Гл. ас. д-р Александър Найденов*

СЪВРЕМЕННИ МЕТОДИ ЗА ОСИГУРЯВАНЕ НА КОНФИДЕНЦИАЛНОСТ НА СТАТИСТИЧЕСКАТА ИНФОРМАЦИЯ

Направен е критичен преглед на най-популярните статистически идеи и методи, свързани с опазването на статистическата тайна и контрола върху разкриването на конфиденциални статистически данни. Разгледани са и въпроси, които са от съществено значение както за производителите, така и за потребителите на статистическа информация, но все още не са дискутирани широко в специализираната българска научна литература.[1]

JEL: C18; C81; C82

Основен приоритет на всяка статистическа институция при производството и разпространението на официални данни е осигуряването на конфиденциалност на статистическата информация и на контрол върху евентуалното идентифициране на отделните статистически единици. Във връзка с практическото приложение на защитата на конфиденциалността в световната практика е създаден голям набор от статистически методи.

Потребителите на статистическа информация винаги са се стремили да получават данни бързо, евтино и във възможно най-детайлен вид (Cox и Zayatz, 1995). Статистическата информация обаче е резултат от своеобразен „производствен“ процес, който не се различава съществено от този при „физическия“ тип продукти. Съвременното разбиране за този процес намира израз в модела GSBPM (Generic Statistical Business Process Model),[2] който описва производството на статистическа информация като производство, включващо следните *осем*[3] основни *етапа*:

- ●специфициране на нуждите на потребителите от определен тип статистическа информация;
- ●проектиране на статистическото изследване;
- ●изграждане и апробиране на статистическия инструментариум;
- ●статистическо наблюдение;
- ●обработка на събраната информация;

* УНСС, катедра „Статистика и иконометрия“, anaidenov@gmail.com

[1] Chief Assistant Prof. Alexander Naidenov, PhD. CONTEMPORARY METHODS FOR STATISTICAL DISCLOSURE CONTROL. *Summary:* A critical review is made of the most popular statistical methods and ideas associated with keeping statistical secret and control on disclosure of confidential statistical data. Some of the most important issues for the statistical data producers and consumers, which are still not widely discussed in the specialized Bulgarian scientific literature, are considered too.

[2] GSBPM, версия 5.0, декември 2013 г.

[3] Според някои източници етапите могат да бъдат девет, като се добавя етапът на архивиране на статистическата информация (Eurostat, 2012).

●анализ на обработената информация;

●разпространение на статистическата информация;

●оценка на различните етапи на „производствения“ процес и търсене на възможности за неговото усъвършенстване.

Въпреки че удовлетворяването на нуждите на потребителите от определен тип статистическа информация е приоритет на всяка статистическа служба, трябва да се имат предвид и *рисковете*, които крие предоставянето на тази информация (Trewin, 2006). Съгласно европейското законодателство[4] и използвания етичен кодекс на статистическата практика (Eurostat, 2011), предоставянето на статистически данни трябва да бъде в такава форма, която да съблюдава изискванията за осигуряване на тяхната конфиденциалност.[5] В случая всеки статистически институт е отговорен за това, че когато предоставя статистическа информация - било в обобщен (табличен) вид, или под формата на микроданни,[6] трябва да се приложат съответните методи, чрез които да се осъществи *контрол върху евентуалното разкриване на идентичността на отделната статистическа единица* (от англ. statistical disclosure control) (вж. Hundepool et al., 2012).

Съгласно Закона за статистиката от 2008 г. съблюдаването на конфиденциалността на статистическите данни като израз на опазването на статистическата тайна е базата, върху която всяка статистическа служба гради доверието на своите респонденти. Последните като основен източник на информация биха били по-склонни да предоставят достоверна информация, ако са напълно уверени, че статистическият институт ще я използва само за статистически цели. Посочените регламенти допускат за целите на науката и научните изследвания да се предоставят не само обобщени, но и индивидуални данни, но отново осигурявайки конфиденциалността им (Eurostat, 2002).

Осигуряването на конфиденциалността на статистическите данни като *част от шестия етап на производствения процес*, описан в GSBPM-модела, е отговорна и трудна задача (Vale, 2010). Нейното решаване е свързано с изграждането на т.нар. *сценарии за разкриване на конфиденциални данни* (от англ. disclosure risk scenarios), още преди да започнат да се прилагат конкретните статистически методи в това отношение (вж. Krenzke et al., 2014). В тези сценарии се прави подробно описание на потенциалните потребители на статистическата информация и на възможните волни и неволни действия по нейното използване за идентифициране на отделните статистически единици.[7] За всеки сценарий се дефинира и т.нар. *риск от разкриване* (от англ. disclosure risk), чрез

[4] Регламенти на ЕК № 1588/90, 322/97, 831/2002, 223/2009 и 557/2013.

[5] Под конфиденциални данни се имат предвид данни, които дават възможност да бъде идентифицирана отделната статистическа единица. Под осигуряване на конфиденциалност се разбира възпрепятстването на опитите за разкриване на самоличността на отделната статистическа единица.

[6] Данни за отделната статистическа единица.

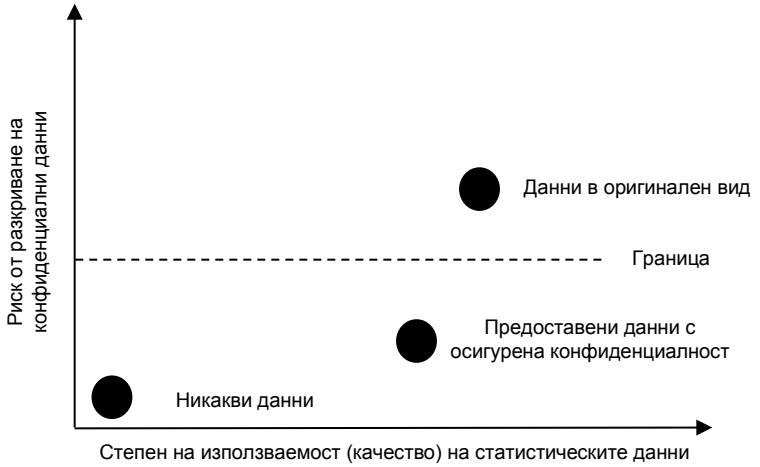
[7] В специализираната литература (например Hundepool et al., 2012) лицата, които волно нарушават конфиденциалността на статистическите данни, се наричат още нарушители (от англ. intruders).

който се измерва вероятността за идентифициране на отделната единица. По такъв начин вниманието на „производителя“ на определен тип статистическа информация се насочва към онези единици или отделни клетки на дадено разпределение, при които съществува най-висок риск от идентифициране (Kounine и Bezzi, 2008).

На следващ етап с помощта на съответните специализирани софтуерни продукти и в зависимост от типа информация (табличен или микроданни) се прилагат методите за осигуряване на конфиденциалност. За така произведената статистическа информация с осигурена конфиденциалност по конкретна методология се изчисляват т.нар. *информационни загуби* (от англ. information loss), като целта е да се измери доколко „крайният продукт“ от приложението на методите за конфиденциалност се различава от статистическата информация в първоначалното ѝ състояние, т.е. в оригиналния ѝ вид (Doyle, 2001). Взимайки предвид информационните загуби, може да се прецени кой от възможните методи е най-ефективен, т.е. едновременно изпълнява условията за предоставяне на „използваема“ статистическа информация и за предотвратяване на идентифицирането на единиците. Изборът на най-ефективния метод не е лесна задача и е свързан с това да се определи най-доброто съотношение риск/използваемост (фиг. 1).

Фигура 1

Съотношение риск/използваемост на статистическата информация



Източник. Hundepool et al, 2012.

От фиг. 1 е ясно, че при отказ от предоставянето на каквато и да било статистическа информация няма риск от разкриване на конфиденциалността, но пък и самите данни „липсват“. Колкото повече данни се предоставят, толкова повече рискът расте, докато се достигне до варианта на данните в техния оригинален вид, даващ пълни възможности за идентификация на статистически-

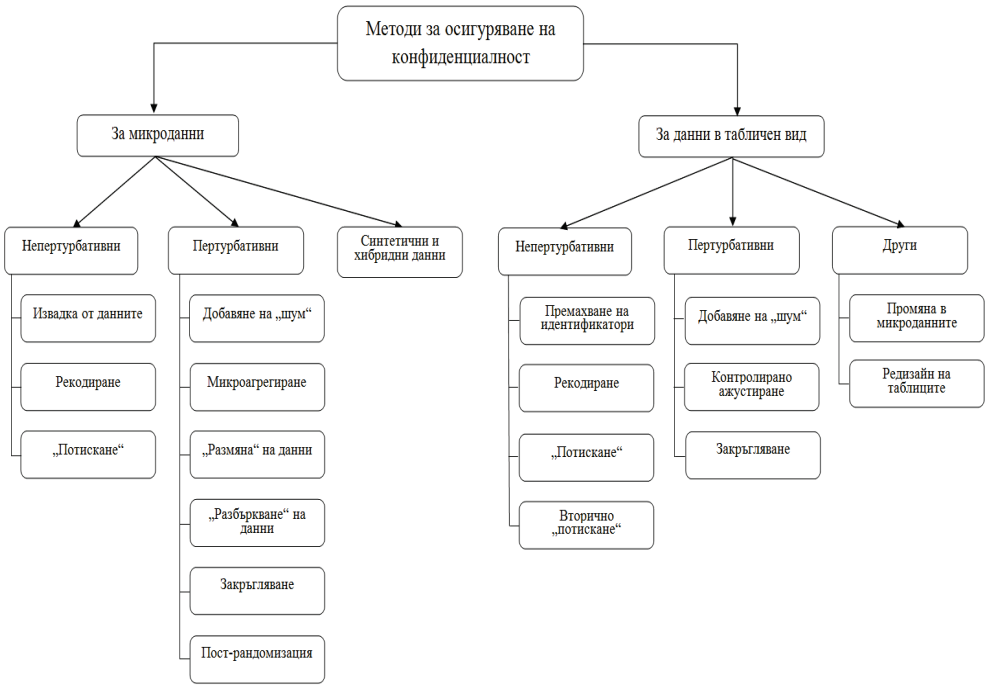
те единици (Duncan et al., 2001). Логично, съотношението риск/използваемост на предоставяните на потребителите данни трябва да се намира някъде между тези два екстремни случая, като същевременно е необходимо да се съблюдава и определена максимална граница, която е дефинирана от много и различни фактори като политиката на статистическата институция по отношение на предоставянето на данни, прилаганите стандарти, регламенти и др.

Съвременни методи за осигуряване на конфиденциалност на статистическата информация

Както вече беше посочено, официалната статистическа служба на всяка страна трябва непрекъснато да балансира между това колко „качествена“ статистическа информация ще предостави на потребителите и какъв е рискът, който тя поема при евентуалното разкриване на конфиденциална информация. С оглед да се подпомогне дейността по контрол върху разкриването на подобна информация през последните 20 години е разработен и в статистическата практика (извън България) се прилага голям набор от методи, които най-общо могат да бъдат класифицирани и обобщени в една схема (фиг. 2).

Фигура 2

Видове методи за контрол върху разкриването на конфиденциална статистическа информация



Приложението на методите за осигуряване на конфиденциалност на статистическата информация до голяма степен зависи от формата на разполагаемите данни - микроданни или/и таблици (Templ et al., 2014; Hundepool et al., 2012).

Ако информацията е във вид на *микроданни*, т.е. данни за отделната единица, европейското, вкл. и българското законодателство (вж. НСИ, 2015) позволяват предоставянето й на потребители, но само в нейния анонимизиран вид, и то *единствено* за научни и изследователски цели. Анонимизираният вид се отнася за такива данни, за които не може да бъде установена идентификацията на отделната статистическа единица. Основните методи за анонимизация са разделени в три групи: непертурбативни, пертурбативни и осигуряващи синтетични/хибридни данни.

Непертурбативните методи не променят оригиналните статистически данни по същество, а до известна степен „замаскират" конфиденциалната информация. Тук са включени методи, които обикновено се използват в комбиниран вид, като:

- *Премахване (изтриване) на идентификационни променливи*, които могат директно или индиректно да доведат до разкриване на самоличността на статистическата единици. Такива идентификатори са например: наименование на фирма, БУЛСТАТ, име на лице, ЕГН, точен адрес на лицето и др.
- *Излъчване на извадка* от оригиналната база от данни - на случаен принцип се избират част от единиците от вече създадения файл с данни, като по този начин не може да се възстанови първоначалният вид на изучаваната съвкупност.
- *Рекодиране*[8] на първоначалните данни - промяна на стойностите, описващи отделните значения на признаците на изследваните единици, в такива от по-обобщен характер. Например, ако се разполага с база данни за предприятията, участвали в едно изследване, то броят на заетите във всяко от тях (например 37) може да се замени с интервала, в който попада този брой (от 10 до 49). По такъв начин потребителят ще разполага с ориентировъчна информация за отделното предприятие, но няма да може да го идентифицира точно по броя на заетите в него.
- *„Потискане"* на конфиденциална информация - премахване на „чувствителна" информация за предприятия или лица, при които е налице висок риск от идентификация. Такъв риск например съществува за големите предприятия в определена община или област, които са единствени по рода си и поради това са широкоизвестни.

Когато в оригиналните данни се внесе известна доза „фалшифициране",[9] тогава се говори за *пертурбативни методи*. Към тях се включват:

[8] Терминът „рекодиране" се отнася до промяната в кодовете на дадена променлива съобразно определени принципи.

[9] Тук под „фалшифициране" се има предвид промяна в истинските стойности на променливите, описващи интересуващите ни признаци на единиците, като се добави нарочно определен размер грешка или отклонение в тях.

• *Добавяне на „шум“ в данните* - прибавяне или изваждане на определено случайно число (случайно отклонение) от реалните стойности на променливите. При добавянето на тези случайни отклонения конкретните специалисти-статистици преценяват дали те да бъдат взаимно корелирани, или не; дали да се извършат линейни, или нелинейни трансформации в данните; дали добавянето на „шум“ да има адитивен, или мултипликативен характер и т.н.

• Когато в микроданните се формират групи от единици, в които има определен минимален брой случаи и отсъстват доминиращи единици, тогава се говори за т.нар. *микроагрегация*. Приложението на този метод цели „обезличаването“ на отделната статистическа единици чрез „маскирането“ ѝ в група от подобни единици. Това агрегиране може да се извърши за признаци, измерени както на силните, така и за слабите скàли. Основна цел в този случай е да се постигне максимална хомогенност на единиците в отделните групи, като същевременно се максимизира различието между групите. Така се минимизира и загубата на информация. Микроагрегирането може да се извърши, като предварително се дефинират размерите на отделните групи или автоматично - чрез оптимизационни процедури. Пример за подобно агрегиране е заместването на оригиналните стойности на печалбите на предприятията с интервали, в които се намират тези печалби.

• *„Размяната“ на данни* се отнася до взаимното разместване на значенията на интересуващия ни признак между две единици от изследваната съвкупност, които имат сходни характеристики по определени дефиниционни признаци. По такъв начин профилът на дадена фирма се променя, но с близък до него, което възпрепятства пряката идентификация на тази единица.

• *„Разбъркването“ на данните* е аналогично на „размяната“ с разликата, че тук се използват специални рандомизационни процедури. Последните са патентовани от Muralidhar и Sarathy (Muralidhar и Sarathy, 2006) и могат да бъдат използвани само с тяхно изрично разрешение.

• В случаите, когато реалните стойности на дадена променлива се закръгляват до най-близкото цяло число (например 5123,4 на 5100), се прилага методът *закръгляване*. Смята се, че липсата на точна стойност на дадена характеристика създава затруднения при разкриването на конфиденциална статистическа информация.

• *Пострандомизацията* е техника за контрол върху разкриването на конфиденциална статистическа информация, която е приложима основно върху категорийни признаци. При нея всяка стойност на една или повече категорийни променливи се променя с определена вероятност, като се съблюдава запазването на характеристиките на вероятностното им разпределение.

В статистическата практика посочените методи не се прилагат изолирано, а по-скоро в комбиниран вид. Например популярен метод е MASSC, при който се комбинират микроагрегиране, разбъркване на данни, ползване на подизвадка от данните и калибриране на резултатите.

Когато данните трябва да се анонимизират за научни и експериментални цели, често в практиката се прилагат и техники за създаването на т.нар. *синтетични* или *изкуствени* бази от данни. Въз основа на оригиналните микроданни се генерират нови, които се доближават до първите, но представляват случайни извадки от стойности от стохастичните разпределения на променливите в оригиналните данни. В този случай се допуска и „приписването“ (от англ. imputation) на определени стойности на дадени променливи на мястото на липсващите такива, съгласно определени правила и принципи. Синтетичните файлове с данни могат да бъдат два вида: *изцяло синтетични* и *хибридни*. При последните част от данните във файла са оригинални, а други са изкуствени, формирани на базата на оригиналните.

Практическото приложение на описаните методи за осигуряване на конфиденциалност на микроданните обикновено се осъществява посредством различен от широкоспектърните *специализиран статистически софтуер* - SAS, SPSS, Stata и т.н. За целта екипи от учени са разработили програмни продукти като μ-ARGUS, sdcMicro и IVEware, всеки от които прилага на практика поголямата част от разгледаните методи, но не и пълния им набор. По тази причина различните национални статистически институти сами решават кой софтуерен продукт да използват в зависимост от методите, които са избрали да прилагат, а някои от тях разработват продукти изцяло за вътрешни нужди.

За разлика от методите, използвани при микроданните, защитата на конфиденциалната информация *в табличен вид* се прилага върху вече обобщена първична статистическа информация във вид на едномерни, двумерни или многомерни разпределения, т.е. под формата на таблици. В този случай методите отново се делят на *непертурбативни* и *пертурбативни*. Към първите спадат методите рекодиране, „потискане“ и вторично „потискане“, а към вторите - добавяне на „шум“, контролирано ажустиране и закръгляване. Към двете основни групи се добавят и някои други методи като редизайн (преформатиране) на таблиците и промяната в микроданните.

Идеите, стоящи зад методите, прилагани върху таблични данни, не се различават от тези при микроданните, но имат и своите специфични особености. Ето защо ще обърнем внимание на техните характеристики, като на първо място ще разгледаме *непертурбативните* методи.

- При *рекодирането,* известно още като *глобално рекодиране*, се извършва обединяване на няколко сходни категории на даден категориен признак в една. Например съгласно действащата класификация на НСИ КИД-2008 (Класификация на икономическите дейности от 2008 г) могат да бъдат обединени всички предприятия, занимаващи се с добив и преработка на суровини, с тези, които имат производствен и строителен характер, в една обща категория „производство“.
- *„Потискането“*, както и при микроданните, се отнася до заличаването на данните в дадена клетка от таблицата, за които се смята, че биха могли да доведат до разкриването на конфиденциална информация за единиците, фор-

миращи стойността на тази клетка. Например, ако се разглежда разпределението на жителите от определено населено място по възраст и някакво рядко заболяване, е възможно в някоя от клетките да попаднат едно или две лица. Тъй като останалите граждани на населеното място вероятно познават това лице/лица, те биха могли да разкрият от останалите взаимно свързани таблици информация за него, която не е била широко известна до момента и която по своята същност е конфиденциална. Поради това данните от тази клетка се изтриват. *Първичната проверка за конфиденциалност* обаче не е достатъчна. Необходимо е да се направи и т.нар. *вторична проверка*, при която да се установи дали чрез изваждане на стойностите от останалите клетки на таблицата от общия ред или колона не би могла да се „отгатне“ заличената стойност в клетката от първичната конфиденциалност. Ако това е така, се потискат (премахват) и стойностите на други клетки от таблицата (обикновено тези с най-ниски стойности), така че по никакъв начин да не бъде възможно възстановяването на стойностите в „потиснатите“ клетки. Това премахване се нарича *вторично „потискане“*.

При *пертурбативните методи* при табличния тип данни най-популярни в статистическата практика по осигуряване на конфиденциалност са:

• *Добавяне на „шум“* или случайно отклонение към стойностите на клетките от таблицата, като се запазят маргиналните разпределения (сумите в таблицата). При добавянето се извършва изваждане или прибавяне на случайно число към всяка клетка, като по този начин изкуствено се деформира първоначалното разпределение на единиците по даден признак.

• При *контролираното ажустиране* всички клетки от изходната таблица се подлагат на модификация така, че резултативната таблица да бъде „достатъчно“ различна от първоначалната (под „достатъчно“ различна се има предвид, че стойностите в резултативната таблица са „деформирани“ в определени първоначално дефинирани граници). Това „деформиране“ обикновено се осъществява с помощта на линейното програмиране и е обвързано с контролиране на степента на загуба на информация.

• Аналогично на подхода при микроданните и тук *закръгляването* се свързва със *заместването* на реалните стойности в клетките на дадена таблица с такива, които са близки до оригиналните с определена степен на точност. Например, ако числото в дадена клетка е 733, то може да бъде закръглено на 730 или дори на 700. Закръгляването се извършва така, че общата сума на клетките в дадена колона и ред да не се различава особено от тези, които са получени в оригиналните таблици.

Най-сигурни резултати от приложението на статистическите методи за контрол върху разкриването на конфиденциални таблични статистически данни се получават, когато микроданните, от които са съставени, *предварително* са подложени на анонимизиране. Тези *промени в данните* са познати в литературата като предтабулационни методи и за тях важи всичко казано за анонимизирането на микроданните (Anbazhagan et al., 2012).

Разбира се, първоначално излъчените разпределения във вид на таблици невинаги са достатъчно адекватни от гледна точка на статистическата конфиденциалност. В такъв случай е възможно да трябва да се извърши *редизайн на таблиците*. При наличието на проблеми с конфиденциалността при голяма териториална дезагрегация може да се наложи да се премине например към по-висока такава - от ниво община на ниво област.

Както е известно, всяко статистическо изследване и резултатите от него са свързани с генерирането на огромен брой статистически разпределения (таблици). Осигуряването на конфиденциалност на всяка от тях поотделно не е толкова трудоемко, но задачата обикновено се усложнява значително поради наличието на т.нар. *свързани разпределения*. Последните се отнасят до таблици, имащи общи признаци (антетки), въз основа на които са построени. Това от своя страна допринася за контролиране на възможните сценарии за разкриване на конфиденциална информация както за отделната таблица, така и за всички останали, свързани с нея. Тази задача сама по себе си е непосилна без използването на специализиран статистически софтуер, който се разработва както от отделните статистически служби за собствена употреба (например SAS модул за Статистическия офис на Канада), така и с по-общо предназначение, например т-ARGUS.

Трябва да се отбележи, че когато информацията в разпределенията (таблиците) се отнася не до честотата на срещания признак сред единиците на дадена съвкупност, а до величината на интересуващата ни характеристика, тогава в допълнение се прилагат и т.нар. *правила за „чувствителност“* (от англ. sensitivity rules). Последните се свързват с дефинирането на минимални изисквания, на които трябва да отговарят единиците, стоящи в основата на формирането на стойността на дадена клетка от таблицата. Най-популярните правила са *минимална честота* (брой единици в клетка), *доминантното правило* (принос на най-голямата единица в клетката) и *p%-тото* правило (разликата между стойността в клетката и дела на двете най-големи единици).

*

„Производството“ на статическа информация е процес, който цели да създаде баланс между две противодействащи си „сили“ – от една страна, потребителите и техните желания да придобият възможно най-детайлна информация, а от друга, респондентите, които гласуват своето доверие на официалния статистически орган, разчитайки на нормативното основание за опазване на статистическата тайна. Всяка официална статистическа институция трябва сама да дефинира границите, в които може да „наруши“ даденото от нея „обещание“ за осигуряване на конфиденциалност на индивидуалните данни, като същевременно се съобрази с възприетия етичен кодекс, нормативните изисквания и съществуващите статистически методи. В своята нелека „битка“ статистиците са подпомогнати както от научната общност – при разработването на нови техники за анонимизация, така и от софтуерните специалисти – при програмното осигуряване на процеса на контрол върху разкриването на конфиденциална статистическа информация.

В крайна сметка обаче въпреки помощта всеки статистически институт избира методите за осигуряване на конфиденциалност в зависимост от разполагаемите човешки и финансови ресурси. Основните методи и подходи за контрол върху разкриването на конфиденциални статистически данни вече са се превърнали в регулярна практика в много страни, но все още не са достатъчно познати в България.

Използвана литература:

НСИ (2015). Правилник за предоставяне на анонимизирани индивидуални данни за научни и изследователски цели, http://www.nsi.bg/bg/node/575/

Anbazhagan, K., R. Sugumar, M. Mahendran, R. Natarajan (2012). An Efficient Approach for Statistical Anonymization Techniques for Privacy Preserving Data Mining. - International Journal of advanced Research in Computer and Communication Engineering.

Cox, L., L. Zayatz (1995). An agenda for research in statistical disclosure limitation, Environmental Protection Agency.

Doyle, P., J. Lane, J. Theeuwes, L. Zayatz (2001). Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Elsevier Science.

Duncan, G., S. Keller-McNulty, S. Stokes (2001). Disclosure risk vs. Data Utility: The R-U Confidentiality Map. Technical Report LA-UR-01-6428, Statistical Sciences Group.

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

Kounine, A., M. Bezzi (2008). Assessing Disclosure Risk in Anonymized Datasets. Carnegy Mellon University.

Krenzke, T., J. Li, L. Li (2014). An Evaluation of the Impact of Missing Data on Disclosure Risk Measures, Survey Research Methods.

Muralidhar, K., R. Sarathy (2006). Data Shuffling: a new masking approach for numericall data. Management science.

Templ, M., B. Meindl, A. Kowarik, and S. Chen (2014). Introduction to Statistical Disclosure Control (SDC). IHSN Working Paper N 007.

Trewin, D. et al. (2006). Principles and guidelines of good practice for managing statistical confidentiality and microdata access. UNECE.

Vale, S. (2010). Exploring the relationship between DDI, SDMX and the Generic Statistical Business Process Model. UNECE.

Eurostat (2002). Commission Regulation (EC) N 831/2002 of 17 May 2002 implementing Council Regulation (EC) N 322/97 on Community Statistics, concerning access to confidential data for scientific purposes.

Eurostat (2011). European Statistics Code of Practice. Statistical Programme Committee.

Eurostat (2012). Documentation of Statistical Production Processes (Methodologycal Guidelines), version 6.5.

14.IX.2015 г.