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.

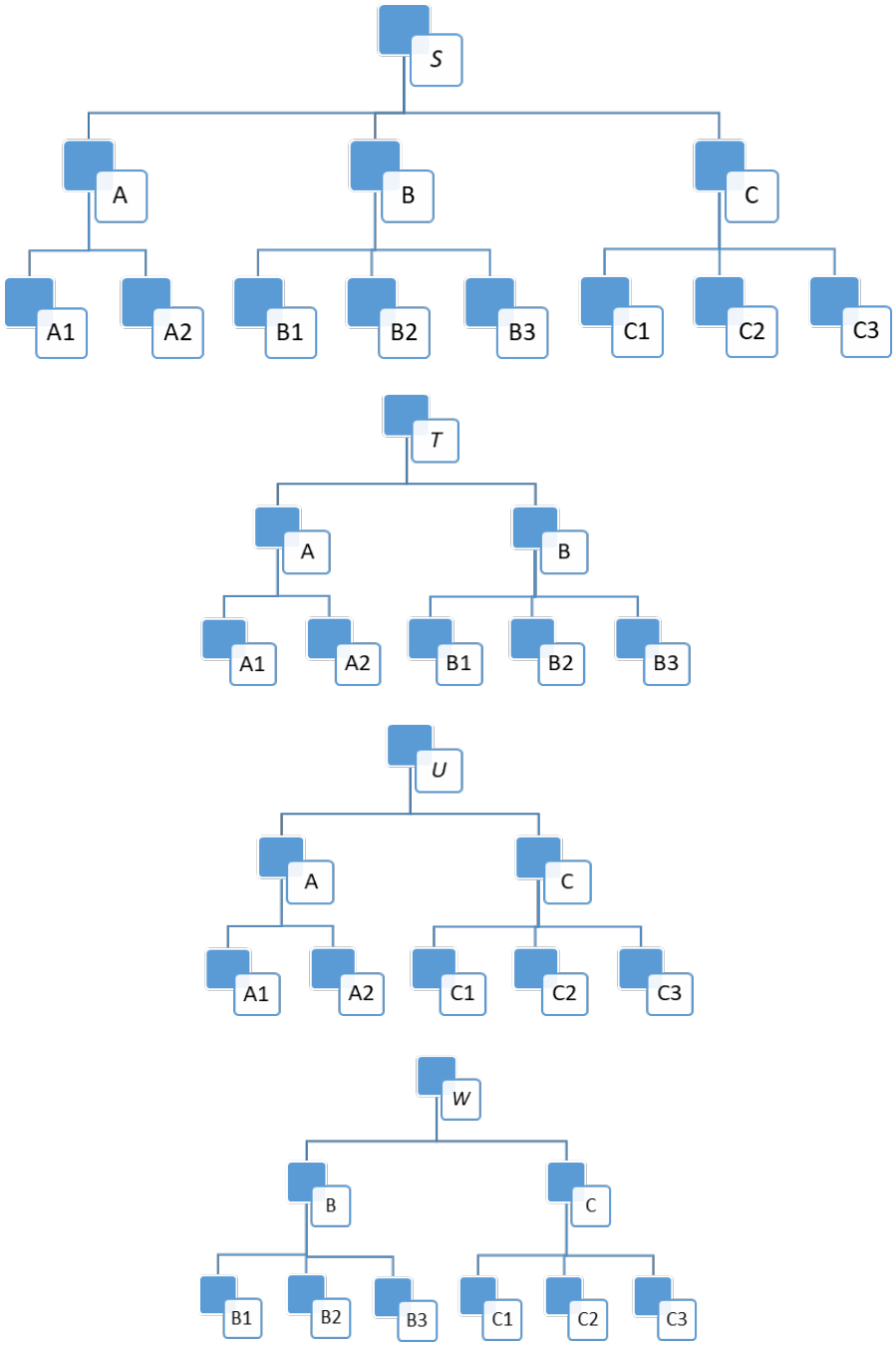
Tabl. 3.5. Liczba biernych zawodowo na obszarze X

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).

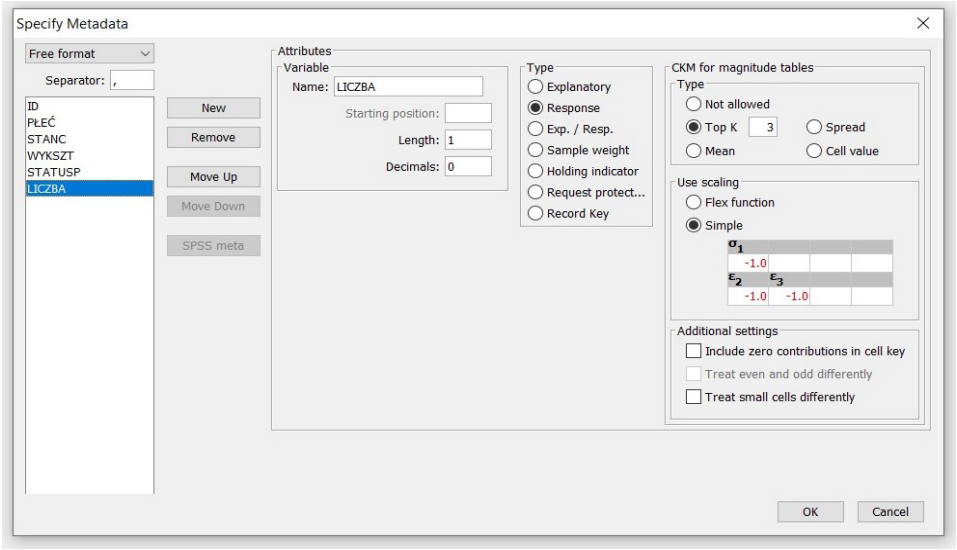
Z liniowej zależności pomiędzy komórkami w tablicy oraz z tego, że X_{11}, X_{21}, X_{13}, $X_{23} \geq 0$, otrzymujemy:

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

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*).

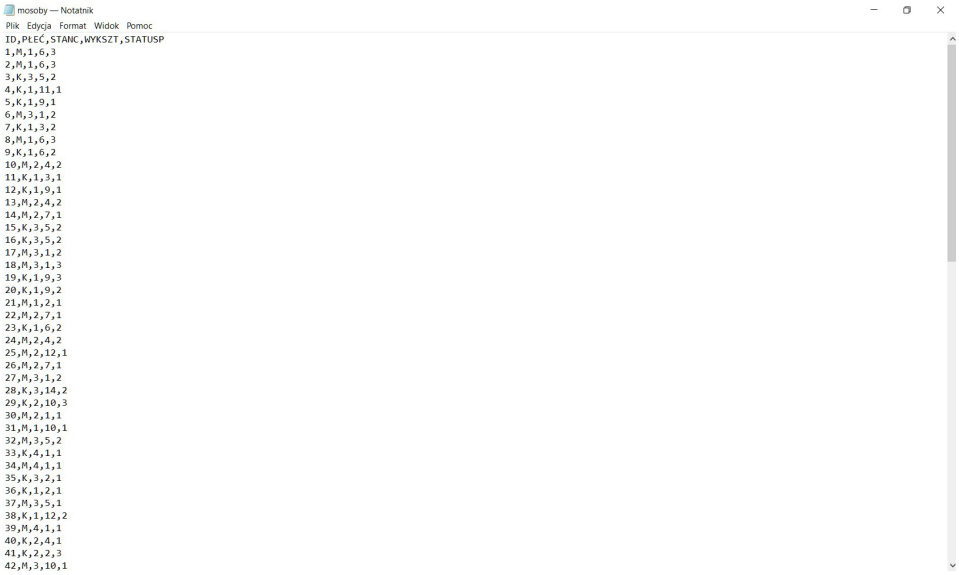
Strata informacji w programie τ-Argus odgrywa pewną rolę we wtórnym ukrywaniu komórek tablicy: pozwala ona rozróżnić komórki, pozostawienie wartości których będzie priorytetowe, od tych, których zawartość powinna raczej być ukryta. Ocena straty informacji jest dokonywana za pomocą funkcji kosztu usunięcia 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.


mosoby — Notatnik
Plik Edycja Format Widok Pomoc
ID,PŁEĆ,STANC,WYKSZT,STATUSP
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
27,M,3,1,2
28,K,3,14,2
29,K,2,10,3
30,M,2,1,1
31,M,1,10,1
32,M,3,5,2
33,K,4,1,1
34,M,4,1,1
35,K,3,2,1
36,K,1,2,1
37,M,3,5,1
38,K,1,12,2
39,M,4,1,1
40,K,2,4,1
41,K,2,2,3
42,M,3,10,1

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.

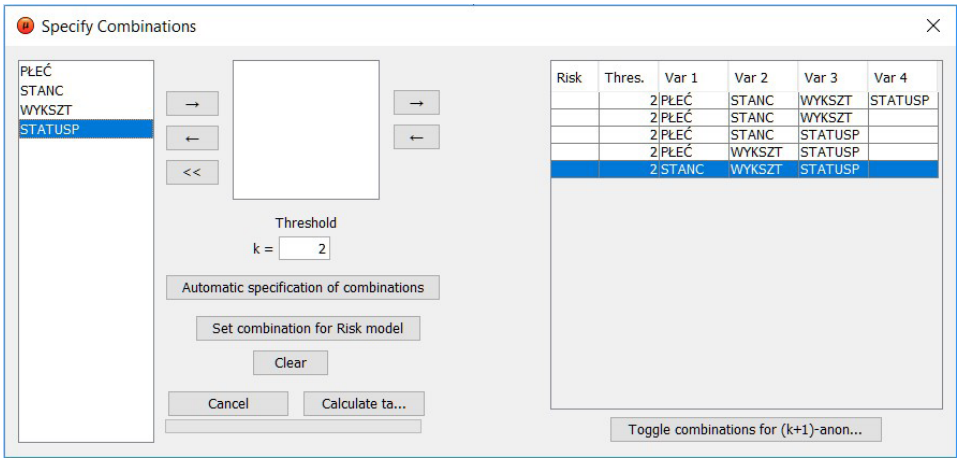
Dla każdej zmiennej można też ustalić tutaj poziom identyfikacji (*Identification level*): 0 – jednostka nie może być zidentyfikowana na podstawie danej zmiennej, 1 – zmienna identyfikuje najbardziej, 2 – zmienna identyfikuje bardziej, 3 – zmienna identyfikuje)[20], priorytet w zakresie ukrywania (*Priority 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.

- 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-

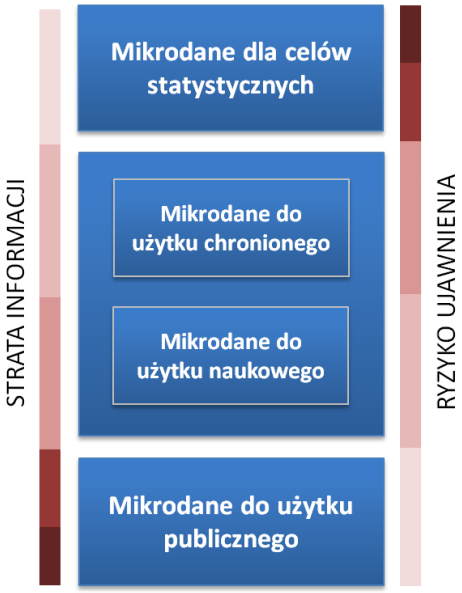
[27] Udostępnianie użytkownikom zewnętrznym różnych kopii zbioru danych jednostkowych z określonej edycji badania statystycznego, przygotowanych specjalnie wedle ich zamówienia, jest dyskusyjną i wątpliwą etycznie praktyką. Należy bowiem pamiętać, że porównanie kilku różnych wersji mikrodanych może skutkować 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.