Възможности за усъвършенстване на методите и подходите за осигуряване на конфиденциалност на статистическата информация

Александър Найденов*

Резюме: Предоставянето на официална статистическа информация от националните статистически институти на потребителите винаги е придружено от известно „напрежение“. От една страна, производителите на статистически продукти целят да предоставят възможно най-качествената информация, съблюдавайки нормативните изисквания, свързани с опазването на статистическата тайна. От друга страна, потребителите желаят все по-детайлна информация за явленията и дори за отделните статистически единици, което е предпоставка за разкриване на конфиденциална статистическа информация. Националният статистически институт на България спазва всички необходими законови предписания, осигурявайки висока степен на защита на личните данни на своите респонденти, като същевременно осигурява продукт с високо качество. Настоящата статия разглежда една малко дискутирана, но много важна материя, касаеща защитата на конфиденциалността на статистическите данни. В допълнение е направен подробен преглед на настоящата ситуация от гледна точка на методите и подходите, използвани за осигуряването ѝ в НСИ, като са дадени и някои препоръки за усъвършенстването на тези методи и по-ефективното използване на разполагаемите ресурси.

* Александър Найденов е доктор, главен асистент в катедра „Статистика и иконометрия“ на УНСС, e-mail: anaidenov@gmail.com

Ключови думи: статистически методи за осигуряване на конфиденциалност, микроданни, данни в табличен вид, възможности за усъвършенстване.

JEL: C83, C18.

1. Въведение

Гледните точки на официалния производител на статистическа информация – националната статистическа служба на всяка държава, и потребителите на тази информация – граждани, фирми, учени и много други, винаги са били в своеобразен „сблъсък“. От една страна, всеки национален статистически институт гарантира чрез съответния нормативен акт[1], че използваните от него конфиденциални данни[2] за статистически цели няма бъдат предоставяни на трети лица и ще представя информацията в обобщен вид (напр. на равнище област). По този начин всяка статистическа служба гради необходимата степен на доверие у респондентите (лица или фирми), което от своя страна, е основа за предоставянето на информация въз-

[1] В България това е регламентирано в Закона за статистиката, Глава шеста „Опазване на тайната“ (НСИ, 2008).

[2] Под конфиденциални данни се има предвид данни, които дават възможност да бъде идентифицирана отделната статистическа единица. В текста е предпочетен терминът „конфиденциалност“ пред българския аналог „поверителност“, за да се акцентира на неговото статистическо интерпретиране.

можно най-идентична с реалността. От друга страна, потребителите на статистическа информация желаят да получат възможно най-детайлните данни (напр. на равнище населено място), включително и данни за отделните статистически единици. Това често пъти е наложено не само от обективната необходимост провежданият от съответните изследователи анализ да предостави резултати на възможно най-детайлно ниво, но и от спецификата на използваните статистически методи.

За да се урегулират отношенията „потребител-производител“ на официална статистическа информация в рамките на Европейския съюз, както и тези отношения между отделните национални статистики и Евростат[3], са имплементирани няколко основни европейски регламента, както следва в хронологичен ред:

- *Регламент №1588/90* относно предоставянето на данни, обект на статистическата конфиденциалност, на статистическия офис на Европейския съюз (Eurostat, 1990).
- *Регламент №322/97* относно принципите за ръководене на националните статистики, процеса на производство на статистическа информация и правилата за осигуряване на конфиденциалност на статистическите данни (Eurostat, 1997).
- *Регламент №831/2002* относно предоставяне на достъп до конфиденциални статистически данни за научни цели (Eurostat, 2002).
- *Регламент №223/2009* относно въвеждането на новата Европейска статистическа система, като основа за партньорство между националните статистически институти, отговорни за събирането, производството и разпространението на статистическа информация (Eurostat, 2009).
- *Регламент №557/2013* относно европейските статистически институти и достъпа до конфиденциални статистически данни за научни цели (Eurostat, 2013).

От гледна точка на българското законодателство, предоставянето на конфиденциални статистически данни се регламентира от два основни нормативни документа, които по същество са базирани на гореизброените европейски регламенти: „Законът за статистиката“ от 2008 г. и актуализираният „Правилник за предоставяне на анонимизирани индивидуални данни за научни и изследователски цели“.

Гореизложените документи, от една страна, гарантират опазването на статистическата тайна[4], а от друга, предоставят възможност само научни и изследователски организации да използват конфиденциална статистическа информация, съблюдавайки строги правила[5] и процедури.

В България процедурата по предоставяне на анонимизирани индивидуални данни за научни и изследователски цели най-общо включва следните *етапи*:

- *Първи*. Подаване на заявка от научната организация до Националния статистически институт (НСИ) за предоставяне на конфиденциални статистически данни[6] и попълване на подробен въпросник, касаещ кандидатстващата институция, проекта, за който се искат данните, начините за обработка на данните и много други;
- *Втори*. Обработка на заявката и издаване на становище от страна на НСИ по отношение на възможностите за предос-

[3] Статистическата служба на Европейския съюз.

[4] Съгласно чл. 25, ал. 1 от Закона за статистиката под *статистическа тайна* се разбира: „Получаваните и събираните при статистическите изследвания индивидуални данни са статистическа тайна и могат да се използват само за статистически цели. Индивидуалните данни, получени за целите на статистическите изследвания, не могат да се ползват като доказателства пред органите на изпълнителната и съдебната власт.“

[5] Подробности относно правилата за достъп могат да бъдат открити в Правилник за предоставяне на анонимизирани индивидуални данни за научни и изследователски цели (НСИ, 2015).

[6] Тук се има предвид най-вече микроданни.

тавяне на исканите данни. Обикновено становището цели да информира потенциалния потребител за това в какъв вид и на каква цена[7] могат да му бъдат предоставени тези данни. При обработката на заявката и изготвянето на становището често пъти се изисква допълнителна информация за кандидатстващата организация с оглед нейното детайлно проучване и потвърждаване на научно-изследователската ѝ специализация.

- *Трети*. След съгласие от страна на потребителя с предложените условия, се пристъпва към свикване на специализирана комисия за достъп до анонимизирани данни (КДАД), съблюдавайки издаденото становище и допълнителната информация, която гласува дали данните могат да бъдат предоставени и в какъв вид. При положителна оценка на заявката, съобразявайки се с нормативните изисквания, се информира Председателят на НСИ, който от своя страна взема решение дали да разреши предоставянето на исканите данни.
- *Четвърти*. От своя страна, при получаването на разрешение от Председателя на НСИ, потребителите от научната организация сключват договор с НСИ за предоставянето на исканите данни, съблюдавайки строги изисквания за съхранение и използване на анонимизираните данни. С оглед гарантиране опазването на статистическата тайна, изследователите, на които предстои да бъдат предоставени анонимизирани статистически данни, подписват специални клетвени декларации.
- *Пети*. След писменото разрешение на Председателя и съгласието на научната организация, специалистите, отговарящи за исканите от потребителя данни, прилагат съответните методи за контрол върху разкриването на конфиденциалността на статистическите данни. Тези методи имат за цел да осигурят анонимност на индивидуалните статистически данни, като същевременно предотвратят евентуалното волно или неволно разкриване на конфиденциалността им от изследователи или нарушители, които биха могли да притежават допълнителна информация за отделните единици.
- *Шести*. След изпълнение на гореописаните стъпки се извършва предоставяне на анонимизираните статистически данни на съответната научна организация в електронен формат на диск или чрез електронна поща.

Докато така описаната процедура се свързва по-скоро с предоставянето на индивидуални статистически данни (микро-данни), статистическата практика се сблъсква с проблеми и при осигуряването на конфиденциалност при предоставянето на данни в обобщен вид (обикновено табличен). В тези случаи, върху данните отново се прилагат специални методи за контрол върху разкриването на конфиденциалността на статистическите данни. Това се налага поради факта, че някои статистически единици притежават твърде специфични характеристики (признаци), които в комбинация с други в двумерни разпределения, дори и в обобщен табличен вид, биха разкрили самоличността на даден индивид или фирма, което всъщност е недопустимо от законова и морална гледна точка.

Както ще стане ясно по-нататък, прилаганите от НСИ методи за осигуряване на конфиденциалност на статистическите данни, както на индивидуално ниво, така и на ниво разпределения, изискват сериозни ресурси – специалисти, софтуер и хардуер. Използваните в момента методи, обаче, в някои случаи са неефективни и пораждат естествена необходимост от тяхното усъвършенстване.

[7] Цената на данните се определя в зависимост от необходимите човеко-часове за нейната подготовка или в зависимост от броя на единиците и променливите, които се съдържат в предоставяния файл с данни.

Последното, разбира се, зависи в голяма степен от възможностите на НСИ да обучи съответни специалисти и да закупи необходимия софтуер и хардуер. В настоящата статия са разгледани както методите, използвани в Националния статистически институт към настоящия момент, така и възможностите за усъвършенстване на тези методи на един следващ етап, с оглед повишаване на ефективността на влаганите ресурси.

2. Методи и подходи за осигуряване на конфиденциалност на статистическата информация, използвани в Националания статистически институт на Република България

Базирайки се на опита на останалите членки на Европейския съюз и използвайки популярни в статистическата практика методи за осигуряване на конфиденциалност на статистическите данни, Националният статистически институт предоставя качествена, надеждна, достоверна и навременна информация за ключови аспекти от социално-икономическия живот на България, същевременно опазвайки статистическата тайна.

От гледна точка на формата, под която НСИ предоставя статистическа информация, може да се говори за два типа данни: *индивидуални* (още наречени микроданни) и данни във вид на *разпределения* (обикновено в табличен вид). Данните в табличен вид осигуряват информация за съвкупностите като цяло (население, предприятия и други) относно важни техни характеристики (брой лица в домакинството, брой предприятия по сектори и т.н.). Микроданните, от своя страна, съдържат конкретните значения на интересуващите ни признаци за всяка отделна статистическа единица.

В статистическата практика по-популярно е предоставянето на *информация в табличен (обобщен) вид*. Това е продиктувано не само от спецификата на статистическия подход, но и от необходимостта от съблюдаването на конфиденциалността на статистическата информация. Съгласно Закона за статистиката, чл. 25, ал. 1, т. 3: „Националният статистически институт и органите на статистиката и техните служители не могат да разгласяват или предоставят статистическа информация, която обобщава данните за по-малко от три статистически единици или за съвкупност, в която относителният дял на стойността на изучаван параметър на една единица е над 85 на сто от общата стойност на този параметър за всички единици от съвкупността." (НСИ, 2008). По тази причина, при обработката и обобщаването на първичните данни в НСИ се прилагат два типа ограничения: *прагово* (от англ. threshold) правило и *доминантно* (от англ. dominance) правило или още наречено *(n,k)* [8] правило (Handerpool et al, 2012).

Праговото правило гарантира, че ако стойността на дадена клетка от една статистическа таблица се формира от 1 или 2 статистически единици, тази клетка ще остане скрита (маскирана) и няма да може да бъде използвана от трети лица (включително и от другата единица от клетката, ако те са 2) за разкриване на самоличността на единицата или единиците, формиращи стойността на тази клетка. Това правило се прилага както в демографската статистика, така и при бизнес наблюденията в Националния статистически институт.

В практиката на *демографската статистика,* осигуряването на конфиденциалност на табличните статистически данни преминава през следните етапи:

- *Първи етап.* Избират се признаците, по които ще се групират (разпределят) единиците на интересуващата ни съвкупност и се извършва групировка на единиците по избраните признаци.

За целите на настоящата статия ще бъде представен хипотетичен пример с ин-

[8] Където *n* означава брой единици, които осигуряват *k* процента от стойността на съответния статистически признак.

формация за единиците в община АБВ, разпределени по населени места и етническа принадлежност, в следния вид:

Таблица 1. *Разпределение на населението от област АБВ по населени места и етническа принадлежност (група).*

Населено място	Общо	Етническа група	
		Група 1	Група 2
Общо	*10000*	*4933*	*5067*
НМ1	499	247	252
НМ2	1001	518	483
НМ3	1510	791	719
НМ4	506	307	199
НМ5	1872	865	1007
НМ6	1933	1444	489
НМ7	6	5	**1**
НМ8	43	14	29
НМ9	2426	606	1820
НМ10	204	136	68

При проверка за т.нар. *първична конфиденциалност*[9], от изходната таблица става ясно, че в едно от населените места има опасност от нарушение на праговото правило, защото има само едно лице от населено място №7, което се е самоопределило като принадлежащо към етническа група №2. Ако обаче информацията от клетката просто бъде заличена (напр. чрез метода на „потискане“ (от англ. suppression) на данните в клетките[10]), то чрез изваждане на сумата от данните на останалите клетки от общото, би могло да се разкрие стойността на тази привидно „скрита“ клетка. По тази причина се прилага и т.нар. *вторична конфиденциалност*. При нея освен клетката, която крие потенциална „опасност“ от разкриване на конфиденциална информация, се „потискат“ и други клетки от таблицата. Най-често допълнителните клетки, които се „потискат“, са тези, които притежават най-ниски стойности по даден признак от таблицата (в примера това е населено място №8).

Таблица 2. *Разпределение на населението от област АБВ по населени места и етническа принадлежност, с осигурена първична и вторична конфиденциалност чрез праговото правило и метода за „потискане“ на клетките.*

Населено място	Общо	Етническа група	
		Група 1	Група 2
Общо	*10000*	*4933*	*5067*
НМ1	499	247	252
НМ2	1001	518	483
НМ3	1510	791	719
НМ4	506	307	199
НМ5	1872	865	1007
НМ6	1933	1444	489
НМ7	..	..	..
НМ8	..	..	..
НМ9	2426	606	1820
НМ10	204	136	68

- *Втори етап.* Всички „рискови“ клетки се „потискат“ и се оформя окончателната таблица с осигурена първична и вторична конфиденциалност. Първичната конфиденциалност обикновено се осъществява автоматично с помощта на специално разработен за целта скрипт на софтуерния продукт SPSS, а вторичната – ръчно от специалист-статистик.

В този случай таблица 1 придобива вида, представен в таблица 2.

[9] Под *първична конфиденциалност* се разбира проверка за конфиденциалността на първичните статистически разпределения, тоест, тези, които са получени непосредствено от обобщаването на индивидуалните данни.
[10] Методът се свързва със скриването на информацията от дадена клетка на таблицата и заместването ѝ с двоеточие на мястото на реалните данни.

- *Трети етап.* Проверява се за *свързана конфиденциалност* на данните, тоест за наличието на конфиденциалност между всички свързани помежду си таблици[11], които се получават в резултат на проведеното изследване. Целта е данните от дадена таблица да не доведат до разкриването на конфиденциалността в друга таблица, която е свързана с нея.
- *Четвърти етап.* Аналогично на предходните етапи, касаещи проверката за конфиденциалност, и тук се „потискат" клетките с най-ниски стойности и тези, които не отговарят на праговото правило, в зависимост от заличената информация при първичната и вторичната конфиденциалност.
- *Пети етап.* Таблиците, съдържащи данни с осигурени първична, вторична и свързана конфиденциалност, се публикуват.

Като *положителни страни* на този подход за осигуряване на конфиденциалността могат да се посочат:

- ✓ лесно приложение, което не изисква специализирани статистически познания;
- ✓ липса на необходимост от специализиран софтуер за приложението на първичната конфиденциалност.

Същевременно подходът притежава и някои *недостатъци*:

- ✓ поради наличието на свързани таблици, конфиденциалността трябва да бъде осигурена във всички тези таблици едновременно, което е ресурсоемко;
- ✓ при вторичната и свързаната конфиденциалности често пъти е налице значителна загуба на данни, която понякога достига до 90% от информацията.

В *допълнение* към отрицателните черти трябва да се спомене, че към момента в НСИ свързаните таблици в демографската статистика се обработват ръчно, което е свързано с изразходването на много човеко-часове, особено когато става дума за огромен брой двумерни и многомерни разпределения, каквито са налице например при обработка на резултатите от Преброяване на населението.

При обобщаване на данните от *бизнес наблюденията*, освен праговото правило (нар. още конфиденциалност тип „А"), се добавя и второто правило – доминантното (нар. още конфиденциалност тип „В"). След като е извършена проверка на разпределенията за наличие на поне 3 единици във всяка клетка, тоест е осигурена първичната конфиденциалност тип „А", се пристъпва към вторичната проверка за конфиденциалност от тип „В". При последната, съгласно Закона за статистиката, се прилага правилото (1,85), тоест информацията в дадена клетка подлежи на „потискане", ако една статистическа единица (предприятие) формира 85 и повече процента от стойността на интересуващия ни признак за цялата изучавана съвкупност или подсъвкупност (напр. едно предприятие формира 85% от приходите на строителните предприятия по видове строителство за обществения сектор). За разлика от демографската статистика, в този случай се налага приложението на това правило, тъй като много често в практиката се срещат фирми, които са доминиращи на даден пазар или в даден сектор и това ги излага на риск те да бъдат идентифицирани, освен в случаите, когато те са дали писмено съгласие за публикуването на данните им.

Също както при демографската статистика, и при бизнес наблюденията, след като бъде определен типът на конфиденциалността (А или В) на статистическата информация, се преминава през горепосочените стъпки. Извършва се „потискане" на данните в клетките на таблиците, както на първично, така и на вторично ниво. След като таблиците придобият необходимата степен на конфиденциалност, включително и свързана, се публикуват официално. И тук първичната конфиденциалност е автома-

[11] Под свързани таблици се разбират таблици, които имат общи признаци (антетки).

тизирана софтуерно, а вторичната и свързаната се извършва ръчно от специалисти.

За разлика от данните в обобщен табличен вид, осигуряването на *конфиденциалност на микроданните* се осъществява на ниво отделна статистическа единица. Както стана ясно в частта „Въведение", въз основа на съответните нормативни документи, НСИ има правно основание да предоставя анонимизирани индивидуални статистически данни за научни и изследователски цели. Това означава, че след като дадена институция е припозната като научна или изследователска организация и след като тя бъде ободрена да получи исканите от нея микроданни, последните преминават през специална процедура наречена *анонимизация*. Тази процедура в Националния статистически институт на България се свързва с два типа промени в променливите, описващи признаците на единиците от дадена съвкупност:

- *Премахване на т.нар. директни и индиректни идентификатори*, тоест от файла с данни се изтриват променливи като: имената на лицето, ЕГН, настоящ адрес, населено място, възраст и други, чрез които може лесно да се разкрие идентификацията на дадено лице. Това се отнася и до наименование на предприятието, адрес на регистрация/седалище, информация за контакти, област, брой заети, сектор (до 4-ти знак на КИД-2008) и други, които биха могли да доведат до пряката идентификация на дадено предприятие.
- *Рекодиране*[12] *на стойностите на променливите* в стойности на по-високо равнище на агрегация. В този случай вместо оригиналните стойности на променливите за всяка единица се предоставят нови стойности, които отговарят на оригиналните, но не дават възможност за пряка идентификация на статистическата единица. Например: ако се предоставят микроданни от бизнес статистиката за съвкупност от предприятия, то вместо сектора на предприятието на ниво четвърти знак (КИД-2008), може да се посочи само сектора на равнище А3 на класификацията, тоест: селско, горско и рибно стопанство, индустрия и услуги. Друг популярен подход е да се използва групиране на индивидуалните значения на признаците в интервали. Например: вместо да се предоставят данни за броя на заетите във всяко предприятие, съответната променлива се рекодира в променлива със стойности, които по своята същност представляват интервали: до 9 заети, от 10 до 49 заети, от 50 до 249 заети и 250 и повече заети.

Както стана ясно от гореизложеното, защитата на конфиденциалността, както на индивидуалните статистически данни, така и на обобщените данни, е трудна и отговорна задача. От една страна, статистическата служба трябва да защитава интересите на своите респонденти, „скривайки" част или цялата информация за тях, а от друга, тази служба трябва да осигури качествена статистическа продукция, тоест продукция, която да изпълни своята основна функция – да даде възможност за взимане на адекватни управленски решения. В тази връзка трябва да се търсят възможности за усъвършенстване на съществуващите методи и подходи за осигуряване на конфиденциалността, с оглед по-ефективното използване на наличните ресурси.

3. Възможности за усъвършенстване на използваните методи и подходи за осигуряване на конфиденциалност на статистическата информация

Редица европейски страни прилагат аналогични на българските правила при съблюдаването на конфиденциалността на статистическата информация. За разли-

[12] По терминът „рекодиране" се има предвид промяната в кодовете на дадена променлива.

ка от българския Закон за статистиката, обаче, чуждестранните нормативни актове дефинират в по-общи граници тези правила, като дават възможност на специалистите в съответното статистическо направление: демографско, социално, бизнес или друго, сами да конкретизират стойностите на праговото и доминантното правило или други минимални изисквания, с оглед спецификата на тяхната дейност.

Следователно една от възможните посоки за развитие при осигуряване на конфиденциалност на статистическата информация е *промяна в дефинирането на ограниченията заложени в чл. 25 от Закона за статистиката,* така че да предоставят по-голяма гъвкавост при определяне на правилата, касаещи тази конфиденциалност. Например в ал. 2, т. 3 на същия член, текстът може да бъде заменен със „...статистическа информация, която не отговаря на минималните изисквания за конфиденциалност". По този начин се предоставя възможност всяка дирекция, отдел или конкретно изследване само да дефинира адекватно минималните изисквания за конфиденциалност, без да се съобразява със спецификата на останалите изследвания.

Също така, от голяма полза би било и въвеждането в практиката на НСИ на *проверка на риска за разкриване на конфиденциална информация* чрез реидентификация[13] на единиците от потенциални нарушители, преди провеждането на процедурите по анонимизация на микроданните. В статистическата теория, касаеща конфиденциалността, съществуват богат набор от измерители за определяне на степента на риск за разкриване на конфиденицални статистически данни за различните видове признаци (променливи). Такива измерители са базирани на биномното и поасоновото разпределение. Самите правила – прагово и доминантно – не е необходимо да бъдат променяни като методика за определяне на минимални изисквания за осигуряване на първична конфиденциалност. Това се потвърждава и от практиката на голяма част от европейските статистически институти, които използват тези правила, както стана ясно, в един по-широк контекст.

Сериозни промени се налагат в още две направления: прилаганата *методология* за осигуряване на конфиденциалност и използвани *софтуерни продукти* (вкл. хардуер). Масово използваната техника за „потискане" на клетките, които съдържат конфиденциална информация, както стана ясно по-горе, води до разход на ресурси и огромни загуби на данни, особено при свързани таблици. Това е конкретният случай при осигуряването на конфиденциалност на данните от демографската статистика. Възможни решения на този проблем при *табличните данни* се крият в наличието на голям набор от други техники и методи за практическо приложение на първична и вторична конфиденциалност. Някои от тези методи са:

- *Преработване (редизайн) на първичните разпределения*. Това означава, че таблиците, в които има опасност от разкриване на конфиденциална информация, могат да бъдат преработени по такъв начин, че две и повече значения на обобщаващия признак да се обединят в едно. Например: ако става дума за признака възраст в интервали и има малък брой единици във възрастовия интервал 85-89 години за дадено населено място, то би могло този интервал да бъде обединен с предишния (80-84 г.) и следващия (90 и повече г.). По този начин тези лица ще бъдат „скрити" сред останалите и няма да може да се разкрие тяхната идентичност. Трябва обаче да се внимава с приложението на тази методика, защото таблиците могат да станат толкова „обобщени", че да се загуби тяхната информативност.

[13] Под реидентификация се разбира разкриване на идентификацията на дадена статистическа единица въпреки приложените методи за анонимизация.

- *Закръгление на числата в клетките*. В този случай числата във всички клетки на таблицата се закръглят до най-близкото цяло число при определена база. Например: ако базата е 5, то числата 34, 29 и 101 ще бъдат закръглени съответно на 35, 30, 100, като това ще се отрази и в сумарните редове и колони на таблицата, към която принадлежат тези числа.
- *Контролирано ажустиране*[14] *на клетките на таблицата*. Този метод е част от цяла „фамилия" методи за осигуряване на конфиденциалност, наречени *пертурбативни*. В този случай клетките на вече готовото разпределение се променят, като към всяка клетка се добавя или изважда определено случайно число („шум"), което се намира в предварително дефинирани граници. Тук се спазва условието по такъв начин да бъдат дефинирани интервалите за генериране на случайни числа, така че резултативната таблица да се доближава максимално до „истинската", без да разкрива реалните данни.
- *Добавяне на „шум" в микроданните.* При този метод преди данните да бъдат обобщени в групировка по даден признак, в променливата, касаеща този признак, се добавя „шум". Това означава, че към всяка стойност на тази променлива се изважда или добавя определено случайно число, така че характеристиките (средна, мода, медиана и т.н.) на разпределението на единиците по този признак да останат непроменени. По този начин конфиденциалността е осигурена още преди самото създаване на таблиците.

Осигуряването на *конфиденциалност на микроданните* под формата на анонимизация, може да се осъществи освен чрез гореизложените методи на премахване и рекодиране на променливите, така и чрез следните:

- *Излъчване на извадка от микроданните*. Тук се има предвид това, че дори данните да са получени в резултат на извадково изследване, би могло да се направи извадка от тези данни и по този начин няма да бъде възможно пълното идентифициране на единиците от изследваната съвкупност.
- *Заместване на „чувствителни" данни с липсващи стойности*. В случаите, когато във файла с данни се съдържа информация, която би довела до лесната реидентификация, то тези „чувствителни" данни се изтриват и на тяхно място се оставят липсващи стойности.
- *Добавяне на „шум" в данните*. Както стана ясно по-горе, масивът с данни би могло да бъде „деформиран" в определени граници, като към стойностите на променливите се добавят или извадят определени случайни числа, така че да се запазят стойностите на характеристиките на разпределенията, за които се отнасят тези променливи. Към тази група се отнасят както методи, които „променят" всяка променлива поотделно, така и методи, при които се извършват промени едновременно в няколко свързани помежду си променливи.
- *„Разбъркване" и „размяна" на данни.* При тези методи се извършва размяна на стойностите на дадена променлива между самите единици, като по този начин се запазва „целостта" на файла с данни, но отделните единици стават неидентифицируеми.
- *Пост-рандомизационни методи.* При тях стойността на дадена променлива се изменя съгласно предварително дефиниран механизъм, базиран на дадено теоретично (вероятностно) разпределение.
- *Синтетични и полу-синтетични микроданни.* Вместо оригиналните микроданни от дадено изследване се предоставят данни, които представляват симулация на изходните данни. Когато става дума

[14] Под „ажустиране" се има предвид промяна в стойностите на клетките, съобразно дадени научно-обосновани принципи.

за изцяло симулирани данни, тогава базата данни се нарича синтетична, а когато само част от данните са симулирани – полу-синтетична или хибридна.

С оглед изпълнение на една от основните цели на статистическата дейност – да осигурява качествена информация, е необходимо да се дефинира доколко произведените данни, с определена степен на конфиденциалност, отговарят на това условие (Templ et al, 2014). За тази цел в световната практика се използват т.нар. *измерители за информационни загуби* (от англ. information loss measures). Такива измерители могат да бъдат изчислени както за всяка таблица като цяло, така и за всяка клетка, която е била обект на „промяна". Тук трябва да се има предвид, че подробното описание на информационните загуби би могло да доведе до индиректно разкритие на конфиденциалността на статистическата информация.

От изложеното дотук става ясно, че приложението на тези методи за двумерни и многомерни разпределения, които са свързани помежду си е трудоемка задача, особено за големи масиви от данни, които са характерни за статистическата практика. По тази причина има разработени редица *специализирани софтуерни продукти за осигуряване на конфиденциалност*, както на микроданните, така и на тези в табличен вид.

В предходен период в България са експериментирани част от тези продукти, но те не са навлезли в практиката на НСИ. Това налага адаптирането на по-широкоспектърни програмни продукти за целите на конфиденциалността (като SPSS, например), което от своя страна, налага прилагането само на някои по-елементарни методи за анонимизация. Би било от голяма полза, ако в бъдеще българската статистика се възползва от значително усъвършенстваните продукти като μ-ARGUS (за анонимизиране на микроданни) и τ-ARGUS (за таблични резултати). Тези продукти имат и някои алтернативи с отворен код като например: sdcMicro и sdcTable, които обаче не притежават богатата функционалност на предходните. Софтуерните продукти от семейството ARGUS са разработени и поддържани от голям екип европейски специалисти в сферата на конфиденциалността и информационните технологии и имплементират по-голямата част от познатите методи за осигуряване на защита на статистическите данни (Handerpool et al, 2012).

От голяма полза за самите специалисти-статистици от практиката би било разработването на *наръчник* и *ясни правила* за прилагането на методи за анонимизация и осигуряване на конфиденциалност на микроданните и табличните разпределения. Това би дало възможност за приемственост между различните поколения експерти и би осигурило яснота при осигуряването на тази конфиденциалност.

Заключение

Съблюдаването на статистическата тайна е от първостепенно значение за всеки официален статистически орган. Въпреки настояването от страна на потребителите на статистическа информация за предоставяне на все по-детайлна такава, европейската и националните нормативни уредби задължават основния „производител" на данни да направи всичко възможно да осуети евентуалните опити на някои потребители да разкрият конфиденциална информация за отделните статистически единици. Прилаганите методи в български условия са съобразени с европейската статистическа практика, но е налице и необходимост от по-ефективно използване на разполагаемите ресурси. Последното би могло да бъде постигнато чрез промени в нормативната база, създаване на добри статистически практики в сферата на конфиденциалността, приложението на специализирани софтуерни продукти и съвременни методи за анонимизация.

Цитирани източници:

НСИ, 2008. Закон за статистиката. Интернет адрес: http://www.nsi.bg/bg/node/553

(NSI, 2008. Zakon za statistikata. Internet address: http://www.nsi.bg/bg/node/553)

НСИ, 2015. Интернет страница на Националния статистически институт, касаеща Правилника за предоставяне на анонимизирани индивидуални данни за научни и изследователски цели. Адрес: http://www.nsi.bg/bg/node/575/

(NSI, 2015. Internet stranitsa na Natsionalnia statisticheski institut, kasaeshta Pravilnika za predostavyane na anonimizirani individualni danni za nauchni i izsledovatelski tseli. Adres: http://www.nsi.bg/bg/node/575/)

Eurostat, 1990. Council Regulation (EC) No 1588/90 of 11 June 1990 on the transmission of the data subject to statistical confidentiality to the Statistical Office of the European Communities.

Eurostat, 1997. Council Regulation (EC) No 322/97 of 17 February 1997 on Community Statistics.

Eurostat, 2002. Commission Regulation (EC) No 831/2002 of 17 May 2002 implementing Council Regulation (EC) No 322/97 on Community Statistics, concerning access to confidential data for scientific purposes.

Eurostat, 2009. Regulation (EC) No 223/2009 of the European Parliament and of Council of 11 March 2009 on European statistics.

Eurostat, 2013. Commission Regulation (EU) No 557/2013 of 17 June 2013 implementing Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics as regards access to confidential data for scientific purposes and repealing Commission Regulation (EC) No 831/2002 Text with EEA relevance.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer and P. de Wolf, 2012. Statistical Disclosure Control. Wiley.

Templ, M., B. Meindl, A. Kowarik and S. Chen, 2014. Introduction to Statistical Disclosure Control (SDC), IHSN Working Paper No. 007.