wybieramy pozycję Open Microdata i w pierwszym wierszu od góry wskazujemy lokalizację pliku z danymi. Klikamy OK i przechodzimy do pozycji Specify→Metadata. Tutaj wybieramy opcję Free format i ustawiamy metadane, czyli dane o tych danych (rys. 5.2).



Rys. 5.2. Wprowadzanie metadanych w programie τ-Argus

Objaśnienia: ID – identyfikator, PŁEĆ – płeć, STANC – stan cywilny prawny, WYKSZT – wykształcenie, STATUSP – status na rynku pracy. Zmienne te objaśniono w przykładzie 3.4. Dodatkowo umieszczono tutaj zmienną LICZBA – jedynkową zmienną techniczną.

Źródło: Dane fikcyjne.

Warto wspomnieć, że jedynym specjalistycznym programem statystycznym, którego pliki z danymi można szybko ładować do programu τ-Argus, jest SPSS (ang. *Statistical Package for the Social Sciences*; obecnie to rozwinięcie nie jest używane). Wtedy mikrodane ładujemy z pliku z rozszerzeniem SAV, w edycji metadanych zaś wskazujemy format SPSS.

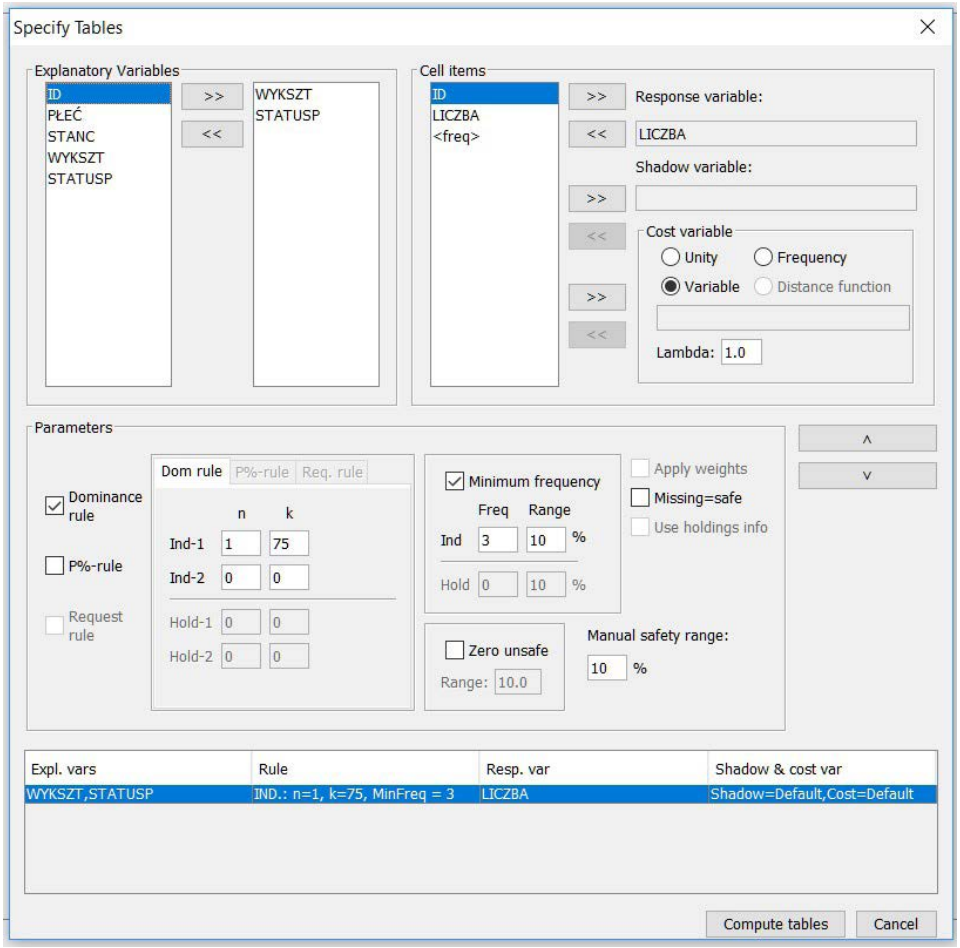
Wróćmy do naszego przykładu. W kolejnym kroku należy wskazać, które zmienne są objaśniające (*Explanatory*), a które są zmiennymi odpowiedzi (*Response*) – to znaczy zmiennymi ilościowymi (numerycznymi, ciągłymi) definiującymi komórki tablicy (w przypadku najprostszych tablic będzie to zmienna składająca się z samych jedynek, umożliwiająca zliczanie obserwacji należących do każdej komórki). Można też określić, która zmienna zawiera wagi wynikające

z przyjętego schematu losowania próby do badania (*Sample weight*), a która definiuje pewne grupy rekordów z określonych powodów rozpatrywane łącznie (*Holding indicator*). Da się także wskazać zmienną określającą arbitralnie, czy dany rekord podlega ochronie, czy też nie (*Request protection*). Dostępna jest ponadto opcja wskazania odległości dla wagi ukrywania (*Distance for suppression weight*) – chodzi tutaj o sytuację, gdy koszt ukrycia każdej komórki zależy od liczby kroków ukrywania. W tej opcji da się wskazać koszt dla każdego kroku, nie większy niż 5. Po prawej stronie można też wskazać oznaczenie braków danych (*Missings*) i kod dla wartości ogółem (*Code for total*), a niżej – listy kodów (*Codelist*) i hierarchię jednostek (*Hierarchy*), jeśli są inne niż automatycznie odczytywane na podstawie pliku z danymi.

W dalszej kolejności przechodzimy do opcji umożliwiającej konstruowanie tablic (Specify→Tables). Tutaj wskazujemy, na podstawie których zmiennych objaśniających tworzymy stosowną tablicę i której zmiennej odpowiedzi do tego celu używamy. Można tutaj też wskazać zmienne „cienia" (*Shadow variable*), czyli zmienne arbitralnie definiujące reguły bezpieczeństwa (np. charakterystyki jednostek wnoszących największy wkład do określonych komórek), jak również zmienne opisujące koszt (*Cost variable*) w różnych wariantach. Niżej natomiast ustala się reguły wskazujące komórki niebezpieczne z punktu widzenia poufności danych. Można tu zastosować regułę dominacji (*n*, *k*) (*Dominance rule*), regułę *p*% (*P%–rule*) oraz specjalne opcje praktykowane w niektórych krajach (*Request rule*). Da się też tutaj określić minimalną częstość w bezpiecznej komórce (*Minimum frequency*), a także to, czy brakujące dane można uznać za bezpieczne (*Missing=safe*), a zera – za niebezpieczne (*Zero unsafe*), lub ustalić arbitralnie zakres bezpieczeństwa (*Manual safety range*). Po dokonaniu odpowiednich wyborów klikamy strzałkę skierowaną w dół (po prawej stronie okienka), co powoduje umieszczenie zbiorczych danych o tablicy w dolnym panelu, a następnie wybieramy opcję Compute tables.

Na rysunku 5.3 ukazano przykład ustalenia takich danych w rozpatrywanym na wcześniejszych rysunkach przykładzie z użyciem klasycznych kryteriów, czyli minimalnej liczebności równej 3 i maksymalnego największego wkładu wynoszącego 75%. Liczba rekordów to 100. Tablicę taką można też zapisać w inny sposób: wybieramy z głównego menu Output→Save Table, a następnie ustalamy opcję na Intermediate format (dający rozszerzenie TAB) i wskazujemy odpowiedni katalog w dolnym panelu oraz klikamy przycisk Write table. Odczyt tablicy przy ponownym uruchomieniu programu następuje po wybraniu polecenia File→Open Table (po czym trzeba przejść jeszcze do opcji Specify→Tables i ustawić w okienku Safety rules na przykład opcję Use given status – jeśli nie chcemy na tym etapie wprowadzać zmian).

W wyniku tych działań otrzymamy tablicę, w której niebezpieczne komórki zostają zaznaczone kolorem czerwonym (rys. 5.4). Klikając którąkolwiek liczbę,

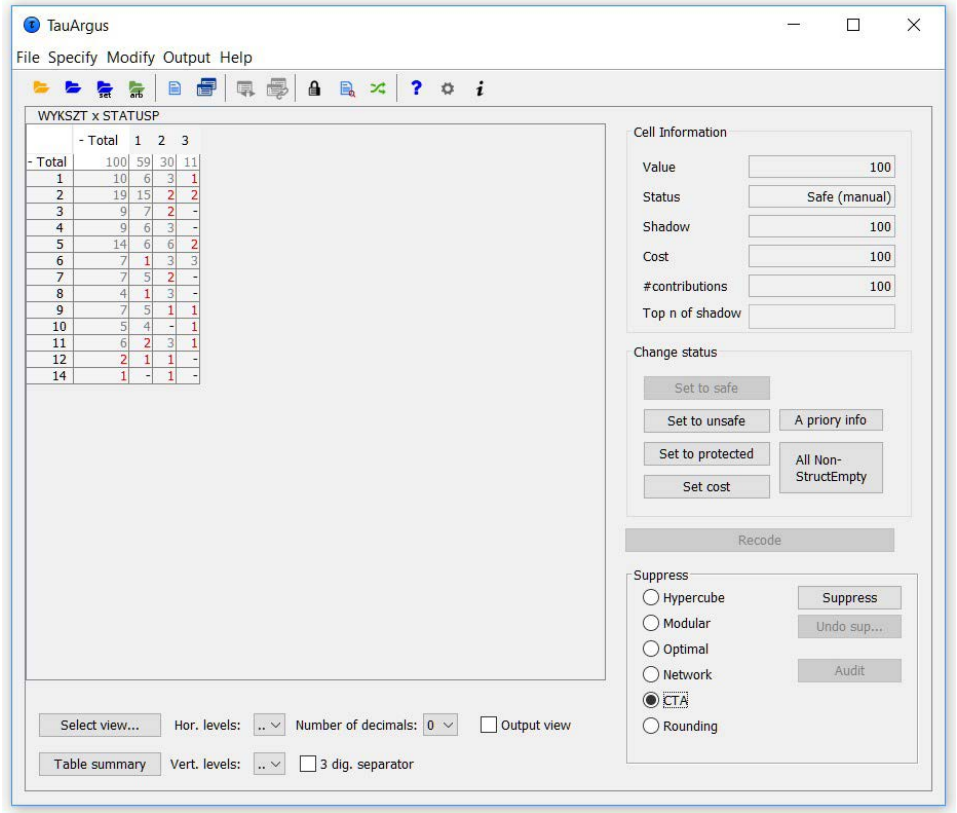


Rys. 5.3. Określanie tablicy w programie *τ*-Argus

Objaśnienia jak do rysunku 5.2.

Źródło: Dane fikcyjne.

ujrzymy po prawej stronie stosowne informacje o danej komórce (a zatem jej wartość, status: bezpieczna/niebezpieczna – *safe/unsafe*, ze wskazaniem, z jakiego powodu ma określony status: arbitralnego ustalenia odpowiedniej reguły, sposobu traktowania zer względnie ze względu na to, czy jest wybrana do ukrywania wtórnego, czy może być kandydatką do tegoż ukrywania albo czy jest pusta), wartość zmiennej cienia, koszt, liczbę jednostek wnoszących wkład w daną komórkę (*#contributions*) oraz największe wartości tegoż wkładu (*Top n of shadow*). Poniżej znajdują się opcje umożliwiające zmianę statusu komórki oraz kosztu jej ukrywania. Jeszcze niżej umieszczono opcję Recode umożliwiającą przekodowanie



Rys. 5.4. Przykład tablicy z danymi niebezpiecznymi w programie τ-Argus

Objaśnienia jak do rysunku 5.2.

Źródło: Dane fikcyjne.

zmiennych w celu ochrony wskazanych danych niebezpiecznych oraz – alternatywnie – możliwości ukrywania tychże danych za pomocą metod:

- hiperkostki (*Hypercube*) – zapewnia, że pojedynczy respondent nigdy nie pojawi się jako jedyny element tylko jednej hiperkostki,
- modularnej (*Modular*) – mającej zastosowanie do tablic hierarchicznych: dzieli się taką tablicę na kilka podtablic niehierarchicznych, wprowadza ich ochronę, a następnie tworzy z nich finalnie chronioną tablicę,
- optymalnej (*Optimal*) – chroni tablicę hierarchiczną jak pojedynczą, bez dzielenia jej na mniejsze podtablice,
- przepływów sieciowych (*Network*) – jest stosowana tylko do dwuwymiarowych tablic z jednym wymiarem hierarchicznym; oparto ją na ciągu podproblemów o krótkiej ścieżce rozwiązywania, pozwalającej na wygenerowanie dopuszczalnej szachownicy ukrywania,

– kontrolowanego dopasowania tablic (*CTA*) – niebezpieczne wartości w komórkach zostają zastąpione przez ich górny lub dolny poziom ochrony, a pozostałe komórki zostają zmodyfikowane w taki sposób, aby zachować odpowiednie sumy wartości,
– zaokrąglania (*Rounding*).

Na samym dole tego okienka znajdują się przyciski umożliwiające zamianę wierszy i kolumn (Select view…) oraz wyświetlenie podsumowania całej tablicy (Table summary). Można też ograniczyć widok tylko do wartości brzegowych (Hor. levels, Vert. levels) lub ustalić dokładność prezentacji, czyli liczbę miejsc dziesiętnych dla danych ilościowych (Number of decimals).

Wykonajmy zatem SDC metodą CTA. Po wybraniu tej opcji i kliknięciu przycisku Suppress pojawia się pytanie, czy użytkownik woli wersję ekspercką (*Do you prefer to use the expert version?*). Ze względu na możliwość uzyskania lepszej jakości wyników zaakceptujmy to, a w pojawiającym się okienku wybierzmy *Type CTA: Std CTA (better solutions)*, *Solver: Cplex* i *Format: Check all file*, inne opcje pozostawiając w ustawionym automatycznie kształcie. Zatwierdziwszy czynność klawiszem *Start optimization*, po zamknięciu okienka otrzymamy tablicę uwidocznioną na rysunku 5.5.

WYKSZT x STATUSP

	- Total	1	2	3
- Total	100	59	30	11
1	10	6	4	0
2	19	15	1	3
3	9	8	1	-
4	9	6	3	-
5	14	6	7	1
6	7	0	4	3
7	7	4	3	-
8	4	0	4	-
9	7	5	0	2
10	5	5	-	0
11	6	1	3	2
12	3	3	0	-
14	0	-	0	-

Rys. 5.5. Przykład tablicy z zastosowaniem kontroli CTA w programie τ-Argus

Objaśnienia jak do rysunku 5.2.

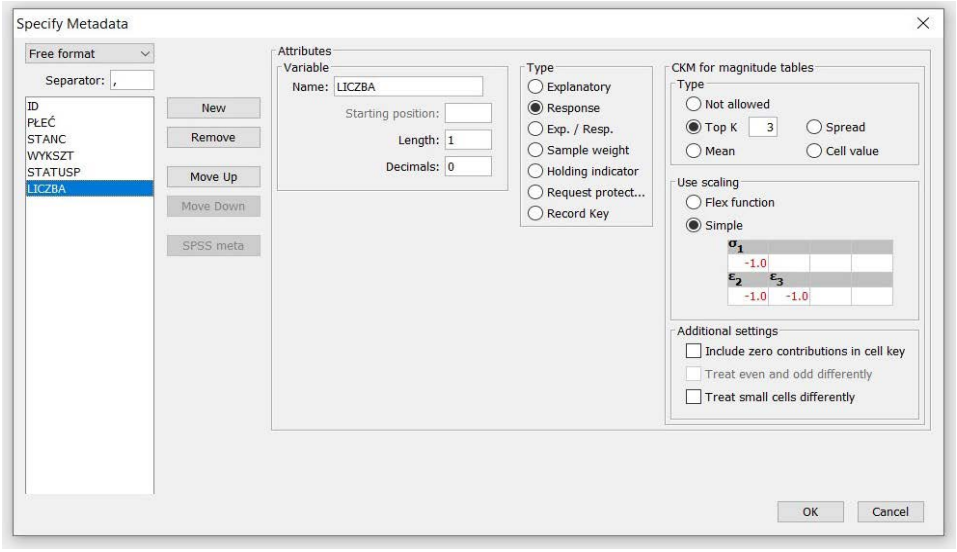
Źródło: Dane fikcyjne.

Niebieski kolor wskazuje zmiany dokonane w elementach uznanych za niebezpieczne. Po porównaniu zmodyfikowanej tablicy z tablicą wyjściową widniejącą na rysunku 5.4 widać, że w wypadku danych będących wcześniej oznaczonymi

jako niebezpieczne da się jednak zaobserwować najistotniejsze przekształcenia. Nastąpiły tutaj przesunięcia między kategoriami. Ostatecznie, choć jeszcze w sześciu komórkach pozostają jedna lub dwie jednostki, to właśnie z powodu owych przesunięć ryzyko odtworzenia danych jednostkowych zostało zminimalizowane. Tak więc nowa tablica jest bezpieczna.

Finalną tablicę można wyeksportować w różnych formatach (Output→Save Table), m.in. do formatu CSV z danymi rozdzielanymi przecinkami – zarówno w postaci klasycznej (*CSV Format*), jak i w wersji dla tablicy przestawnej w Excelu z uwzględnieniem statusu komórek (*CSV for pivot table*). Można też zapisać plik w klasycznej postaci tekstowej (*Code–value*), choć też w układzie przestawnym i z możliwością uwzględnienia m.in. statusu komórek, ponadto w formacie SBS obejmującym określone metadane wymagane w praktyce przez Eurostat (*SBS format*), we wspomnianym już wyżej formacie TAB umożliwiającym dalsze prace w omawianym programie, a także w formacie JJ umożliwiającym określanie połączeń między tablicami hierarchicznymi i strukturami niezbędnymi dla niektórych optymalizacji SDC.

Począwszy od wersji 4.1.9, program τ-Argus ma wbudowaną możliwość wykorzystania metody kluczy komórkowych do tablic wielkości. Pojawia się ona w ustawieniach metadanych, gdy wskażemy zmienną odpowiedzi. Ukazano to na rysunku 5.6.



Rys. 5.6. Przykład możliwości skorzystania z metody kluczy komórkowych w programie τ-Argus

Objaśnienia jak do rysunku 5.2.

Źródło: Dane fikcyjne.

Można tutaj określić:

- czy zerowe udziały jednostek w komórkach powinny być uwzględnione podczas wyznaczania kluczy komórkowych, czy też nie (*Include zero contributions in cell key*),
- czy szum powinien być skalowany z użyciem K największych udziałów w wartości komórki lub średniego wkładu, różnicy pomiędzy maksymalnym a minimalnym wkładem do wartości komórki albo samej wartości komórki (*Top K*, *Mean*, *Spread*, *Cell value*, odpowiednio) czy też wcale (*Not allowed*),
- czy komórki obejmujące nieparzystą liczbę jednostek powinny być zakłócane odrębnie od komórek, na które składa się parzysta liczba jednostek (*Treat even and odd differently*),
- czy do skalowania szumu powinna być użyta tzw. funkcja flex (*Flex function*; jej idea polega na tym, że ustalony komponent zakłóceń musi zależeć od wartości komórek, w związku z czym użytkownicy powinni określić zakresy odnoszące się do pożądanych wielkości dla dużej i małej wartości komórki – zob. np. Meindl i Enderle (2019)), czy stosowne parametry definiujące pożądane wielkości użytkownik ustali arbitralnie (*Simple*),
- czy małe komórki mają być traktowane podobnie jak komórki częstościowe (*Treat small cells differently*).

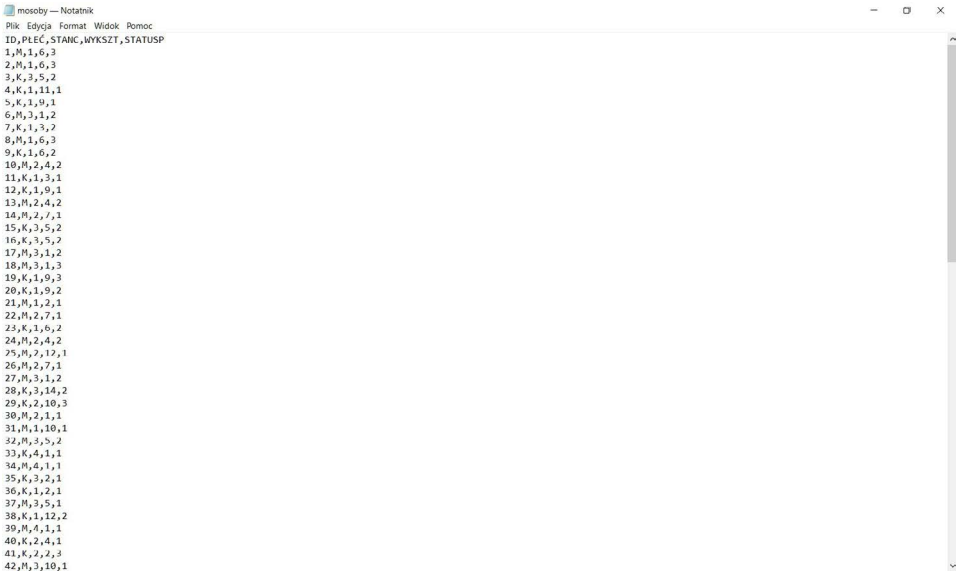
Strata informacji w programie τ-Argus odgrywa pewną rolę we wtórnym ukrywaniu komórek tablicy: pozwala ona rozróżnić komórki, pozostawienie wartości których będzie priorytetowe, od tych, których zawartość powinna raczej być ukryta. Ocena straty informacji jest dokonywana za pomocą funkcji kosztu usunięcia obrazującej wartość informacyjną danej komórki. Im większa owa wartość, tym strata, która powstałaby na skutek ukrycia jej wartości, jest znaczniejsza. Funkcję kosztu definiuje się na kilka sposobów: jako sumę udziałów poszczególnych jednostek w wartości komórki, częstość jednostek objętych komórką czy poprzez arbitralne przypisanie wartości. I tutaj jednak nie ma możliwości dokonania porównań danych wejściowych z danymi wyjściowymi (po zastosowaniu SDC).

Program oraz podręcznik obejmujący wszystkie – także te nieomówione tutaj szerzej – możliwości programu dostępny jest pod adresem https://research.cbs.nl/casc/tau.htm lub https://github.com/sdcTools/tauargus/releases.

5.2. Program μ-Argus

Program ten służy do przeprowadzania kontroli ujawniania mikrodanych. Ma zatem zastosowanie wówczas, gdy chcemy chronić teoretycznie zanonimizowane (czyli pozbawiane kluczowych identyfikatorów) dane jednostkowe przed możliwością odtworzenia informacji dla konkretnych jednostek.

Podobnie jak w przypadku programu τ-Argus, dane można importować z pliku w formacie DAT (przygotowanego jak wskazano w części 5.1.) czy SAV (pochodzącym z SPSS). Jednak μ-Argus oferuje dodatkowo także opcje bezpośredniego importu z pliku w formacie CSV (z danymi rozdzielanymi przecinkami). Tak więc można przygotować taki plik np. poprzez eksport stosownych danych z Excela lub programu statystycznego (o ile tam jest to możliwe). Po dokonaniu eksportu należy skontrolować, czy plik CSV faktycznie zawiera dane rozdzielone przecinkami, a separatorem dziesiętnym jest kropka[18], oraz czy nazwy zmiennych znajdują się w pierwszym wierszu. Tak przygotowany plik będzie miał postać uwidocznioną na rysunku 5.7.



Rys. 5.7. Przykład właściwie przygotowanego pliku z danymi do wykorzystania w programie μ-Argus

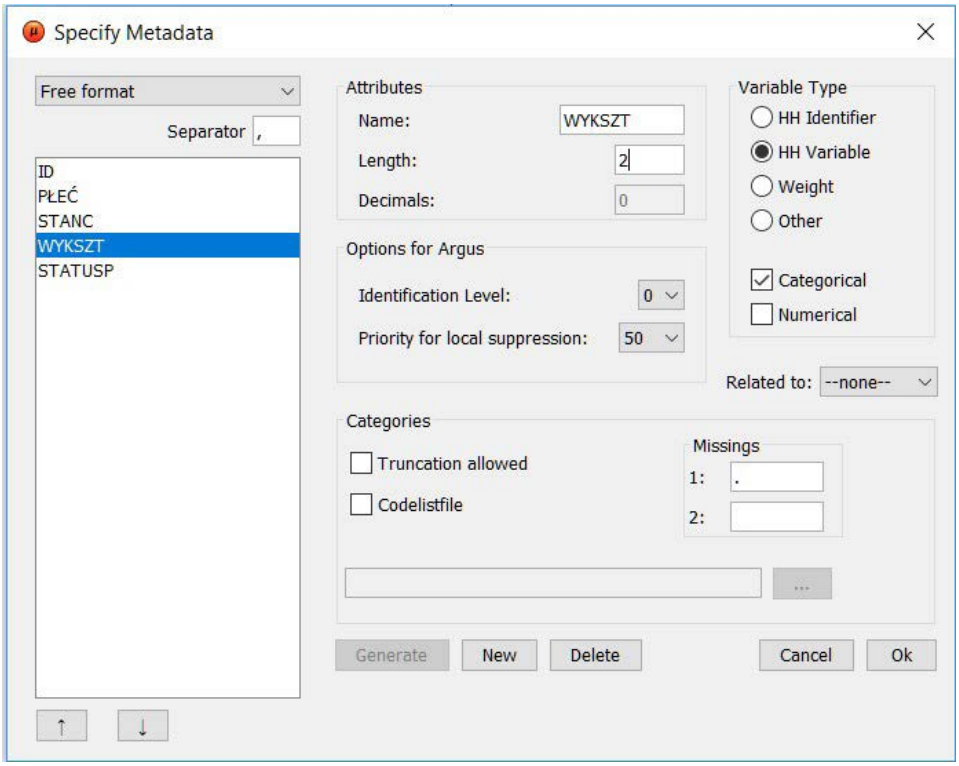
Objaśnienia jak do rysunku 5.2.

Źródło: Dane fikcyjne.

Uruchamiamy program μ-Argus i w menu File wskazujemy opcję Open micro data. W panelu Microdata: wskazujemy lokalizację pliku z danymi oraz jego format i klikamy przycisk OK. Następnie z pozycji menu Specify wybieramy opcję Metadata. Ustalamy wariant formatowania na Free with meta, po czym klikamy przycisk Generate, wskazujemy przecinek jako separator i zatwierdzamy ten wy-

[18] Jak już wspomniano w części 5.1, m.in. Excel czasem jako separatory wpisuje średniki – wówczas najpierw należy zamienić przecinki będące separatorem miejsc dziesiętnych na kropkę, a następnie średniki na przecinki.

bór przyciskiem OK. W wyniku tych działań nazwy zmiennych zostaną pobrane z pliku z danymi. Przestawiamy teraz wariant formatowania na Free format i dla każdej zmiennej wskazujemy odpowiednią długość[19], a jeżeli jest to zmienna ilościowa – także liczbę miejsc dziesiętnych. Ustalamy też w razie potrzeby, które informacje są identyfikatorami jednostki nadrzędnej – np. gospodarstwa domowego dla osób (*HH Identifier*), a które są zmiennymi dotyczącymi jednostki nadrzędnej (tzn. przyjmują jednakową wartość dla wszystkich jednostek podrzędnych wchodzących w skład tej nadrzędnej), wagami lub innymi (*HH Variable, Weight, Other,* odpowiednio). W wypadku zmiennych wskazujemy, czy dana zmienna jest jakościowa (kategorialna – *Categorical*), czy też ilościowa (numeryczna – *Numerical*).



Rys. 5.8. Wprowadzanie metadanych w programie *μ*-Argus

Objaśnienia jak do rysunku 5.2.

Źródło: Dane fikcyjne.

[19] Chodzi tu o maksymalną liczbę znaków, za pomocą których zapisywana jest dana informacja. Na przykład, jeżeli identyfikator będzie liczbą z zakresu od 1 do 100, to jego długość wyniesie 3, jeśli zaś wartości zmiennej (takiej jak np. wykształcenie) są kodowane od 1 do 14, to jej długość jest równa 2.

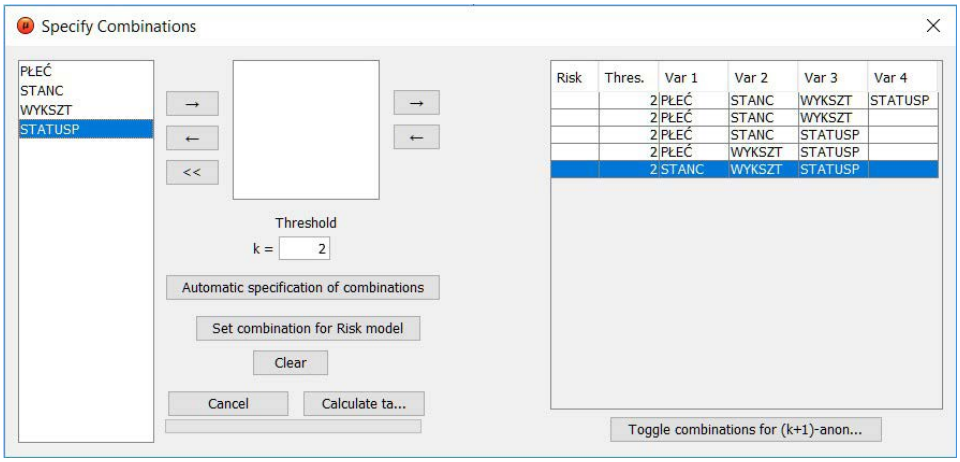
Dla każdej zmiennej można też ustalić tutaj poziom identyfikacji (*Identification level*): 0 – jednostka nie może być zidentyfikowana na podstawie danej zmiennej, 1 – zmienna identyfikuje najbardziej, 2 – zmienna identyfikuje bardziej, 3 – zmienna identyfikuje)[20], priorytet w zakresie ukrywania (*Priority for local suppression*), powiązania między zmiennymi (*Related to:*) oraz dopuszczenie obcięcia wartości do liczby całkowitej (*Truncation allowed*). Można również podać specjalny plik w formacie CDL z listą kodów zmiennej (*Codelistfile*) oraz znaki stosowane dla braków danych (*Missings*). Strzałkami u dołu można też zmienić kolejność zmiennych. Rysunek 5.8 ukazuje finalny widok okna metadanych.

Po kliknięciu klawisza OK pojawia się pytanie o zapisanie metadanych w pliku (*Metadata has been changed. Save changes to file*?). Jeśli chcemy zachować te metadane, klikamy Yes i wskazujemy docelową lokalizację oraz nazwę pliku (program przedstawia pewną sugestię w tym zakresie). Plik z metadanymi ma format RDA. Po wykonaniu tych czynności wchodzimy jeszcze raz do pliku CSV (przy użyciu np. Notatnika Windows) i usuwamy z niego wiersz z nazwami zmiennych, tak aby plik zaczynał się od pierwszego wiersza z danymi. Wtedy wracamy do programu μ-Argus i wybieramy z menu opcję Specify→Combinations.

W oknie, które wówczas się ukaże (rys. 5.9), trzeba wskazać kombinacje wartości zmiennych mogące potencjalnie prowadzić do ujawnienia informacji jednostkowych. Mogą one zostać wygenerowane automatycznie (opcja *Automatic specification of combinations*), jednak tylko na podstawie zmiennych, dla których w metadanych poziom identyfikacji został ustalony jako 1, 2 lub 3 (a zatem gdy zadeklarowano, że w jakimś stopniu zmienne stwarzają ryzyko ujawnienia chronionych danych). Wówczas w naszym przykładzie zmienne, dla których ów poziom wynosi 1, są wiązane ze zmiennymi o poziomie od 1 do 2, te zaś – ze zmiennymi o poziomie identyfikacji od 1 do 3 itd. Jeśli natomiast zagrożeń wcześniej nie zadeklarowano, to powiązania trzeba wskazać ręcznie. W tym celu w kolumnie po lewej stronie zaznaczamy dane zmienne i za pomocą strzałki przenosimy je do środkowego okienka, po czym kolejną strzałką po prawej stronie owej kolumny zatwierdzamy. Te czynności powtarzamy dla innych kombinacji, które zamierzamy rozpatrywać. Na rysunku 5.9 ukazano przykład tworzenia takich powiązań dla trzech i więcej zmiennych spośród rozpatrywanych w poprzednio określonym zbiorze danych.

Polecenie *Threshold* ustala maksymalną wartość komórki w tablicy, która jest uznawana za niebezpieczną z punktu widzenia ochrony danych wrażliwych. Tutaj ustalono ją tradycyjnie na 2. Można na tym ekranie także ustawić tablicę dla nowego modelu ryzyka (Set combination for risk model). Przycisk Toggle combinations for (k + 1)-anonymity umożliwia z kolei przełączenie tradycyjnej

[20] W najnowszych wersjach programu dodano także poziomy 4 i 5. Są one stosowane przede wszystkim wówczas, gdy do sprawdzenia pod kątem zachowania poufności jest duża liczba kombinacji wartości zmiennych.



Rys. 5.9. Określanie kombinacji wartości zmiennych w programie *μ*-Argus

Objaśnienia jak do rysunku 5.2.

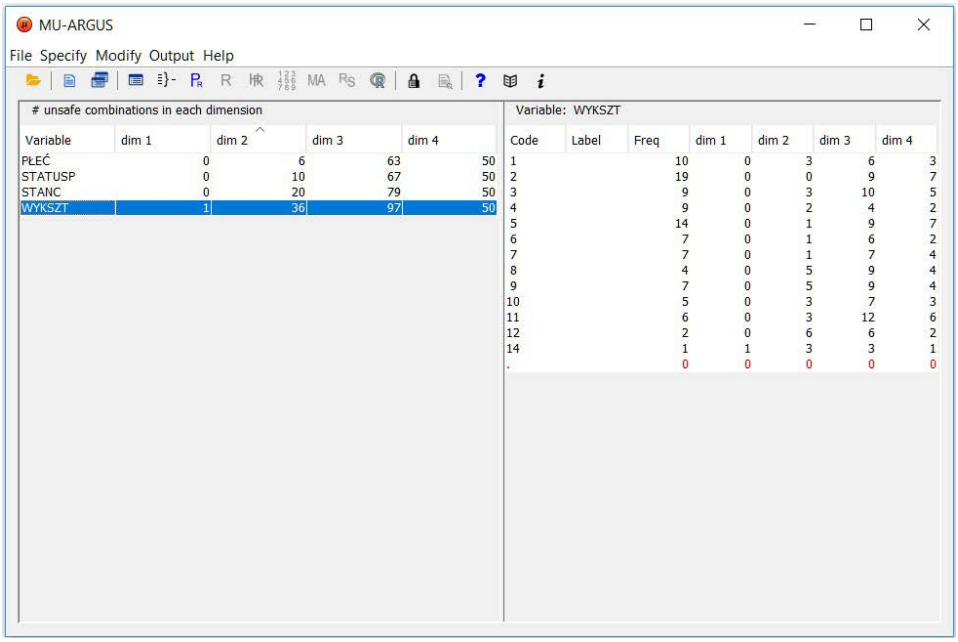
Źródło: Dane fikcyjne.

opcji wykrywania wrażliwości na regułę (k + 1)-anonimowości, według której informacja dla danej jednostki nie może być inna niż odpowiednia informacja dla przynajmniej k innych jednostek znajdujących się w bazie (gdzie k to liczba naturalna mniejsza niż liczba jednostek ogółem w bazie). Po ustawieniu wszystkich parametrów klikamy przycisk Calculate table. Ukazuje się wówczas zbiorcze podsumowanie liczby niebezpiecznych jedno-, dwu-, trzy- i czterowymiarowych kombinacji zawierających wartości poszczególnych zmiennych. Kliknięcie określonej zmiennej w lewym „podoknie" powoduje wyświetlenie się szczegółów na temat występowania poszczególnych jej wartości w niebezpiecznych kombinacjach z wartościami innych zmiennych (rys. 5.10).

Mając ustalone dane i wrażliwe kombinacje wartości zmiennych, można przystąpić do właściwej kontroli ujawniania danych. Program *μ*-Argus (poprzez wybór z menu opcji Modify) oferuje następujące metody w tym zakresie[21]:

- Przekodowywanie (*Global Recode*). Obejmuje między innymi obcinanie miejsc dziesiętnych dla danych ilościowych oraz pobieranie listy kodów dla danej zmiennej z pliku w formacie DLL. Rekodowane zmienne są zapisywane w pliku w formacie GRC.
- PRAM (*PRAM Specification*). Ustala się arbitralnie prawdopodobieństwo pozostawienie danej kategorii zmiennej jakościowej bez zmian.
- Określenie indywidualnego ryzyka (*Individual Risk Specification*). Wyznacza się tu ryzyko odtworzenia przez osobę nieupoważnioną danych wraż-

[21] Program uaktywnia opcje możliwe do zastosowania dla aktualnie rozpatrywanych danych.



Rys. 5.10. Sumaryczne informacje o niebezpiecznych kombinacjach wartości zmiennych w programie μ-Argus

Objaśnienia jak do rysunku 5.2.

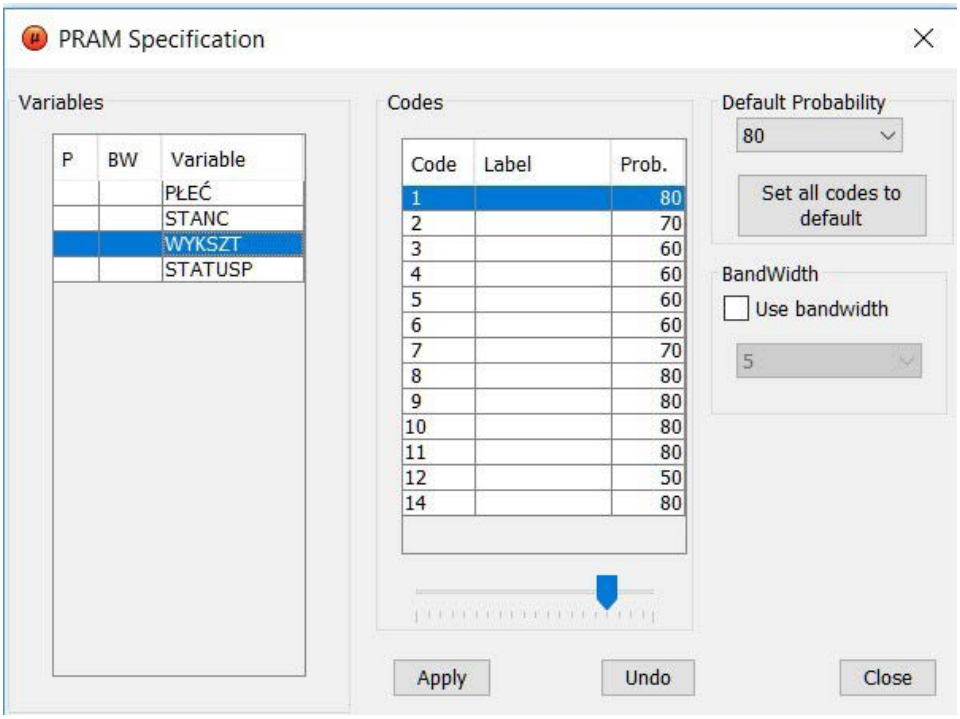
Źródło: Dane fikcyjne.

liwych w sytuacji, gdy ma ona do dyspozycji archiwum z identyfikatorami jednostki dla całej populacji oraz (zanonimizowanymi) danymi z próby badawczej. Ryzyko to jest wyrażone jako górna granica odpowiedniego prawdopodobieństwa prawidłowego połączenia owych rekordów. Tutaj – wzorując się na koncepcji Benedettiego i Franconi (1998) – definiuje się ryzyko indywidualne jako $r_i = E\left(\frac{1}{F_k} \mid f_k\right) = \sum_{h \geq f_k} \frac{1}{h} \cdot P\{F_k = h \mid f_k\}$, gdzie F_k to liczba rekordów w populacji, dla których wartości określonego zestawu zmiennych kluczowych są takie same jak dla danego rozkładu z próby, f_k zaś to liczebność próby. Wskazany rozkład warunkowy modeluje się za pomocą podejścia nadpopulacyjnego, tzn. przy założeniu, że $F_k|\pi_k$ to niezależne obserwacje z rozkładu Poissona o parametrach (N, π_k), a $f_k|F_k$ są również niezależnymi obserwacjami z rozkładu dwumianowego o parametrach (F_k, p_k). Wówczas $F_k|f_k$ ma ujemny rozkład dwumianowy z prawdopodobieństwem sukcesu p_k i liczbą sukcesów f_k. Prawdopodobieństwa p_k estymuje się na podstawie analizowanych danych metodą największej wiarygodności. Rozkład ryzyka można przedstawić w postaci graficznej; określenie indywidualnego

ryzyka może być podstawą np. do ukrywania określonych informacji lub ich przekodowywania.
- Określenie ryzyka dla gospodarstw domowych (*Household Risk Specification*). Polega na określeniu poziomu ryzyka, gdy w mikrodanych występuje identyfikator gospodarstwa domowego. Określenia „gospodarstwo domowe" autorzy programu używają ze względu na najpopularniejszy kierunek jego zastosowania, *de facto* jednak można w tym kontekście rozpatrywać dowolną grupę jednostek ustaloną w pewien określony metodologicznie sposób.
- Modyfikacja zmiennych numerycznych, czyli ilościowych (*Modify Numerical Variables*). Możliwość modyfikacji zmiennych ilościowych poprzez zaokrąglanie, przekodowanie górne i dolne lub dodawanie szumu.
- Mikroagregacja (*Numerical Micro Aggregation*).
- Wymiana rang (*Numerical Rank Swapping*).
- Generowanie danych syntetycznych (*Synthetic data*) – symulowanych, hybrydowych, którymi można zastąpić dane wrażliwe (zmienne takie wskazuje się podczas ustalania metadanych opcją Identification Level). Generowanie takie może zostać przeprowadzone albo za pomocą modelu ekonometrycznego (kombinacją liniową wartości zmiennych wrażliwych i niewrażliwych z wyrazem wolnym i losowym szumem), albo przy użyciu generatora z mikroagregacją (przy czym w uzyskanych skupieniach wartości zmiennych zastępuje się odpowiednimi wielkościami symulowanymi). W każdym z tych przypadków uzyskane dane zastępcze zachowują średnią i kowariancję oryginalnych zmiennych (więcej szczegółów na ten temat podają m.in. Hundepool i in. (2012)).

W rozpatrywanym wcześniej przypadku wybierzmy zatem metodę PRAM. Ustalamy w niej prawdopodobieństwa pozostawienia danej kategorii bez zmian jak w przykładzie 3.8, przy czym dla brakujących tam, a tu występujących kategorii przydzielamy prawdopodobieństwo 0,8 (rys. 5.11). Następnie wskazujemy kursorem myszki zmienne, które będą poddane PRAM (tzw. zmienne PRAM–med) i klikamy każdorazowo klawisz Apply (zmienne zostaną zaznaczone na czerwono). Tutaj możemy zaznaczyć w ten sposób wszystkie. Następnie klikamy Close.

Efekt zastosowania metody zobaczymy, wybierając z menu Output opcję Make protected file. Możemy tam zaznaczyć opcję unikającą ukrywania (*No suppression*). W innym przypadku niżej wskazywane są priorytety w ukrywaniu dla poszczególnych zmiennych lub użycie entropii (ukrywana jest wtedy zmienna z najniższą wartością funkcji entropii). Po prawej stronie można też wybrać sposób postępowania z identyfikatorem (*HH Identifier*): utrzymać w bezpiecznym pliku (*Keep in safe file*), zmienić na kolejne numery (*Change into sequence number*) lub usunąć (*Remove from safe file*). Można również zapisać rekordy w kolejności losowej (*Write records in random order*). Po kliknięciu klawisza Make file i wska-



Rys. 5.11. Ustawienia PRAM w programie μ-Argus

Objaśnienia jak do rysunku 5.2.

Źródło: Dane fikcyjne.

zaniu odpowiedniego katalogu zostaje wygenerowany plik w formacie SAF. Tak naprawdę jest to plik CSV z danymi rozdzielanymi przecinkami bez nagłówka, ale z wartościami w cudzysłowach.

Można też wygenerować i zapisać w formacie HTML szczegółowy raport z przeprowadzanych działań SDC.

W wersji 5.1.6 i późniejszych rozpatrywanego programu, w opcji Modify jest również możliwość przeprowadzenia celowanej wymiany rekordów (Targeted Record Swapping). Po kliknięciu tego wariantu ukazuje się stosowne okienko, gdzie można ustalić zmienne służące do wyznaczenia podobieństwa rekordów (Similar), zmienne hierarchiczne (Hierarchy), zmienne służące do oceny ryzyka (Risk) oraz zmienne przenoszone (Carry) – czyli zmienne, które są dodatkowo zakłócane w przypadku dokonywania wymian w obrębie określonej hierarchii (w celu zachowania spójności wartości na poszczególnych jej poziomach). Można też określić próg k-anonimowości (Threshold (k-anonymity)), oczekiwany odsetek wymian (Swaprate), inicjator generatora liczb losowych (Seed) oraz identyfikator gospodarstwa domowego (Household ID).

W programie μ-Argus istnieje możliwość wyznaczania entropii jako miary straty informacji. Jest to jednak formuła inna niż prezentowana wcześniej (p. wzór (4.10)). Opiera się ona mianowicie na częstości występowania danej wartości w zmiennych jakościowych i ma postać:

$$H(x) = -\frac{1}{n} \sum_{x \in K(X)} f(x) \log_2 \frac{f(x)}{n}, \tag{5.1}$$

gdzie $K(X)$ to zbiór możliwych wartości zmiennej X, a $f(x)$ – częstość występowania wartości x tejże zmiennej dla każdego $x \in K(X)$. Co ważniejsze, ta miara entropii jest w zasadzie wykorzystywana tylko do (opcjonalnego) wypracowywania balansu między niektórymi metodami SDC, np. przekodowywaniem a lokalnym ukrywaniem danych. Na podstawie jej wartości – o ile użytkownik wybierze tę możliwość – program na finalnym etapie generowania wynikowego zbioru po SDC dodatkowo ukryje pewne dane, o ile wystąpi jeszcze ryzyko ujawnienia danych wrażliwych. Ukryta zostanie wówczas wartość, dla której entropia dana wzorem (5.1) jest najmniejsza.

Tak więc ocena straty informacji ma – raczej ograniczone – zastosowanie tylko do optymalizacji niektórych rodzajów SDC dla mikrodanych. Nie jest ona przeprowadzana finalnie, to znaczy poprzez odpowiednie porównania danych wejściowych i wyjściowych oraz charakterystyk ich rozkładów i współzależności.

Więcej szczegółów na temat programu μ-Argus oraz sam program można znaleźć pod adresem https://research.cbs.nl/casc/mu.htm lub https://github.com/sdcTools/muargus/releases (tutaj znaleźć można też opcje z wbudowanym kompilatorem Java, tzw. wersje *bundle*).

5.3. Narzędzia środowiska R

W ostatnich latach narzędzia SDC pojawiły się także w środowisku R. Jest to zarówno oprogramowanie służące do analiz statystycznych i wizualizacji ich efektów, jak i specyficzny język programowania. Środowisko to ma swoją stronę centralną (https://cran.r-project.org/) oraz strony zwierciadlane w różnych krajach świata. W Polsce do niedawna taką zwierciadlaną stronę prowadził Zakład Klimatologii i Ochrony Atmosfery Instytutu Geografii i Rozwoju Regionalnego Uniwersytetu Wrocławskiego. Obecnie nasz kraj nie ma, niestety, żadnej. W związku z tym polscy użytkownicy korzystają ze stron zlokalizowanych w krajach ościennych lub w Wielkiej Brytanii. Środowisko R powstało na uniwersytecie w Auckland (The University of Auckland) w Nowej Zelandii. Nazwa pochodzi od pierwszych liter imion twórców: Roberta Gentlemana i Rossa Ihaki. Oprogramowanie do obsługi środowiska funkcjonuje w różnych systemach operacyjnych: Windows, (Mac) OS X i Linux. Jest całkowicie bezpłatne. W ramach omawiane-

go środowiska korzystający z niego tworzą własne pakiety narzędziowe służące do rozwiązywania konkretnych problemów, po czym – gdy okazują się one powszechnie przydatne – umieszczają je w ogólnodostępnych zasobach systemu[22]. Zasoby te obejmują obecnie kilkanaście tysięcy pakietów, które – w zależności od własnych potrzeb – użytkownik może sobie zainstalować i z nich korzystać. Posługiwanie się środowiskiem R ułatwiają nakładki edycyjne, które umożliwiają sprawne edytowanie zapisów w języku R oraz uruchamianie poszczególnych fragmentów zapisu i ich ewentualną korektę. Najpopularniejszą z takich nakładek jest R–Studio (https://www.rstudio.com), powszechnie znana jest też nakładka Tinn–R (https://sourceforge.net/projects/tinn-r/).

Zainteresowanym środowiskiem R i najczęściej używanymi jego narzędziami statystycznymi warto polecić książki autorstwa Biecka (2014) oraz pod redakcją Gatnara i Walesiaka (2009; 2011).

Do przeprowadzania kontroli ujawniania danych służą dwa pakiety tego środowiska:

- `sdcTable` – umożliwiający przeprowadzanie SDC w wypadku danych tabelarycznych (autorstwa Bernharda Meindla ze Statistics Austria),
- `sdcMicro` – pozwalający na kontrolę ujawniania mikrodanych (stworzony przez dr. Matthiasa Templa, niegdyś pracownika Statistics Austria i Departamentu Statystyki i Rachunku Prawdopodobieństwa Wiedeńskiego Uniwersytetu Technologicznego w Wiedniu, wykładowcy ZHAW School of Engineering[23] w Winterthur w Szwajcarii, a obecnie FHNW[24] w Olten, wspomnianego już wcześniej Bernharda Meindla oraz dr. Alexandra Kowarika ze Statistics Austria).

Zanim ukażemy charakterystykę możliwości powyższych pakietów, warto wspomnieć, że w środowisku R można sprawnie importować dane z arkusza Excel. Obecnie służy do tego pakiet `xlsx`. Trzeba jednak nadmienić, że jego instalacja i prawidłowe funkcjonowanie wymaga uprzedniego zainstalowania pakietu `rJava` umożliwiającego obsługę niezbędnych skryptów Java. Do importowania plików z Excela służy komenda `read.xlsx`. Szczegóły dotyczące stosowania tej procedury oraz innych rozwiązań rzeczonego pakietu można znaleźć w jego dokumentacji pod adresem https://cran.r-project.org/web/packages/xlsx/xlsx.pdf.

[22] Warto jednak nadmienić, że udostępnienie takie wymaga przeprowadzenia stosownej procedury weryfikacyjnej. Oznacza to, że aby pakiet był ogólnodostępny, musi przejść specjalne testy. Dodatkowym atutem przemawiającym za jego umieszczeniem w publicznych zasobach R jest wcześniejsze opublikowanie – poddanego dogłębnej recenzji – artykułu naukowego opisującego ów pakiet.

[23] Skrót ZHAW oznacza Zürcher Hochschule für Angewandte Wissenschaften (Zurychską Szkołę Nauk Stosowanych).

[24] Skrót FHNW oznacza Fachhochschule Nordwestschweiz (Wyższą Szkołę Zawodową Północno-Zachodniej Szwajcarii), mającą swoje placówki m.in. w Bazylei (Basel), Muttenz, Olten i Solurze (Solothurn).

Proces kontroli ujawniania danych realizowany przy użyciu pakietu `sdcTable` składa się z trzech zasadniczych kroków. Pierwszym jest ***przygotowanie danych***. Ładowane informacje mogą mieć dwojaką postać:

- mikrodanych, na podstawie których konstruowane są odpowiednie tablice zbiorcze,
- danych zagregowanych, w których oprócz mikrodanych zawarte są także odpowiednie predefiniowane agregacje tychże danych – na przykład oprócz danych dla każdej osoby według płci czy wieku można tam znaleźć rekord zawierający liczbę osób danej płci lub w danym przedziale wiekowym albo sumę określonych danych indywidualnych (np. kwot wydatków na określony cel); zbiór z takimi danymi zawiera także zmienną wskazującą odpowiednie częstości obserwacji.

W każdym z tych dwóch przypadków należy ustalić hierarchię sumowania. Może to być tylko klasyczna suma wszystkich wielkości lub także podsumy według określonej hierarchii. Deklarację hierarchii składa się poprzez przypisanie każdemu kodowi danej zmiennej jakościowej odpowiedniej pozycji hierarchicznej. W praktyce wygląda to tak, że tworzymy listę wymiarów zawierającą dokładnie dwie kolumny, z których pierwsza określa poziomy hierarchii, a druga zawiera etykiety odpowiednich kategorii. Poziomy hierarchii sumowania określa liczba powtórzeń symbolu „@": pojedynczy symbol oznacza sumę ogółem (najwyższy poziom), „@@" – kategorie drugiego szczebla, które wskazują składniki tej sumy, „@@@" – poziom trzeciego stopnia, czyli składniki sumy wartości dla kategorii drugiego szczebla itd. Na przykład dla płci lista wymiaru powstaje poprzez komendy:

```
> dimV1 <- matrix(nrow=0, ncol=2)
> dimV1 <- rbind(dimV1, c('@','Razem'))
> mat <- matrix(nrow=2, ncol=2)
> mat[,1] <- rep('@@',2)
> mat[,2] <- c('K','M')
> dimV1 <- rbind(dimV1, mat)
> print(dimV1)
```

i ma postać:

```
[,1] [,2]
[1,] "@" "Razem"
[2,] "@@" "K"
[3,] "@@" "M"
```

przy czym „K" oznacza kobietę, „M" – mężczyznę.

Jeśli dane są drugiego ze wskazanych wcześniej typów, tzn. zawierają predefiniowane agregacje, to trzeba je także uwzględnić w konstrukcji list wymiarowych.

Kolejny krok polega na utworzeniu obiektu klasy `sdcProblem`. W przypadku „czystych" mikrodanych odbywa się to np. w taki sposób:

```
>listawym<-list(V1=dimV1, V2=dimV2, V3=dimV3)
>danep<-makeProblem(data=dane,dimList=listawym,
dimVarInd=1:3,freqVarInd=NULL,numVarInd=4:5,weightInd=NULL,
sampWeightInd=NULL)
```

W pozycji `dimVarInd` podaje się numery kolumn, w których znajdują się zmienne wskazujące wymiary, a w pozycji `numVarInd` – numery kolumn ze zmiennymi ilościowymi (numerycznymi). Opcje `weightInd` i `sampWeightInd` dotyczą odpowiednio informacji o wagach próbkowych oraz wartości tych wag, których używa się w miejsce częstości (np. w przypadku uogólnień danych z badania reprezentacyjnego na całą populację). Natomiast gdy dane zawierają agregaty, wówczas uwzględnia się także kolumnę częstości, wpisując jej miejsce w kolejności w pliku, np. `freqVarInd=4,` i wskazując odpowiednie kolumny dla pozostałych zmiennych w `dimVarInd` oraz `numVarInd`.

Przygotowany w ten sposób obiekt można poddać ***wstępnemu ukrywaniu*** (ang. *primary suppression*). Służy do tego polecenie `primarySuppression`, w którym wskazuje się rodzaj ukrywania oraz jego kryterium. Na przykład, jeśli chcemy ukrywać częstości nie większe od 2, to wpisujemy:

```
> supdata<-primarySuppression(dane,type="freq",maxN=2).
```

Takie proste ukrywanie jednak – jak wiadomo – nie zawsze jest skuteczne, gdyż na podstawie nieukrytych informacji można nierzadko wydedukować te chronione. Jeśli takie sytuacje występują, to trzeba przeprowadzić ***ukrywanie wtórne*** (ang. *secondary suppression*). Pakiet `sdcTable` umożliwia dokonanie tego przy użyciu następujących metod:

- OPT – zapewnia całościową ochronę danych tablic, wykorzystując algorytm cięcia i gałęzi (ang. *cut and branch algorithm*, zob. np. Mitchell (2002); rozwiązanie to stosuje się jednak tylko dla niewielkich rozmiarowo przypadków;
- HiTaS (ang. *Hierarchical Tables Suppression*) – podział rozwiązywanego problemu na mniejsze podproblemy i zastosowanie algorytmu góra – dół (ang. *top – down algorithm*) do podtablic, które – zgodnie z określoną hierarchią – są chronione w szczególny sposób; jak podają Hundepool i in. (2006), znając (dzięki ukrywaniu pierwotnemu) wszystkie niebezpieczne komórki tablic, ukrywanie wtórne przeprowadza się w ten sposób, że każda podtablica tablicy głównej jest chroniona i że różne tablice nie mogą być połączone w celu cofnięcia ochrony innych (pod)tablic,
- HYPERCUBE – ochrona opiera się tutaj na heurystycznej ochronie całościowej poprzez chronienie podtablicy bazującej na znajdowaniu i ukrywaniu

struktur geometrycznych (kostek n-wymiarowych) niezbędnym do ochrony wrażliwych komórek wskazanych podczas ukrywania pierwotnego,
– SIMPLEHEURISTIC – odmiana procedury heurystycznej, która może być sprawnie zastosowana do bardzo dużych rozmiarowo problemów.

Ochrona danych może być przeprowadzona np. komendą:

```
>danechr<-protectTable(supdata,method="HITAS")
>danechrinfo<-getInfo(danechr,type='finalData')
```

Drugi wiersz pozwala na wygenerowanie informacji o zakresie chronionych danych. Tak przygotowane dane można wydrukować lub zapisać (np. w Excelu, korzystając z procedury `write.xlsx` pakietu `xlsx`).

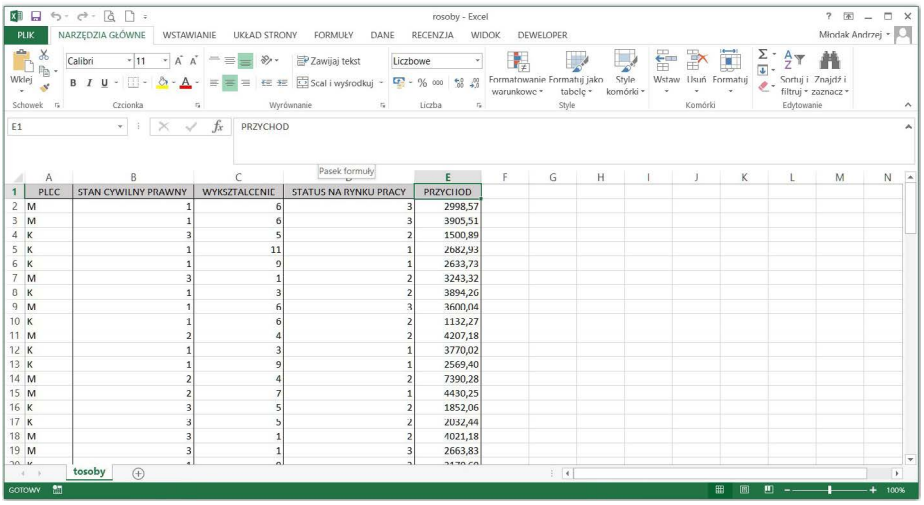
Pakiet `sdcTable` zapewnia również możliwości utworzenia plików wsadowych ze skryptem dla τ-Argus oraz ich uruchomienia w tymże programie. Więcej szczegółów na temat tego pakietu można znaleźć w jego dokumentacji (zob. https://cran.r-project.org/web/packages/sdcTable/sdcTable.pdf lub https://cran.r-project.org/web/packages/sdcTable/vignettes/sdcTable.pdf). Pisze też o tym Meindl (2011). Sporo praktycznych przykładów zastosowania pakietu podają Minami i Abe (2017).

Przykład 5.1 egzemplifikuje możliwość zastosowania pakietu `sdcTable`.

Przykład 5.1. Załóżmy, że w arkuszu Excel zgromadzono dane o osobach, o których była mowa w częściach 5.1 i 5.2, oraz że wzbogacono je o dane dotyczące miesięcznego przychodu każdej z tych osób. Wyglądają one zatem tak jak na rysunku 5.12.

Zakładamy, że żadna zmienna nie ma podkategorii i że interesują nas liczebności czy sumy określonych kategorii lub wartości numerycznych ogółem, odpowiednio. W konsoli roboczej R (RGui) z menu File wybieramy opcję ChangeDir i ustawiamy lokalizację bieżącego katalogu, w którym znajdują się analizowane pliki z danymi. Upewniwszy się, że odpowiednie pakiety są zainstalowane (Packages→Load package…), uruchamiamy skrypt:

```
library(xlsx)
danew<-read.xlsx(file="rosoby.xlsx",sheetIndex=1,
sheetName="tosoby",header=TRUE)
dim.plec<-data.frame(levels=c('@','@@','@@'),
codes=c('Total','K','M'),stringsAsFactors=FALSE)
dim.stan<-data.frame(levels=c('@',rep('@@',4)),
codes=c('Total','1','2','3','4'),stringsAsFactors=FALSE)
dim.wyk<-data.frame(levels=c(,@',rep(,@@',13)),
codes=c('Total','1','2','3','4','5','6','7','8','9','10','11',
'12','14'),stringsAsFactors=FALSE)
dim.stat<-data.frame(levels=c('@',rep('@@',3)),
codes=c('Total','1','2','3'),stringsAsFactors=FALSE)
```



	A	B	C	D	E
1	PLEC	STAN CYWILNY PRAWNY	WYKSZTALCENIE	STATUS NA RYNKU PRACY	PRZYCIOD
2	M	1	6	3	2998,57
3	M	1	6	3	3905,51
4	K	3	5	2	1500,89
5	K	1	11	1	2682,93
6	K	1	9	1	2633,73
7	M	3	1	2	3243,32
8	K	1	3	2	3894,26
9	M	1	6	3	3600,04
10	K	1	6	2	1132,27
11	M	2	4	2	4207,18
12	K	1	3	1	3770,02
13	K	1	9	1	2569,40
14	M	2	4	2	7390,28
15	M	2	7	1	4430,25
16	K	3	5	2	1852,06
17	K	3	5	2	2032,44
18	M	3	1	2	4021,18
19	M	3	1	3	2663,83

Rys. 5.12. Przykładowe mikrodane do konstrukcji tablic

Objasnienia: PRZYCHOD – miesięczny przychód w zł. Pozostałe objaśnienia jak do rysunku 5.2

Źródło: Dane fikcyjne.

```
dimList<-list(dim.plec,dim.stan,dim.wyk,dim.stat)
names(dimList)<-c('PLEC','STAN.CYWILNY.PRAWNY.',
'WYKSZTALCENIE','STATUS.NA.RYNKU.PRACY')
dimVarInd<-c(1:4)
numVarInd<-5

library(sdcTable)
pmikrodane<-makeProblem(data=danew,dimList=dimList,
dimVarInd=dimVarInd,freqVarInd=NULL, numVarInd=numVarInd,
weightInd=NULL,sampWeightInd=NULL)
pukryte<-primarySuppression(pmikrodane,type='freq',
maxN=2)
print(pukryte)
ukryte<-data.frame(pukryte.sdc=getInfo(pukryte,
type='sdcStatus'))
print(ukryte)
chronione<-protectTable(pukryte,method='HYPERCUBE')
print(getInfo(chronione,type='finalData'))
print(chronione)
```

Pierwsza część tego algorytmu formuje hierarchiczną strukturę danych i wskazuje, które zmienne określają wymiary (czyli są zmiennymi jakościowymi) oraz zmienną ilościową (numeryczną). Ostatecznie informacje zawarte są w obiektach `dimList` (lista hierarchiczna), `dimVarInd` (numery kolumn, w których znajdują się zmienne jakościowe według kolejności w pliku) oraz `numVarInd` (numer ko-

lumny ze zmienną ilościową (numeryczną)). Druga część to ukrywanie pierwotne (`primarySuppression`) typu częstościowego, oparte na zasadzie k-anonimizacji z $k = 2$. Przynosi nam to taki rezultat:

```
> print(pukryte)
The object is an 'sdcProblem' with 840 cells in 4 dimen-
sion(s)!

The dimensions are:
 - PLEC (2 levels; 3 codes; of these being 1 aggregates)
 - STAN.CYWILNY.PRAWNY. (2 levels; 5 codes; of these being
   1 aggregates)
 - WYKSZTALCENIE (2 levels; 14 codes; of these being 1 ag-
   gregates)
 - STATUS.NA.RYNKU.PRACY (2 levels; 4 codes; of these
   being 1 aggregates)
Current suppression pattern:
 - Primary suppressions: 241
 - Secondary suppressions: 0
 - Publishable cells: 599
```

Wynik ten podsumowuje, ile poziomów agregacji i kategorii ma każda zmienna, oraz wskazuje, ile z 840 możliwych kombinacji owych kategorii podlega pierwotnemu ukrywaniu (tutaj jest ich 241). Dokładny wykaz statusu każdej kombinacji otrzymamy w wyniku zastosowania instrukcji wydruku informacji o rezultatach ukrywania pierwotnego. Oto pierwsze 16 pozycji takiego wydruku:

```
> ukryte<-data.frame(pukryte.sdc=getInfo(pukryte,type='sdcStatus'))
> print(ukryte)
    pukryte.sdc
1             s
2             s
3             s
4             s
5             s
6             s
7             s
8             u
9             s
10            s
11            u
12            u
13            s
14            s
15            u
16            z
```

Symbolem „u" oznaczono tutaj kombinacje, które zostały ukryte w procesie ukrywania pierwotnego, gdyż muszą być chronione (gdyby w analizowanej tablicy wystąpiły komórki ochronione ukrywaniem wtórnym, zostałyby one oznaczone

jako „x"). Symbol „s" oznacza, że informacja dla danej kombinacji może zostać opublikowana, a „z" – że informacja nie musi być ukrywana.

Ukrywanie wtórne następuje dzięki funkcji `protectTable` wykorzystującej metody kostek wielowymiarowych (HYPERCUBE). Jego wynik jest następujący:

```
> chronione<-protectTable(pukryte,method='HYPERCUBE')
The algorithm is now starting run 1
The algorithm is now starting run 2
The algorithm is now starting run 3
> print(getInfo(chronione,type='finalData'))
      PLEC STAN.CYWILNY.PRAWNY. WYKSZTALCENIE STATUS.NA.RYNKU.PRACY Freq
  1: Total                Total         Total                  Total  100
  2: Total                Total         Total                      1   59
  3: Total                Total         Total                      2   30
  4: Total                Total         Total                      3   11
  5: Total                Total             1                  Total   10
 ---
836:     M                    4            12                      3    0
837:     M                    4            14                  Total    0
838:     M                    4            14                      1    0
839:     M                    4            14                      2    0
840:     M                    4            14                      3    0
      PRZYCHOD sdcStatus
  1: 355454.41         x
  2: 204599.94         x
  3: 118087.54         x
  4:  32766.93         x
  5:  37711.55         x
 ---
836:      0.00         s
837:      0.00         s
838:      0.00         s
839:      0.00         s
840:      0.00         s
```

Jest to fragment zestawienia ukazującego częstość występowania poszczególnych kombinacji kategorii zmiennych jakościowych oraz status ochrony zmiennej ilościowej (numerycznej – `Przychod`). Oznaczenia tego statusu są takie jak wyżej. Symbol „x" wskazuje, że dana informacja została objęta ochroną w wyniku ukrywania wtórnego. W tym wypadku wrażliwość danych na ujawniania jest zatem bardzo duża. Na zakończenie można uzyskać podsumowanie struktury uzyskanego obiektu wynikowego (`summary(chronione)`) czy szczegóły na jej temat (`print(chronione)`).

Drugi ze wspomnianych wcześniej pakietów, `sdcMicro`, służy – jako się rzekło – do przeprowadzania kontroli mikrodanych. Przeprowadzanie SDC jest tutaj proste, wymaga od użytkownika pakietu jedynie zadeklarowania, które z analizowanych zmiennych są kluczowe, które zawierają wagi (np. wynikające ze schematu losowania próby w badaniu reprezentacyjnym), a które są ilościowe (numeryczne, ciągłe). Można tutaj – podobnie jak w wypadku pakietu `sdcTable` – importować dane z arkusza Excel, ale – wykorzystując procedurę `readMicrodata` – wczytywać także pliki w formatach *.sas7bdat (SAS), *.sav (SPSS), *.dta (STATA),

*.rdata (R), RDF i CSV. Jej użycie wymaga jednak posługiwania się interfejsem `sdcApp`.

Pakiet `sdcMicro` oferuje narzędzia SDC omówione w częściach 5.1 i 5.2, wykorzystuje też pewne skrypty z programu μ-Argus. Można tutaj zatem obliczać częstość wystąpień poszczególnych kategorii zmiennych jakościowych (funkcja `freqCalc`) i na ich podstawie wyznaczyć poziom indywidualnego ryzyka ujawniania (`indivRisk`). W zakresie samej kontroli dostępne zaś są:

- metoda PRAM (`pram`),
- ukrywanie lokalne (`localSuppression`),
- nakładanie szumu (skorelowanego, o ograniczonej korelacji, opartego na wykrywaniu wielowymiarowych obserwacji odstających; funkcja `addNoise`),
- wymiana rang (`rankSwap`),
- mikroagregacja (`microaggregation`).

Mamy też tutaj kilka nowości takich jak:

- nakładanie szumu metodą ROMM (ang. *random orthogonal matrix masking*), czyli maskowania losową macierzą ortogonalną: zakłócanie postaci $\mathbf{Z} = \mathbf{AX}$, gdzie $\mathbf{A}$ jest macierzą losową, ortogonalną, czyli taką, że $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ – zob. np. Templ (2017) (funkcja `addNoise`, parametr `ROMM`),
- mikroagregacja z wykorzystaniem grupowania obserwacji; w każdej z grup uzyskanych analizą skupień obserwacje są sortowane według pierwszej głównej składowej (funkcja `microaggregation`, parametr `clustppca`),
- tasowanie (ang. *shuffling*) – zakłócanie zmiennych modelem regresji względem zmiennych bezpiecznych $\mathbf{Y} = \boldsymbol{\beta}\mathbf{S} + \boldsymbol{\varepsilon}$, gdzie $\mathbf{Y}$ to zmienne zakłócane, $\mathbf{S}$ – zmienne bezpieczne, $\boldsymbol{\beta}$ – wektor współczynników, $\boldsymbol{\varepsilon}$ – reszty modelu; ogólne addytywne zakłócanie danych (ang. *general additive data perturbation* – GADP) zapewnia przy tym, że macierz kowariancji zmiennych zakłócanych ze zmiennymi wrażliwymi oraz macierz kowariancji zmiennych wrażliwych i niewrażliwych są identyczne (funkcje: `shuffle`, `shuffle2` – CGADP, czyli metoda GADP oparta na wielowymiarowej dystrybuancie łącznego rozkładu, funkcja `robGadp` – odporny wariant GADP, funkcja `robShuffle` – odporne tasowanie),
- mikroagregacja metodą RMDM (ang. r*obust Mahalanobis distance based microaggregation*) – algorytm MDAV (o którym była mowa w części 3.2) oparty na medianie Webera (L_1–median) jako środku zbioru danych oraz odległości Mahalanobisa (funkcja `microaggregation`, parametr `rmd`).

Wynikowy zbiór mikrodanych z dokonanymi stosownymi modyfikacjami wynikającymi z zastosowanych metod kontroli ujawniania danych można zapisać w formacie Excel (polecenie `write.xlsx` pakietu `xlsx`) lub – za pomocą wymagającej także nakładki `sdcApp` – procedury `writeSafeFile` w formatach *.rdata (dane R), *.sav (SPSS), *.dta (STATA), *.csv (CSV) oraz *.sas7bdat (SAS).

Więcej szczegółów na temat tych procedur można znaleźć w dokumentacji pakietu (https://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf). Ciekawy przegląd możliwości pakietu wraz z egzemplifikacjami podał też Templ (2008; 2017).

Praktyczne zastosowanie rozpatrywanego pakietu ukazano w przykładzie 5.2.

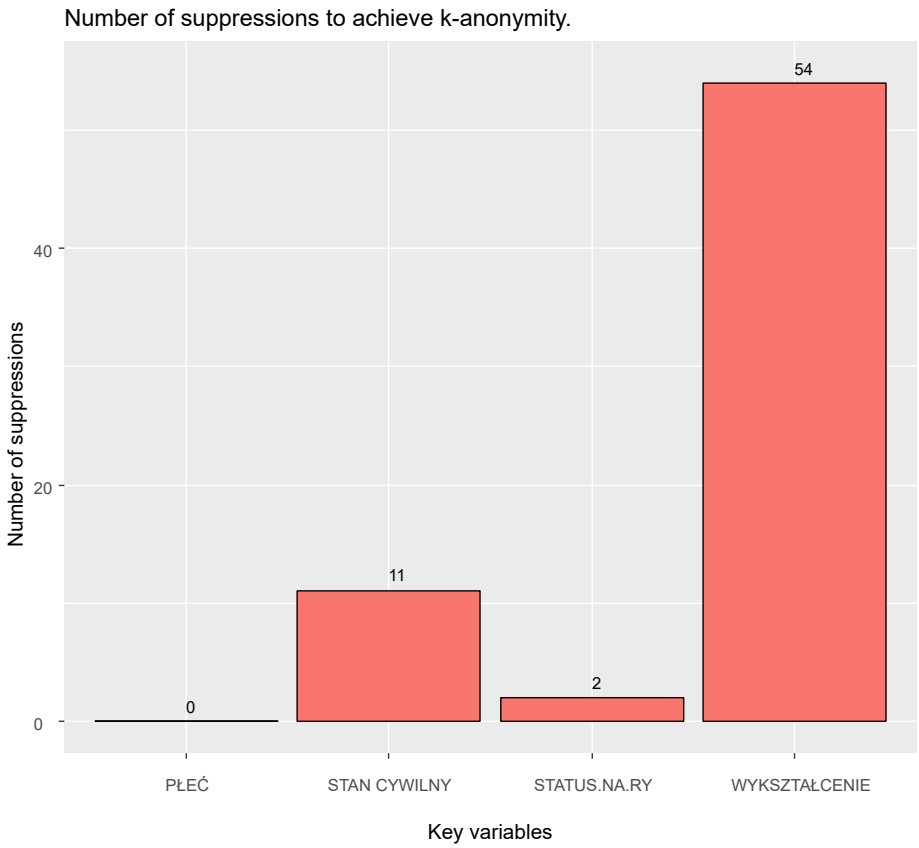
Przykład 5.2. Będziemy posługiwać się mikrodanymi wskazanymi w przykładzie 5.1 (w tym na rys. 5.12). Zastosujemy ukrywanie lokalne z $k = 3$. W tym celu uruchamiamy następujący skrypt:

```
library(xlsx)
danew<-
read.xlsx(file="rosoby.xlsx",sheetIndex=1,sheetName="tosoby",header=TRUE)
library(sdcMicro)
kv<-which(colnames(danew)%in%c('PLEC','STAN.CYWILNY.PRAWNY.',
'WYKSZTALCENIE','STATUS.NA.RYNKU.PRACY'))
czestosc<-freqCalc(danew,keyVars=kv,w=NULL)
ryzyko<-indivRisk(czestosc)
print(ryzyko)
methods(class=indivRisk)
danes<-createSdcObj(danew,keyVars=kv,numVars='PRZYCHOD')
print(danes)
danes<-localSuppression(danes,k=3)
print(danes)
plot(danes)
danes@risk$individual
danes@risk$global
danek<-extractManipData(danes)
write.xlsx(danek,file="bezpie.xlsx")
```

Algorytm ten definiuje zmienne kluczowe spośród zmiennych jakościowych (tutaj wszystkie cztery), wyznacza indywidualne ryzyko dla takich zmiennych, tworzy obiekt typu `sdcMicroObj` oraz poddaje go lokalnemu ukrywaniu i prezentuje jego wyniki, zapisując plik z zabezpieczonymi danymi. Wstępna ocena zagrożeń przynosi następujące rezultaty:

```
> print(ryzyko)
method=approx, qual=1
 --------------------------
0 obs. with high risk> methods(class=indivRisk)
[1] print
see '?methods' for accessing help and source code
> danes<-createSdcObj(danew,keyVars=kv,numVars='PRZYCHOD')
> print(danes)
Infos on 2/3-Anonymity:

Number of observations violating
 - 2-anonymity: 50 (50.000%)
 - 3-anonymity: 64 (64.000%)
 - 5-anonymity: 100 (100.000%)
```

Nie zaobserwowano zatem rekordów z wysokim ryzykiem, jednak w 64 z nich występuje ryzyko ujawnienia w ujęciu 3-anonimizacji. Podsumowanie wyników ukrywania lokalnego ukazano na rysunku 5.13, na którym widnieją liczby ukrytych wartości poszczególnych zmiennych.



Rys. 5.13. Podsumowanie ukrywania lokalnego dokonanego w pakiecie `sdcMicro`

Objaśnienia jak do rysunku 5.12.

Źródło: Opracowano z wykorzystaniem pakietu `sdcMicro`. Dane fikcyjne.

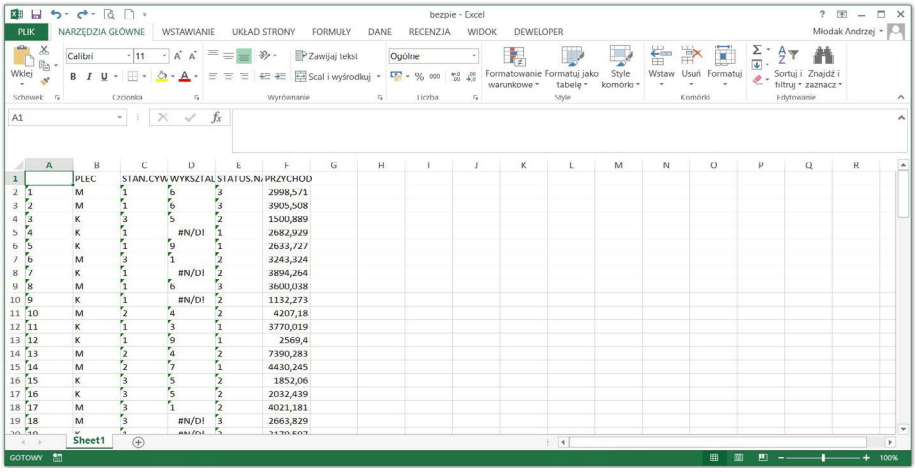
Bardziej szczegółowe rezultaty dostępne są po wydrukowaniu stosownych elementów (dla oszczędności miejsca tutaj przedstawiono tylko fragment ukazanej całej sturekordowej listy):

```
> danes<-localSuppression(danes,k=3)
> print(danes)
Infos on 2/3-Anonymity:

Number of observations violating
  - 2-anonymity: 0 (0.000%) | in original data: 50 (50.000%)
  - 3-anonymity: 0 (0.000%) | in original data: 64 (64.000%)
  - 5-anonymity: 11 (11.000%) | in original data: 100 (100.000%)
--------------------------------------------------------------
> danes@risk$individual
             risk fk Fk
  [1,] 0.25000000  4  4
  [2,] 0.25000000  4  4
  [3,] 0.12500000  8  8
  [4,] 0.06250000 16 16
  [5,] 0.16666667  6  6
  [6,] 0.14285714  7  7
  [7,] 0.09090909 11 11
  [8,] 0.25000000  4  4
  [9,] 0.09090909 11 11
 [10,] 0.25000000  4  4
-----------------------
 [90,] 0.06250000 16 16
 [91,] 0.11111111  9  9
 [92,] 0.09090909 11 11
 [93,] 0.11111111  9  9
 [94,] 0.07142857 14 14
 [95,] 0.14285714  7  7
 [96,] 0.05882353 17 17
 [97,] 0.12500000  8  8
 [98,] 0.09090909 11 11
 [99,] 0.09090909 11 11
[100,] 0.11111111  9  9
> danes@risk$global
$risk
[1] 0.1304218

$risk_ER
[1] 13.04218

$risk_pct
[1] 13.04218

$threshold
[1] 0

$max_risk
[1] 0.01
```

W pierwszej części tego wydruku otrzymujemy podsumowanie wprowadzonych zmian. Wynika z niego, że w wyniku ukrywania lokalnego wyeliminowano całkowicie możliwość identyfikacji jednostki według zasady 2- i 3-anonimowości. Jedynie w 11% pewne zagrożenie może wynikać z niezachowania zasady 5-anonimowości. Druga część wyników zawiera szacunki ryzyka ujawnienia danych – indywidualne i globalne. `Risk` oznacza oszacowanie ryzyka ujawnienia danych dla danego rekordu. Symbolami `fk` i `Fk` oznaczono tutaj liczby rekordów o takiej samej kombinacji wartości zmiennych kluczowych jak dane w próbie i w populacji odpowiednio (tutaj te dwa ujęcia są tożsame). Kolejne miejsca na tym wydruku zajmują oszacowanie globalnego ryzyka i parametry z tym związane, a zatem:

- `$risk`: ryzyko globalne (suma ryzyk indywidualnych),
- `$risk_ER`: oczekiwana liczba reidentyfikacji (identyfikacji danej jednostki na podstawie obecnych danych oraz danych z innych źródeł),
- `$risk_pct`: ryzyko globalne w procentach,
- `$threshold`: ustalony próg ryzyka ujawnienia obserwacji dla danego maksymalnego globalnego ryzyka,
- `$max_risk`: wejściowe, automatycznie wprowadzone, maksymalne ryzyko globalne.

Procedura ukrywania zastępuje ukryte dane symbolem „NA", który po zapisaniu pliku w Excelu zmienia się na „#N/D!" (rys. 5.14).



Rys. 5.14. Wynikowy plik z bezpiecznymi danymi w Excelu

Objaśnienia jak do rysunku 5.12.

Źródło: Opracowano z wykorzystaniem pakietu `sdcMicro`. Dane fikcyjne.

Pakiet `sdcMicro` przewiduje też pewien wachlarz możliwości w zakresie oceny straty informacji. Dotyczy on efektów niektórych metod zakłóceniowych (mikroagregacja, dodawanie szumu oraz wymiana rang). Zawarty jest w wyjścio-

wych parametrach funkcji `summary` dla obiektu klasy „micro". Zakres miar straty informacji obejmuje następujące wskaźniki:

- `amean` – przeciętna względna wartość bezwzględnych różnic między średnimi arytmetycznymi (por. wzór (4.9)),
- `amedian` – przeciętna względna wartość bezwzględnych różnic między medianami,
- `aonestep` – przeciętna względna wartość bezwzględnych jednokrokowych odchyleń od mediany (jeśli obserwacja jest poza przedziałem med(X) ± (3/1,486) mad(X), to podstawia się w jej miejsce dolną lub górną granicę powyższego przedziału, zależnie od tego, czy jest ona mniejsza od pierwszej, czy też większa od drugiej z nich, a następnie oblicza się sumę względnych wartości bezwzględnych różnic między średnimi dla takich zmiennych),
- `devvar` – przeciętna względna wartość bezwzględnych różnic między wariancjami,
- `amad` – przeciętna względna wartość bezwzględnych różnic między medianowymi odchyleniami bezwzględnych (mad),
- `acov` – przeciętna względna wartość bezwzględnych różnic między kowariancjami,
- `arcov` – przeciętna względna wartość bezwzględnych różnic między kowariancjami odpornymi (w ujęciu minimalizacji wyznacznika macierzy kowariancji obserwacji),
- `acor` – przeciętna względna wartość bezwzględnych różnic między współczynnikami korelacji,
- `arcor` – przeciętna względna wartość bezwzględnych różnic między odpornymi współczynnikami korelacji (w ujęciu minimalizacji wyznacznika macierzy kowariancji obserwacji),
- `acors` – przeciętna względna wartość bezwzględnych różnic korelacji rangowych,
- `adlm` – przeciętna bezwzględna różnica między współczynnikami regresji liniowej (bez wyrazu wolnego),
- `adlts` – przeciętna bezwzględna różnica między współczynnikami regresji wyznaczonymi metodą najmniejszych uciętych kwadratów (bez wyrazu wolnego),
- `apcaload` – przeciętna bezwzględna różnica między ładunkami głównych składowych,
- `apppacaload` – przeciętna bezwzględna różnica między ładunkami głównych składowych w ujęciu odpornym (wedle podejścia rzutu skierowanego),
- `atotals` – przeciętna względna wartość bezwzględnych różnic między sumami ogółem,
- `pmtotals` – przeciętna względna różnica między sumami ogółem.

Ważne jest, żeby do przeprowadzenia SDC zastosować instrukcję `valTable`. Warto jednak zauważyć, że instrukcja ta ma zastosowanie tylko do zmiennych ilościowych oraz że jest oparta w zasadzie na pewnych domyślnych formach odpowiednich metod SDC, podczas gdy inne szczegółowe procedury pozwalają na swobodny dobór różnorodnych ich parametrów. Powyższe mierniki zaliczają się do różnych grup według klasyfikacji zaprezentowanej w rozdziale 4. Wskaźniki `amean`, `amedian`, `aonestep`, `atotals` i `pmtotals` można określić jako miary zakłócenia rozkładu, `devvar`, `amad`, `acov`, `arcov` – jako miary wpływu na wariancję szacunków (choć kowariancje częściowo obrazują także związki między zmiennymi), pozostałe zaś – jako miary wpływu na siłę związku.

Podamy obecnie egzemplifikację zastosowania funkcji `valTable` pakietu `sdcMicro` środowiska R do oceny rozmiaru straty informacji (przykład 5.3).

Przykład 5.3. Posłużymy się oryginalnymi danymi z przykładu 3.6. Zastosowane zostaną następujące metody SDC: PCA (parametr pca; oparta na sortowaniu według pierwszej głównej składowej), MDAV (mdav) oraz RMDM (wskazywana parametrem rmd). W celu oszacowania straty informacji ustawiamy katalog roboczy i uruchamiamy skrypt:

```
library(xlsx)
danew<-read.xlsx(file="mosobyp.xlsx",sheetIndex=1,header=TRUE)
library(sdcMicro)
v<-valTable(danew, method=c("pca","mdav","rmd"))
print(v[1:14])
```

Otrzymujemy następujący wydruk:

```
[1] "method pca will be applied"
[1] "method mdav will be applied"
[1] "method rmd will be applied"
[1] "calculating summary statistics"
  method amean amedian aonestep devvar  amad  acov  acor acors   adlm apcaload
1    pca     0   0.511    0.092  3.544 0.734 1.772 3.228 1.692 97.986    0.967
2   mdav     0   0.476    0.035  1.973 0.830 0.986 1.660 0.819 65.328    0.699
3    rmd     0   0.284    0.071  0.546 0.806 0.273 0.378 0.873 34.753    0.197
  apppcaload atotals pmtotals
1      3.658       0        0
2      0.938       0        0
3      2.082       0        0
>
```

Z przykładu wynika, że zastosowanie SDC nie zmienia wartości średnich ani sum ogółem, natomiast przynosi pewne straty na wariancji i współzależności zmiennych. Najlepsza pod tym względem wydaje się metoda RMDM. Oprócz tego w pakiecie `sdcMicro` dostępne są następujące miary:

– miara IL1 dla zmiennych ilościowych dana wzorem $\mathrm{IL1} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\left|x_{ij} - x_{ij}^{*}\right|}{\sqrt{2}S_j}$, gdzie: S_j – odchylenie standardowe oryginalnej cechy ilościowej X_j, x_{ij} – oryginalna wartość cechy X_j dla jednostki i, x_{ij}^{*} – wartość cechy X_j dla jednostki i po SDC, n – liczba jednostek, m – liczba zmiennych,
– różnica w wartościach własnych (ang. *difference of eigenvalue*s): miara oparta na zrelatywizowanych bezwzględnych odległościach między wartościami własnymi macierzy kowariancji zmiennych ilościowych.

Analizy empiryczne prowadzone przez autorów książki wykazały, że miara IL1 jest wrażliwa na odchylenie standardowe – dla zmiennych słabo zróżnicowanych może przyjmować nadmierne wartości. Jeżeli natomiast niektóre ze zmiennych ilościowych w bazie dotyczą tylko określonych subpopulacji (np. tylko osób w wieku poprodukcyjnym), to sporo informacji program rozpoznaje jako brakujące, co powoduje problemy z obliczeniem drugiej z analizowanych miar.

Począwszy od wersji 5.6.1. pakietu, zestaw dostępnych miar straty informacji rozszerzono o miary zaproponowane przez polskich statystyków – wyrażoną wzorem (4.1) z odległościami wyrażonymi zależnościami (4.2), (4.3) i (4.6) (w tym z możliwością określania straty informacji zarówno globalnie, jak i na poszczególnych zmiennych) oraz wyrażoną wzorem (4.12) miarę straty korelacji dla zmiennych ilościowych. Są to odpowiednio funkcje `IL_variables` oraz `IL_correl`. Uwzględniają one wszystkie typy zmiennych i nie powodują wielu problemów, o których była mowa wyżej.

Pakiet `cellKey` stanowi bibliotekę narzędzi środowiska R służących do przeprowadzania nakładania *ex post* zakłóceń na dane tabelaryczne metodą kluczy komórkowych (ang. *cell-key*). Autorem pakietu jest Bernhard Meindl ze Statistik Austria. Narzędzie to udostępniono na platformie GitHub (pod następującym adresem: https://github.com/sdcTools/cellKey). Ponieważ jest ono ciągle doskonalone i cały czas podlega testowaniu, nie zostało jeszcze włączone do repozytorium CRAN całego środowiska. Można zatem z niego korzystać tylko na platformie GitHub, a instalacja wymaga zastosowania podanej tam procedury. W razie potrzeby niezbędne staje się także wskazanie danych serwera proxy. Jak wszystkie narzędzia środowiska R, pakiet jest ogólnodostępny (*open source)*, a jego wymagania systemowe są takie same jak w całym środowisku R. Program powstał przede wszystkim dla potrzeb spisowych, ale tak naprawdę jest nieco bardziej ogólny: może być stosowany w zasadzie do dowolnych (ważonych) tablic wielkości i (ważonych) tablic częstości w sposób opisany w pracy Thompson i in. (2013).

Przypomnijmy, że – jak podają np. Dove i in. (2017) czy Dove i in. (2018) – metoda kluczy komórkowych jest metodą posttablicową. Każdemu rekordowi przyporządkowuje się najpierw liczbę losową, po czym tworzy się docelową tablicę częstości. Z kolei dla każdej komórki tej tablicy klucze rekordów wchodzą-

cych w jej skład sumuje się, a następnie wyznacza resztę z dzielenia owej sumy przez liczbę rekordów w populacji. Jest to właśnie klucz komórkowy. Kolejny krok polega na wygenerowaniu tablicy zakłóceń, której wiersze odpowiadają poszczególnym możliwym wartościom komórek, kolumny – ich kluczom komórkowym, a komórki zawierają odpowiednie wielkości szumu. Wielkość wskazaną przez taką pomocniczą tablicę dla danej wielkości komórki oryginalnej tablicy i odpowiadającego jej klucza komórkowego dodaje się do owej oryginalnej wartości jako szum.

Tak naprawdę dla usprawnienia procesu testowania i modernizacji biblioteka dotycząca tej metody składa się z dwóch pakietów środowiska R. Pierwszym z nich jest pakiet `cellKey`, zawierający zasadniczy algorytm opisywanej metody, drugim zaś – `ptable`, służący do generowania tablicy zakłóceń (ang. *perturbation table* lub skrótowo *p–table*). Tablica ta jest konstruowana z zastosowaniem podejścia największej entropii wykorzystywanego do obliczenia prawdopodobieństw przejścia uwzględniającego pewne progi pożądanego natężenia szumu (Marley i Leaver, 2011; Giessing, 2016). Pakiet `cellKey` używa tejże tablicy do wyznaczenia zakłóconych wartości tablic częstości i wielkości. W przypadku złożonych hierarchii jednostek korzysta się także z pakietu `sdcHierarchies`.

Dane do konstruowania tablic wczytuje się w tradycyjny w środowisku R sposób, np. z arkusza Excel, korzystając z odpowiedniego pakietu (np. `xlsx` lub `readxl`). Następnie tworzy się – również klasycznymi metodami – dodatkowe zmienne zerojedynkowe, umożliwiające zliczanie rekordów o odpowiednich cechach (np. kobiet lub osób o przeciętnym wynagrodzeniu miesięcznym brutto wyższym niż 4000 zł). Poszczególne funkcje pakietu `cellKey` dotyczą kolejnych etapów procedury. W takiej też kolejności je przedstawimy. Są to zatem funkcje:

- `ck_generate_rkeys` – losująca klucze rekordów i dodająca je do zbioru z danymi,
- `hier_create` – dająca możliwość ustalenia hierarchii jednostek w zbiorze danych (funkcja z pakietu `sdcHierarchies`),
- `ck_setup` – tworząca obiekt, który można potem modyfikować, zawierający wszystkie elementy niezbędne do nakładania zakłóceń (dane, wskazanie zmiennych zerojedynkowych do zliczania i zmiennych ilościowych – numerycznych, czyli ciągłych, hierarchie, wagi itp.),
- `ck_param_cnts` – ustalająca parametry generowania tablicy zakłóceń dla zmiennych zliczających (funkcja zagnieżdżona w pakiecie `ptable`); parametry: `D` – parametr zakłócenia dla zakłócenia maksymalnego (liczba lub wektor), `V` – parametr zakłócenia dla wariancji (liczba), `type` – określenie formatu tablicy wejściowej: używa się opcji `abs` lub `destatis` (według wersji stosowanych w Australii lub w Niemczech, odpowiednio), `js` – wartość progowa dla blokowania małych częstości (tzn. dodatnie częstości niższe niż wartość progowa nie będą się pojawiać), `pstay` – parametr

opcjonalny: prawdopodobieństwo p ($0 < p < 1$), że oryginalna wielkość pozostanie niezakłócona, `optim` – parametr optymalizacyjny, standardowo przyjmuje wartość 1, `mono` – (logiczny) wektor określający parametr optymalizacyjny dla warunku monotoniczności,

- `ck_param_nums` – ustalająca parametry generowania tablicy zakłóceń dla zmiennych ilościowych (funkcja zagnieżdżona w pakiecie `ptable`); obejmuje m.in. parametry: `D` – parametr zakłócenia dla zakłócenia maksymalnego (liczba lub wektor), `type` – określenie rodzaju wielkości w komórce użytych podczas zakłócania (np. k największych udziałów, średnia arytmetyczna, rozstęp, suma), `mult_params` – funkcja zgięcia (ang. *flex function*) określona przez `ck_flexparams`, definiująca m.in. maksymalne wielkości zakłóceń dla różnego rodzaju komórek, czy `pos_neg_var` – zapewniająca nieujemność zakłóconych wartości,
- `params_cnts_set` – przypisująca do zmiennych zliczających parametry określone funkcją `ck_param_cnts`; dostęp do niej wymaga podania nazwy zbioru i nazwy zmiennej rozdzielonych znakiem dolara (np. `dane$params_cnts_set`),
- `params_nums_set` – przypisująca do zmiennych ilościowych parametry określone funkcją `ck_param_nums`, odpowiednio; dostęp do niej wymaga podania nazwy zbioru i nazwy zmiennej rozdzielonych znakiem dolara (np. `dane$params_nums_set`),
- `perturb` – nakładająca zakłócenia dla wskazanych zmiennych; dostęp do niej wymaga podania nazwy zbioru i nazwy zmiennej rozdzielonych znakiem dolara (np. `dane$perturb`),
- `freqtab` – naliczająca tablicę częstości z zakłóceniami; dostęp do niej wymaga podania nazwy zbioru i nazwy zmiennej rozdzielonych znakiem dolara (np. `dane$freqtab`),
- `numtab` – naliczająca tablicę wielkości z zakłóceniami; dostęp do niej wymaga podania nazwy zbioru i nazwy zmiennej rozdzielonych znakiem dolara (np. `dane$numtab`),
- `measures_cnts` – ukazująca rozkład nałożonego szumu; dostęp do niej wymaga podania nazwy zbioru i nazwy zmiennej rozdzielonych znakiem dolara (np. `dane$measures_cnts`).

Zapisu tablic wynikowych z zakłóceniami można dokonać za pomocą tradycyjnych narzędzi środowiska R – np. w formacie Excel przy użyciu pakietu `xlsx` lub `writexl`. Więcej informacji na temat możliwości pakietu `cellKey` oraz teoretyczny opis stosowanych w nim rozwiązań podali Meindl i Enderle (2019).

Pakiet `cellKey` ma wbudowane metody oceny straty informacji w zakresie miar zakłócenia rozkładu postaci:

- d_1 – wartość bezwzględna różnicy między wartościami oryginalnymi a zakłóconymi,

- d_2 – względna wartość bezwzględna różnicy między wartościami oryginalnymi i zakłóconymi,
- d_3 – wartość bezwzględna różnicy pomiędzy pierwiastkami kwadratowymi wartości oryginalnych i zakłóconych.

Ponadto dla każdej odległości d_1, d_2 i d_3 podawane tu są:

- `cat` – określona wartość (dla d_1) lub przedział (dla d_2 i d_3),
- `cnt` – liczba rekordów, dla których odległość jest mniejsza lub równa odpowiedniej wielkości `cat`,
- `pct` – udział liczby rekordów, dla których odległość jest mniejsza lub równa odpowiedniej wielkości, w odpowiedniej wartości `cat`.

Pakiet `recordSwapping` jest przeznaczony do stosowania w procesie kontroli ujawniania mikrodanych metody celowanej wymiany rekordów. Jego autorem, wymienianym w winietce pakietu, jest Johannes Gussenbauer, choć w pliku z jego opisem natrafić można jeszcze na dwa inne nazwiska: Bernarda Meindla oraz Alexandra Kowarika. Implementacja procedury została wykonana wyłącznie w języku C++; jest oparta na kodzie SAS. Celowaną wymianę rekordów można wykonać za pomocą funkcji `recordSwap`. Wszystkie inne funkcje w tym pakiecie wywoływane są z wnętrza funkcji `recordSwap` oraz eksportowane jedynie do celów testowych.

Funkcja `recordSwap()` ma następujące argumenty:

- `data` – zbiór mikrodanych zawierający tylko liczby całkowite[25],
- `similar` – wektor zawierający profile podobieństwa, czyli zestawy zmiennych, które należy wziąć pod uwagę przy zamianie gospodarstw domowych,
- `hierarchy` – indeksy kolumn zmiennych odnoszące się do hierarchii geograficznej lub innej w mikrodanych,
- `risk` – indeksy kolumn zmiennych, które będą brane pod uwagę przy szacowaniu ryzyka (argument ten jest używany tylko wtedy, gdy nie podano parametru *k*-anonimowości `th`),
- `hid` – indeks kolumny, który odnosi się do identyfikatora gospodarstwa domowego,
- `th` — liczba całkowita określająca próg gospodarstw domowych wysokiego ryzyka (*k* w zasadzie *k*-anonimowości),
- `swaprate` — zmienna typu double, przyjmująca wartości pomiędzy 0 a 1, określająca odsetek gospodarstw domowych, które powinny zostać zamienione,
- `seed` — liczba całkowita definiująca ziarno dla generatora liczb losowych (w celu odtworzenia).

[25] Jeżeli zmienne są innego typu, to trzeba je zamienić na całkowitoliczbowe, używając na przykład funkcji `as.integer`. Pakiet obsługuje wprawdzie obecnie także używane w `sdcMicro` obiekty klasy `sdcMicroObj`, ale i w tym wypadku taka zamiana formatu jest niezbędna.

Funkcja `recordSwap` wraz z niezbędnymi narzędziami została włączona do pakietu `sdcMicro` w repozytorium CRAN. Pakiet można też zainstalować odrębnie, korzystając z platformy GitHub. Więcej informacji na ten temat oraz instrukcję instalacji można znaleźć pod adresem https://github.com/sdcTools/recordSwapping.

Do konstruowania danych syntetycznych może służyć kilka pakietów. Pierwszym z nich jest pakiet `simFrame` napisany przez Andreasa Alfonsa z Erasmus Universiteit Rotterdam (Holandia). Zawiera on narzędzia i metody do symulacji populacji dla potrzeb związanych z badaniami statystycznymi opartej na danych pomocniczych. Znajdziemy tutaj zatem metody bazujące na modelach, a także optymalizacyjne algorytmy kalibracyjne i kombinatoryczne. W tym kontekście omawiany pakiet może być wykorzystany dla potrzeb tworzenia danych syntetycznych w celu ochrony informacji wrażliwych przed ich ujawnieniem. W zakresie realizacyjnym, biorąc pod uwagę kontrolę ujawniania danych, najbardziej przydatne są następujące funkcje pakietu `simFrame`:

- `generate` – funkcja generująca dane na podstawie danego modelu (rozkładu teoretycznego),
- `draw` – losowanie próby,
- `contaminate` – zanieczyszczanie danych,
- `runSimulation` – uruchomienie eksperymentu symulacyjnego; uwzględnia m.in. operat losowania, próbki arbitralne lub klasę kontrolną dla ich losowania, liczbę powtórzeń losowania, kontrolę zanieczyszczeń lub braków danych itp.,
- `clusterSetup` – zdefiniowanie wielokrotnego losowania ze skupienia,
- `clusterRunSimulation` – uruchomienie eksperymentu symulacyjnego na skupieniu; uwzględnia m.in. operat losowania, próbki arbitralne lub klasę kontrolną dla ich losowania, liczbę powtórzeń losowania, kontrolę zanieczyszczeń lub braków danych itp.,
- `aggregate` – agregacja wyników symulacji, czyli podział danych na podzbiory (jeśli to konieczne) oraz wyznaczenie stosownych statystyk opisowych.

Więcej informacji na temat pakietu można znaleźć na stronie https://cran.r-project.org/web/packages/simFrame/simFrame.pdf.

Pakiet `simPop`, również dostępny w repozytorium CRAN R, służy do generowania danych syntetycznych na podstawie danych z badania statystycznego z uwzględnieniem informacji pomocniczych. Jednym z kierunków tworzenia takowych danych jest ochrona zawartych w nich informacji wrażliwych przed nieuprawnionym ujawnieniem. Autorzy pakietu to:

- Matthias Templ – Fachhochschule Nordwest Schweiz w Olten, Szwajcaria,
- Alexander Kowarik – Statistik Austria,
- Bernhard Meindl – Statistik Austria,

– Andreas Alfons – Erasmus Universiteit Rotterdam, Holandia,
– Mathieu Ribatet – Université de Nantes, Nantes, Francja,
– Johannes Gussenbauer – Statistik Austria.

Podstawową klasą obiektu w pakiecie jest `simPopObj-class`. Obiekt tej klasy zawiera informacje o próbie, populacji i opcjonalnie pewne wartości brzegowe w formie tablicy. Do generowania danych syntetycznych pod kątem kontroli ujawniania danych mają zastosowanie przede wszystkim następujące funkcje:

– `specifyInput` – tworząca standaryzowany obiekt wejściowy klasy `dataObj` zawierający informację o wagach, identyfikatorach i rozmiarach gospodarstw domowych, identyfikatorach osób i – opcjonalnie – o warstwach,
– `simStructure` – symulująca podstawowe zmienne jakościowe definiujące strukturę gospodarstw domowych – np. identyfikator, wiek (wiek uważa się tutaj za zmienną jakościową), płeć czy głowa gospodarstwa – dla populacji na podstawie ponownego próbkowania z danych pozyskanych w badaniu; można też *sensu stricto* skategoryzować wiek, biorąc pod uwagę odpowiedni podział na grupy,
– `simCategorical` – symulująca zmienne jakościowe w danych dla populacji; w celu prawidłowego przeprowadzenia symulacji uprzednio należy zasymulować strukturę gospodarstwa domowego za pomocą funkcji `simStructure`,
– `getBreaks` – wyznaczająca punkty przełamania (granic przedziałów wartości) w kategoryzacji zmiennych ciągłych lub półciągłych[26] z wykorzystaniem kwantyli (ewentualnie ważonych),
– `simContinuous` – symulująca zmienne ilościowe (ciągłe) w danych dla populacji z wykorzystaniem wielomianowych modeli log-liniowych łączonych z liczbami losowymi z finalnych kategorii lub (dwukrokowymi) modelami regresyjnymi z błędem losowym; w celu prawidłowego przeprowadzenia symulacji uprzednio należy zasymulować strukturę gospodarstwa domowego za pomocą funkcji `simStructure`, podobnie jak inne predyktory jakościowe,
– `simComponents` – symulująca komponenty zmiennych ilościowych (np. dochodu) dla danych populacyjnych poprzez ponowne próbkowanie odpowiednich frakcji danych z badania; w celu prawidłowej realizacji tej czynności zmienna ilościowa musi być uprzednio podzielona na składowe, a każda zmienna jakościowa definiująca warunki – zasymulowana,
– `simRelation` – symulująca zmienne jakościowe dla danych populacyjnych z uwzględnieniem relacji pomiędzy członkami gospodarstwa domo-

[26] To zaczy przyjmujących dla znacznej liczby jednostek wartości zerowe.

wego: w celu prawidłowej realizacji tej czynności należy najpierw zasymulować strukturę danych w populacji przy użyciu `simStructure`.

W wyniku symulacji otrzymujemy obiekt, w którym zapisana jest zarówno wyjściowa próbka (`samp`), jak i zasymulowana populacja (`pop`). Można na nich, przy użyciu typowych statystyk opisowych, dokonywać stosownych obliczeń. Szczegółową dokumentację pakietu znaleźć można na poświęconej mu stronie CRAN (https://cran.r-project.org/web/packages/simPop/simPop.pdf). Dokładny opis narzędzi teoretycznych oraz prezentację zastosowania poszczególnych funkcji podał m.in. Templ (2017).

Trzeci z rozpatrywanych tutaj pakietów to `synthpop`. Stanowi on narzędzie służące do tworzenia syntetycznych wersji mikrodanych zawierających poufne informacje – tworzenia w taki sposób, aby były one bezpieczne podczas udostępniania danych użytkownikom dla analiz eksploracyjnych. Chodzi tutaj przede wszystkim o zastąpienie oryginalnych wrażliwych wartości przez dane syntetyczne przy minimalnym zniekształceniu informacji zawartych w zbiorze. Zmienne (jakościowe lub ilościowe) zostają zsyntezowane pojedynczo za pomocą modelowania sekwencyjnego. Powtórzenia są generowane poprzez losowanie z warunkowych rozkładów dopasowanych do danych oryginalnych z wykorzystaniem parametrycznych lub klasyfikacyjnych modeli drzew regresyjnych. Poprzez odpowiedni dobór parametrów użytkownik może wpłynąć na ryzyko ujawnienia danych wrażliwych i analityczną jakość uzyskanych danych syntetycznych.

Pakiet jest dostępny w zasobach repozytorium CRAN środowiska R. Autorami pakietu są:

- Beata Nowok – Institute of Geography, School of GeoSciences, University of Edinburgh (Szkocja, Wielka Brytania),
- Gillian M. Raab – Institute of Geography, School of GeoSciences, University of Edinburgh (Szkocja, Wielka Brytania),
- Joshua V. Snoke – RAND Corporation, Pittsburgh Office, Pittsburgh, PA (Stany Zjednoczone),
- Chris Dibben – Institute of Geography, School of GeoSciences, University of Edinburgh (Szkocja, Wielka Brytania).

Dane można wczytać w tradycyjny sposób, przedstawiony wcześniej w tym rozdziale. Dodatkowo wbudowana funkcja `read.obs` pozwala na import zbiorów danych z plików zewnętrznych w formatach: *.sav (SPSS), *.dta (Stata), *.xpt (SAS), *.csv (dane w pliku rozdzielane przecinkami) oraz *.tab (dane rozdzielane tabulatorami) i *.txt (plik tekstowy z ogranicznikami). Generowanie danych syntetycznych odbywa się tu przede wszystkim przy użyciu funkcji `syn` – generowanie zbiorów danych syntetycznych. Umożliwia ona generowanie danych syntetycznych między innymi przy użyciu metody FCS, podejścia CART (także w odniesieniu do szeregów czasowych), normalnej regresji liniowej (w tym w wa-

riancie zachowującym rozkłady brzegowe czy po transformacji logarytmicznej albo pierwiastkowej zmiennej zależnej), regresji logistycznej, regresji wielomianowej, predykcyjnego dopasowania średniej, próbek bootstrapowych czy modeli logarytmiczno-liniowych. Można tutaj także określać porządek syntetyzacji. Służy do tego macierz predykcyjna. Jest to macierz kwadratowa o rozmiarze równym liczbie zmiennych, definiująca zestaw predyktorów do zastosowania dla każdej zmiennej docelowej w wierszu; każdy element takiej macierzy ma wartość 0 lub 1, przy czym wartość 1 oznacza, że zmienna odpowiadająca danej kolumnie jest używana jako predyktor odpowiedniej zmiennej określonej przez wiersz. Procedura pozwala ponadto uwzględnić ograniczenia wartości wynikające z definicji określonych zmiennych czy predefiniowane grupy jednostek (na podstawie wskazanych zmiennych grupujących), zastosować wygładzanie dla szeregów czasowych oraz wykorzystać ewentualne warstwowanie próby.

Dane syntetyczne są tu generowane jako obiekt `synds`. Eksport owych danych do formatu Excela dokonuje się w sposób klasyczny dla środowiska R. Ponadto rozpatrywany pakiet zawiera funkcję `write.syn` umożliwiającą ich eksport do plików innych typów: *.sav (SPSS), *.dta (Stata), *.sas7bdat (SAS), *.csv, *.tab, *.rda, *.RData i *.txt. Pełną dokumentację omówionego powyżej pakietu można pobrać pod adresem https://cran.r-project.org/web/packages/synthpop/synthpop.pdf. Szerzej o możliwościach pakietu pisali na przykład Nowok i in. (2016).

Dobrym narzędziem syntetyzacji danych może być także pakiet `mice`. Jego autorem jest zespół 18 naukowców z różnych krajów (Stef van Buuren, Karin Groothuis-Oudshoorn, Gerko Vink, Rianne Schouten, Alexander Robitzsch, Patrick Rockenschaub, Lisa Doove, Shahab Jolani, Margarita Moreno-Betancur, Ian White, Philipp Gaffert, Florian Meinfelder, Bernie Gray, Vincent Arel-Bundock, Mingyang Cai, Thom Volker, Edoardo Costantini, Caspar van Lissa) pod kierunkiem Stefa van Buurena z TNO (nl. Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek – Holenderska Organizacja Stosowanych Badań Naukowych). Pakiet powstał właściwie w celach realizacji imputacji wielokrotnej, ale ustawiając odpowiednio stosowne parametry procedury `mice` można tutaj także generować dane syntetyczne. Procedura ta umożliwia m.in. stosowanie metody FCS, CART, Bayesowskiej regresji liniowej, regresji liniowej (w tym ignorującej błędy modelu czy z bootstrappem), regresji liniowej lasso oraz regresji logistycznej w kilku opcjach. Więcej na temat pakietu można przeczytać pod adresem https://cran.r-project.org/web/packages/mice/mice.pdf.

Porównanie przeprowadzone przez autorów niniejszej książki na zbiorze mikrodanych pochodzącym z amerykańskiego badania wspólnotowego (American Community Survey), obejmującym 1 035 201 rekordów i 33 zmienne, ukazało, że generowanie danych syntetycznych metodą FCS i predykcyjnego dopasowania średnich jest bardziej efektywne w wypadku pakietu `mice` (trwało to wówczas 40 minut) niż `synthpop` (zajmuje wiele godzin).

Na zakończenie prezentacji warto nadmienić, że pakiet sdcMicro umożliwia dokonanie syntetyzacji metodą IPSO. Istnieją też pakiety: semTools (https://cran.r-project.org/web/packages/semTools/semTools.pdf), który umożliwia symulację mikrodanych z użyciem macierzy kowariancji, skośności i kurtozy z danych oryginalnych, oraz zaproponowany przez Emmanuela Lazaridisa z Data Cycle Limited w Truro (Wielka Brytania) pakiet sdcTarget, którego zasoby umożliwiają sprawne wyznaczanie macierzy podstawień w SDC (jeden z elementów MASSC, zob. część 2.3.1). Szczegóły podane są w dokumentacji pakietu: https://cran.microsoft.com/snapshot/2016-08-21/web/packages/sdcTarget/index.html.

5.4. Aplikacja sdcApp()

Aplikacja sdcApp() to przygotowany w języku Shiny graficzny interfejs użytkownika (GUI) dla pakietu sdcMicro w środowisku R. Żeby z niej skorzystać, wystarczy – po uprzednim pobraniu z repozytorium CRAN bądź z innego źródła (np. z platformy GitHub), zainstalowaniu i wczytaniu z biblioteki wspomnianego pakietu (wraz ze wszystkimi wymaganymi przez niego pakietami zależnymi) – wpisać w konsoli programu polecenie sdcApp() i je uruchomić. Aplikacja ta umożliwia przeprowadzenie procesu kontroli ujawniania mikrodanych, począwszy od przetworzenia zbioru danych jednostkowych na wejściu zgodnie z wymaganiami pakietu, a na eksporcie zasobów gotowych do udostępnienia dla zewnętrznego użytkownika skończywszy.

Po uruchomieniu procedury w oknie domyślnej przeglądarki internetowej otworzy się GUI aplikacji. Składa się on z siedmiu poniższych kart, spośród których jako pierwsza jest wymieniona startowa:

- O\Pomoc (*About\Help*) – w karcie znajdują się informacje o aplikacji i dane kontaktowe autorów, ale jest też opcja ustawienia ścieżki dla danych wynikowych, zatrzymania i restartu aplikacji,
- Mikrodane (*Microdata*) – karta ta umożliwia wczytanie mikrodanych oraz ich wyświetlanie i przetworzenie w celu przygotowania pod wymagania pakietu sdcMicro przed przystąpieniem do przeprowadzenia procesu SDC,
- Anonimizacja (*Anonymize*) – w karcie tej formułuje się zadanie kontroli ujawniania mikrodanych (poprzez przypisanie zmiennym odpowiednich ról oraz ustawienie parametrów), ocenia się ryzyko ujawnienia informacji poufnych w wejściowych zasobach, zapewnia ochronę tajemnicy statystycznej poprzez wywoływanie wybranych metod SDC, jak również ocenia ich wpływ na zapewnianą poufność oraz użyteczność zbiorów danych jednostkowych do udostępnienia,

- Ryzyko\użyteczność (*Risk\Utility*) – karta pozwala na przeprowadzenie pogłębionej analizy (w postaci raportów, tablic oraz wizualizacji) ryzyka ujawnienia informacji poufnych oraz straty informacji poniesionej wskutek zastosowania metod maskujących,
- Eksport danych (*Export Data*) – z poziomu tej karty możliwe jest wygenerowanie raportu z przebiegu procesu SDC oraz zapisanie mikrodanych, w których zapewniona będzie ochrona tajemnicy statystycznej,
- Reproduktywność (*Reproducibility*) – w tym miejscu możliwe jest wygenerowanie skryptu z kodem zapisanym w składni języka R, który umożliwia powtórne wywołanie raz przeprowadzonego z użyciem aplikacji `sdcApp()` procesu SDC w sposób w pełni automatyczny; karta ta umożliwia też zapisanie lub wczytanie zdefiniowanego wcześniej zadania SDC,
- Cofnij (*Undo*) – opcja pozwala na cofnięcie ostatniej czynności wykonanej w procesie SDC; można jej użyć kilkukrotnie, tak aby cofnąć efekty więcej niż jednej czynności (od końca).

Opcje dostępne w każdej karcie mogą się różnić, m.in. w zależności od przeprowadzonych czynności czy charakterystyki wejściowego zbioru danych jednostkowych. Każda karta składa się z dwóch paneli: nawigacyjnego po lewej stronie oraz głównego w centralnej części karty. Ponadto niektóre karty mają po prawej stronie dodatkowy panel zawierający informacje podsumowujące bieżące zadanie kontroli ujawniania mikrodanych.

Wśród licznych zalet aplikacji `sdcApp()` należy wymienić:

- dostępność przejrzystego i czytelnego interfejsu,
- możliwość uruchomienia GUI aplikacji w różnych przeglądarkach internetowych,
- intuicyjność w posługiwaniu się nią,
- możliwość importowania oraz eksportowania plików o formatach obsługiwanych przez większość popularnych pakietów statystycznych (tj. R, SAS, SPSS czy STATA),
- możliwość przygotowania wejściowych mikrodanych zgodnie z wymaganiami pakietu `sdcMicro`,
- dostępność wszystkich najpopularniejszych metod ochrony tajemnicy statystycznej, takich jak: anonimizacja, lokalne ukrywanie danych, PRAM, przekodowanie (globalne bądź górne/dolne), dodawanie szumu, mikroagregacja oraz wymiana rang,
- posiadanie dla każdej metody maskowania własnego kreatora uruchamianego w nowym okienku; parametryzacja danego podejścia polega na ustawieniu wartości poszczególnych parametrów poprzez ich wybór z listy bądź przesuwanie suwakiem (z określonymi wartościami granicznymi, możliwymi do przyjęcia przez każdy z nich; domyślna wartość jest ustawiona na początku),

- niezwłoczne uzyskanie informacji o skutkach wywołanej metody (w tym przede wszystkim o jej wpływie na ryzyko ujawnienia informacji poufnych oraz jej wkładzie w ponoszoną stratę informacji),
- opatrzenie każdej miary opisem ułatwiającym interpretację jej wartości,
- możliwość cofnięcia ostatnich czynności wykonanych w procesie SDC,
- możność przeprowadzenia oceny zapewnianej poufności oraz użyteczności w sposób kompleksowy, jak również analizy porównawczej zasobów wejściowych z wyjściowymi (w postaci tabelarycznej lub wizualnej),
- wygenerowanie skryptu języka R, który pozwoli na automatyzację przeprowadzania raz zaplanowanego procesu kontroli ujawniania mikrodanych,
- możliwość zgłaszania autorom pakietu `sdcMicro` i aplikacji `sdcApp()` wszelkich napotkanych błędów oraz sugestii co do ewentualnych dalszych kierunków rozwoju tego narzędzia,
- liczne grono użytkowników pakietu i aplikacji, dostępność materiałów pomocniczych i dydaktycznych oraz publikacji naukowych (w wielu językach), fora dyskusyjne, zloty i spotkania użytkowników, szkolenia off-line i on-line,
- ciągły rozwój narzędzia oraz częste wydawanie uaktualnień.

Słabą stroną charakteryzowanego narzędzia jest dłuższy czas przetwarzania w porównaniu z wykonywaniem analogicznego zadania bezpośrednio z poziomu konsoli programu R. Niemniej jednak ten aspekt zależy od mocy obliczeniowych komputera, którym dysponuje analityk na potrzeby przeprowadzenia badania empirycznego.

5.5. Inne możliwości informatyczne

Inne programy statystyczne nie mają na ogół odrębnych procedur związanych z kontrolą ujawniania danych. Można jednak pewne czynności z tym związane wykonać z wykorzystaniem aktualnie istniejących możliwości, choć oczywiście jest to bardziej praco- i czasochłonne.

W środowisku SAS na przykład można spróbować przeprowadzić SDC, używając procedur PROC UNIVARIATE i PROC FREQ. Za ich pomocą da się przeanalizować brzegowe i łączne rozkłady rozpatrywanych zmiennych w celu ustalenia, gdzie ewentualnie konieczne byłoby ukrycie lub przekodowanie pewnych informacji (w tym łączenie odpowiednich kategorii). Można też nałożyć szum rozmywający za pomocą makroprocedury odchyleniowej (ang. *jitter macro*, zob. np. http://www.datavis.ca/sasmac/jitter.html). Zaproponował ją prof. Michael Friendly z York University w Ontario w Kanadzie. Za pomocą procedury PROC MI z kolei możliwe staje się utworzenie syntetycznych zbiorów danych ze stosowną instrukcją dla użytkownika, jak użyć procedury PROC MIANALYZE, aby umieścić zmodyfikowane i oryginalne dane ponownie razem.

Niewielkie rozmiarowo ukrywanie tablic w środowisku SAS można przeprowadzić na dwa sposoby:

- optymalną metodą opartą na wykorzystaniu procedury PROC OPTMODEL do wyznaczenia najmniej kosztownego schematu wtórnego ukrywania komórek niezbędnych dla ochrony danych wrażliwych (zob. np. *Optimal Cell Suppression in SAS - Final Macro* pod adresem https://web.archive.org/web/20200221223611/http://daynebatten.com/2015/05/optimal-cell-suppression-sas-macro/%20),
- metodą heurystyczną, która stosuje prosty algorytm umożliwiający znalezienie akceptowalnego (ale niekoniecznie optymalnego) rozwiązania; odpowiednia makroprocedura (zob. https://web.archive.org/web/20200221223611/http://daynebatten.com/2015/04/small-cell-suppression-heuristic/) działa dobrze na zbiorach posortowanych za pomocą procedury PROC SORT.

Autorem obu wskazanych wyżej makr jest Dayne Batten, badacz naukowy w firmie Republic Wireless (Raleigh, Karolina Północna, Stany Zjednoczone).

Poniżej podano kilka przykładów wykorzystania narzędzi informatycznych do wyznaczenia miar straty. Wykorzystując język wsadowy środowiska SAS, można bez większego trudu zaprogramować wyznaczanie podstawowej miary zakłócenia rozkładu, którą jest miara opisana wzorem (4.1). Przykład 5.4 to właśnie ilustruje.

Przykład 5.4. Rozważmy dane wykorzystywane w części 5.2 (rys. 5.7). Załóżmy, że SDC wykonano metodą PRAM, z parametrami wskazanymi na rysunku 5.11. Mamy zatem dwa zbiory danych: wyjściowy (mosoby) oraz po dokonaniu SDC (mosobyb). Algorytm obliczania miary straty oparty na formule (4.1) z odległościami cząstkowymi określonymi zależnościami (4.2) – dla zmiennych PŁEĆ (płeć), STANC (stan cywilny prawny) i STATUSP (status na rynku pracy) – oraz (4.3) – dla zmiennej WYKSZT (wykształcenie) jest następujący:

```
libname sdc 'D:\SAS_Files\sasusers\mlodaka\smo';

data sdc.mosoby;
set sdc.mosoby;
if 'PŁEĆ'n='M' then PLEC=1; else PLEC=0;
run;

data sdc.mosobyb;
set sdc.mosobyb;
if 'PŁEĆ'n='M' then PLEC=1; else PLEC=0;
run;

proc iml;
use sdc.mosoby;
read all var {STANC WYKSZT STATUSP PLEC} into dane;
use sdc.mosobyb;
read all var {STANC WYKSZT STATUSP PLEC} into daneb;
n=nrow(dane);
m=ncol(dane);
lambdaz=j(m,1,0);
do j=1 to m;
```

```
 do i=1 to n;
  if j=2 then do;
    lambdaz[j]=lambdaz[j]+(abs(dane[i,j]-daneb[i,j])/(n*13));
  end;
  else do;
   if dane[i,j]^=daneb[i,j] then lambdaz[j]=lambdaz[j]+(1/n);
  end;
 end;
end;
nwiersz={'STANC' 'WYKSZT' 'STATUSP' 'PLEC'};
print lambdaz[rowname=nwiersz format=commax6.4 label='Strata informacji dla
poszczególnych zmiennych'];
lambda=lambdaz[+]/m;
print lambda[format=commax6.4 label='Strata całościowa'];
quit;
```

Stratę informacji dla poszczególnych zmiennych stanowiły odpowiadające tym zmiennym komponenty sumy (4.1), tzn. strata informacji na zmiennej X_j wynosiła $\lambda_j = \sum_{i=1}^{n} d(x_{ij}, x_{ij}^*)/n, j = 1, 2, \ldots, m$. W wyniku zastosowania tego algorytmu otrzymano taki wydruk:

Strata informacji dla poszczególnych zmiennych	
STANC	0,4200
WYKSZT	0,1338
STATUSP	0,2700
PLEC	0,1400

Strata całościowa
0,2410

Wynika stąd, że największą stratę generuje zmienna dotycząca stanu cywilnego prawnego, najmniejszą zaś – wykształcenie. Ogólnie strata informacji jest relatywnie niska.

Kolejny przykład dotyczy oceny wpływu zastosowania kontroli ujawniania danych na siłę związku między zmiennymi na podstawie porównania testów niezależności i współczynników zbieżności dla tablic kontyngencji.

Przykład 5.5. Na podstawie danych z przykładu 5.4 konstruujemy tablice kontyngencji statusu na rynku pracy (STATUSP) względem płci (PLEC). W tym celu używamy poniższego kodu SAS:

```
proc freq data=sdc.mosoby;
   tables statusp*plec/ chisq relrisk nopercent norow nocol;
   title 'Status na rynku pracy według płci';
run;

proc freq data=sdc.mosobyb;
   tables statusp*plec/ chisq relrisk nopercent norow nocol;
   title 'Status na rynku pracy według płci po SDC';
run;
```

Otrzymujemy następujące wydruki:

Status na rynku pracy według płci
The FREQ Procedure

Table of STATUSP by PLEC			
STATUSP	**PLEC**		
Frequency	**0**	**1**	**Total**
1	29	30	59
2	17	13	30
3	5	6	11
Total	51	49	100

Statistics for Table of STATUSP by PLEC

Statistic	DF	Value	Prob
Chi-Square	2	0.6014	0.7403
Likelihood Ratio Chi-Square	2	0.6029	0.7397
Mantel-Haenszel Chi-Square	1	0.0194	0.8891
Phi Coefficient		0.0776	
Contingency Coefficient		0.0773	
Cramer's V		0.0776	

Sample Size = 100

Status na rynku pracy według płci po SDC
The FREQ Procedure

Table of STATUSP by PLEC			
STATUSP	**PLEC**		
Frequency	**0**	**1**	**Total**
1	30	21	51
2	14	14	28
3	9	12	21
Total	53	47	100

Statistics for Table of STATUSP by PLEC

Statistic	DF	Value	Prob
Chi-Square	2	1.6628	0.4354
Likelihood Ratio Chi-Square	2	1.6664	0.4347
Mantel-Haenszel Chi-Square	1	1.6408	0.2002
Phi Coefficient		0.1289	
Contingency Coefficient		0.1279	
Cramer's V		0.1289	

Sample Size = 100

Wydruki zawierają tablice kontyngencji obu rodzajów (Table of STATUSP by PLEC) oraz statystyki weryfikujące kwestię zależności rozpatrywanych zmiennych (Statistics for Table of STATUSP by PLEC). Widać zatem, że w obu wypadkach testy chi-kwadrat Pearsona (Chi-Square), ilorazu wiarygodności chi-kwadrat (Likelihood Ratio Chi-Square) oraz Mantela-Haenszela (Mantel-Haenszel Chi-Square) wskazują na brak podstaw do odrzucenia hipotezy zerowej o niezależności obu zmiennych. W wypadku danych ukształtowanych w wyniku zastosowania SDC wartości tych testów (Value) są wyraźnie wyższe (o ponad 1,06, a dla testu Mantela-Haenszela o ponad 1,6), co w konsekwencji daje niższe poziomy istotności ex post (Prob) – o 0,3049, 0,3050 oraz 0,6889, odpowiednio. Tym samym wyrazistość niezależności jest tutaj nieco mniejsza. Znajduje to odzwierciedlenie także w przypadku współczynników zbieżności: φ (Phi Coefficient), kontyngencji (Contingency Coefficient) oraz V Cramera (Cramer's V), które po zastosowaniu SDC zwiększają się, odpowiednio, o 0,0513, 0,0506 oraz 0,0513 i tym samym zaznaczają nieznacznie mniejszą skalę niezależności rozpatrywanych zmiennych, bez naruszania jednak istoty zjawiska.

6

Rozwiązania organizacyjne kontroli dostępu do danych

W ostatniej części opracowania poruszono ważne dla ochrony danych problemy organizacyjne związane z udostępnianiem zasobów statystycznych i jego kontrolą w celu uniemożliwienia ujawnienia informacji wrażliwych. Ukazano także zagadnienie ostatecznej kontroli danych wynikowych zamieszczanych w publikacjach. Omówiono tutaj zatem argumenty przemawiające za potrzebą udostępniania wynikowych danych statystycznych, w tym odpowiednio przygotowanych danych jednostkowych, specyfikę tegoż udostępniania w wypadku danych przeznaczonych do celów naukowych oraz rodzaje i formy udostępnianych danych. Zaznaczono wyraźnie formalne wymogi, które muszą spełnić osoby czy instytucje ubiegające się o dostęp do mikrodanych, oraz warunki ochrony, które instytucja udostępniająca dane (najczęściej jest to urząd statystyczny, jednak może to być w praktyce każdy gestor) powinna zapewnić, aby w maksymalny możliwy sposób zabezpieczyć wrażliwe informacje przed ich przejęciem przez osoby do tego nieuprawnione.

6.1. Specyfika podejścia do udostępniania danych osobom ze środowiska naukowego

Podstawową i kluczową rolą krajowych urzędów statystycznych jest tworzenie oficjalnych statystyk. Starają się one również sprostać oczekiwaniom użytkowników informacji statystycznych i udostępniać im coraz bardziej szczegółowe dane wynikowe, na które ci zgłaszają popyt. To obserwowane rosnące zapotrzebowanie na mikrodane jest zgłaszane przez osoby ze środowiska naukowego, jak również przez podmioty z sektora prywatnego i publicznego, w tym ze środowiska rządowego. Bywa również tak, że o dostęp do mikrodanych ubiegają się studenci, uczniowie czy inne zainteresowane osoby. Zarówno wśród krajowych urzędów statystycznych, jak i po stronie odbiorców istnieje zgodność co do tego, że publikowanie wyników badań statystycznych w formie zagregowanej nie jest wystarczające. Jednym z tego powodów jest niemożność wykorzystania w pełni przez

gestorów danych potencjału źródeł informacji w postaci mikrodanych. Co więcej, w wyniku na przykład znaczniejszego wykorzystania zasobów administracyjnych czy danych z innych źródeł – w tym ogólnodostępnych (np. big data) – rośnie zarówno liczba źródeł mikrodanych, jak i ich wielkość. Zachodzące zmiany technologiczne są nieodłącznie związane z tworzeniem rosnącej liczby zbiorów mikrodanych pochodzących z różnych źródeł, zapewniają jednak coraz to nowe metody do ich przetwarzania, analizy i wizualizacji. Ograniczone możliwości (w tym kadrowe) instytucji statystycznych nie pozwalają na pełne ich wykorzystanie w badaniach i analizach. Utworzenie określonych zasobów na podstawie badań ankietowych jest kosztowne zarówno z perspektywy krajowych urzędów statystycznych czy innych gestorów, jak również z punktu widzenia obciążenia pojedynczego respondenta (lub innej jednostki dostarczającej danych). Dodatkowo zakres prac naukowo-badawczych prowadzonych przez użytkowników informacji statystycznych jest uwarunkowany postacią udostępnionych im zasobów, gdyż dobór metod statystycznych bądź ekonometrycznych, a także innych metod do przetwarzania, analizy, wizualizacji i tabelaryzacji, może wymagać na przykład dysponowania zbiorami danych jednostkowych.

Było to istotnym przyczynkiem do rozpoczęcia udostępniania bardziej szczegółowych danych wynikowych w mniej zagregowanej lub w niezagregowanej postaci, w szczególności zaś mikrodanych, również użytkownikom zewnętrznym. Osoby te niekoniecznie muszą pozostawać z gestorem danych w stosunku pracy, choć i takie rozwiązania są na arenie międzynarodowej praktykowane.

O ile publikowanie tablic statystycznych czy innej postaci wyników analiz ma wieloletnią tradycję, o tyle udostępnianie mikrodanych przez krajowe urzędy statystyczne lub innych gestorów nadal jest dość nowym zjawiskiem. Duncan i in. (2011) podają, że pierwsze zbiory mikrodanych były udostępnione do publicznego użycia w latach 60. XX wieku. Pochodziły one w pierwszej kolejności z amerykańskiego spisu powszechnego. Natomiast pierwsze systemowe wprowadzenie mikrodanych do użytku publicznego (które zgromadzono podczas spisu powszechnego przeprowadzonego w 1971 r.) było dziełem Statistics Canada. Umotywowano je po pierwsze udzieleniem szerokiego dostępu do społecznej i ekonomicznej bazy danych wiedzy o społeczności Kanady, a po drugie – dostarczeniem podstawy do analiz naukowych, której nie dają dane zagregowane. Przez te lata udało się jednak wypracować pewne podejścia w zakresie udostępniania bardziej szczegółowych danych wynikowych – w tym w postaci zbiorów danych jednostkowych – które z sukcesem są stosowane przez wybrane krajowe urzędy statystyczne. Przede wszystkim te, które na arenie międzynarodowej wiodą prym w rozwoju kontroli ujawniania danych statystycznych. Udostępniane mikrodane mogą być zasobami dostępnymi dla każdego, jednak z ważnych powodów grono ich odbiorców może również być ograniczone. Hundepool i in. (2012) zwrócili uwagę, że pomimo wypracowania w ramach kontroli ujawniania danych sta-

tystycznych wyrafinowanych metod ochrony poufności, które z powodzeniem mogą zostać użyte do przygotowania ogólnodostępnych mikrodanych, często nie jest możliwe zaspokojenie potrzeb badaczy i naukowców poprzez swobodne udostępnienie im zasobów ukształtowanych po zastosowaniu tychże metod. Zobowiązanie do zapewnienia ochrony poufności nie zawsze pozwala bowiem na publiczne udostępnienie zbiorów danych jednostkowych odznaczających się wystarczającym stopniem szczegółowości (są to sytuacje, gdy z powodu nadmiernej liczby zmiennych ryzyko ujawnienia pozostaje na znacznym poziomie). Z tych powodów zdecydowano się zastosować inną, uzupełniającą filozofię podejścia w dostępie do mikrodanych, tzn. udostępnianie restrykcyjne – tylko dla wybranych grup użytkowników i według ściśle przestrzeganych kryteriów. W zamian za to tacy użytkownicy zyskują dostęp do mikrodanych bardziej szczegółowych. W związku z tym należy podkreślić, że przywilej dostępu do tych bardziej szczegółowych informacji statystycznych najczęściej jest ograniczony i dotyczy jedynie osób reprezentujących środowisko naukowe oraz mających afiliację stosownej jednostki naukowej. Czyni się tak ze względu na niewątpliwe walory poznawcze rezultatów ich prac naukowo-badawczych.

Zapewnienie dostępu do zbiorów danych jednostkowych użytkownikom zewnętrznym w celach naukowo-badawczych niesie liczne korzyści, wśród których należy wyróżnić przede wszystkim następujące:

- przynajmniej częściowe zaspokojenie potrzeb odbiorcy informacji statystycznych, co może wyeliminować konieczność przeprowadzenia dodatkowych badań,
- umożliwienie osobom ze środowiska naukowego przeprowadzenia kompleksowych analiz statystycznych i ekonometrycznych, które wnoszą istotny wkład w publiczną debatę, naukę i podejmowanie decyzji,
- pełniejsze wykorzystanie posiadanych zasobów i bardziej kompleksowe poznanie badanego zjawiska,
- umożliwienie budowania nowych kierunków analiz, m.in. opartych na integracji danych (o ile dopuszcza się możliwość jej przeprowadzenia),
- budowanie zaufania jako potwierdzenie rzetelności publikowanych danych zagregowanych,
- podkreślenie istotnej funkcji pełnionej przez statystykę publiczną lub innego odpowiedniego gestora,
- zwiększenie współpracy podmiotów publicznych, prywatnych oraz środowiska naukowego,
- zapewnienie krytycznego spojrzenia na metodyczne aspekty badań statystycznych,
- dążenie do poprawy jakości prowadzonych badań statystycznych,
- wspieranie międzynarodowych porównań i tworzenie wzorców, pomoc w projektowaniu badań ankietowych,

– zapewnienie harmonizacji pomiędzy krajowymi urzędami statystycznymi, dbałości o spójność i porównywalność, zacieśnienie współpracy (Thomas i in., 2011).

Pomimo restrykcyjnego udzielania dostępu do danych jednostkowych – jedynie wąskiemu gronu odbiorców i tylko dla celów naukowo-badawczych – wyzwaniem pozostaje jednak opracowanie zasad i form udostępniania, które dalej będą w wystarczającym stopniu chronić poufność respondentów. Pełniejszy dostęp do mikrodanych stawia bowiem zupełnie inne wyzwania przed ochroną tajemnicy statystycznej niż dostęp ograniczony. Problemem jest nieprzewidywalność wyników analiz opracowywanych przez użytkowników tych zasobów. Celem udostępnienia mikrodanych jest umożliwienie owym użytkownikom prowadzenia analiz i prac naukowo-badawczych, które nie mają z góry określonego schematu. Odbiorcy ci realizują swoje prace w warunkach swobody badawczej. Ogólnie rzecz ujmując, środowisko, w którym będą się odbywać prace prowadzone przez uprzywilejowanych użytkowników zewnętrznych z wykorzystaniem mikrodanych, można zdefiniować jako środowisko naukowe, gdzie profesjonaliści mają dość szeroką swobodę działania, a wyniki ich studiów są trudne do przewidzenia. Jak twierdzi Ritchie (2007), środowisko to wymusza zmianę reguł ochrony poufności względem tych tradycyjnych, wypracowanych w wieloletniej praktyce udostępniania danych wynikowych w zagregowanej postaci.

W zbiorze zasad i praktyk opracowanych przez statystyków europejskich w ramach działań ONZ (UNECE, 2007) podkreśla się, że ochrona poufności mikrodanych oraz dostępu do nich stała się sprawą ponadnarodową. Rosnąca współpraca międzynarodowej społeczności akademickiej oraz dokonywanie międzynarodowych porównań wyników analiz i badań prowadzi do poszukiwania dostępu do mikrodanych przez naukowców w różnych krajach na podobnych zasadach. Istniejące w tym względzie różnice mogą utrudnić dokonywanie porównań i wspólne projekty. Różnice te wynikają z odmiennych regulacji prawnych, tradycji społecznej współpracy oraz możliwości wsparcia środowiska akademickiego funkcjonujących w poszczególnych krajach. Stąd też dążenie do wypracowania jednolitych zasad i reguł postępowania w dostępie do mikrodanych. Należy dodać, że jeżeli pliki z danymi jednostkowymi mają zostać udostępnione na arenie międzynarodowej, to będzie się to wiązać z koniecznością rozważenia dodatkowych kwestii prawnych i organizacyjnych. Może to także zwiększać obawy opinii publicznej w określonym kraju co do poufności danych zebranych przez krajowe urzędy statystyczne.

Duncan i in. (2011) określili cztery kryteria potrzebne do wypracowania zasad dostępu do mikrodanych:

Wskazanie, kto może mieć dostęp do mikrodanych

Wymaga to formalnego określenia zaufanych użytkowników, podmiotów i instytucji. Częścią społeczności badaczy będą pracownicy naukowi wyższych

uczelni, ale do tej grupy można zaliczyć także analityków i badaczy z organizacji pozarządowych oraz instytutów międzynarodowych i krajowych. Szczegółowe wymogi udostępniania mikrodanych stosowane przez krajowe urzędy statystyczne różnią się między sobą – zarówno w odniesieniu do grup badaczy, jak i do osób indywidualnych. Statistics Canada uzależnia na przykład uzyskanie dostępu od wkładu badacza w rozwój i popularyzację dyscyplin wspieranych przez Komitet Nauk Społecznych i Humanistycznych (ang. Social Sciences and Humanities Research Council – SSHRC). Innym wymogiem w tym względzie może być uzasadnienie, że badacz nie jest w stanie wykonać analizy na podstawie ogólnodostępnych zbiorów mikrodanych, a jeszcze innym – że badanie przyniesie również korzyść z punktu widzenia programu badań krajowego urzędu statystycznego.

Określenie miejsca, gdzie dostęp może być uzyskany

Wymóg zapewnienia ochrony poufności sprawia, że gestor danych chce mieć kontrolę nad działaniami użytkownika w tym zakresie. W związku z tym najdogodniejszą opcją lokalizacji udostępniania jest miejsce, nad którym posiada władztwo: czy to w postaci własności, czy też praw zarządzania. Powoduje to kłopoty z punktu widzenia użytkownika związane z kosztami podróży, godzinami otwarcia itp. US Census Bureau w 1982 r. ustanowiło Centrum Badań Ekonomicznych (ang. Center for Economic Research – CES), w którym udostępniano mikrodane ekonomiczne z zakresu przetwórstwa przemysłowego. W 1994 r. biuro to powołało regionalne Centrum Danych Naukowych (ang. Research Data Center – RDC). Z pomocą Narodowej Fundacji Nauki w ciągu następnych lat uruchomiono liczne punkty RDC na uniwersytetach w całych Stanach Zjednoczonych, jak również w Europie. Rozwój technologiczny powoduje poszukiwanie możliwości udostępniania zdalnego, co wyeliminowałoby wiele niedogodności związanych z wymogiem miejsca. Takie rozwiązania są już stosowane w Danii, Szwecji, Holandii i w Słowenii.

Określenie, jakiego typu analizy są dozwolone.

W niektórych krajach (takich jak Stany Zjednoczone) regulacje prawne warunkują dostęp korzyścią, jaką może odnieść z tego instytucja statystyczna. Reguła ta nie jest jednak zbyt szeroko rozpowszechniona. Użytkownicy są zobowiązani do zapewnienia ochrony poufności danych zawartych w wynikach przeprowadzonych przez siebie analiz. W Kanadzie oczekuje się, że udostępnianie mikrodanych przyniesie społeczne korzyści, takie jak:

- zbudowanie lepszej perspektywy dla poznania społeczności Kanady,
- utworzenie bazy do badań społecznych w większych i mniejszych ośrodkach na obszarze kraju,
- rozszerzenie współpracy między Statystyką Kanadyjską a społecznością akademicką,
- stworzenie możliwości szkolenia dla przyszłych badaczy społecznych.

Określenie, w jaki sposób można uzyskać dostęp do mikrodanych

Przy określaniu tych kryteriów trzeba wziąć pod uwagę wszystkie wymienione wyżej warunki.

Trzeba pamiętać, że fundamentalną zasadą prowadzenia badań statystycznych przez krajowe urzędy statystyczne jest to, że wszelkie dane zbierane od osób fizycznych lub prawnych podlegają ścisłej ochronie pod rygorem zachowania poufności. W związku z tym wszelkie kryteria, pod którymi mikrodane zostaną udostępnione, muszą być zgodne z tą zasadą. Kryteria te wyrażono w ramach współpracy statystyków pod egidą ONZ (UNECE, 2007) w następującej postaci:

- *Jest dopuszczalne, aby zbiory danych jednostkowych zebrane przez statystykę publiczną były wykorzystywane do analiz statystycznych i rozwoju badań naukowych, o ile poufność jednostek statystycznych zostanie zachowana.*

Udostępnienie takich zasobów użytkownikom spoza statystyki publicznej nie narusza więc fundamentalnej zasady prowadzenia badań przez krajowe urzędy statystyczne. Decyzję w tym zakresie podejmują jednak instytucje statystyczne w poszczególnych krajach, a względy, którymi się kierują przy podjęciu decyzji o udostępnieniu mikrodanych, mogą być różne. Za odmową udostępnienia – całości lub części – zbioru danych jednostkowych mogą przemawiać też inne względy, jak np. jakość danych.

- *Mikrodane mogą być wykorzystane tylko do celów naukowo-badawczych.*

Wyklucza to między innymi ich udostępnienie do celów administracyjnych (związanych na przykład z postępowaniem administracyjnym lub sądowym).

- *Udostępnienie powinno mieć podstawy prawne oraz wypracowane reguły zachowania poufności.*

Praca badaczy na mikrodanych wiąże się z opracowywaniem licznych statystyk wynikowych, modeli statystycznych, porównań czy danych wynikowych w zasadzie w dowolnej innej postaci. Wymaga to potwierdzenia, że wyniki tych prac nie naruszają poufności. Stosowne regulacje prawne powinny być uzupełnione poprzez procedury administracyjne i techniczne, które są czytelne i sprzyjają postrzeganiu konieczności używania mikrodanych zgodnie z zasadami ochrony poufności.

- *Zasady udostępniania mikrodanych oraz ich użycia są dostępne publicznie.*

Celem tego jest upowszechnienie w świadomości publicznej, że udostępnianie mikrodanych służy celom społecznym i jest uzasadnione z punktu widzenia interesu publicznego. Ważne staje się tu przekonanie, że kryteria udostępniania są obiektywne i transparentne.

Na koniec warto jeszcze przytoczyć pięć problematycznych zagadnień wymienionych przez Ritchiego (2007) jako te, które trzeba uwzględnić w dyskusji nad skutecznym działaniem w ochronie poufności mikrodanych:

- rosnąca intensywność współpracy naukowej społeczności międzynarodowej oraz brak wypracowanych wspólnych standardów poufności dla udostępniania danych pomiędzy krajami,
- wzrastająca liczba badań podkreślająca znaczenie naukowych analiz jako takich; próba przeniesienia metod ochrony poufności dla danych zagregowanych lub anonimizacji dla wyników tych analiz może być nieefektywna lub szkodliwa,
- wykraczanie przez zakres analiz naukowych poza tradycyjne modele stosowane w dotychczasowych metodach ochrony poufności,
- rosnąca liczba wystąpień o dostęp do mikrodanych, również w lokalizacjach poza siedzibą krajowego urzędu statystycznego, wymaga wypracowania transparentnych metod w zakresie zarządzania dostępem,
- problem weryfikacji metod ochrony mikrodanych: w przeciwieństwie do metod ochrony poufności dla danych zagregowanych, które są poddawane gruntownemu testowaniu, doskonaleniu oraz wymianie doświadczeń, metody opracowane dla mikrodanych są zwykle wypracowywane wewnętrznie i nie są poddawane niezależnym ocenom zewnętrznym. Brakuje też wspólnych „dobrych praktyk".

Chociaż od ukazania się powyższej publikacji minęło już ponad piętnaście lat, to pomimo prężnego rozwoju dziedziny kontroli ujawniania danych statystycznych nadal nie wszystkie problemy zasygnalizowane przez Ritchiego (2007) udało się w pełni rozwiązać.

6.2. Typy udostępnianych mikrodanych

Wśród zasobów, do których dostęp jest udzielany użytkownikom zewnętrznym przez krajowe urzędy statystyczne, a które charakteryzują się większą szczegółowością oraz większym zakresem dostępnych informacji niż dostępne w publikacyjnych lub wynikowych tablicach statystycznych czy w różnej postaci wynikach analiz, wyróżnić należy przede wszystkim zbiory danych jednostkowych, a także kostki danych. Pierwsze są udostępniane w niezagregowanej postaci lub co najwyżej po zagregowaniu jednostek statystycznych podrzędnych w jednostki statystyczne nadrzędne (oczywiście dopuszcza się tu również możliwość agregowania, czyli zmniejszania szczegółowości poszczególnych zmiennych w zbiorze). Wymaga to jednak występowania w wejściowych mikrodanych hierarchicznego układu jednostek statystycznych. Na przykład hierarchię taką w danych spisowych mogłyby tworzyć następujące jednostki statystyczne: osoby, rodziny, gospodarstwa domowe, mieszkania oraz budynki. W pewnych okolicznościach możliwa jest sytuacja, że wskazane byłoby udostępnienie zbioru danych jednostkowych na poziomie gospodarstw domowych, a nie osób. Mielibyśmy w takim wypadku

do czynienia z agregacją jednostek statystycznych. Drugie wymienione wcześniej struktury – tzn. kostki danych – charakteryzują się bardziej zagregowaną postacią niż mikrodane, lecz nadal odznaczają się większym stopniem szczegółowości niż tablice statystyczne. Dlatego na samym końcu tego podrozdziału wspomniano również o nich.

Ze względu na kilka aspektów, którymi są: cel wykorzystania, grupa użytkowników, a także stopień i formy ochrony informacji poufnych, mikrodane można sklasyfikować jako (UNECE, 2007; Eurostat i FRIBS Task Force, 2017):

- mikrodane dla celów statystycznych,
- mikrodane dla celów naukowych:
  - mikrodane do użytku chronionego,
  - mikrodane do zastosowań naukowych,
- mikrodane do użytku publicznego.

Klasyfikację tę zobrazowano również na rysunku 6.1.

Rys. 6.1. Klasyfikacja typów udostępnianych mikrodanych według celu ich wykorzystania, grupy ich użytkowników oraz stopnia i formy ochrony informacji poufnych

Mikrodane dla celów statystycznych (ang. *microdata for statistical purposes*) to poufne zasoby gromadzone w toku prowadzenia badań statystycznych lub pozyskiwane z innych źródeł (np. ze źródeł administracyjnych czy big data), którym bezwzględnie powinna być zapewniona ochrona tajemnicy statystycznej. Zasoby te są dostępne jedynie wybranym pracownikom statystyki publicznej i służą przede wszystkim do tworzenia i obliczania oficjalnych danych statystycznych oraz opracowywania informacji wynikowych prezentowanych i rozpowszechnianych w różnej postaci. Przekazuje się je również wewnątrz Europejskiego Systemu Statystycznego w ramach chronionej wymiany danych. Nigdy natomiast nie są one przedmiotem udostępnienia użytkownikom zewnętrznym lub nawet użytkownikom wewnętrznym, jeżeli nie wymagają tego ich zadania służbowe.

W wypadku tych zasobów ryzyko ujawnienia informacji poufnych jest bardzo wysokie. Wynika to najczęściej z dostępności identyfikatorów wśród zmiennych

w takich zbiorach danych jednostkowych. Jeżeli nawet przeprowadzono anonimizację lub pseudonimizację owych baz, to tutaj ochrona poufności na tym zwykle się kończy. Nadal więc może dojść do identyfikacji jednostek statystycznych czy respondentów na podstawie kombinacji wartości zmiennych quasi-identyfikatorów, gdyż nie są one zabezpieczane żadną inną metodą kontroli ujawniania mikrodanych. W wypadku zbiorów tego typu najczęściej nie podejmuje się nawet próby oceny ryzyka ujawnienia informacji poufnych, a to ze względu na wyłączny dostęp do nich dla jedynie wyselekcjonowanego grona pracowników statystyki publicznej.

Mikrodane dla celów naukowych (ang. *microdata for scientific purposes*) to takie zasoby, na których został przeprowadzony proces kontroli ujawniania mikrodanych i w związku z tym mogą one zostać udostępnione wybranym użytkownikom zewnętrznym do celów naukowo-badawczych. Dzieli się je na **mikrodane do użytku chronionego** (ang. *secure use files*) i na **mikrodane do zastosowań naukowych** (ang. *scientific use files*). Cechą wspólną obu tych typów poufnych zbiorów danych jednostkowych jest to, że nie są one powszechnie dostępne, a wręcz przeciwnie – udostępnia się je jedynie wąskiemu gronu użytkowników do zrealizowania założonych celów naukowo-badawczych. Są to licencjonowane zbiory mikrodanych, do których dostęp udzielany jest dopiero po wcześniejszym zawarciu z ubiegającym się o niego odpowiedniej umowy lub porozumienia. Warunki licencjonowania w tym zakresie różnią się w zależności od kraju i zależą m.in. od krajowych oraz międzynarodowych uwarunkowań prawnych czy od regulacji wewnętrznych i przyjętych przez gestora danych zasad etycznych. Ponadto zapisy umowy lub porozumienia licencyjnego zależą również od typu udostępnianych mikrodanych oraz formy udzielanego dostępu. Na ogół jednak obejmują one takie elementy jak:

- określenie warunków przechowywania udostępnianych zasobów i korzystania z nich,
- ustalenie, że mikrodane będą użyte tylko do celów naukowo-badawczych,
- wskazanie listy osób, które powinny mieć wyłączny dostęp do zbiorów danych jednostkowych,
- zapis, że próby identyfikacji jednostek statystycznych są niedozwolone,
- wymóg, że wszelkie kopie zbiorów danych jednostkowych zostaną zwrócone lub zniszczone po zakończeniu projektu,
- zakaz podejmowania prób łączenia (innymi słowy – użycia metod statystycznej integracji danych) udostępnionych mikrodanych z danymi pozyskanymi z innych źródeł lub nawet od tego samego gestora danych,
- to, że gestor danych może również wymagać udzielenia mu zgody na wykorzystanie wszelkich wyników analiz, które opracowano na podstawie udostępnionych przez niego zasobów.

Zasoby pierwszego wymienionego typu, tj. mikrodane do użytku chronionego, są dostępne dla osób ze środowiska naukowego w siedzibie gestora danych,

w innej wyznaczonej przez niego lokalizacji lub w trybie zdalnego dostępu. Na podstawie poufnych danych jednostkowych użytkownik zewnętrzny opracowuje tablice statystyczne oraz różnej postaci wyniki analiz, a sprawdzeniem, czy można je bezpiecznie udostępnić poza to środowisko chronione i czy nie zostaną tym samym ujawnione informacje poufne, zajmują się oddelegowani przez gestora danych pracownicy odpowiedniej jednostki.

Mikrodanym do użytku chronionego towarzyszy wysokie ryzyko ujawnienia informacji poufnych. Przed ich udostępnieniem w ściśle kontrolowanym środowisku przeprowadza się co prawda ich anonimizację lub pseudonimizację – co wyklucza możliwość bezpośredniej identyfikacji jednostek statystycznych – lecz zestaw zmiennych pośrednio identyfikujących respondenta nadal może pozwolić na jego zidentyfikowanie, bowiem inne metody kontroli ujawniania mikrodanych zwykle nie są tutaj stosowane.

Mikrodane do zastosowań naukowych są udostępniane osobom prowadzącym prace naukowo-badawcze np. na płycie DVD, na innym nośniku danych, czy też w postaci hiperłącza umożliwiającego pobranie zasobów z serwera. Często mają one postać dopasowanej do potrzeb konkretnego zamówienia bazy danych jednostkowych[27]. Kontrolą ujawniania wyników analiz pod kątem poufności zajmują się wyłącznie osoby prowadzące prace naukowo-badawcze. Może się ona odbywać z uwzględnieniem przekazanych wraz z zasobami wytycznych, instrukcji lub podręcznika. Osoby reprezentujące gestora danych nie uczestniczą w tej czynności, ale instytucja udostępniająca zasoby może jednak wymagać od użytkownika np. przesłania wszelkich wyników analiz przed ich opublikowaniem.

Mikrodanym do zastosowań naukowych towarzyszy niskie, zredukowane ryzyko ujawnienia informacji poufnych. Na zbiorach takich – oprócz przeprowadzenia anonimizacji lub pseudonomizacji, które wykluczają możliwość bezpośredniej identyfikacji respondenta – stosowane są również inne metody kontroli ujawniania mikrodanych (z optymalnie ustalonymi wartościami odpowiednich parametrów), które redukują możność pośredniej identyfikacji tychże respon-

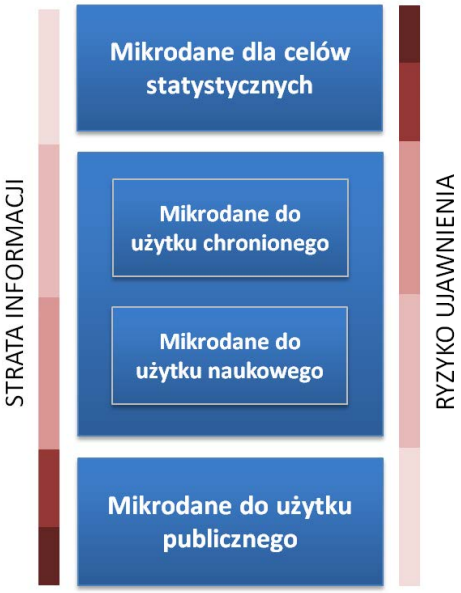
[27] Udostępnianie użytkownikom zewnętrznym różnych kopii zbioru danych jednostkowych z określonej edycji badania statystycznego, przygotowanych specjalnie wedle ich zamówienia, jest dyskusyjną i wątpliwą etycznie praktyką. Należy bowiem pamiętać, że porównanie kilku różnych wersji mikrodanych może skutkować wykryciem niespójności w danych (które mogą być skutkiem użycia metod ochrony tajemnicy statystycznej, lecz które niewątpliwie wpłynęłyby na wizerunek i zaufanie do gestora danych), ale przede wszystkim wzrostem ryzyka ujawnienia informacji poufnych i w konsekwencji doprowadzić do identyfikacji respondentów w udostępnionych zasobach. Część metod stosowanych w procesie kontroli ujawniania mikrodanych – głównie tych, które są oparte na rachunku prawdopodobieństwa – zwraca bowiem inne rezultaty przy kolejnych ich wywołaniach – nawet jeżeli wartości ich parametrów ustawiono w ten sam sposób. Ze względu na to wspomniane porównanie może prowadzić do ujawnienia metody wykorzystanej w procesie oraz jej parametryzacji, a to – dla doświadczonego użytkownika – wystarczy, by móc cofnąć proces SDC i odtworzyć oryginalne wartości zmiennych dla wszystkich (lub przynajmniej dla części) obserwacji.

dentów. W szczególności używa się tutaj wybranych metod maskujących (niezakłóceniowych bądź zakłóceniowych) na quasi-identyfikatorach – w celu redukcji ryzyka ujawnienia tożsamości respondenta, a także na zmiennych wrażliwych – w celu zminimalizowania ryzyka ujawnienia jego atrybutu. Identyfikacja nadal może zostać przeprowadzona na podstawie kombinacji wartości zmiennych kluczowych, czyli w sposób pośredni, ale tylko dla jednostek o rzadkich wartościach tych charakterystyk. Zakres zapewnianej ochrony poufności jest więc tutaj większy niż w wypadku mikrodanych do użytku chronionego, lecz nadal są to zasoby o charakterze poufnym, do których dostęp mogą mieć wyłącznie osoby upoważnione.

Mikrodane do użytku publicznego (ang. *public use files*) to odpersonalizowane zasoby powszechnie dostępne dla każdego zainteresowanego nimi użytkownika na przykład na stronie internetowej gestora danych. Ich pobranie może co najwyżej wymagać m.in. podania adresu e-mail w formularzu, jednak nie wiąże się z tym konieczność zawarcia umowy czy wniesienia opłaty. Zasoby takie, ze względu na skalę zmian dokonanych metodami kontroli ujawniania mikrodanych, zazwyczaj nie są podstawą do formułowania wniosków i wyłuskiwania prawidłowości statystycznych o badanej populacji, a jedynie mogą posłużyć do celów szkoleniowych oraz do przygotowania skryptów programów przed uzyskaniem dostępu do mikrodanych dla celów naukowych i rozpoczęciem pracy na nich z wykorzystaniem stosownych narzędzi informatycznych.

Identyfikacja jednostki statystycznej nie jest w tych zbiorach możliwa, bowiem ryzyko ujawnienia informacji poufnych, oczywiście przy ustalonych założeniach co do jego oceny, jest eliminowane w procesie kontroli ujawniania mikrodanych na etapie przygotowania tych zasobów do udostępnienia. Osiąga się to poprzez przeprowadzenie ich anonimizacji bądź pseudonimizacji przed przystąpieniem do kontroli ujawniania mikrodanych lub na samym początku tego procesu, tzn. przed zastosowaniem innych metod ochrony poufności, jak również przez użycie innych metod z restrykcyjnie dobranymi wartościami przyjmowanych przez nie parametrów. Jest to konieczne, ponieważ do dyspozycji użytkownika zewnętrznego mogą pozostawać licznie występujące w przestrzeni publicznej, ale również dostępne mu z innych źródeł o charakterze prywatnym, zbiory danych jednostkowych zawierające identyfikatory oraz pseudo-identyfikatory. Z powodu licznych ograniczeń w użytkowaniu, preferencje w zakresie udostępniania mikrodanych zmieniają się z udostępniania powszechnego na korzyść udostępniania bardziej restrykcyjnego dla wybranych użytkowników lub grup użytkowników. W krajach, w których udostępnia się mikrodane, są one cenione przez środowisko naukowe.

Na rysunku 6.2 podsumowano typy udostępnianych mikrodanych ze względu na poziom zapewnianej ochrony tajemnicy statystycznej (innymi słowy – na redukcję ryzyka ujawnienia informacji poufnych) oraz na ich użyteczność (czyli na ponoszoną w procesie kontroli ujawniania mikrodanych stratę informacji).



Rys. 6.2. Typy udostępnianych zbiorów danych jednostkowych według poziomu ryzyka ujawnienia informacji poufnych oraz straty informacji

Uwaga: Odcienie koloru brązowego po lewej i prawej stronie oznaczają intensywność znaczenia straty lub ryzyka odpowiednio w danym przypadku (im kolor intensywniejszy, tym znaczenie większe).

Na koniec warto jeszcze wspomnieć o innej postaci danych wynikowych, które bywają udostępniane użytkownikom zewnętrznym. Chodzi tutaj o dane w pewnym stopniu zagregowane, a jednak odznaczające się większą szczegółowością niż publikacyjne lub wynikowe tablice statystyczne. Mowa o **kostkach danych**. Są to udostępniane wielowymiarowe struktury o bardzo dużym stopniu szczegółowości. Mogą one być przygotowane według wcześniej zdefiniowanych schematów lub wygenerowane na specjalne życzenie użytkownika. Stopień szczegółowości w kostkach może być tak znaczny, że metody ich ochrony są porównywalne z tymi, które stosuje się do mikrodanych. Pionierem w udostępnianiu danych wynikowych w takiej postaci było Centralne Biuro Statystyczne Holandii. Stopień szczegółowości kostek nie jest tak wysoki jak w wypadku mikrodanych, co może być postrzegane jako wada. Inną niedogodnością jest to, że przygotowanie takiej wielowymiarowej tablicy w innym niż w ustalony schemacie wymaga najczęściej uiszczenia dodatkowej opłaty. Należy również wspomnieć, że nie każde narzędzie informatyczne zapewnia możliwość pracy na danych o takiej strukturze. Jedną z zalet kostek jest natomiast to, że proces ochrony poufności pozostaje całkowicie pod kontrolą gestora danych, a inną – także dogodna forma udostępniania – za pośrednictwem Internetu.

6.3. Formy udostępniania

Najpowszechniejszymi miejscami udostępniania mikrodanych są **centra danych naukowych** – miejsca, w których pod kontrolą udostępniającego badacz może korzystać z dostępu do mikrodanych. Stanowiska komputerowe w RDC są wyposażone w odpowiednie oprogramowanie umożliwiające analizy z dostępem do zbiorów mikrodanych. Komputery zostają odpowiednio zabezpieczone przed np. nieuprawnionym kopiowaniem, a praca na nich odbywa się pod nadzorem pracowników urzędu lub innego gestora danych. Z pomieszczenia RDC nie można nic wynieść. Jest ono również odizolowane od „świata zewnętrznego" – brak w nim dostępu do Internetu czy możliwości podłączenia jakichkolwiek urządzeń zewnętrznych do stanowiska. Nie ma tu też możności skopiowania, zapisania na własnym nośniku czy wydrukowania otrzymanych wyników przed ich sprawdzeniem pod kątem ochrony tajemnicy statystycznej (chyba że użytkownikowi udostępniono dane po kontroli SDC lub dane syntetyczne). Naukowiec, przed udostępnieniem mu danych, musi spełnić liczne wymogi. Wśród wymogów, które warunkują ów dostęp, mogą być:

- uzasadnienie, że wyniki prowadzonej analizy służą dobru publicznemu,
- sprecyzowanie, w jaki sposób wyniki będą publicznie dostępne,
- przedstawienie świadectwa afiliacji naukowej,
- zgoda na nadzór ze strony instytucji udostępniającej.

Główne elementy krytyki laboratoriów ze strony społeczności naukowej to konieczność stawienia się w miejscu często odległym od miejsca zatrudnienia czy zamieszkania oraz praca z narzędziami, które mogą być traktowane przez naukowca jako nieprzyjazne lub niewystarczająco znane. Problemem w tym zakresie może też być koszt utrzymania. Ta forma udostępniania jest od wielu lat stosowana przez urzędy statystyczne. Niektóre z nich – jak np. Statystyka Kanadyjska – udostępniły RDC w wielu miejscach w kraju. We współpracy z wyższymi uczelniami Statystyka Kanadyjska utworzyła bowiem miejsca dostępu nie tylko we własnych obiektach, ale również na terenie uczelni. Udostępniane są tam mikrodane z badań społecznych – pod kontrolą urzędu, z zachowaniem kontroli ujawniania danych prowadzonej według określonych przez niego zasad. Program RDC jest otwarty dla badaczy spoza statystyki publicznej – tzn. badaczy, którzy nie są pracownikami Statystyki Kanadyjskiej. Do tej grupy należą akademicy, pracownicy rządu federalnego i rządów lokalnych oraz absolwenci wyższych uczelni. Jako wady owych centrów wymienia się wysokie koszty zarządzania i utrzymania. Niedogodnością jest też konieczność poddania każdej pracy badawczej sprawdzeniu pod względem ochrony poufności. Wymaga to czasu i prowadzi do opóźnień w publikacji.

Centralne Biuro Statystyczne Holandii utworzyło w 1998 r. Centrum Badań Mikrodanych Gospodarczych (ang. Centre for Research on Economic Micro-

data – CEREM) na terenie uczelni. Działanie to było odpowiedzią na zgłaszane przez środowiska naukowe potrzeby dostępu do mikrodanych z badań gospodarczych oraz mikrodanych fiskalnych dotyczących dochodów. Centrum było pójściem o krok dalej w stosunku do RDC, które działało w siedzibie CBS od 1994 r. W 2002 r. w CBS – jako specjalny departament – utworzono Centrum Statystyki dla celów Polityki Społeczno-Gospodarczej (ang. Centre for Policy Related Statistics – CPS). Chodziło mianowicie o umożliwienie realizacji analiz statystycznych w zakresie prowadzonych działań w dziedzinie polityki społecznej oraz o ocenę ich efektów. Od 2005 r. CPS nadzoruje całość zarządzania dostępem do mikrodanych. Dostęp do mikrodanych jest możliwy dla wyższych uczelni i instytutów naukowych, ale również Eurostatu oraz krajowych urzędów statystycznych Unii Europejskiej.

Centra zdalnego dostępu (ang. *remote access facilities* – RAF) z kolei oferują nowe możliwości technologiczne, niwelują niektóre niedogodności związane z udostępnianiem mikrodanych na miejscu oraz prowadzą do poszukiwania również innych sposobów wsparcia środowisk naukowych odnośnie do analiz z wykorzystaniem mikrodanych. Centra RAF oferują dwie formy pracy – zdalne przetwarzanie i zdalny dostęp.

W wypadku **zdalnego przetwarzania** badacz ma pełny dostęp do metadanych zbiorów mikrodanych. Na ich podstawie przygotowuje skrypt do wykonania. Skrypt ów zostaje przesłany do urzędu statystycznego lub innego udostępniającego dane gestora, gdzie znajdują się zbiory mikrodanych. Skrypt zostaje sprawdzony i wykonany w urzędzie. Wynik przetwarzania jest następnie przekazywany badaczowi. Ponieważ proces poprawy błędów może być żmudny i długotrwały, stosuje się także udostępnienie sztucznych zbiorów mikrodanych, które odtwarzają strukturę rzeczywistych. W ten sposób możliwe jest testowanie np. poprawności składni skryptów względem danego środowiska napisanych w języku SAS. Rozwiązanie to zawiera niedogodności zarówno dla naukowca, jak i dla instytucji statystycznej. Dla naukowca kłopotliwy jest tutaj bowiem możliwy długi cykl przetwarzania, dla instytucji statystycznej zaś lub innego gestora danych – czasochłonność kontroli skryptów pod kątem zarówno ochrony poufności, jak i ich przetwarzania (Hundepool i in., 2012).

Zdalny dostęp umożliwia dostęp do mikrodanych bez konieczności stawienia się w określonej lokalizacji. Rozwój technologiczny Internetu pozwala na ustanowienie bezpiecznego połączenia poprzez sieć komputerową. Użytkownik–badacz nawiązuje bezpieczne połączenie komputerowe z dowolnej lokalizacji z wykorzystaniem wirtualnej sieci prywatnej (ang. *virtual private network* – VPN). Połączenie jest szyfrowane, a do jego nawiązania potrzebne są informacje uwierzytelniające użytkownika. Udostępniane dane pozostają na serwerach urzędu statystycznego lub odpowiedniego innego gestora danych. Przewaga tej opcji w stosunku do zdalnego przetwarzania skryptów polega na tym, że użytkownik widzi na swoim

ekranie efekt swojego skryptu. Natomiast nie jest możliwe ani pobranie żadnych danych na swój komputer, ani ich drukowanie, ani nawet kopiowanie zawartości ekranu. Mikrodane mogą również być całkiem ukryte, a skrypty przesłane przez badacza – wykonywane przez system. Kontroli podlegają zarówno kody wejściowe, jak i dane wynikowe. Żeby zapobiec nieautoryzowanemu dostępowi do danych, przeprowadzana jest tu także (np. w sposób biometryczny) identyfikacja badacza (zob. Eurostat i FRIBS Task Force, 2017).

Jednym z narzędzi stosowanych w tym rozwiązaniu może być oprogramowanie Citrix. Jest ono również wykorzystywane w wewnętrznej sieci korporacyjnej polskiego Głównego Urzędu Statystycznego do prac analitycznych w środowisku SAS. Możliwość pobrania raportu przez badacza musi być poprzedzona złożeniem stosownej prośby do gestora danych, a następnie poddana kontroli ochrony poufności. Na tej podstawie jest podejmowana decyzja o przesłaniu raportu. Dopiero wtedy badacz może raport pobrać i zapisać lub wydrukować. Zarówno opracowanie Hundepoola i in. (2012), jak i podręcznik ONZ (UNECE, 2007) wskazują, że procedura taka wiąże się z odpowiedzialnością – i to nie tylko pracownika, który recenzuje pracę badacza i wydaje decyzję, ale także instytucji. Kilka krajów – m.in. Kanada, Australia i Dania – zdecydowało się wprowadzić stosowne rozwiązania prawne, które regulują te kwestie. Doświadczenia tamtejszych statystyków wyniesione z wdrażania odpowiednich przepisów są ogólnie pozytywne. Niestety, w tych rozwiązaniach nie można w całości wyeliminować ryzyka identyfikacji. Wskazane jest też tutaj zawarcie dodatkowego zobowiązania ze strony badacza (który powinien być świadomy warunków dostępu) oraz przeprowadzenie szkolenia dla użytkowników korzystających ze zdalnego dostępu.

Warto zauważyć, że RAF w praktyce nie musi oznaczać jednego punktu dostępowego. Bazy danych, którymi interesuje się użytkownik, mogą bowiem być rozproszone u różnych gestorów, w związku z czym możliwe jest udostępnienie owych informacji przez każdego z nich w różnym czasie i miejscu. Stąd w istocie jest to system dostępowy.

Jeszcze innym rozwiązaniem, które zapewni naukowcowi dostęp do mikrodanych, jest czasowa umowa o pracę – **tymczasowe zatrudnienie**. Badacz podlega wtedy tym samym procedurom co inni pracownicy statystyki publicznej lub innego gestora, między innymi składa ślubowanie o zachowaniu poufności albo oświadczenie o podobnej treści. Cel takiej umowy jest sensowny i uzasadniony, jeżeli pracownik naukowy wypełnia również obowiązki wynikające ze statutowej działalności danej instytucji. W innym wypadku rozwiązanie tego rodzaju może być oceniane krytycznie i postrzegane jako sztuczne, co bywa przyczyną obniżenia zaufania do owej udostępniającej dane instytucji. Na przykład podczas wdrażania nowego rozwiązania dla instytucji statystycznej uzasadnione i korzystne może się również okazać zatrudnienie osoby z przydatnymi, unikatowymi kwalifikacjami, które nadadzą temu rozwiązaniu dodatkowy, wartościowy walor.

W wypadku populacji przedsiębiorstw mamy do czynienia z asymetrycznością rozkładów, co rzutuje na **udostępnianie mikrodanych z badań podmiotów gospodarczych**. Chociaż asymetria występuje również w populacjach osób i gospodarstw domowych, to jednak częściej obserwuje się ją właśnie w populacji przedsiębiorstw. Dodatkowo największe firmy włączane są do prób z prawdopodobieństwem równym jeden[28]. Ponadto w niektórych krajach bazy danych przedsiębiorstw są częściej udostępniane publicznie aniżeli zbiory innego typu. W związku z tym możliwe byłoby użycie techniki łączenia rekordów dla identyfikacji jednostek czy odtworzenia danych wrażliwych. Dodatkowe ograniczenia w zakresie udostępniania mikrodanych pomiędzy krajami mogą narzucić też reguły konkurencji. Z powyższych powodów udostępnianie mikrodanych może być realne tylko dla najmniejszych przedsiębiorstw (UNECE, 2007). Najlepszym sposobem ich udostępniania mogą się okazać centra danych – z powodu możliwości przeprowadzenia w nich również kontroli łączenia źródeł w zakresie bezpieczeństwa danych. Centralne Biuro Statystyczne Holandii na przykład, jak wcześniej wspomniano, udostępnia mikrodane z badań przedsiębiorstw w ramach Centrum Statystyki dla celów Polityki Społeczno-Gospodarczej.

[28] W praktyce ze względu na występujące odmowy udziału w badaniach, prawdopodobieństwo to może być nieznacznie mniejsze od jeden.

Podsumowanie

Zaprezentowane w niniejszej pracy metodologiczne i matematyczne fundamenty kontroli ujawniania danych obecnie coraz częściej są kluczowym elementem opracowywania wyników badań statystycznych i ich upowszechniania. Tym samym kontrola ujawniania danych w coraz większym stopniu staje się nieodzownym etapem badania statystycznego.

Jak wspomniano we wprowadzeniu, zapotrzebowanie na mikrodane dla celów naukowo-badawczych rośnie coraz intensywniej. Posługiwanie się bowiem takimi zasobami zwiększa istotnie potencjalną wszechstronność wynikowych analiz oraz daje możliwość projektowania i przeprowadzania eksperymentów symulacyjnych dla oceny na przykład efektywności określonych narzędzi statystycznych i ekonometrycznych. W ten sposób właśnie przeważa nad – z natury rzeczy ograniczonym – arbitralnie opracowanym klasycznym przekazem opisowo-tabelaryczno-graficznym. Można nawet zaryzykować stwierdzenie, że mikrodane w tym kontekście staną się z biegiem czasu bardziej pożądane niż gotowe publikacje, i to także dla potrzeb ambitniejszych prac dyplomowych (licencjackich, magisterskich czy doktorskich). Stosowanie opisanych tutaj metod (ukrywania, nakładania szumu, postrandomizacji i innych) powinno zatem się stać w nieodległej przyszłości standardem pracy każdego gestora danych jednostkowych (czy to pochodzących z badań, czy ze źródeł administracyjnych), a zatem nie tylko statystyki publicznej, ale także np. Zakładu Ubezpieczeń Społecznych, Kasy Rolniczego Ubezpieczenia Społecznego (rejestry ubezpieczonych), Ministerstwa Finansów i urzędów skarbowych (rejestr podatkowy POLTAX), Centralnej Ewidencji Pojazdów i Kierowców itp.

Efektywna kontrola ujawniania danych wymaga wzmożonego wysiłku oraz wykorzystania właściwych i odpowiednio wydajnych narzędzi informatycznych. Powinna ona obejmować nie tylko wszystkie możliwe kombinacje wariantów i wartości w udostępnianych zbiorach, które potencjalnie mogłyby prowadzić do uzyskania lub odtworzenia danych wrażliwych przez nieuprawnioną osobę, ale także najbardziej prawdopodobne zewnętrzne zasoby informacji, którymi taka osoba może dysponować, i ich ewentualne powiązania z przekazywanymi jej wynikami. Oczywiście wymaga to zaangażowania odpowiednio wykwalifikowanych specjalistów – zarówno w zakresie SDC, jak i metodologii konkretnych

badań statystycznych – oraz wprowadzenia właściwych rozwiązań organizacyjnych i technologicznych.

Te same postulaty i oczekiwania dotyczą także danych tabelarycznych i wynikowych statystyk sumarycznych. Tutaj również kontrola ujawniania danych powinna być przeprowadzana starannie i wszechstronnie, a dobór wykorzystanych w tym celu metod – optymalny. Dopasowanie najefektywniejszego sposobu zabezpieczania danych wrażliwych jest szczególnie istotne w wypadku tablic powiązanych hierarchicznie. Wówczas wśród kwestii do rozstrzygnięcia znajduje się także problem, czy oprzeć SDC na podziale na tablice niehierarchiczne (a następnie w odniesieniu do nich stosować np. zaokrąglanie bądź kontrolowane dopasowanie CTA), czy też traktować je w tym ujęciu jako tablice proste, a hierarchię odtworzyć na końcu. Najważniejszym wyznacznikiem skuteczności obranego rozwiązania jest ryzyko ujawnienia oraz strata informacji wynikła z zastosowania SDC. Wielkości te powinny być możliwie jak najmniejsze. Są one też podstawą każdego raportu z przeprowadzonej kontroli ujawniania danych.

Warto zaznaczyć, że każdy z tych dwóch wskaźników jakościowych ma nieco inne przeznaczenie. Ocena ryzyka ujawnienia służy bowiem przede wszystkim wewnętrznym celom gestora danych i zespołu dokonującego kontroli ich ujawniania. Dlatego powinna być poufna. Szacunek straty ma z kolei istotne znaczenie również dla użytkownika danych. Pozwala mu bowiem ocenić błędy estymacji prowadzonej w ramach jego przedsięwzięcia badawczego. Jest to bardzo przydatne na przykład w sytuacji, gdy udostępniane dane pochodzą z badania reprezentacyjnego i obejmują także stosowane w nim wagi, które są podstawą szacowania wartości określonych zmiennych dla całej populacji czy też estymacji dla małych obszarów. Stąd oczekiwana strata informacji na skutek przeprowadzonej kontroli ujawniania danych powinna być każdorazowo przekazywana korzystającemu z nich użytkownikowi.

Stosunkowo najmniej kłopotu sprawia stosowanie SDC w wypadku statystyk opisowych i zaawansowanych analiz. Używane w tym celu dane pochodzą na ogół z zasobów poddanych wcześniej kontroli ujawniania na poziomie mikrodanych czy tabulogramów, zatem ryzyko ujawniania informacji wrażliwych jest tutaj minimalne. Czasem jednak tego typu czynności bazują na zbiorach niepoddanych SDC, które są niewidoczne dla użytkownika, ale dostępne np. dla stosowanego programu komputerowego – a ten na ich podstawie dokonuje odpowiednich obliczeń i generuje ustalone przez korzystającego z rzeczonych zasobów tablice bądź wykresy. Dzieje się tak przede wszystkim w kompleksowych i wszechstronnych bazach danych takich jak Bank Danych Lokalnych prowadzony przez GUS czy też bazy danych udostępniane przez Eurostat. W największym chyba stopniu ryzyko ujawnienia danych wrażliwych dotyczy np. wielowymiarowych kostek danych OLAP, w których zawarte kombinacje wartości zmiennych mogą być bardzo różnorodne, a ich analiza pod kątem SDC wymaga zazwyczaj szczególnie dużo wysiłku.

Wymienione wyżej argumenty przemawiają za tym, aby kwestia kontroli ujawniania danych (w tym doboru stosownych metod i reguł w tym zakresie) stała się immanentną składową procesu projektowania metodologii każdego badania statystycznego oraz wykorzystania źródeł administracyjnych, a także była traktowana z taką samą pieczołowitością jak inne czynności wykonywane podczas realizacji całego procesu badawczego.

Słownik pojęć

Analityk danych → ***użytkownik uprawniony***.

Anonimizacja – przede wszystkim oznacza usunięcie ze zbioru mikrodanych zmiennych – identyfikatorów oraz tych spośród quasi-identyfikatorów, które w danym układzie stały się identyfikatorami w całej lub w znacznej części rekordów rozpatrywanego zbioru. W odróżnieniu od pseudonimizacji jest procesem nieodwracalnym. Pojęcie anonimizacji bywa także stosowane w szerszym kontekście, do ujęcia całości działań prowadzonych w ramach procesu kontroli ujawniania danych.

Atrybut danych → ***zmienna.***

Atrybuty niewrażliwe (*zmienne wynikowe niebędące poufne*) – zmienne niemające takiego stopnia poufności jak atrybuty wrażliwe (np. miejscowość zamieszkania lub zawód); potencjalnie mogą się stać pseudoidentyfikatorami.

Atrybuty wrażliwe (*poufne zmienne wynikowe*) – atrybuty, które są uznawane za należące wyłącznie do sfery prywatnej oraz poddane ochronie prawnej (wyznanie, stan zdrowia, poglądy polityczne, dochód osobisty itp.).

Barnardyzacja – metoda kontroli ujawniania danych tabelarycznych polegająca na dodawaniu do / odejmowaniu od wszystkich wartości komórek liczby jeden ze stosunkowo małym prawdopodobieństwem. Zerowe wartości komórek nie są zmieniane. Wartości brzegowe koryguje się odpowiednio w celu zachowania addytywności, a liczba zmienionych komórek jest nieznana użytkownikom tablicy.

Błąd nieefektywności – błąd w udostępnieniu danych polegający na nieuzasadnionym wstrzymaniu możliwości publikacji rezultatów analizy, która nie narusza reguł poufności danych.

Błąd poufności – błąd w udostępnieniu danych polegający na zaakceptowaniu do publikacji wyników analizy, która narusza reguły poufności.

Celowana wymiana rekordów (ang. *targeted record swapping* – TRS) – specyficzna metoda wymiany rekordów polegająca na wykorzystaniu hierarchicznej struktury mikrodanych (zwłaszcza w kontekście geograficznym). Wymiana danych następuje pomiędzy jednostkami wyższego poziomu niż podstawowy. Bazuje ona na określaniu grup rekordów o największym ryzyku ujawnienia (wyznaczanym z zastosowaniem reguły *k*-anonimowości) na każdym poziomie hierarchii i wymianie danych z zakresu zmiennych definiujących określone poziomy dokonywanej między grupami rekordów najbliższymi według odległości wyznaczanej na podstawie wartości określonych zmiennych (tzw. zmiennych podobieństwa).

Centrum Badań Ekonomicznych (ang. *Center for Economic Research* – CES) – miejsce utworzone przez US Census Bureau w 1982 r., w którym udostępniano mikrodane ekonomiczne z zakresu przetwórstwa przemysłowego.

Centrum danych naukowych (*research data centre* – RDC) – miejsce, w którym pod kontrolą udostępniającego (i po spełnieniu okre-

ślonych wymogów formalnych) badacz może korzystać z dostępu do mikrodanych na stanowiskach komputerowych wyposażonych w odpowiednie oprogramowanie i stosownie zabezpieczonych przed np. nieuprawnionym kopiowaniem, jak również pozbawionych dostępu do Internetu czy możliwości podłączenia jakichkolwiek urządzeń zewnętrznych.

Centrum zdalnego dostępu → ***system zdalnego dostępu.***

Czasowa umowa o pracę – forma dostępu do poufnych danych statystycznych polegająca na tymczasowym zatrudnieniu wnioskodawcy w instytucji udostępniającej dane (np. w służbach statystyki publicznej). Podlega on wówczas tym samym procedurom co inni pracownicy statystyki, między innymi składa ślubowanie o zachowaniu poufności danych.

Częstość → ***liczność.***

Dane administracyjne – dane zawarte w rejestrach administracyjnych.

Dane jednostkowe – suma danych jednostkowych identyfikowalnych i danych jednostkowych nieidentyfikowalnych.

Dane jednostkowe identyfikowalne – dane statystyczne zawierające informacje dotyczące konkretnego podmiotu gospodarki narodowej albo osoby fizycznej, identyfikujące bezpośrednio ten podmiot albo osobę według nazwy, imienia i nazwiska, adresu lub publicznie dostępnego numeru identyfikacyjnego oraz pozwalające na pośrednią identyfikację tego podmiotu albo osoby z użyciem innych środków niż środki pozwalające na bezpośrednią identyfikację, z wyłączeniem środków wymagających nadmiernych kosztów, czasu lub działań.

Dane jednostkowe nieidentyfikowalne – dane statystyczne zawierające informacje dotyczące konkretnego podmiotu gospodarki narodowej albo osoby fizycznej niepozwalające na bezpośrednią ani pośrednią identyfikację tego podmiotu albo osoby.

Dane panelowe – dane, które łączą w sobie zarówno wymiar przestrzenny, jak i czasowy, np. stopa bezrobocia w gminach województwa łódzkiego w latach 2005–2021 (stan w dniu 31 grudnia).

Dane poufne – dane umożliwiające bezpośrednią lub pośrednią identyfikację jednostek statystycznych, co skutkuje ujawnieniem informacji indywidualnych. Żeby określić, czy możliwa jest identyfikacja danej jednostki statystycznej, należy wziąć pod uwagę wszystkie przewidywalne środki, które mogą być użyte przez osobę trzecią do zidentyfikowania jednostki statystycznej.

Dane statystyczne – ilościowe i jakościowe, jednostkowe i zagregowane informacje opisujące złożone zjawisko w rozpatrywanej populacji.

Dane syntetyczne – dane generowane przy użyciu mechanizmów stochastycznych na podstawie oryginalnej struktury i definicji zmiennych, zachowujące wartość analityczną zbioru wyjściowego (czyli związki między zmiennymi i kształty ich rozkładów), a jednocześnie cechujące się maksymalnie wysokim poziomem ochrony przed identyfikacją jednostki i ujawnieniem informacji wrażliwych.

Dane tabelaryczne – wynikowe informacje statystyczne opracowane w formie tablic częstościowych lub wielkościowych.

Dane zagnieżdżone w ilustracji – sytuacja polegająca na tym, że to, co widać na rysunku statystycznym (wykresie, kartogramie, piktogramie), co prawda może być samo w sobie przekazem bezpiecznym z punktu widzenia ochrony informacji wrażliwych, jednakże przekaz taki powstał z wykorzystaniem potencjalnie wrażliwych mikrodanych, na podstawie których stosowny program komputerowy dokonał stosownych obliczeń. Ostatecznie wyniki takich obliczeń ukazywane są na ilustracji, a same dane źródłowe są „podczepione" do niej w – niekoniecznie widocznym – pliku.

Dodawanie szumu (ang. *noise addition*) – zakłóceniowa metoda SDC polegająca na doda-

waniu do oryginalnych danych wrażliwych specjalnie zdefiniowanych zakłóceń w celu zniekształcenia owych informacji uniemożliwiającego odtworzenie ich faktycznej postaci, jednak przy minimalizowaniu negatywnych skutków dla jakości odpowiednich informacji zagregowanych dla populacji (a najlepiej zachowaniu ich bez zmian). Szum taki modeluje się często za pomocą odpowiednich zmiennych losowych. Ich wartości dodaje się do wartości odpowiednich zmiennych (opcja addytywna, biały szum) lub przez te wartości się mnoży (opcja multyplikatywna). Szum można nakładać tylko dla zmiennych ilościowych.

Domniemany identyfikator → ***pseudoidentyfikator.***

Drzewo klasyfikacji i regresji (ang. *classification and regression tree* – CART) – podejście wykorzystywane w konstruowaniu danych syntetycznych: model CART dzieli zbiór potencjalnych regresorów na względnie jednorodne z predykcyjnego punktu widzenia podzbiory. Szereg takich podziałów przedstawia się za pomocą struktury drzewiastej, której liście odpowiadają owym podzbiorom.

Dynamiczna baza danych – baza danych statystycznych, której użytkownik ma możliwość zadawania pytań według ustalonego formatu. Pytania te są następnie przetwarzane, a narzędzia bazy generują odpowiedzi na nie. Zakres stosowanych tu metod i możliwości zależy od rodzaju danych i innych uwarunkowań związanych z udostępnianiem tych danych.

Głębokie uczenie – metoda konstruowania danych syntetycznych należąca do klasy metod uczenia maszynowego oparta na narzędziach sztucznej inteligencji. Pozwala efektywnie określać modele predykcyjne dla bardzo dużych zbiorów danych. W jej ramach można wykorzystywać generatywną sieć kontradyktoryjną (ang. *generative adversarial network* – GAN), która próbuje poznać podstawową strukturę oryginalnych danych poprzez generowanie nowych danych (a dokładniej – nowych próbek) z tego samego rozkładu statystycznego co dane oryginalne, bazując na generatorze i na weryfikującym uzyskane przez niego dane dyskryminatorze.

Identyfikacja – działanie polegające na ustaleniu związku między udostępnionymi danymi uznawanymi za poufne a konkretnym podmiotem, którego one dotyczą.

Identyfikacja bezpośrednia – identyfikacja jednostki statystycznej według jej nazwy lub adresu bądź według publicznie dostępnego numeru identyfikacyjnego.

Identyfikacja poprzez dopasowanie/łączenie rekordów – sytuacja wtórnej identyfikacji, gdy intruz ma dostęp do rejestru publicznego i próbuje dopasować te informacje w celu zidentyfikowania przebadanych w badaniu reprezentacyjnym jednostek.

Identyfikacja pośrednia – identyfikacja jednostki statystycznej z użyciem wszelkich innych środków niż środki identyfikacji bezpośredniej.

Identyfikacja spontaniczna – wtórna identyfikacja jednostki wynikająca z tego, że intruz ma szczegółową, osobistą wiedzę o jednej lub kilku indywidualnych jednostkach i przypadkowo je rozpoznaje w próbie obejmującej wyłącznie jednostki poddane badaniu.

Identyfikacja wtórna – sytuacja gdy określona jednostka zostaje zidentyfikowana przez intruza na podstawie rozmaitych posiadanych przez niego informacji takich jak wiedza o zależnościach występujących między rekordami a zmiennymi, zewnętrzne dane dotyczące rozpatrywanych jednostek lub całej ich populacji czy związki udostępnianych mu danych z opublikowanymi wynikami innych badań statystycznych.

Identyfikator – jednoznaczny, unikatowy wyróżnik osoby lub podmiotu w bazie danych (np. numer PESEL).

Imputacja – sztuczne uzupełnienie braków danych występujących w rozpatrywanej bazie

przeprowadzane w sposób losowy albo – najczęściej – na podstawie dostępnych w tejże bazie informacji, z wykorzystaniem grupowania rekordów względem określonych cech, pobieranie zastępczych wartości od najbardziej podobnych rekordów z kompletnymi danymi, ekonometrycznych modeli zależności itp. Przyczynia się nierzadko do wzmocnienia ochrony poufności czy konstrukcji danych syntetycznych.

Indywidualne ryzyko ujawnienia – ryzyko wtórnej identyfikacji konkretnej jednostki.

Intruz – użytkownik, którego celem działania jest naruszenie prywatności innych osób lub poufności określonych danych.

Jednostka statystyczna – podstawowa jednostka objęta obserwacją (osoba fizyczna, gospodarstwo domowe, podmiot gospodarczy czy inny podmiot), do której odnoszą się dane.

Jednostka zagrożona ryzykiem ujawnienia – taka jednostka, że możliwe jest jej wyodrębnienie, a niemożliwe okazuje się pomylenie jej z innymi jednostkami.

Kalibracja wag – korekta wag w badaniu reprezentacyjnym w kierunku zachowania zgodności sum oszacowań wartości określonych kluczowych zmiennych jednostek niższego rzędu dla populacji z odpowiednimi oszacowanymi wartościami dla jednostek rzędu wyższego. Może być bardzo przydatna do minimalizowania straty informacji w SDC.

Kombinacja wartości – ciąg wartości rozpatrywanych zmiennych, w którym każda zmienna jest reprezentowana przez co najwyżej jedną wartość. Kombinacją jest zatem także zestaw wartości poszczególnych badanych zmiennych dla danej jednostki, a więc stosowny rekord w bazie danych.

Komórki z ryzykiem pierwotnym – komórki tablic o małych częstościach w wypadku tablic częstości lub dodatkowo o wysokiej dominacji pojedynczych jednostek w wypadku tablic wielkości.

Komórki z ryzykiem wtórnym – komórki tablic, w wypadku których występuje wtórne ryzyko ujawnienia informacji wrażliwych.

Koncepcje unikatowości – koncepcje zagrożenia identyfikacją jednostki; dwie najpopularniejsze to koncepcja unikatowości kombinacji wartości quasi-identyfikatorów (dla zmiennych kategorialnych: kombinacje występujące tylko raz identyfikują jednostkę) oraz koncepcja unikatowości wartości w sąsiedztwie oryginalnych wartości (dla zmiennych ciągłych: w określonym sąsiedztwie danej wartości nie występują inne).

Kontrola ujawniania danych (KUD; ang. *statistical disclosure control* – SDC) – postępowanie polegające na dokonaniu przed udostępnieniem danych statystycznych ich weryfikacji w celu wyeliminowania – lub przynajmniej absolutnego zminimalizowania – ryzyka ujawnienia bądź odtworzenia informacji o jednostkach statystycznych przez użytkowników udostępnianych zasobów.

Kontrolowane dopasowanie tablic (ang. *controlled tabular adjustment* – CTA) – metoda SDC dla danych tabelarycznych polegająca na publikacji tablicy, w której wartości komórek z ryzykiem pierwotnym zostają zmienione w ten sposób, że wartości te są odpowiednio odległe od wartości rzeczywistych, tzn. odległości owe należą do pewnego ustalonego przedziału poufności. Metoda CTA opiera się na programowaniu liniowym. Wartości komórek w docelowej tablicy są zmieniane na bezpieczne. W celu zachowania addytywności zmienione muszą być również wartości niektórych komórek bezpiecznych. Czyni się to tak, aby ingerencja w dane była możliwie najmniejsza.

Korekta cykliczna – ochrona komórek tablicy przed ujawnieniem informacji wrażliwych przeprowadzana w sposób cykliczny. Polega ona na tym, że wartości brzegowe nie są zmieniane, a wartości komórek wewnętrznych zostają zmodyfikowane w sposób losowy z zachowaniem addytywności.

Kostki danych – wielowymiarowe tablice o bardzo dużym stopniu szczegółowości. Tablice takie mogą być udostępniane według wcześniej zdefiniowanych schematów lub wygenerowane na specjalne życzenie użytkownika. Stopień szczegółowości w kostkach może być tak znaczny, że metody ochrony są porównywalne ze stosowanymi do mikrodanych.

Laboratorium danych – miejsce, w którym badacz lub inny użytkownik pod kontrolą udostępniającego może korzystać z dostępu do mikrodanych. Stanowiska komputerowe w laboratorium są wyposażone w odpowiednie oprogramowanie umożliwiające analizy z dostępem do zbiorów mikrodanych. Komputery są odpowiednio zabezpieczone np. przed nieuprawnionym kopiowaniem, a praca na nich odbywa się pod nadzorem pracowników urzędu lub odpowiedniego innego gestora.

Licencjonowane zbiory mikrodanych – zbiory danych zanonimizowanych, których udostępnienie wybranym użytkownikom następuje po wcześniejszym zawarciu z nimi odpowiedniej umowy lub porozumienia. Szczegółowość danych w zbiorach licencjonowanych może być większa niż w zbiorach danych jednostkowych dostępnych dla wszystkich.

Liczność (*częstość*) – liczba jednostek należących do danej kategorii wyznaczonej przez kombinację wartości odpowiednich powiązanych zmiennych jakościowych.

Lokalne ukrywanie danych (ang. *local suppression*) – niezakłóceniowa metoda kontroli ujawniania. Polega na usuwaniu pewnych wartości niektórych zmiennych dla konkretnych jednostek w celu uniknięcia ich identyfikacji. Efektem lokalnego ukrywania jest zwiększenie liczby rekordów, dla których kombinacja określonych wartości pewnych innych zmiennych, uznanych za kluczowe, jest taka sama. Podejście to ma zastosowanie przede wszystkim do danych jakościowych.

Łączenie statystyczne (ang. *record linkage*) – znajdowanie w bazie danych rekordów odpowiadających tej samej jednostce, ale pochodzących z różnych źródeł (np. badań statystycznych czy rejestrów administracyjnych). Stosowane (a nawet konieczne) do łączenia zbiorów jednostek, które opierają się na tym samym identyfikatorze jednostki (np. numerze PESEL dla osób czy REGON bądź NIP w wypadku podmiotów gospodarczych). Jest to łączenie danych mogących pochodzić od różnych gestorów, ale uzyskanych z zastosowaniem jednolitej metody.

Mapa ryzyko-użyteczność (R–U) – wypracowanie metod, które minimalizują ryzyko ujawnienia danych i jednocześnie maksymalizują ich użyteczność. Zarówno użyteczność, jak i ryzyko są wyrażone jako miary liczbowe, na podstawie których instytucja udostępniająca dane przyjmuje maksymalny dopuszczalny poziom prawdopodobieństwa ich ujawnienia.

Maskowanie niezakłóceniowe (ang. *non–perturbative masking*) – rodzina metod SDC prowadzących do tego, że wrażliwe dane stają się, w różny sposób, niewidoczne dla zewnętrznego użytkownika. Tak więc w finalnym udostępnianym zbiorze albo określona informacja jednostkowa figuruje w dokładnej postaci, albo jej nie ma wcale.

Maskowanie zakłóceniowe (ang. *perturbative masking*) – rodzina metod SDC polegających na zakłócaniu wrażliwych wartości zmiennych w celu uniemożliwienia dokładnego ich odtworzenia przez nieuprawnionego użytkownika przy jednoczesnej minimalizacji strat informacyjnych.

MASSC (ang. *micro agglomeration, substitution, subsampling and calibration*) – zakłóceniowa metoda SDC, będąca połączeniem czterech kroków: mikroaglomeracji, podstawiania, podpróbkowania i kalibracji. Pozwala to na równoczesną kontrolę ryzyka ujawnienia i straty informacji z powodu zastosowania odpowiednich mechanizmów SDC.

Metadane – dane o danych, czyli informacje dotyczące definicji pojęć występujących w danych, okresów i jednostek, do których dane

się odnoszą, użytych metod ich wyznaczenia, ewentualnych wyjątków od reguł przyjętych w określeniu zmiennych czy braków danych, wyjaśnienia znanych przyczyn odstępstw i ubytków itp.

Metadane jakościowe – metadane opisujące różne wymiary jakości statystyk wynikowych (np. ich aktualność czy dokładność).

Metadane metodologiczne – metadane opisujące metody wykorzystane do utworzenia zbioru danych (np. metody doboru próby, pozyskiwania wyników czy ich przygotowania i obróbki).

Metadane pojęciowe – metadane opisujące użyte pojęcia i ich praktyczną implementację, pozwalające użytkownikowi zrozumieć, co statystyki mierzą, i ich przydatność.

Metadane referencyjne – metadane, które opisują zawartość i jakość danych statystycznych (np. pojęciowe, metodologiczne czy jakościowe).

Metadane strukturalne – metadane, które identyfikują strukturę danych (np. nazwy kolumn w mikrodanych lub wymiary w kostkach statystycznych) lub strukturę powiązanych metadanych (np. jednostki miary).

Metoda kluczy komórkowych (ang. *cell-key method* – CKM) – posttablicowe podejście kontroli ujawniania danych polegające na dodawaniu do każdej komórki tablicy szumu losowanego z zastosowaniem określonego mechanizmu z ustalonego rozkładu. Istotne są tutaj klucze przyporządkowywane rekordom w zbiorze mikrodanych będących podstawą konstrukcji tablicy – liczby losowe z rozkładu jednostajnego $U(0, 1)$, które dodaje się odpowiednio do naliczania tablicy i uzyskuje w efekcie klucz komórki oraz predefiniowane prawdopodobieństwa przejścia (tzw. *p*-tablica – ang. *p-table*) służące do wyznaczania szumu jako funkcji kluczy komórek i wartości komórek.

Metoda MDAV (ang. *multivariate microaggregation based on maximum distance to average vector*) – metoda wielowymiarowej mikroagregacji opartej na maksymalnej odległości od wektora średnich, w której punktem wyjścia określania grup skrajnych rekordów – najmniejszych i największych wedle ustalonego porządku – jest wyznaczenie średniego rekordu (tzn. sztucznego rekordu składającego się ze średnich arytmetycznych badanych zmiennych) i określenie rekordu najbardziej od niego odległego.

Metoda optymalna – metoda ochrony poufności danych w tablicy hierarchicznej poprzez traktowanie jej jak pojedynczej, bez dzielenia na mniejsze podtablice, co zwiększa wydajność działań SDC.

Metoda postrandomizacyjna (ang. *the post-randomization method – PRAM*) – probabilistyczna metoda SDC generująca określone zakłócenia. Wartości zmiennych jakościowych dla pewnych rekordów zostają tutaj zamienione na inne z wykorzystaniem specyficznego mechanizmu probabilistycznego, a konkretnie – macierzy przejść Markowa. Metoda PRAM łączy w sobie dodawanie szumu, ukrywanie danych oraz przekodowywanie.

Metoda RMDM (ang. *robust Mahalanobis distance based microaggregation*) – szczególny wariant mikroagregacyjnej metody MDAV oparty na medianie Webera (jako środku zbioru danych) oraz odległości Mahalanobisa.

Metoda ROMM (ang. *random orthogonal matrix masking*) – metoda dodawania szumu w ochronie mikrodanych, polegająca na zakłócaniu oryginalnych wartości przy użyciu losowej macierzy ortogonalnej.

Metoda w pełni warunkowej specyfikacji (ang. *the fully conditional specification* – FCS) – metoda konstrukcji danych syntetycznych. Dane k-wymiarowe (gdzie k jest liczbą naturalną) uzyskuje się tu poprzez k-krotne losowanie z rozkładów jednowymiarowych. Każda zmienna jest syntetyzowana osobno za pomocą odpowiedniego modelu regresyjnego opartego na dekompozycji wielowymiarowego rozkładu

łącznego na szereg jednowymiarowych rozkładów warunkowych każdej zmiennej względem jej poprzedniczek.

Metody kontroli ujawniania danych statystycznych – metody zmniejszenia ryzyka ujawnienia informacji na temat jednostek statystycznych polegające zazwyczaj na ograniczaniu ilości lub modyfikacji uwalnianych danych,

Metody posttablicowe – metody SDC, które służą do ochrony poufności danych po naliczeniu tablicy. Oprócz ochrony metody te mają na celu również ocenę ryzyka naruszenia poufności oraz maksymalizację użyteczność danych po ich zastosowaniu.

Metody pretablicowe – rodzina metod SDC dla danych tabelarycznych polegających na tym, że ochronę poufności stosuje się jeszcze przed naliczeniem tablic. W związku z tym metody te są niezależne od konkretnego naliczenia danej tablicy. Do tej kategorii można zaliczyć też każdą z metod stosowanych do ochrony mikrodanych.

Miary wpływu na siłę związku – miary straty informacji spowodowanej kontrolą ujawniania danych oceniające zachowanie kierunków i siły związków między badanymi zjawiskami, które odzwierciedlają zebrane dane. Najpowszechniej do tego celu używa się współczynnika korelacji, porównując w różny sposób odpowiednie jego wartości przed i po SDC.

Miary wpływu na wariancję szacunków – miary straty informacji spowodowanej kontrolą ujawniania danych oceniające wpływ zmian dokonywanych w wyniku zastosowania SDC na zmienność rozpatrywanych wielkości statystycznych. Opierają się głównie na porównaniu wariancji czy odchyleń standardowych w danych oryginalnych i w danych po SDC.

Miary zakłócenia rozkładu – miary straty informacji spowodowanej kontrolą ujawniania danych w kontekście zakłócenia rozkładu oryginalnych zmiennych wskutek zastosowania SDC. Opierają się na unormowanych różnicach między odpowiednimi wartościami w zbiorze danych oryginalnych oraz w zbiorze danych zniekształconych.

Mikroaglomeracja (ang. *microagglomeration*) – zakłóceniowa metoda SDC polegająca na dokonaniu podziału zbioru rekordów na skupienia o podobnym ryzyku ujawnienia danych wrażliwych. W tym celu wykorzystuje się odpowiednie quasi-identyfikatory. Następnie dokonuje się probabilistycznego podstawienia w celu zakłócenia oryginalnych danych (np. poprzez generowanie szumu czy macierzy przejść Markowa – podobnie jak w podejściu PRAM). W dalszej kolejności za pomocą probabilistycznego losowania podpróbek ukrywa się pewne zmienne bądź nawet całe rekordy (ukrywanie odbywa się zatem z określonym prawdopodobieństwem). Finalnie wreszcie dokonuje się stosownej kalibracji wag użytych w losowaniu zasadniczej próby do badania w celu zachowania jakości oszacowań odpowiedniej dla zmiennych wyjściowych, których precyzja ma istotne znaczenie dla użytkowników danych.

Mikroagregacja (ang. *microaggregation*) – rodzina zakłóceniowych narzędzi SDC zapewniających ochronę poufności danych w ujęciu makro poprzez odpowiednie działania na poziomie mikro. U podstaw stosowania mikroagregacji leżą fundamentalne reguły publikacyjne, dopuszczające publikowanie zbiorów mikrodanych, gdy zawierają one co najmniej k rekordów, a żaden z nich nie dominuje pod danym względem (tzn. jego udział w danej wielkości ogółem dla grupy nie jest większy niż $p\%$). Parametry naturalne k i p ($0 < p < 100$) są zazwyczaj arbitralnie ustalone. Ich wspólną cechą jest podział obiektów na wewnętrznie jednorodne, co najmniej k-elementowe, podzbiory. Dla każdej takiej grupy obliczane są średnie wartości zmiennych z danymi wrażliwymi – i tymi średnimi zastępuje się oryginalne wartości.

Mikrodane – dane jednostkowe zgromadzone w zbiorze, którego zasób informacyjny pochodzi z rejestru administracyjnego lub z badania statystycznego. Opisują one poszczególne jednostki,

których obsługą w określonym przez jego zadania zakresie zajmuje się dany gestor rejestru lub które były objęte określonym badaniem.

Mikrodane do użytku chronionego (ang. *secure use files*) – rodzaj mikrodanych poufnych do celów naukowych, wobec których nie zastosowano innych niż anonimizacja metod kontroli ujawniania danych statystycznych. Dostępne są one dla osób ze środowiska naukowego w siedzibie gestora danych (np. Eurostatu czy krajowego urzędu statystycznego) lub w trybie zdalnego dostępu (bez komunikacji z internetem oraz bez możliwości drukowania, kopiowania czy zapisywania na zewnętrznym nośniku). Sprawdzaniem, czy nie zostaje ujawniona informacja wrażliwa, zajmują się pracownicy odpowiedniej jednostki.

Mikrodane do zastosowań naukowych (ang. *scientific use files*) – rodzaj mikrodanych poufnych do celów naukowych, wobec których zastosowano zaawansowane metody kontroli ujawniania danych statystycznych, tak by ograniczyć do odpowiedniego poziomu i zgodnie z obecnymi najlepszymi praktykami ryzyko zidentyfikowania jednostki statystycznej. Udostępniane są one badaczom np. na płycie DVD w postaci zanonimizowanej, poddanej stosownej kontroli bazy danych, co umożliwia prowadzenie badania przez naukowca w jego miejscu pracy. Płyta taka jednak musi być przechowywana w bezpieczny sposób, a wszelkie wyniki niefinalne – chronione hasłem. Analiza danych wynikowych jest przeprowadzana przez osoby prowadzące badanie w ich miejscu pracy, na podstawie wytycznych przesłanych wraz z mikrodanymi.

Mikrodane do użytku publicznego (ang. *public use files*) – pliki zawierające zanonimizowane dane jednostkowe, w których identyfikacja jednostki statystycznej nie jest możliwa lub w których ryzyko takiej identyfikacji jest zaniedbywalnie małe. Dane takie są powszechnie dostępne.

Mikrodane poufne – mikrodane będące danymi poufnymi.

Mikrodane poufne do celów naukowych (ang. *microdata for scientific purposes*) – mikrodane udostępniane w formie do użytku chronionego lub naukowego użytkownikom prowadzącym określone prace naukowo-badawcze na podstawie stosownych umów lub porozumień określających zasady bezpieczeństwa poufności obowiązujące użytkownika w posługiwaniu się takimi danymi.

Mikrodane poufne do celów statystycznych (ang. *microdata for statistical purposes*) – mikrodane, które wykorzystuje się do tworzenia oficjalnych danych statystycznych, a także które są przekazywane wewnątrz Europejskiego Systemu Statystycznego w ramach chronionej wymiany danych.

Mikrodane proste – mikrodane, w których pomiędzy zawartymi w nich zmiennymi nie występują istotne powiązania uzasadniające nakładanie podczas przetwarzania ograniczeń czy warunków.

Mikrodane złożone – mikrodane, w których pomiędzy zawartymi w nich zmiennymi występują określone, mające znaczenie, powiązania, które można wykorzystać poprzez nakładanie podczas przetwarzania odpowiednich ograniczeń czy warunków.

Model oparty na zasadach (ang. *principles-based model*) – metoda weryfikacji danych wynikowych pod kątem ryzyka ujawnienia informacji poufnej, minimalizująca ryzyko popełnienia zarówno błędu poufności, jak i błędu nieefektywności. Model dostarcza wskazówek do sprawdzania danych wynikowych, charakteryzuje się bardziej szczegółowym spojrzeniem na dane wynikowe, a nie stosowaniem prostych reguł. Może być stosowany łącznie z zasadą kciuka.

Ocena ryzyka wtórnej identyfikacji na podstawie charakterystyk populacji (*ocena ryzyka wtórnej identyfikacji na podstawie kluczy w populacji i z wykorzystaniem modeli statystycznych lub heurystycznych do oszacowania pożądanych wielkości*) – ocena ryzyka oparta na

wnioskowaniu. W sytuacji gdy w mikrodanych wszystkie quasi-identyfikatory mają charakter jakościowy, ryzykiem ujawnienia jest funkcja komórek w tabeli kontyngencji skonstruowanej przez krzyżowe zestawienie quasi-identyfikatorów w populacji.

Ocena ryzyka wtórnej identyfikacji na podstawie kluczy w populacji i z wykorzystaniem modeli statystycznych lub heurystycznych do oszacowania pożądanych wielkości → ***ocena ryzyka wtórnej identyfikacji na podstawie charakterystyk populacji.***

Ocena ryzyka wtórnej identyfikacji na podstawie kluczy w próbie → ***ocena ryzyka wtórnej identyfikacji na podstawie udostępnianych mikrodanych.***

Ocena ryzyka wtórnej identyfikacji na podstawie teorii łączenia rekordów → ***ocena ryzyka wtórnej identyfikacji na podstawie zewnętrznej bazy danych.***

Ocena ryzyka wtórnej identyfikacji na podstawie zewnętrznej bazy danych (*ocena ryzyka wtórnej identyfikacji na podstawie teorii łączenia rekordów*) – ocena stosowana w wypadku zmiennych ilościowych, bazująca na idei rzadkości w sąsiedztwie rekordu (incydentalnego występowania danych obserwacji w rekordach bardzo podobnych do danego). Rzadkość występowania danych obserwacji w sąsiedztwie rekordu ocenia się zaś z wykorzystaniem technik łączenia rekordów. W wersji *a posteriori* ryzyko to mierzy się odsetkiem poprawnych dopasowań przy łączeniu oryginalnych danych z danymi po zastosowaniu metod zakłócających w ogólnej liczbie rekordów.

Ocena ryzyka wtórnej identyfikacji na podstawie udostępnianych mikrodanych (*ocena ryzyka wtórnej identyfikacji na podstawie kluczy w próbie*) – ocena zakładająca, że intruz dysponuje wiedzą o tym, które jednostki z populacji generalnej wzięły udział w badaniu. Ryzyko dotyczy sytuacji, w której wszystkie quasi-identyfikatory mają charakter jakościowy, jak również gdy w zbiorze mikrodanych nie zostały zastosowane żadne metody zakłócające. Ryzyko ujawnienia odnosi się do komórek tabeli kontyngencji skonstruowanej na podstawie krzyżowego zestawienia quasi-identyfikatorów. Jeżeli liczebności w tej tabeli dla kombinacji kategorii quasi-identyfikatorów są mniejsze niż przyjęty próg, kombinacje te należy uznać za zagrożone.

Ochrona uzupełniająca → ***ochrona wtórna.***

Ochrona wtórna (*uzupełniająca*) – rodzina technik SDC służących do ochrony danych tabelarycznych, prowadzących do uniemożliwienia ujawnienia komórek z ryzykiem pierwotnym.

Opracowywanie danych – działania zmierzające do utworzenia, usprawnienia i udoskonalenia metod, standardów i procedur statystycznych stosowanych przy tworzeniu i rozpowszechnianiu danych statystycznych, jak również do projektowania nowych danych i wskaźników.

Paradane – informacje objaśniające naukowe procesy poznawcze i interpretację danych statystycznych, np. sformalizowany opis metod wykorzystania owych danych liczbowych czy charakterystyka metodologicznych podstaw ich analizy. Paradane są ściśle powiązane z metadanymi i na nich się opierają. Przede wszystkim stanowią dokumentację procesu badawczego, który doprowadził do osiągnięcia danego rezultatu (należą do niej m.in. opis badania lub rejestru, z którego pozyskano określone dane, techniki ich pozyskania czy problemy wynikłe podczas tych czynności). Do paradanych zalicza się także charakterystyki badanych jednostek niewchodzące w zakres badania, ale pozyskiwane automatycznie podczas jego realizacji i tę realizację opisujące (np. liczba wejść respondenta na stronę z formularzem badania, liczba wprowadzonych przez niego zmian i modyfikacji danych, czas wypełniania formularza itp.).

Parowanie statystyczne (ang. *statistical matching*) – łączenie różnych zbiorów danych w sytuacji, gdy każdy z nich obejmuje odmienne jednostki (należące jednak do tej samej popu-

lacji docelowej), ale wszystkie zawierają pewien zestaw wspólnych (podstawowych) zmiennych. Te wspólne zmienne są używane do wzajemnego wiązania podobnych jednostek w zbiorach danych. Można wykorzystać do tego celu podejścia: losowe, oparte na odległości rekordów (w zakresie owych wspólnych zmiennych) lub modele parametryczne.

Podejście hiperkostki (ang. *hypercube*) – metoda SDC dla tablicy hierarchicznej lub łączonej polegająca na tym, że – podobnie jak dla podejścia modułowego – tablica zostaje podzielona na proste podtablice, a następnie poszukuje się rozwiązania dla każdej z podtablic w sposób iteracyjny. W odróżnieniu od podejścia modułowego, dla każdej z podtablic znalezione rozwiązanie nie musi być rozwiązaniem optymalnym, co z kolei z reguły prowadzi do bardziej restrykcyjnego ukrycia komórek, niż byłoby to konieczne.

Podejście modułowe (HiTaS) – metoda SDC dla tablicy hierarchicznej lub łączonej polegająca na tym, że dzieli się ją na tablice bez hierarchii i poszukuje optimum ukrycia dla każdej wydzielonej tablicy z osobna. Łącząc wyniki w odpowiedni sposób, uzyskuje się rozwiązanie dla całej tablicy, które nie musi być optymalne. Celem tego podejścia jest szybsze znalezienie rozwiązania, które może jednak prowadzić do większej straty informacji.

Podpróbkowanie (ang. *subsampling*) – metoda maskowania niezakłóceniowego. Oznacza to udostępnianie pewnej określonej próbki rekordów spośród figurujących w bazie danych zgromadzonych w trakcie badania statystycznego. Próbka ta może być dobrana w sposób losowy albo nielosowy.

Poufne dane do celów naukowych – dane, które umożliwiają tylko pośrednią identyfikację jednostek statystycznych, przyjmujące formę plików do zastosowań chronionych lub plików do zastosowań naukowych.

Poufne zmienne wynikowe → ***atrybuty wrażliwe.***

Próbkowanie wtórne (ang. *resampling*) – zakłóceniowa metoda SDC polegająca na tym, że z obserwacji danej zmiennej X losujemy niezależnie p próbek $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_p$, z których każda ma rozmiar n (gdzie n to liczba rekordów). Stąd $\boldsymbol{s}_1 = (x_{li_1}, x_{li_2}, \ldots, x_{li_n})$, $i_k \in \{1, 2, \ldots, n\}$, $k = 1, 2, \ldots, n$, $l = 1, 2, \ldots, p$. Następnie wartości należące do każdej z tych próbek sortujemy w ten sam sposób (np. w kolejności niemalejącej) i otrzymujemy posortowane próbki $s_l = (x_{li_{(1)}}, x_{li_{(2)}}, \ldots, x_{li_{(n)}})$, $l = 1, 2, \ldots, n$. Obserwacje zmiennej X zastępujemy finalnie średnimi arytmetycznymi wartości odpowiedniej rangi z tych próbek.

Przedział poufności – metoda ochrony informacji wrażliwych w tablicach polegająca na uniemożliwieniu poznania wartości komórki wrażliwej w przedziale, którego granice są wyznaczone przez przyjęte reguły poufności. Granice przedziału poufności wyznaczane są dla komórek w ramach ukrycia pierwotnego.

Przekodowanie (*restrukturyzacja*, ang. *recoding*) – zmiana sposobu kodowania określonych wrażliwych zmiennych. Jeśli zmienne mają charakter jakościowy, polega ono na połączeniu kilku kategorii w jedną – bardziej zgrubną i o większej liczbie należących do niej jednostek, co pozwala ukryć informację wrażliwą. Dla zmiennej ilościowej zaś przekodowanie polega na zastąpieniu owej zmiennej przez jej odpowiednik w postaci jakościowej. Metoda ma zastosowanie zarówno do mikrodanych, jak i do tablic.

Przekodowanie górne i dolne (ang. *top and bottom coding*) – rodzaj przekodowywania stosowany do zmiennych, których wartości mogą być uporządkowane (a zatem wyrażone są na skali porządkowej, różnicowej lub ilorazowej). Nowe, zgrubne kategorie są tworzone tylko dla wartości najwyższych i najniższych.

Przepływy sieciowe (ang. *network flows*) – metoda SDC dla tablicy hierarchicznej lub łączonej możliwa do zastosowania w tablicach dwuwymiarowych z co najwyżej jedną zmienną hierarchiczną. Sieć reprezentuje relacje agregacji według kolumn i wierszy w tablicy. Zapewnia

relację addytywności względem wartości brzegowych i wartości globalnej.

Pseudoidentyfikator (*quasi-identyfikator*, *zmienna kluczowa*, *domniemany identyfikator*) – zmienna w zbiorze danych, która – w połączeniu z innymi dostępnymi źródłami danych, do których ma dostęp użytkownik owego zbioru – może prowadzić do ujawnienia tożsamości respondenta.

Pseudonimizacja – przetworzenie danych osobowych w taki sposób, by nie można ich było już przypisać konkretnej osobie, której dane dotyczą, bez użycia dodatkowych informacji, pod warunkiem że takie dodatkowe informacje są przechowywane osobno i są objęte środkami technicznymi i organizacyjnymi uniemożliwiającymi ich przypisanie zidentyfikowanej lub możliwej do zidentyfikowania osobie fizycznej. W odróżnieniu od anonimizacji jest procesem odwracalnym.

Punkt dostępu – środowisko fizyczne lub wirtualne wraz z jego strukturą organizacyjną, w którym udzielany jest dostęp do poufnych danych do celów naukowych.

Quasi-identyfikator → ***pseudoidentyfikator.***

Reguła dominacji (*n*, *k*) – reguła publikacji danych tabelarycznych uniemożliwiająca dokonanie odkrycia przybliżonego. Określa ona, że komórka jest obarczona ryzykiem pierwotnym, jeśli n największych firm w komórce tablicy reprezentuje więcej niż k% całkowitej wartości cechy X w komórce, czyli gdy $x_1 + \ldots + x_n > \frac{k}{100} X$.

Zakłada się, że wartości cechy wszystkich jednostek reprezentowanych w tablicy są nieujemne. W wypadku zastosowania tej reguły minimalna liczebność w komórce wynosi $100n/k$.

Reguła koncentracji *p*% – reguła publikacji danych tabelarycznych uniemożliwiająca dokonanie odkrycia przybliżonego. Zgodnie z nią przyjmuje się, że komórka stwarza ryzyko pierwotne, jeśli całkowita wartość w komórce po odjęciu dwóch jednostek o kolejno największym udziale w komórce jest mniejsza niż p% ($p \in (0, 100)$) odpowiedniej wartości dla jednostki o największym udziale, tzn. gdy $X - x_2 - x_1 < \frac{p}{100} x_1$. Dla tej reguły zakłada się również nieujemność wartości cech jednostek, które wchodzą w skład komórek. W wypadku reguły p% minimalna liczebność jednostek składających się na wartość komórki wynosi 3.

Reguła minimalnej liczby respondentów reprezentowanych przez komórkę tablicy – reguła publikacji danych tabelarycznych oparta na tym, że mała liczba respondentów może stwarzać możliwość naruszenia poufności informacji wrażliwych. W skrajnym przypadku komórka może reprezentować jednego respondenta. Komórkę uznaje się zatem za wrażliwą, jeśli liczba respondentów reprezentowana w komórce jest mniejsza niż n, gdzie n jest pewną arbitralnie ustaloną liczbą naturalną. Bardzo często przyjmuje się, że $n = 3$.

Reguła *p* / *q* – reguła publikacji danych tabelarycznych określająca, że komórka jest zagrożona ryzykiem pierwotnym, jeśli respondent należący do niej (i znający do q% pozostałych w niej udziałów) może oszacować dane wielkości dla innego respondenta z precyzją p% prawdziwej wartości. Reguła ta ma zastosowanie wyłącznie w wypadku tablic wielkości.

Reguła progu – reguła SDC bazująca wyłącznie na informacji z próby, wedle której podejrzewa się, że kombinacje quasi-identyfikatorów (klucze) posłużą intruzowi do wtórnej identyfikacji respondenta (np. gdy – ze względu na wartości klucza – respondent jest rzadki w populacji generalnej). Dla każdego klucza ustala się wartość progu wrażliwości. Klucz uznaje się za bezpieczny, jeżeli jego wartość występuje częściej niż ustalony dlań próg. W przeciwnym razie klucz taki należy chronić ze względu na ryzyko wtórnej identyfikacji.

Rejestry administracyjne – zbiory danych gromadzonych przez gestorów publicznych

i niepublicznych w związku z wykonywaniem przez owe podmioty określonych zadań w systemie administracji rządowej lub samorządowej bądź w związku z prowadzeniem działalności gospodarczej, a zatem głównie w celach innych niż statystyczne. Obejmują one wykazy, listy oraz spisy: podmiotów (osób fizycznych, osób prawnych, jednostek organizacyjnych nieposiadających osobowości prawnej); obiektów materialnych; procesów ekonomicznych lub technologicznych; zdarzeń społecznych, ekonomicznych, technicznych, ekologicznych i innych, których rejestrowanie i ewidencjonowanie jest niezbędne jednostkom do realizacji ich funkcji w społeczeństwie lub gospodarce.

Rekord unikatowy – rekord taki, że kombinacja wartości zmiennych zawarta w nim nie występuje nigdzie więcej w rozpatrywanej bazie danych.

Restrukturyzacja → ***przekodowanie***.

Rozpowszechnianie danych – działanie, którego celem jest udostępnianie danych statystycznych i analiz statystycznych użytkownikom.

Różnicowanie prywatności (ang. *differential privacy* – DP) – reguła ochrony danych wrażliwych oparta na ocenie możliwości identyfikacji jednostki niewystępującej w danej bazie na podstawie informacji pochodzących z owej bazy i ewentualnych innych źródeł. Stanowi ona swoistą alternatywę dla tradycyjnych ram kontroli ujawniania. Pozwala na efektywną ocenę metod SDC i udostępniania danych poprzez porównanie tego, co zastosowanie danego podejścia może ujawnić o jednostce, gdy informacje na jej temat są obecne w udostępnianym zbiorze danych, z tym, co może się stać na jej temat jawne, gdy w bazie ona nie występuje lub nie podała do niej stosownych informacji. Oczywiście, im mniejszy wkład jednostki w dany zbiór, tym ryzyko ujawnienia informacji wrażliwej drogą DP jest mniejsze.

Ryzyko hierarchiczne – ryzyko ujawnienia związane z hierarchiczną strukturą danych: jeżeli osoba, której udało się naruszyć poufność, dokonała poprawnej identyfikacji choćby jednej jednostki niższego poziomu należącej do jednostki poziomu wyższego, to skutkiem wiedzy o tej przynależności może być identyfikacja pozostałych jednostek poziomu niższego wchodzących w skład tejże jednostki poziomu wyższego. Ryzyko hierarchiczne wyznacza się zarówno na poziomie indywidualnym, jak i globalnym.

Ryzyko pierwotne – ryzyko bezpośredniego poznania wartości cechy respondenta przez osoby do tego nieuprawnione.

Ryzyko ujawnienia – prawdopodobieństwo zajścia zdarzenia polegającego na tym, że dany podmiot pozna charakterystykę innego podmiotu na podstawie udostępnionych danych, naruszając tym samym jego autonomię w zakresie prawa do prywatności.

Ryzyko wewnętrzne – ryzyko potencjalnej identyfikacji jednostki tylko na podstawie danych udostępnionych użytkownikowi przez gestora.

Ryzyko wtórne (*ryzyko wtórnej identyfikacji*) – ryzyko ujawnienia informacji wrażliwej po wyeliminowaniu ryzyka pierwotnego. Ujawnienie takie może nastąpić np. poprzez wykorzystanie związków i zależności pomiędzy poszczególnymi informacjami czy danymi w bazie lub w tablicy. W tym drugim wypadku zakłada się, że częścią tablicy są także jej wartości brzegowe. Ryzyko wtórne można określić równoważnie jako prawdopodobieństwo prawidłowego dopasowania danych do konkretnej jednostki. Zależy ono także od niezależnych informacji, którymi (wedle podejrzeń przeprowadzającego kontrolę ujawniania) dysponuje intruz.

Ryzyko wtórnej identyfikacji → ***ryzyko wtórne.***

Ryzyko zewnętrzne – ryzyko potencjalnej identyfikacji jednostki, gdy użytkownik ma dostęp także do alternatywnych źródeł danych, które może skutecznie powiązać z zasobem udostępnionym mu przez gestora.

Scenariusz ujawnienia – hipotetyczny sposób możliwego naruszenia poufności chronionych danych przez użytkownika. Określa się go na podstawie wiedzy o planowanych do udostępnienia zasobach oraz o tym, do jakich innych zbiorów danych jednostkowych ów użytkownik może mieć dostęp, a także w jaki sposób by z nich skorzystał do naruszenia poufności. Scenariusz, który niesie ze sobą największe ryzyko ujawnienia, nazywa się najgorszym możliwym scenariuszem.

Skala ilorazowa – rodzaj skali pomiarowej; wyrażone na niej dane mogą służyć do rozróżniania i uporządkowania jednostek pod określonym względem oraz można na nich wykonywać wszystkie operacje arytmetyczne (dodawanie, odejmowanie, mnożenie i dzielenie) – np. przeciętne miesięczne wynagrodzenie czy liczba ludności przypadającej na jednego lekarza. Przyjmuje się tutaj milcząco istnienie zera bezwzględnego, które w praktyce zwykle nie występuje.

Skala interwałowa → *skala różnicowa*.

Skala nominalna – rodzaj skali pomiarowej; wyrażone na niej dane mogą służyć jedynie do rozróżniania jednostek: można stwierdzić, czy dwie jednostki są w zakresie danej cechy tożsame, czy różne (taka sytuacja występuje np. w wypadku płci lub gminy zamieszkania). Wartości obserwacji wyrażonych na takiej skali nie da się uporządkować, a jedyną relację pomiędzy nimi stanowi relacja równości. Liczb nie można tu zatem porównywać ani wykonywać na nich żadnych operacji algebraicznych – są one bowiem tylko etykietami, oznaczeniami.

Skala pomiarowa – rodzina funkcji przekształcających zbiorowość statystyczną w przestrzeń liczb rzeczywistych. Każdej jednostce przyporządkowuje się odpowiednią liczbę odzwierciedlającą rodzaj lub nasilenie określonego zjawiska.

Skala porządkowa – rodzaj skali pomiarowej; wyrażone na niej dane mogą służyć do rozróżniania i uporządkowania jednostek pod określonym względem (np. poziom wykształcenia czy stopień niepełnosprawności osób). Zakłada się tutaj bowiem istnienie w zbiorze danych określonego porządku spełniającego klasyczne założenia spójności (jeżeli dwa obiekty są różne, to jeden jest mniejszy od drugiego), antysymetrii (jeżeli jeden obiekt jest mniejszy od drugiego, to nie może być równocześnie na odwrót) i przechodniości (jeśli jeden obiekt jest mniejszy od drugiego, a ów drugi od trzeciego, to pierwszy jest mniejszy od trzeciego). Podobnie jak w przypadku skali nominalnej, na danych wyrażonych za pomocą skali porządkowej nie można wykonywać żadnych operacji arytmetycznych.

Skala przedziałowa → *skala różnicowa*.

Skala różnicowa (*przedziałowa, interwałowa*) – rodzaj skali pomiarowej; wyrażone na niej dane mogą służyć do rozróżniania i uporządkowania jednostek pod określonym względem oraz mogą być dodawane i odejmowane, ale ich mnożenie i dzielenie jest niewykonalne. O skali różnicowej pomiaru mówimy, jeśli liczbowy opis danej cechy przewiduje występowanie obserwacji wyrażonych zarówno liczbami ujemnymi, jak i dodatnimi (oraz zera). Może to dotyczyć np. przyrostu naturalnego na 1000 ludności, salda migracji, wskaźnika rentowności przedsiębiorstwa itp.

Statystyczne zaciemnianie zachowujące informacje (ang. *information preserving statistical obfuscation* – IPSO) – metoda generowania danych syntetycznych przy jednoczesnym zachowaniu wartości określonych statystyk i esencji wniosków statystycznych. Istotą tego podejścia jest syntetyzacja subbazy zmiennych poufnych w taki sposób, aby w modelu regresji wielorakiej tych zmiennych względem zmiennych niebędących poufnymi syntetyczne wartości zmiennych poufnych dawały takie same (lub bardzo zbliżone) oszacowania parametrów i błędów standardowych (i macierzy kowariancji) jak wówczas, gdyby zastosowano oryginalny zbiór danych.

Strata informacji – ubytek zasobu informacyjnego zbioru danych statystycznych na skutek

ukrycia lub zniekształcenia pewnych zawartych w nim danych dokonanego w wyniku zastosowania kontroli ujawniania danych. Bywa uważana za pojęcie subiektywne, choć jej skalę w wymiarze obiektywnym można dość precyzyjnie oceniać w różny sposób – poprzez miary zakłócenia rozkładu (oparte na metrykach odległości pomiędzy rzeczywistymi a zmienionymi wartościami), przez wpływ na wariancję szacunków (gdy pod uwagę są brane różnice między wariancjami dla przeciętnych wartości określonych poziomów agregacji) lub całej tablicy, za pomocą jednoczynnikowej analizy wariancji ANOVA dla wybranej zmiennej zależnej względem wybranych niezależnych zmiennych jakościowych czy poprzez wpływ na siłę związku (wówczas wykonuje się test niezależności pomiędzy wymiarami, które tworzą dane/tablicę).

System zdalnego dostępu (ang. *remote access facility* – RAF) – system zapewniający zdalny dostęp do danych statystycznych. Oferuje nowe możliwości technologiczne, niwelujące niektóre niedogodności związane z udostępnianiem mikrodanych w laboratorium danych. Prowadzi również do poszukiwania innych sposobów wsparcia środowisk naukowych odnośnie do analiz z wykorzystaniem mikrodanych. Centrum RAF może mieć jedną z dwóch form: zdalnego przetwarzania (wówczas badacz ma pełny dostęp do metadanych zbiorów mikrodanych, na których możliwe jest przeprowadzenie założonej analizy; na ich podstawie przygotowuje skrypt do wykonania, który następnie zostaje sprawdzony i wykonany przez pracowników urzędu statystycznego lub innego gestora danych, który je udostępnia; badacz otrzymuje jego wynik) i zdalnego dostępu (umożliwiającego dostęp do mikrodanych bez konieczności stawienia się w określonej lokalizacji – głównie dzięki nawiązaniu bezpiecznego połączenia poprzez sieć komputerową z dowolnego miejsca).

Szczególny wyjątek – rekord, który jest unikatowy zarówno w zestawie zmiennych, jak również w określonym podzestawie tego zestawu.

Tablice częstości – tablice, w których każda komórka reprezentuje liczność (częstość), czyli liczbę jednostek należących do danej kategorii, tzn. do wymiaru tabeli, który dana komórka obrazuje. Tablice te są typowe dla badań społecznych.

Tablice hierarchiczne – tablice zawierające hierarchiczny układ jednostek według określonego przekroju, w których sumy pewnych wartości dla jednostek położonych niżej w tej hierarchii muszą być równe odpowiednim wartościom dla określonych jednostek znajdujących się w owej hierarchii wyżej.

Tablice łączone – tablice przedstawiające dane, które można połączyć według przynajmniej jednej wspólnej zmiennej występującej w tychże tablicach.

Tablice wielkości – tablice, w których każda komórka reprezentuje wartość sumaryczną dla cechy ilościowej dla jednostek, które dana komórka reprezentuje.

Tasowanie (ang. *shuffling*) – metoda ochrony zmiennych wrażliwych w mikrodanych polegająca na zakłócaniu ich wartości modelem regresji tychże zmiennych względem zmiennych bezpiecznych.

Tworzenie informacji statystycznych – wszelka działalność związana ze zbieraniem, przechowywaniem, przetwarzaniem, zestawianiem i analizą danych, działaniami niezbędnymi do zestawienia danych statystycznych.

Ujawnienie atrybutu – sytuacja, w której na podstawie udostępnionych danych możliwe jest poznanie dodatkowej charakterystyki podmiotu lub grupy podmiotów oraz ich identyfikacja.

Ujawnienie atrybutu grupy – uzyskanie na podstawie danych tabelarycznych dodatkowych informacji o zidentyfikowanej grupie jednostek lub o tym, że zidentyfikowana grupa jakiegoś atrybutu nie posiada. Niebezpieczeństwo ujawnienia danych wrażliwych w tej formie występuje zwłaszcza wówczas, gdy wszystkie lub prawie wszystkie jednostki należą do jednej kategorii.

Ujawnienie atrybutu jednostki – sytuacja, gdy na podstawie danych tabelarycznych, z powodu niewielkich wartości w komórkach tablicy w pierwszej kolejności dochodzi do identyfikacji jednostki – na przykład wartość w komórce lub wartość brzegowa wynosi jeden. Następnie poprzez inne publikacje oparte na tym samym źródle danych zostają uzyskane dodatkowe informacje o zidentyfikowanej jednostce.

Ujawnienie cechy – sytuacja, gdy wskutek udostępnienia zbioru mikrodanych zostaje ujawniona wrażliwa informacja o konkretnej indywidualnej jednostce. Jeżeli w udostępnionych mikrodanych znajdują się poufne lub wrażliwe zmienne, to ujawnienie tożsamości skutkuje ujawnieniem cechy.

Ujawnienie danych – poznanie przez podmiot (osobę lub instytucję) faktów na temat innego podmiotu (osoby lub instytucji), do znajomości których nie uprawnia go ani obowiązujące prawo, ani zasady wynikające z innych norm zawodowych czy etycznych.

Ujawnienie dedukcyjne – sytuacja, gdy na podstawie udostępnionych mikrodanych możliwe jest dokładniejsze określenie wartości pewnych cech indywidualnej jednostki, niż byłoby to możliwe w inny sposób – można je określić z wysoką pewnością na podstawie statystycznych zależności.

Ujawnienie dokładne – sytuacja, gdy do danego obszaru lub domeny należą dwie jednostki, co prowadzi do identyfikacji i ujawnienia atrybutu jednej z nich przez drugą, która zna swój udział w wartości agregatu.

Ujawnienie poprzez wnioskowanie – sytuacja, gdy na podstawie publikowanych danych jest możliwe ustalenie szacunku dotyczącego charakterystyki podmiotu z większą dokładnością, niż zakłada to publikacja. Szacunek może być wykonany za pomocą metod wnioskowania statystycznego o wysokim stopniu precyzji.

Ujawnienie przez łączenie – uzyskanie na podstawie danych tabelarycznych dodatkowych informacji poprzez wykorzystanie relacji łączących dwie lub więcej tablic.

Ujawnienie przybliżone – sytuacja, w której informację wrażliwą można wtórnie odkryć w sposób przybliżony, ale z odpowiednio wysoką precyzją. Na przykład, jeśli liczba respondentów w komórce tablicy wynosi więcej niż trzy (sięga nawet kilkudziesięciu), to mimo dużej liczby jednostek udział jednego z respondentów w wartości komórki może być znaczny, wynosić np. 95%. Mając wiedzę o dominacji jednej jednostki w komórce, np. firmy reprezentującej dany rodzaj działalności w określonym regionie, można z całkiem dokładnym przybliżeniem ustalić wartość cechy dominującej firmy.

Ujawnienie tożsamości – rodzaj identyfikacji wtórnej, gdy rekord opisujący daną jednostkę może zostać rozpoznany w udostępnionym zbiorze mikrodanych w wyniku porównania informacji dla indywidualnej jednostki w próbie z pozostającą do jego dyspozycji odrębną listą jednostek, zawierającą indywidualne zmienne identyfikujące.

Ukrywanie komórek tablicy – najpopularniejsza metoda stosowana w celu ochrony danych w tablicach: wartości wszystkich komórek wskazanych jako wrażliwe zostają zastąpione ustalonym symbolem.

Użyteczność → ***strata informacji***.

Użytkownik uprawniony – osoba, której celem działania jest analiza danych lub prowadzenie badań oraz która nie ma intencji naruszania prywatności innych osób ani poufności określonych informacji statystycznych.

Wagi → ***zmienne ważące.***

Wartość brzegowa – w tablicy wielkości suma wartości cechy jednostek dla określonego podzbioru zmiennych klasyfikacyjnych.

Wykorzystanie do celów statystycznych – wykorzystanie wyłącznie do celów opracowywa-

nia i tworzenia wyników oraz analiz statystycznych.

Wymiana danych (ang. *data swapping*) – zakłóceniowa metoda SDC polegająca na przekształceniu bazy danych poprzez zamianę między rekordami objętych ochroną wartości danej zmiennej. Czyni się to w taki sposób, aby zachować odpowiednie wielkości agregatowe lub częstości dla poziomów wyższego rzędu. Metodę można stosować do każdego rodzaju danych

Wymiana rang (ang. *rank swapping*) – zakłóceniowa metoda SDC, którą można stosować zarówno do danych wyrażonych na skali porządkowej, jak i na skalach silniejszych. Polega ona na uporządkowaniu wartości zmiennej X w kolejności rosnącej. Następnie każda zrangowana w ten sposób wartość zmiennej X jest zamieniana z inną wartością losowo wybraną spośród tych wartości, których rangi zawierają się w pewnym ograniczonym przedziale – na przykład spośród tych, których rangi nie różnią się od rangi danej wartości więcej niż o p% całkowitej liczby rekordów, gdzie $p \in (0, 100)$ jest ustalonym parametrem.

Wymiar tablicy – liczba zmiennych klasyfikacyjnych znajdujących się w tablicy.

Zaokrąglanie (ang. *rounding*) – zakłóceniowa metoda ochrony wrażliwych wielkości. Oryginalne wartości zastępuje się ich wersjami zaokrąglonymi. Dla danej zmiennej wartość zaokrąglenia jest wybrana zazwyczaj ze zbioru punktów zaokrągleń definiującego zestaw zaokrągleń. Metodę można stosować dla zmiennych ilościowych oraz dla tablic.

Zaokrąglanie kontrolowane – rodzaj zaokrąglania polegający na wykorzystaniu programowania liniowego w celu zmiany wartości komórek o bardzo niewielką wartość wraz z zachowaniem addytywności. Możliwe jest tu przyjęcie ustalonego progu ochrony dla komórek chronionych. Komórki są wówczas zaokrąglane tak, aby zmieniona wartość różniła się od rzeczywistych wartości co najmniej o założoną wartość progową, wyrażoną np. w procentach. Metoda może być stosowana w mikrodanych i w tablicach; w tym drugim przypadku trudność implementacji i złożoność obliczeniowa zwiększają się, gdy rozmiary tablicy rosną. Problemem staje się też wówczas zachowanie poufności w wypadku tablic łączonych.

Zaokrąglanie losowe – rodzaj zaokrąglania, w którym każda wartość zmiennej lub komórki tablicy jest zaokrąglana niezależnie od innych, lecz w ten sposób, że większe prawdopodobieństwo ma zdarzenie polegające na tym, że owa wartość będzie zmieniona na wartość bliższą wielokrotności podstawy. Schematy przypisanych prawdopodobieństw mogą być różne, ale nie powinny być obciążone, to znaczy że wartość oczekiwana różnicy między komórką zmienioną a źródłową wynosi zero.

Zaokrąglanie standardowe – rodzaj zaokrąglania, w którym wartość zmiennej lub komórki w tablicy jest zaokrąglona do najbliższej wielokrotności przyjętej podstawy. Na ogół za podstawę zaokrąglania przyjmuje się 3 lub 5.

Zasada kciuka (ang. *rule-of-thumb*) – metoda weryfikacji danych wynikowych pod kątem ryzyka ujawnienia informacji poufnej, zapobiegająca popełnieniu błędów poufności. Błędy nieskuteczności są tu akceptowane. Opiera się na ścisłych regułach (zasady: progu, stopni swobody, ujawniania danych o grupach, przewagi), które mogą być stosowane w mniej lub bardziej zautomatyzowany sposób; mogą zostać zastosowane nawet bez obszernej wiedzy i doświadczenia z zakresu kontroli ujawniania danych statystycznych.

Zasada progu (ang. *threshold rule*) – jedna z podzasad zasady kciuka, odnosząca się do liczby jednostek. Wedle niej w każdej komórce (punkcie danych itp.) we wszystkich zestawieniach tabelarycznych i podobnych danych wynikowych powinno się znajdować co najmniej 10 jednostek nieprzeważonych.

Zasada przewagi – podzasada zasady kciuka, wedle której w zestawieniach tabelarycz-

nych lub podobnych jednostka o największym udziale wartości w komórce nie powinna mieć większego udziału w ogólnej wartości tej komórki niż 50%.

Zasada stopni swobody – podzasada zasady kciuka, która mówi, że wszystkie dane wynikowe uzyskane w drodze modelowania powinny mieć co najmniej 10 stopni swobody (liczba stopni swobody to liczba obserwacji pomniejszona o liczbę parametrów oraz inne ograniczenia wynikające z modelu) oraz że do budowy tego modelu powinno być wykorzystanych także co najmniej 10 jednostek.

Zasada ujawniania danych o grupach – podzasada zasady kciuka stanowiąca, że aby zapobiec ujawnieniu grupy, w zestawieniach tabelarycznych i podobnych wartość żadnej komórki nie powinna przekraczać 90% odpowiedniej sumy brzegowej (wierszowej lub kolumnowej). Niedopuszczalna jest również sytuacja, gdy pewne zmienne wykorzystane w konstrukcji tabeli definiują grupę jednostek, a inne zmienne ujawniają ważną informację dla każdego z członków grupy.

Zasada *k*-anonimowości – zasada, zgodnie z którą dla liczby naturalnej $k > 1$ w zbiorze danych każda kombinacja wartości k pseudoidentyfikatorów występuje dla przynajmniej k rekordów.

Zasada *l*-różnorodności – zasada, zgodnie z którą dla liczby naturalnej $l > 1$ zbiór danych powinien zawierać przynajmniej l różnych wrażliwych wartości każdej cechy takich, że są one najczęściej przyjmowanymi wartościami, a ich częstości są identyczne lub prawie identyczne.

Zasada *t*-bliskości – zasada, zgodnie z którą rozkład cechy wrażliwej w każdej badanej klasie jest bliski rozkładu tejże cechy w całej populacji (tzn. różnica pomiędzy tymi rozkładami nie przekracza ustalonego progu t).

Zbieranie danych – badania oraz wszelkie inne formy pozyskiwania informacji z różnych źródeł, w tym ze źródeł administracyjnych.

Zbiory częściowo syntetyczne – zbiory danych, w których tylko niektóre informacje zostały przekształcone na syntetyczne; najczęściej przekształcenie to dotyczy zmiennych będących najistotniejszymi nośnikami informacji wrażliwych.

Zbiory nieme (ang. *dummy files*) – zestawy danych syntetycznych, w których utrzymana jest struktura i reguły logiczne obowiązujące w danych oryginalnych, ale na ogół nie wartość analityczna oryginału.

Zbiory w pełni syntetyczne – zbiory danych, w których wartości wszystkich zmiennych zostały przekształcone na syntetyczne dla zachowania odpowiedniej wartości analitycznej zgromadzonych informacji w porównaniu z oryginalnym zbiorem danych.

Zbiór mikrodanych – baza danych składająca się z rekordów (wierszy), spośród których każdy zawiera zmienne (atrybuty) o indywidualnej jednostce podlegającej analizie.

Zmienna (*atrybut danych*) – cecha jednostek tworzących określoną populację będąca przedmiotem badania statystycznego lub rejestracji, opisana danymi statystycznymi określonego rodzaju dla tychże jednostek.

Zmienna ciągła (*ilościowa, numeryczna*) – zmienna, która przyjmuje wartości ze zbioru nieskończonego wyrażone na różnicowej lub ilorazowej skali pomiarowej. Tym samym jest ciągłą funkcją na zbiorze obiektów.

Zmienna dyskretna – zmienna, która przyjmuje wartości ze zbioru skończonego lub przeliczalnego.

Zmienna grupująca (*zmienna klasyfikująca)* – zmienna jakościowa, według której odbywa się grupowanie jednostek dla celów publikacyjnych i analitycznych. W wypadku tablic zmienna taka dzieli zbiorowość jednostek na poszczególne komórki tablicy.

Zmienna jakościowa → *zmienna kategorialna*.

Zmienna ilościowa → ***zmienna ciągła.***

Zmienna kategorialna (*jakościowa*) – zmienna przyjmująca wartości ze skończonego zbioru kategorii. Innymi słowy, jest to zmienna, której obserwacje wyrażone są na nominalnej lub porządkowej skali pomiarowej.

Zmienna klasyfikująca → ***zmienna grupująca.***

Zmienna kluczowa → ***pseudoidentyfikator.***

Zmienna numeryczna → ***zmienna ciągła.***

Zmienna wynikowa – zmienna, z której zakresu dane zostały zebrane w wyniku badania statystycznego lub pozyskane z rejestru administracyjnego; są one przeznaczone do udostępnienia oraz stanowią podstawę opracowań statystycznych.

Zmienne hierarchiczne – szczególny przypadek zmiennych jakościowych; występują wówczas, gdy odnoszą się do domen, które mają charakter hierarchiczny, zagnieżdżony i które mogą być liniowo uporządkowane według stopnia tego zagnieżdżenia.

Zmienne kartograficzne – zmienne służące do określania topograficznego położenia obiektów w terenie lub ich identyfikacji w inny sposób przy tworzeniu wykresów mapowych.

Zmienne niehierarchiczne – zmienne jakościowe odnoszące się do domen, które nie mają charakteru hierarchicznego.

Zmienne niewrażliwe → ***atrybuty niewrażliwe.***

Zmienne o gospodarstwach domowych – zmienne o identycznych wartościach dla każdej osoby wchodzącej w skład danego gospodarstwa domowego.

Zmienne predefiniujące nadrzędną jednostkę statystyczną – zmienne przyjmujące identyczne wartości dla każdej jednostki statystycznej podrzędnej wchodzącej w skład jednostki statystycznej nadrzędnej w danej hierarchii.

Zmienne regionalne – zmienne przestrzennie lokalizujące respondenta. Mogą dotyczyć zarówno jednostek administracyjnych (makroregion, województwo, region, podregion, powiat, gmina), jak i obszarów funkcjonalnych (np. wyznaczonych na podstawie poziomu urbanizacji czy oddziaływań społeczno-ekonomicznych).

Zmienne ważące (*wagi*) – specjalne zmienne występujące w zbiorach danych pochodzących z badań reprezentacyjnych, wykorzystywane w procesie estymacji do uogólniania wyników z próby na populację generalną. Są opracowywane na podstawie schematu doboru próby, przy uwzględnieniu poprawek związanych z jednostkowymi brakami odpowiedzi i ewentualnych innych kryteriów.

Zmienne wrażliwe → ***atrybuty wrażliwe.***

Zmienne wynikowe niebędące poufne → ***atrybuty niewrażliwe.***

Zmienne z preskupieniami – zmienne o identycznych wartościach dla każdej jednostki wchodzącej w skład danego predefiniowanego skupienia jednostek. Szczególnym przypadkiem tego rodzaju zmiennych są zmienne o gospodarstwach domowych.

Zróżnicowanie prywatności – strategia ochrony danych polegająca na tym, że do zbioru danych dodawane są dodatkowe rekordy, co ma zwiększyć gwarancję prywatności w sensie probabilistycznym.

Bibliografia

American Statistical Association [ASA]. (2016). *Ethical guidelines for statistical practice*. Committee on Professional Ethics of the American Statistical Association. http://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf

Benedetti, R. i Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. W: *Pre-proceedings of New Techniques and Technologies for Statistics* (t. 1, s. 225–232).

Benschop, T. i Welch, M. (2019). S*tatistical disclosure control for microdata: A theory guide*. The World Bank. https://sdctheory.readthedocs.io/en/latest/

Benschop, T., Machingauta, C. i Welch, M. (2022). *Statistical disclosure control: A practice guide*. The World Bank. https://media.readthedocs.org/pdf/sdcpractice/latest/sdcpractice.pdf

Bickel, P. J. i Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, *9*(6), 1196–1217. https://doi.org/10.1214/aos/1176345637

Biecek, P. (2014). *Przewodnik po pakiecie R*, wyd. 3 rozsz. Oficyna Wydawnicza GiS.

Bielak-Jomaa, E. i Lubasz, D. (red.). (2017). *RODO. Ogólne rozporządzenie o ochronie danych. Komentarz*. Wolters Kluwer Polska.

Biemer, P. P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C. i West, B. T. (2017). *Total survey error in practice*, John Wiley & Sons.

BIP [Biuletyn Informacji Publicznej]. (b.d.). PBSSP 2022. https://bip.stat.gov.pl/dzialalnosc-statystyki-publicznej/program-badan-statystycznych/pbssp-2022/

Bond, S., Brandt, M. i de Wolf, P.-P. (2015), *Guidelines for the checking of output based on microdata research*. Data without Boundaries. Guidelines for Output Checking, Improved Methodologies for Managing Risks of Access to Detailed OS Data.. https://ec.europa.eu/eurostat/cros/system/files/dwb_standalone-document_output-checking-guidelines.pdf

Box, G. E. i Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B* (*Methodological*), *26*, 211–252.

Brandt, M., Franconi, L., Guerke, C., Hundepool, A., Lucarelli, M., Mol, J., Ritchi, F., Seri, G. i Welpton, R. (2010). *Guidelines for the checking of output based on microdata research*. ESSNet SDC. http://eprints.uwe.ac.uk/22487

Calviño, A. (2017). A simple method for limiting disclosure in continuous microdata based on principal component analysis. *Journal of Official Statistics*, *33*(1), 1–27. http://dx.doi.org/10.1515/JOS-2017-0002

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. W: *Proceedings of the Section on Survey Research Methods* (s. 41–49). American Statistical Association.

Cox, L. H. (1995). Protecting confidentiality in business surveys. W: B. G. Cox, D. A. Binder, B. Nanjamma Chinnappa, A. Christianson, M. J. Colledge i P. S. Kott (Eds.), *Business survey mthods* (s. 443-473). John Wiley & Sons. https://doi.org/10.1002/9781118150504.ch24

Cox, L. H., Karr, A. F. i Kinney, S. K. (2011). Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act. *International Statistical Review*, *79*(2), 160–183. https://doi.org/10.1111/j.1751-5823.2011.00140.x

Czech, K. (2017). *Dowody i postępowanie dowodowe w międzynarodowym arbitrażu handlowym oraz inwestycyjnym. Zagadnienia wybrane*. Wolters Kluwer Polska.

Daalmans, J. i de Waal, T. (2010). *A general formulation of the secondary cell suppression problem*. Statistics Netherlands.

De Waal, T. i Coutinho, W. (2020). Solving the disclosure auditing problem for secondary cell suppression by means of linear programming. *Transactions on Data Privacy*, *13*(2), 67–90. http://www.tdp.cat/issues16/tdp.a355a19.pdf

De Wolf, P. P. (2007). *Cell suppression in a special class of linked tables*. Joint UNECE/Eurostat Worksession on Statistical Data Confidentiality (Manchester, United Kingdom, 17–19 December 2007). https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2007/12/confidentiality/wp.21.e.pdf

De Wolf, P. P., Gouweleeuw, J. M., Kooiman, P. i Willenborg L.C.R.J. (1999). Reflections on PRAM. W: European Commission, Eurostat. *Statistical data protection. Proceedings of the conference, Lisbon, 25 to 27 March 1998* (s. 337–349). Publications Office.

De Wolf, P. P., Hundepool, A., Giessing, S., Salazar, J.-J., Castro, J. (2014). *τ-ARGUS User's Manual*. Version 4.1. Statistics Netherland. https://research.cbs.nl/casc/Software/TauManualV4.1.pdf

De Wolf, P.-P. i Zeelenberg. K. (2015). Challenges for statistical disclosure control in a world with big data and open data. W: *Proceedings of the 60th World Statistics Congress*, vol. 60. MPRA. https://mpra.ub.uni-muenchen.de/88658/1/MPRA_paper_88658.PDF

Dehnel, G., Młodak, A., Klimanek, T i Szymkowiak, M. (2022). Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality. W*iadomości Statystyczne. The Polish Statistician*, *67*(2), 51–61. https://ws.stat.gov.pl/Article/2022/2/051-061

Domingo-Ferrer, J. (2009). Inference control in statistical databases. W: L. Liu i M. T. Özsu (Eds.), *Encyclopedia of database systems*. Springer.

Domingo-Ferrer, J. i Torra, V. (2003). On the connections between statistical disclosure control for microdata and some artificial intelligence tools. *Information Sciences*, *151*, 153–170. https://doi.org/10.1016/S0020-0255(03)00064-1

Domingo-Ferrer, J. i Torra, V. (2004). Disclosure risk assessment in statistical data protection. *Journal of Computational and Applied Mathematics*, *164–165*, 285–293. https://doi.org/10.1016/S0377-0427(03)00643-5

Domingo-Ferrer, J. i Torra, V. (2005). Ordinal, continuous and heterogeneous *k*-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, *11*(2), 195–212. https://doi.org/10.1007/s10618-005-0007-5

Domingo-Ferrer, J., Mateo-Sanz, J. M. i Torra, V. (2001). Comparing SDC methods for microdata on the basis of information loss and disclosure risk. W: *Pre-proceedings of ETK-NTTS 2001 conference Hersonissos (Crete) 18-22 june 2001*, vol. 2 (s. 807–826).

ISIS, EUROSTAT. https://crises-deim.urv.cat/web/docs/publications/conferences/597.pdf

Domingo-Ferrer, J., Sebé, F. i Castellà-Roca, J. (2004). On the security of noise addition for privacy in statistical databases. W: J. Domingo-Ferrer i V. Torra (Eds.), *Privacy in Statistical Databases. PSD 2004*. Lecture Notes in Computer Science, vol. 3050 (s. 149–161). Springer. https://doi.org/10.1007/978-3-540-25955-8_12

Dove, I., Blanchard, S. i Spicer, K. (2017). *Applying cell-key perturbation to 2021 census output*. The Government Statistical Service, UK. https://gss.civilservice.gov.uk/wp-content/uploads/2017/01/ExN-Disclosure-control-methodology-in-2021-Census-outputs-Spicer-Blanchard-Dove-ONS.docx.

Dove, I., Ntoumos, C. i Spicer, K. (2018). Protecting census 2021 origin-destination data using a combination of cell-key perturbation and suppression. W: J. Domingo-Ferrer i F. Montes (Eds.), *UNESCO Chair in Data Privacy, International Conference, PSD 2018. Valencia, Spain, September 26-28, 2018* (s. 43–55). Springer.

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control. Theory and implementation*. Lecture Notes in Statistics, vol. 201. Springer. https://doi.org/10.1007/978-1-4614-0326-5

Drechsler, J. i Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics* & *Data Analysis*, *55*(12), 3232–3243. https://doi.org/10.1016/j.csda.2011.06.006

Duncan, G. T., Elliot, M. i Salazar-González, J. J. (2011), *Statistical confidentiality. Principles and practice*. Springer.

Duncan, G., Keller-McNulty, S. i Stokes, S. (2001). *Disclosure risk vs. data utility: The r-u confidentiality map*. Technical Report LA-UR-01-6428, Los Alamos National Laboratory, Statistical Sciences Group, Los Alamos, New Mexico. https://www.niss.org/sites/default/files/technicalreports/tr121.pdf

Duncan, G. T. i Roehrig, S. F. (2007). Reconciling information privacy and information access in a globalized technology society. W: G. D. Garson (Ed.), *Modern public information technology systems: Issues and challenges* (s. 72–94). IGI Global.

Dwork, C. i Smith, A. (2009). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, *1*(2), 135–154. https://doi.org/10.29012/jpc.v1i2.570

El Emam, K., Mosquera, L. i Hoptroff, R. (2020). *Practical synthetic data generation. Balancing privacy and the broad availability of data*. O'Reilly Media.

Eurostat. (2011). Europejski kodeks praktyk statystycznych dla krajowych i wspólnotowych organów statystycznych. Przyjęty przez Komitet ds. Europejskiego Systemu Statystycznego 28 sierpnia 2011 r. http://ec.europa.eu/eurostat/documents/3859598/5922337/10425-PL-PL.PDF

Eurostat i FRIBS Task Force (2017), *European business statistics manual – Contents and introduction*. Statistical Office of the European Union and the Framework Regulation Integrating Business Statistics (FRIBS) Task Force. http://ec.europa.eu/eurostat/documents/54610/7779382/EBS-manual-table-of-contents-and-introduction.pdf

Fischetti, M. i Salazar-González, J. J. (2003). Partial cell suppression: A new methodology for statistical disclosure control. *Statistics and Computing*, *13*(1), 13–21. https://doi.org/10.1023/A:1021927424942

Gatnar, E. i Walesiak, M. (red.). (2009). *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN.

Gatnar, E. i Walesiak, M. (red.). (2011). *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*. Wydawnictwo C.H. Beck.

Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG). http://www.gesetze-im-internet.de/bstatg_1987/BStatG.pdf

Giessing, S. (2009). *Techniques for using τ-ARGUS modular on sets of linked tables*. Joint UNECE/Eurostat work session on statistical data confidentiality (Bilbao, Spain, 2–4 December 2009). https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.35.e.pdf

Giessing, S. (2016). Computational issues in the design of transition probabilities and disclosure risk estimation for additive noise. W: J. Domingo-Ferrer i M. Pejic-Bach (Eds.), *Privacy in Statistical Databases. PSD 2016*. Lecture Notes in Computer Science, vol. 9867 (s. 237–251). Springer. https://doi.org/10.1007/978-3-319-45381-1_18

Gjaltema, T., Burnett-Isaacs, K., Raab, G., Sallier, K., Task, C., Chang, J. X., Gauvin, H., Girard, C., Kaloskampis, I., Ramsden, A., Rodriguez, R., Slokom, M. i Thomas S. (2022). *Synthetic data for National Statistical Organization: A starter guide*. United Nations, Economic Commission for Europe (UNECE). https://statswiki.unece.org/display/SDS/Synthetic+Data+Sets+public

Government Statistical Service [GSS]. (2009). *National Statistician's Guidance: Confidentiality of Official Statistics.* Office for National Statistics, UK. https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Confidentiality-of-Official-Statistics-National-Statisticians-Guidance.pdf.

Government Statistical Service [GSS]. (2014). *GSS/GSR disclosure control guidance for microdata produced from social surveys.*

Gregory, A. (2011), *The Data Documentation Initiative (DDI): An introduction for National Statistical Institutes*. Open Data Foundation. https://www.semanticscholar.org/paper/The-Data-Documentation-Initiative-(-DDI-)-%3A-An-for-Gregory/7306a239bdc9f285c8a3001b67e2a56b5bacb5c3

Gregory, A. i Heus, P. (2007). *DDI and SDMX: Complementary, Not Competing Standards*. Open Data Foundation. http://www.opendatafoundation.org/papers/DDI_and_SDMX.pdf

Hansen, S. L. i Mukherjee, S. (2003). A polynomial algorithm for optimal univariate microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, *15*(4), 1043–1044.

Hochfellner, D., Müller, D., Schmucker, A. i Roß, E. (2012). *Data protection at the Research Data Centre. FDZ-Methodenreport 6/2012*. Research Data Centre (FDZ) of the German Federal Employment Agency at the Institute for Employment Research.

Höninger, J., Pattloch, D. i Voshage, R. (2010). *On-site access to micro data: Preserving the treasure, preventing disclosure*. State Statistical Institute Berlin-Brandenburg, Research Data Centre.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nordholt, E., Seri, G. i De Wolf, P.-P. (2006), *Handbook on statistical disclosure control*, Version 1.0. CENEX SDC, Eurostat. https://ec.europa.eu/eurostat/cros/system/files/CENEX-SDC_handbook.pdf

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Schulte Nordholt, E., Seri, G. i De Wolf, P.-P. (2010). *Handbook on statistical disclosure control*. ESSNet SDC. https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., De Wolf, P.-P. (2012). *Statistical disclosure control*. John Wiley & Sons.

ISI Declaration on Professional Ethics, adopted by the ISI Council 22 & 23 July 2010, Reykjavik, Iceland. International Statistical Institute – Permanent Office, The Hague, The Netherlands. https://isi-web.org/images/about/Declaration-EN2010.pdf

Jarosz, M. (2013). Przekazywanie niepomyślnych informacji w praktyce klinicznej. *Onkologia w Praktyce Klinicznej*, *9*(6), 225–229.

Jolliffe, I. T. (2002), *Principal component analysis*, 2nd ed. Springer.

Jordon, J., Yoon, J. i van der Schaar, M. (2019). PATE–GAN: Generating synthetic data with differential privacy guarantees. W: *International Conference on Learning Representations, New Orleans, Louisiana, United States, May 6–May 9, 2019*. https://openreview.net/forum?id=S1zk9iRqF7

Kim, H. J., Drechsler, J. i Thompson, K. J. (2021). Synthetic microdata for establishment surveys under informative sampling. *Journal of the Royal Statistical Society Series A – Statistics in Society*, *184*(1), 255–281. https://doi.org/10.1111/rssa.12622

Komisja Europejska. (b.d.). Które dane personalne uznawane są za wrażliwe?. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_pl

Li, N., Li, T. i Venkatasubramanian, S. (2007). *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. W: *2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey.*

Litwiński, P. (2009). *Ochrona danych osobowych w ogólnym postępowaniu administracyjnym*. Wolters Kluwer Polska.

Machanavajjhala, A., Gehrke, J., Kifer, D. i Venkitasubramaniam. M. (2006). *l*-diversity: Privacy beyond *k*-anonymity. W: *Proceedings of the 22nd International Conference on Data Engineering* (ICDE'06), *Atlanta, GA, USA. 2006*.

Marley, J. i Leaver, V. (2011). A method for confidentialising user-defined tables: Statistical properties and a risk-utility analysis. W: *International Statistical Institute proceedings of the 58th World Statistics Congress 2011, Dublin* (s. 21–26). http://2011.isiproceedings.org/papers/450007.pdf

Mateo-Sanz, J. M. i Domingo-Ferrer, J. (1998). A comparative study of microaggregation methods. *Qüestiió*, *22*(3), 511–526. https://upcommons.upc.edu/bitstream/handle/2099/4090/article.pdf

Meindl, B. (2011). *A computational framework to protect tabular data – R–package sdcTable*. Joint UNECE/EUROSTAT work session on statistical data confidentiality (Tarragona, Spain, 26–28 October 2011). https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/10_Meindl.pdf

Meindl, B. i Enderle, T. (2019). *cellKey – consistent perturbation of statistical tables*. Joint UNECE/Eurostat work session on statistical data confidentiality (Conference of European Statisticians, 29-31 October 2019, the Hague, the Netherlands). https://

unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S7_Austria_and_Germany_cellKey_Meindl_AD.pdf
Minami, K. i Abe, Y. (2017). Statistical disclosure control for tabular data in R. *Romanian Statistical Review*, *65*(4), 67–76.
Mitchell, J. E. (2002). Branch-and-cut algorithms for combinatorial optimization problems. W: *Handbook of Applied Optimization* (s. 65–77). Oxford University Press.
Mivule, K. (2012). Utilizing Noise Addition for Data Privacy, an Overview. W: *Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2012)* (s. 65–71). https://doi.org/10.48550/arXiv.1309.3958
Młodak, A. (2006). *Analiza taksonomiczna w statystyce regionalnej*. Difin.
Młodak, A. (2019). Wykorzystanie miernika kompleksowego w ocenie straty informacji na skutek kontroli ujawniania mikrodanych. *Przegląd Statystyczny*, *66*(1), 7–26.
Młodak, A. (2020). Strata informacji wskutek przeprowadzenia kontroli ujawniania danych wynikowych. *Wiadomości Statystyczne. The Polish Statistician*, *65*(9), 7–27.
Młodak, A., Pietrzak, M. i Józefowski, T. (2022). The trade-off between the risk of disclosure and data utility in SDC: A case of data from a survey of accidents at work. *Statistical Journal of the IAOS*, *38*, 1503–1511.
Nayak, T. K. i Adeshiyan, S. A. (2016). On invariant post-randomization for statistical disclosure control. *International Statistical Review / Revue Internationale de Statistique, 84*(1), 26–42. http://www.jstor.org/stable/44162459
NFJP. (b.d.). Multyplikatywny. W: Narodowy Fotokorpus Języka Polskiego. https://nfjp.pl/lemma/multyplikatywny
Nicolaas, G. (2011). *Survey paradata: A review*. ESRC National Centre for Research Methods Review paper. https://eprints.ncrm.ac.uk/id/eprint/1719/1/Nicolaas_review_paper_jan11.pdf
Nowok, B., Raab, G. i Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, *74*(11), 1–26. https://doi.org/10.18637/jss.v074.i11
Nyczaj, K. (2010). Rejestry administracyjne jako źródło wiedzy statystycznej. *Wiadomości Statystyczne*. *The Polish Statistician*, *55*(11), 9–20.
OECD. (2008). Recommendation of the Council for enhanced access and more effective use of public sector information. https://www.oecd.org/sti/40826024.pdf
OECD. (2014). *OECD Expert Group for International Collaboration on Microdata Access. Final Report*. https://www.oecd.org/sdd/microdata-access-final-report-OECD-2014.pdf
OECD. (2015). *Recommendation of the OECD Council on good statistical practice*. https://www.oecd.org/statistics/good-practice-toolkit/Brochure-Good-Stat-Practices.pdf
OECD. (2019). *Enhancing access to and sharing of data: Reconciling risks and benefits for data re-use across societies*. OECD Publishing. https://doi.org/10.1787/276aaca8-en
OECD. (2021). *Recommendation of the Council concerning access to research data from public funding*. https://www.oecd.org/sti/recommendation-access-to-research-data-from-public-funding.html
Oleński, J. (2005). *Rejestry administracyjne i systemy katastralne w infrastrukturze informacyjnej państwa*. Wydział Nauk Ekonomicznych, Uniwersytet Warszawski.
Oleński, J. (2006). *Infrastruktura informacyjna państwa w globalnej gospodarce*. Wydział Nauk Ekonomicznych, Uniwersytet Warszawski.

Panek, T. i Zwierzchowski, J. (2013). *Statystyczne metody wielowymiarowej analizy porównawczej. Teoria i zastosowania*. Oficyna Wydawnicza Szkoły Głównej Handlowej w Warszawie.

Pietrzak, M., Józefowski, T., Klimanek, T. i Młodak A. (2022). An optimized selection of statistical disclosure control methods – A case study involving microdata from the Polish survey of accidents at work. W: K. Jajuga, G. Dehnel i M. Walesiak (Eds.), *Modern classification and data analysis. SKAD 2021. Studies in classification, data analysis, and knowledge organization* (s. 63–78). Springer.

Poljičak, M. i Stančić, H. (2014). Statistical microdata – Production of safe datasets, transparent presentation of contents and advanced services for users through metadata authorization system. W: *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2014* (s. 1527–1532). https://doi.org/10.1109/MIPRO.2014.6859808

PWN. (1968). Multyplikator. W: S. Skorupka, H. Auderska i Z. Łempicka (red.), *Mały słownika języka polskiego*. PWN.

Reiss, S. P. (1984). Practical data-swapping: The first steps. *ACM Transactions on Database Systems*, *9*(1), 20–37. https://doi.org/10.1145/348.349

Reiss, S. P., Post, M. J. i Dalenius, T. (1982). Non-reversible privacy transformations. W: *Proceedings of the ACM Symposium on Principles of Database Systems* (s. 139–146), Association for Computing Machinery (ACM).

Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, *21*(3), 441–462.

Ritchie, F. (2007). *Disclosure detection in research environments in practice*. W: *Work session on statistical data confidentiality, Manchester, 17–19 December 2007* (s. 399–406). Eurostat. Methodologies and Working papers. Office for Official Publications of the European Communities.

Ritchie, F. i Elliot, M. (2015). *Principles- versus rules-based output statistical disclosure control in remote access environments.* University of the West of England, Economics Working Paper Series, 1501. https://www2.uwe.ac.uk/faculties/BBS/BUS/Research/Economics%20Papers%202015/1501.pdf

RODO. (2016). Rozporządzenie 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) (Dz. Urz. UE L. 119/1).

Rozporządzenie Komisji (UE) nr 1151/2010 z dnia 8 grudnia 2010 r. w sprawie wykonania Rozporządzenia Parlamentu Europejskiego i Rady (WE) nr 763/2008 w sprawie spisów powszechnych ludności i mieszkań w odniesieniu do ustaleń dotyczących raportów jakości i ich struktury oraz formatu technicznego przekazywania danych (Dz. Urz. UE L 324/1). http://eur-lex.europa.eu/legal-content/PL/TXT/PDF/?uri=CELEX:32010R1151&rid=1

Rozporządzenie Komisji (UE) nr 557/2013 z dnia 17 czerwca 2013 r. w sprawie wykonania rozporządzenia (WE) nr 223/2009 Parlamentu Europejskiego i Rady w sprawie europejskiej statystyki w zakresie dostępu do poufnych danych do celów naukowych

i uchylające rozporządzenie Komisji (WE) nr 831/2002 (Dz. Urz. UE L 164/16). http://eur-lex.europa.eu/legal-content/PL/TXT/PDF/?uri=CELEX:32013R0557&rid=11

Rozporządzenie Parlamentu Europejskiego i Rady (WE) nr 223/2009 z dnia 11 marca 2009 r. w sprawie statystyki europejskiej oraz uchylające rozporządzenie Parlamentu Europejskiego i Rady (WE, Euratom) nr 1101/2008 w sprawie przekazywania do Urzędu Statystycznego Wspólnot Europejskich danych statystycznych objętych zasadą poufności, rozporządzenie Rady (WE) – nr 322/97 w sprawie statystyk Wspólnoty oraz decyzję Rady 89/382/EWG, Euratom w sprawie ustanowienia Komitetu ds. Programów Statystycznych Wspólnot Europejskich (Dz. Urz. UE L 87/164). http://eur-lex.europa.eu/legal-content/PL/TXT/PDF/?uri=CELEX:32009R0223&from=EN

Rozporządzenie Wykonawcze Komisji (UE) 2017/881 z dnia 23 maja 2017 r. w sprawie wykonania rozporządzenia Parlamentu Europejskiego i Rady (WE) nr 763/2008 w sprawie spisów powszechnych ludności i mieszkań w odniesieniu do ustaleń dotyczących raportów jakości i ich struktury oraz formatu technicznego przekazywania danych, zmieniające rozporządzenie (UE) nr 1151/2010 (Dz. Urz. UE L 135/6). http://eur-lex.europa.eu/legal-content/PL/TXT/PDF/?uri=CELEX:32017R0881&qid=1516800818760&from=PL

Sharma, N. i Gaud, N. (2015). K-modes clustering algorithm for categorical data. *International Journal of Computer Applications*, *127*(17), 1–6. https://doi.org/10.5120/ijca2015906708

Shlomo, N. i Young, C. (2006). Information loss measures for frequency tables. W: *Monographs of official statistics. Work session on statistical data confidentiality, Geneva, 9–11 November 2005* (s. 277–289). Office for Official Publications of the European Communities.

Signore, M., Scanu, M., & Brancato, G. (2015). Statistical metadata: a unified approach to management and dissemination. *Journal of Official Statistics*, 31(2), 325-347.

Singh, A. C., Yu, F. i Dunteman, G. H. (2004). MASSC: A new data mask for limiting statistical information loss and disclosure. W: P. D. Muñoz, J. Riecan (Eds.), *Monographs of official statistics. Work session on statistical data confidentiality, Luxembourg, 7 to 9 April 2003*. Part 3 (s. 373–394). Eurostat.

SJP. (b.d.). Multiplikatywny. W: *Słownik języka polskiego*. https://sjp.pl/multiplikatywny

Snoke, J., Raab, G. M., Nowok, B., Dibben, C. i Slavković A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society, Series A – Statistics in Society*, *181*(3), 663–688. https://doi.org/10.48550/arXiv.1604.06651

Solanas, A. i Martinez-Balleste, A. (2006). V-MDAV: A multivariate microaggregation with variable group size. W: *17th COMPSTAT Symposium of the IASC (International Association of Statistical Computing), Rome* (s. 917–925). http://vneumann.etse.urv.cat/webCrises/publications/bcpi/compstat06_solanas.pdf

Spicer, K., Tudor, C. i Cornish, G. (2014). *Intruder testing: Demonstrating practical evidence of disclosure protection in 2011 UK Census*. Joint UNECE/Eurostat work session on statistical data confidentiality (Ottawa, Canada, 28-30 October 2013). https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_5_Spicer.pdf

Statistics Act (R.S.C., 1985, c. S-19). http://laws-lois.justice.gc.ca/eng/acts/S-19/FullText.html

Statistics Canada. (2017). *Directive on microdata linkag*e. https://www.statcan.gc.ca/eng/record/policy4-1.

Statistics Netherlands Act – Act of 20 November 2003 enacting a Law governing Statistics Netherlands, Bulletin of Acts, Orders and Decrees of the Kingdom of the Netherlands 2003, 55.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. http://www.jstor.org/stable/1671815

Sullivan, G. R. (1989). *The use of added error to avoid disclosure in microdata releases. Retrospective Theses and Dissertations*. 9186. https://doi.org/10.31274/rtd-180813-9036

Templ, M. (2008). Statistical disclosure control for microdata using the R–package sdcMicro. *Transactions on Data Privacy*, 1, 67–85.

Templ, M. (2017). *Statistical disclosure control for microdata. Methods and applications in R*. Springer International Publishing.

Templ, M., Kowarik, A. i Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, *67*(4), 1–36. https://doi.org/10.18637/jss.v067.i04

Templ, M., Meindl, B. i Kowarik, A. (2021). *Introduction to statistical disclosure control (SDC).* International Household Survey Network, Data-Analysis OG.

Templ, M., Meindl, B., Kowarik, A. i Chen, S. (2014). *Introduction to Statistical Disclosure Control (SDC).* IHSN Working Paper, 007. http://www.ihsn.org/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf

The Statistics and Registration Service Act 2007. https://www.legislation.gov.uk/ukpga/2007/18/pdfs/ukpga_20070018_en.pdf

The World Bank. (b.d.a). Access to information. www.worldbank.org/en/access-to-information

The World Bank. (b.d.b). World Bank Group information research guide. https://researchguides.worldbankimflib.org/c.php?g=924038&p=6659878

The World Bank. (2018). *Managing personal data responsibly: The World Bank Group personal data policy.* http://documents.worldbank.org/curated/en/466121527794054484/pdf/Privacy-Board-Paper-050318-vF-05042018.pdf

Thomas, W., Gregory, A. i Hamilton, A. (2011). *Metadata standards to support controlled access to microdata.* Joint UNECE/Eurostat work session on statistical data confidentiality (Tarragona, Spain, 26-28 October 2011).

Thompson, G., Broadfoot, S. i Elazar, D. (2013). *Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics*. Joint UNECE/Eurostat work session on statistical data confidentiality (Ottawa, Canada, 28–30 October 2013). https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf

UNECE. (2007). *Managing statistical confidentiality & microdata access. Principles and guidelines of good practice.* United Nations. https://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf.

UNECE. (2009). *Principles and guidelines on confidentiality aspects of data integration undertaken for statistical or related research purposes.* United Nations. https://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf

UNECE. (2016). *Generic law on official statistics for Eastern Europe, Caucasus and Central Asia*. United Nations. https://www.unece.org/fileadmin/DAM/stats/publications/2016/ECECESSTAT20163_E.pdf

UNECE. (2018a). *Guidance on modernizing statistical legislation*. United Nations. http://www.unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20183.pdf

UNECE. (2018b). *Guidelines on the use of registers and administrative data for population and housing censuses*. United Nations. https://unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20184.pdf

UNECE. (2021). *Expert Meeting on Statistical Data Confidentiality. 01-03 December 2021. Poznań. Poland*. https://unece.org/statistics/events/SDC2021

UNECE. (2023). Upcoming Statistics Meetings and Events. https://unece.org/info/events/unece-meetings-and-events/statistics

United Nations General Assembly. (2014). Resolution adopted by the General Assembly on 29 January 2014 [without reference to a Main Committee (A/68/L.36 and Add.1)]. 68/261. Fundamental principles of official statistics. https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf

Ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej (t.j. Dz. U. z 2022 r. poz. 459 z późn. zm.)

Ustawa z dnia 2 lipca 2004 r. o swobodzie działalności gospodarczej (Dz. U. z 2004 r. Nr 173 poz. 1807 z późn. zm.).

Ustawa z dnia 10 maja 2018 r. o ochronie danych osobowych (Dz. U. z 2018 r. poz. 1000).

Vale, S. (2010). *Exploring the relationship between DDI, SDMX and the Generic Statistical Business Process Model*. DDI Working Paper. http://dx.doi.org/10.3886/DDIOtherTopics01

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Willenborg, L. (2012) Discussion [of Skinner, C.: Statistical Disclosure Risk: Separating Potential and Harm]. *International Statistical Review*, *80*(3), 375–378. https://doi.org/10.1111/j.1751-5823.2012.00188.x

Willenborg, L. i de Waal, T. (1996). *Statistical disclosure control in practice*. Lecture notes in Statistics, vol. 111. Springer. https://doi.org/10.1007/978-1-4612-4028-0

Willenborg, L. i de Waal, T. (2001). *Elements of statistical disclosure control*. Lecture Notes in Statistics, vol 155. Springer. https://doi.org/10.1007/978-1-4613-0121-9

Woo, Y. M. J. i Slavković, A. (2014). Generalized linear models with variables subject to post randomization method. *Statistica Applicata – Italian Journal of Applied Statistics, 24*(1), 29–56.

Spis rysunków

Spis tablic

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

Confidentiality vs. utility of statistical information. Dilemmas of statistical disclosure control

Summary*

https://doi.org/10.18559/978-83-8211-168-2-summary

Introduction

Comprehensive and precise statistical information is essential in order to plan and carry out development activities in different spheres of the socio-economic reality, to monitor their effects, conduct new studies and improve research tools. These needs are what drives the growing demand for statistical data regarding various aspects of the surrounding reality and enabling data users to analyse underlying phenomena of interest.

Data collected in statistical surveys or maintained in administrative registers contain a wide range of useful information about various characteristics of units, including their direct identifiers. These characteristics are protected by law and are subject to statistical confidentiality. The protection of personal data has become socially relevant since the introduction of Regulation (EE) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). The GDPR has prompted the adoption of appropriate national laws. This is why, before statistical outputs can be released, data holders are obliged to analyse them in order to minimise the risk of disclosing sensitive information that could be used by end users to re-identify respondents. These procedures are

* More in the monograph in Polish: Młodak, A., Pietrzak, M., Klimanek, T., Józefowski, T. and Lańduch, P. (2023). *Poufność a użyteczność informacji statystycznych. Dylematy ochrony udostępnianych danych*. Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu. https://doi.org/10.18559/978-83-8211-168-2

known as statistical disclosure control (SDC). The first and simplest step undertaken during SDC is to remove key identifiers (such as a person's name, surname and personal identification number or a company registration number and its tax registration number).

Nowadays, however, such simple solutions, which have been used for many years, are far from sufficient. As the number of variables describing statistical units increases, the resulting number of possible combinations of categories of nominal or ordinal variables grows very quickly. Therefore, there is a considerable likelihood of the existence of unique combinations, which, in extreme cases, may contain single units that can be re-identified. This risk is particularly likely to occur in microdata (i.e., anonymised unit-level data) or multi-dimensional arrays of data, known as OLAP cubes; however, these are precisely the types of data which are increasingly sought after, especially by scientists and researchers. The risk of disclosure or re-identification can be ever greater if a survey or an administrative register contains information representing continuous variables measured on an interval or ratio scale. In such cases, it is possible to calculate how much particular groups contribute to total values (e.g., the share of employees in the public sector in the total number of employees). One should also consider the possibility that a potential end user might have access to other datasets, which could help them to identify individual units. Moreover, under current regulations, contact information of businesses (including sole proprietors) is publicly accessible, which means that their sensitive data need to be protected in a different way. Consequently, a number of advanced SDC methods have been developed (noise addition, post-randomisation, controlled tabular adjustment, etc.), which are based on mathematical statistics and IT solutions and have become an important methodological tool for official statistics.

SDC methods – description and discussion

Specific methods of statistical disclosure control have been developed for three main types of data:

- microdata,
- tabular data,
- output checking.

The monograph focuses on each of these three applications by describing specific characteristics of each type of data, their sources and fundamental principles of their protection, theoretical and IT tools used for this purpose as well as organisational and technological aspects of releasing statistical data, which are associated with the risk of unit re-identification or disclosure of sensitive information. In addition, it also describes ways of protecting confidentiality that can be applied to outputs of statistical analyses, such as descriptive statistics, estimates

of econometric models or charts. The authors have attempted to present the complexity of SDC and various rules of estimating re-identification risk. A comprehensive assessment of the effects of applying SDC should include an estimate of disclosure risk and the expected information loss due to the suppression or perturbation of sensitive information. The main purpose of SDC is to achieve an optimal trade-off between these two minimisation goals.

An optimal level of confidentiality regarding data collected by NSIs as well as those maintained by other holders can be ensured by employing appropriate SDC methods and effective measures of disclosure risk and information loss resulting from the application of SDC procedures. The choice of methods and their parameters depends on many different factors, such as laws and regulations, type and scope of data to be released in a given survey, organisation of data release or IT tools used for this purpose.

The monograph is an attempt to provide a comprehensive description of all of these aspects, including the goal and definition of statistical disclosure control, its formal and legal principles, particularly current regulations and international recommendations regarding data confidentiality, types of released data (including metadata, i.e., information about definitions of concepts that characterise units, reference periods of data collection and units for which data are collected, measurement methods, possible exceptions to rules of determining variables or missing data, explanations of known causes of deviations or gaps; paradata, i.e., information about the process by which the data were collected, which can be used to explain interpretations of the resulting characteristics of survey units as well as data not included in the survey itself but automatically collected during survey administration, such as the number of times a given respondent has visited the webpage with the survey form, the number of data changes made, the time taken to complete the form, etc. as well as other data). We explain the disclosure process, types of data users from the perspective of SDC, typologies of statistical outputs with respect to the protection of sensitive information and the trade-off between disclosure risk and data utility. One section of the monograph contains an overview of the most important SDC solutions and regulations implemented in different European countries and another one provides a description of the main kinds of microdata and the role of metadata and paradata in the SDC process.

The first of the two main factors that determine the effectiveness of the SDC process is the extent to which the risk of disclosure is minimised, i.e., the probability of identifying a particular statistical unit (its identity) or obtaining new, previously unknown information about it (its attributes) from data released by a national statistical institute or another data holder. Disclosure risk is measured by analysing potential disclosure scenarios in specific settings and considering the following commonly used concepts:

- uniqueness of combinations of quasi-identifiers (keys) that can be used to identify units that are at risk of disclosure – for categorical variables; these are typically *a priori* measures, which are calculated before the application of (non-perturbative or perturbative) masking methods at the implementation stage; according to the uniqueness approach, disclosure risk is measured by applying some basic rules: *k*-anonymity, *l*-diversity or *t*-closeness, which can be combined, in the case of sample surveys, with additional models accounting for the possibility of disclosing sensitive data in the original datasets as well as in the outputs (population estimates created by applying appropriate sampling weights),
- uniqueness of values in the neighbourhood of original values – for continuous variables; in this case *a posteriori* risk measures are used, i.e., those that are calculated by comparing microdata before and after the application of SDC methods.

Disclosure risk can be measured for each record separately and used to provide protection for selected records deemed to be at risk of disclosure; an alternative approach consists in using characteristics of such at-risk records to create a general definition of disclosure risk for the entire dataset. Therefore, the following levels of disclosure risk can be distinguished from the perspective of SDC:

- individual – for a single microdata record,
- global – for a whole microdata set,
- associated with a hierarchical structure of data – the impact of the hierarchical structure of the microdata on disclosure risk at a given level of the hierarchy is measured for a single record or a whole dataset.

Two types of risk are considered in disclosure scenarios:

- internal risk – when a unit can potentially be re-identified only on the basis of data made available to the user;
- external risk – when the user has access to other sources of data, which can be linked with the released dataset; this risk is much harder to measure because there is usually little or no information about other data sources available to the user, except for what can be inferred from the user's place of employment (e.g., if the user works at a labour office, they are likely to have access to the register of unemployed people, which can be linked with microdata from the LFS survey).

Methods of measuring disclosure risk for tabular data differ from those used for microdata or for output checking. In the latter case, disclosure risk is measured in the safe environment where accredited researchers are granted access to confidential microdata and IT tools and where results of their analyses are verified to ensure they are non-disclosive. To facilitate output checking, all output can be classified into a limited number of categories based on its functional form (tables, regression, etc.) and not on the scope of information contained in the data. Each

category is then labelled 'safe' or 'unsafe'. The fact that a particular output has been labelled 'unsafe' does not mean it will not be cleared for release. Similarly, outputs classified as safe will not necessarily be released outside the safe environment. Except certain well defined and limited cases, outputs classified should be released to researchers with no or minimal changes. Outputs are classified as unsafe in their present form if the likelihood of disclosure is assessed to be high, which means they cannot be released without substantial changes. In this case, the risk of disclosure can be assessed in two ways: either according to the rule-of-thumb model, which focuses on preventing confidentiality errors, or by choosing the principles-based model, where the goal is to minimise the risk of disclosure and to maximise data utility. It should be pointed out that values of risk measures should also be confidential and available only to those responsible for SDC.

The monograph presents the most important SDC methods for microdata and tabular data, which can be divided into two main groups:

- Non–perturbative masking – a family of SDC methods that employ various ways of suppressing sensitive information. As a result, the released dataset does not contain information about certain units or the amount of detail is reduced.
- Perturbative masking – a family of SDC methods consisting in distorting sensitive information in order to prevent its reconstruction by unauthorised users while minimising the loss of information.

These methods can be quite simple (e.g., anonymisation or local suppression) or more sophisticated, including methods based on advanced algorithms and mathematical tools (e.g., microaggregation, noise addition, rank swapping, targeted record swapping, controlled tabular adjustment or the cell key method). All of these theoretical tools are discussed in detail. In addition, there is a comparison of basic properties of the available methods for different types of data. The monograph also includes a presentation of SDC methods for data released in statistical publications, especially descriptive statistics, results of various analyses, charts and choropleth maps. A separate section is devoted to methods of and dilemmas associated with synthetic data generation. It should be emphasised that there is no universally accepted hierarchy of SDC methods in terms of the risk of disclosure and information loss – their usefulness depends on specific data and selected parameters.

Statistical disclosure control is inherently associated with the introduction of uncertainty regarding the values of released variables. This uncertainty is the result of suppressing or changing values in a microdata record or a table cell. Consequently, the application of SDC leads to a loss of information contained in the original data, which can negatively affect the quality of released information or calculations and estimates produced by data users. This is why, in addition to data, users should be informed about the expected information loss resulting from the application of SDC. As pointed out in the introduction, minimisation

of this loss is the second optimisation criterion of the SDC process, apart from the minimisation of the risk of unit re-identification and disclosure of sensitive information. This means that the utility of information in microdata, tabular data and analytic outputs released to users should be as close as possible to that of the original data. This requires effective methods of measuring information loss and ways of minimising it. The monograph explains the very concept of information loss and the most important categories of measures of information loss:

- Measures of changes in variable distributions, based on distance metrics between original and modified values. For example, for each geographical unit in the dataset, the distance between original and modified values is calculated and then the distances are averaged.
- Measures of impact of SDC on the variance of estimates, which account for the difference between variances for average values of certain data subsets or the entire dataset (in the cases of tabular data – columns, rows or the whole table) before and after the SDC process. Another approach is to conduct ANOVA for a selected dependent variable with respect to selected independent categorical variables. In this case, information loss is measured by comparing changes in the components of R^2 (by decomposing total variance into within-group and between-group variance) for the original dataset/table based on original data and the dataset/table modified as a result of applying SDC. The application of SDC can lead to heteroscedasticity, in other words, can cause between-group variance to decrease and within-group variance to increase or vice versa.
- Measures of impact of SDC on the strength of relationships, which involves comparing the direction and strength of relationships between certain phenomena with those observed in the original data. In this case, information loss can be measured by calculating correlation coefficients or conducting tests of independence between corresponding variables in selected breakdowns, in other words, for a specific contingency table. Other approaches can also be used.

The monograph provides numerous examples showing how to construct such measures, how to interpret them, which demonstrate their usefulness for different use cases of released data, including measures of estimation precision.

Nowadays, data collection, processing and analysis cannot be performed without the help of appropriate IT tools. This is also true in the case of statistical disclosure control. Although the development of this branch of statistics is only now starting to accelerate, a number of useful programming tools have already been created to facilitate the SDC process involving digital sets of numerical or symbolic data. The monograph presents an overview of the most commonly used IT tools for SDC. The section starts with the presentation of two, probably most well-known, open source programmes: τ-Argus and μ-Argus. Both were developed by

Statistics Netherlands (Centraal Bureau voor de Statistiek) in the course of a few European projects. They are Java-based programmes, available with and without a bundled JRE7 distribution. Other SDC tools have been implemented in the R software and include a number of dedicated packages, such as sdcTable, sdcMicro, recordSwapping, cellKey, etc. The section ends with a brief description of possibilities offered by other SDC tools.

The protection of data confidentiality is also associated with organisational problems related to releasing official statistics and the need to check them to prevent a disclosure of sensitive information. The monograph presents arguments in favour of releasing statistical outputs, including appropriately prepared unit-level data for scientific research purposes (Scientific Use Files), different ways in which access to such files can be granted as well as specific requirements associated with such forms of release. These include formal requirements that persons or institutions applying for access to microdata must satisfy, as well as requirements regarding the level of protection that the data administrator (typically a NSI, but this can be any data holder) should guarantee in order to prevent unauthorised persons from gaining access to sensitive information.

Structure of the monograph

The monograph consists of six chapters. The first one presents general concepts of SDC and relevant definitions, as well as regulations concerning the protection of sensitive information that are in effect in different countries, including Poland. The chapter contains a description of the main types of data that are released and the significance of metadata, paradata, and additional data from the perspective of SDC. The second chapter is devoted to the risk of disclosure and its measurement. It highlights differences between microdata and aggregated data in frequency and magnitude tables or outputs of statistical analyses. The third chapter contains a detailed description of various SDC methods and techniques that can be applied to the three categories of data mentioned above. It presents potential threats to data confidentiality associated with already published descriptive statistics, charts or outputs of statistical analyses and identifies ways in which such data can be used to reconstruct sensitive information. Aspects relating to information loss are discussed in the fourth chapter, which contains the definition of the problem and the main measures of information loss, including original measures developed by the authors. The chapter also analyses the impact of information loss on the quality of estimates produced from data treated with SDC techniques. The fifth chapter provides a detailed discussion of IT tools for SDC, with emphasis on τ-Argus, μ-Argus and two R packages: `sdcTable` and `sdcMicro`.

The sixth chapter focuses on the organisational aspects of providing access to data and the principles that should be followed in this regard. In particular, this

chapter describes different types of microdata that can be released, principles that should be followed in research data centres (RDC) and appropriate security measures, the process of access control and the scope of authorisation. The manograph ends with a conclusion summarising the main principles and recommendations regarding the application of SDC. There is also a glossary of key terms used in the monograph.

Given the general relevance of statistical disclosure control, the monograph is intended to serve as a practical reference guide for methodologists designing statistical surveys and persons responsible for survey quality and the confidentiality of collected information. For this reason, it contains numerous examples and practical discussions of particular aspects that should make the content more accessible.

Keywords: non-perturbative methods, perturbative methods, disclosure risk, information loss, microdata, tabular data, statistical outputs.

Translated by Grzegorz Grygiel

Informacja o autorach

Dr hab. **Andrzej Młodak** – konsultant w Ośrodku Statystyki Małych Obszarów Urzędu Statystycznego w Poznaniu, profesor Akademii Kaliskiej im. Prezydenta Stanisława Wojciechowskiego, Międzywydziałowy Zakład Matematyki i Statystyki.

Mgr **Michał Pietrzak** – absolwent Uniwersytetu Ekonomicznego w Poznaniu, były pracownik Ośrodka Statystyki Małych Obszarów Urzędu Statystycznego w Poznaniu, specjalista z zakresu kontroli ujawniania danych.

Dr hab. **Tomasz Klimanek** – profesor Uniwersytetu Ekonomicznego w Poznaniu, Katedra Statystyki; z-ca Dyrektora Urzędu Statystycznego w Poznaniu.

Dr **Tomasz Józefowski** – pracownik Uniwersytetu Ekonomicznego w Poznaniu, Katedra Statystyki; kierownik Ośrodka Statystyki Małych Obszarów Urzędu Statystycznego w Poznaniu.

Mgr **Paweł Lańduch** – programista w Ośrodku Statystyki Krótkookresowej, Dział Metodologii i Programowania Urzędu Statystycznego w Poznaniu.