AZ ADATFELFEDÉS ELLENI VÉDELEM STATISZTIKAI ESZKÖZEI

ERDEI VIRÁG – HORVÁTH ROLAND

A tanulmány az adatfelfedés elleni védelem statisztikai eszközeit mutatja be, az adatvédelem problémáinak tárgyalása mellett és azok összefüggésében. Ismerteti az adatvédelem európai és magyar jogi alapjait, a tájékoztatási formák bővülését is. Az eszközök, módszerek tárgyalásakor sor kerül a táblázatos- és mikroadatokban lévő adatfelfedési kockázat, majd a táblázatos adatokra vonatkozó védelmi eszközök és a mikroadat védelem különböző módjainak ismertetésére, gyakorlati példákon keresztül.

TÁRGYSZÓ: Adatvédelem. Adatfelfedés elleni statisztikai eszközök

Talán nem szerénység azt állítani, hogy lassan immár 15 éves demokráciánkban az adatvédelem szó mindenkinek ismerősen cseng. Rádió- és tévéműsorok állandó szereplője az adatvédelmi biztos, gyakran újságok vezető híre az adatvédelemmel kapcsolatos valamely aktuális téma. Az utca embere természetesen azt látja a kifejezés mögött, hogy a korábbi mindent tudó állammal szemben napjainkban már inkább a semmit nem tudó állam áll. Érdekvédő szervezetek és jogászok hada áll szemben az állammal, illetve minden egyéb magán illetve hivatalos szervvel, amennyiben az szeretne valami nem jogában állót megtudni rólunk, hiszen személyes adataink védettek, mi rendelkezünk felőlük, és jogi felhatalmazás híján nehezen tudható meg tőlünk bármi is.

A magyarországi demokrácia érésének folyamán a személyes adatok védelme volt az első, amit mindenki megismert, de az évek során az egyre tudatosabban viselkedő az állampolgárok annak is tudatában kezdenek lenni, hogy a közérdekű adatok nyilvánosságához is joguk van, valamint általában az információhoz. Állampolgárként ugyanolyan vehemenciával igényelhetünk információkat, mint amilyen mértékben ragaszkodunk személyes adataink védelméhez. Az Európai Unióhoz történő csatlakozás nyomán Európa és a világ kitágul számunkra. Nő az információigényünk, egyre jobban tisztában vagyunk a jogainkkal és a lehetőségeinkkel. Számítani lehet arra, hogy a csatlakozás hatására az emberek egyre jobban felmérik a lehetőségeiket, és élni is fognak velük, például egyre több információt fognak igényelni.

Az adatgyűjtők, így a statisztikai hivatalok is, hatalmas adatvagyonnal rendelkeznek, mégis egyes becslések szerint ennek csak 30–40 százaléka hasznosul, kerül nyilvánosságra. Ennek egyik fő oka az adatvédelem. Az informatika óriási térnyerése következté-

ben megnőtt az adatfelfedés lehetősége. A statisztikai hivatalnak meg kell felelnie a törvényi adatvédelmi kötelezettségeknek, s ezt annál is inkább meg kell tennie, mivel egy esetleges adatfelfedés nagyban aláásná az adatszolgáltatói bizalmat, és így a statisztikai tevékenységet. A kötelező adatvédelemmel szemben azonban az információszabadság állampolgári joga áll.

Tanulmányunk célja, hogy bemutassuk az adatfelfedés elleni védelem statisztikai eszközeit. Ezen eszközök birtokában válhat lehetővé a minél szélesebb körű biztonságos adatközlés, az adatfelfedés egyidejű elkerülésével.

AZ ADATFELFEDÉS ELLENI VÉDELEM KÖRNYEZETE

A statisztikai és egyéb adatgyűjtések célja az adatok elemzés, feldolgozás utáni nyilvánosságra hozatala. Ez az adatfelvételek végső és legérzékenyebb pontja. A magyar statisztikai törvény kimondja, hogy a statisztikai módszerekkel felvett, feldolgozott, tárolt és elemzett adatok az államhatalmi és a közigazgatási szervek, valamint a társadalom szervezetei és tagjai tájékoztatását szolgálják.

Tájékoztatási kötelezettség, szélesedő lehetőségek

Az állami, központi költségvetésből finanszírozott statisztikai szerveknek a törvényben rögzítetten túl erkölcsi kötelessége is az adatok legteljesebb mértékű közzététele, hiszen az adatokat mi, állampolgárok térítés nélkül szolgáltatjuk, és az állam statisztikákat felhasználó tevékenysége, munkája is a mi érdekünkben történik. A magyar statisztikai törvény szerint a hivatalos statisztikai szolgálathoz tartozó szervek által végrehajtott adatgyűjtések eredményei – az adatvédelemre vonatkozó szabályok betartása mellett – nyilvánosak.[1]

A nyilvánosságra hozatal, a tájékoztatás „kiadványokból és más adathordozókon lévő adatállományokból történő közlésekből áll".[2] A papír alapú tájékoztatás magában foglalja a különböző kiadványokat, évkönyveket, tájékoztatókat, brosúrákat stb., ám napjainkban a gyors és nagy információigény miatt egyre inkább tért nyer az egyéb adatközlés. Ilyenek az internetes adatközlés, a CD-k, és az egyéb nem papír alapú adathordozók, de akár a telefonon keresztül történő adatszolgáltatás is.

Az adatközléseknél különbséget tehetünk aszerint is, hogy azok egy konkrét „legyártott" adatot, táblázatot tartalmaznak, vagy a felhasználó, adatkérő közreműködésével egy állományból egyedi beállítás alapján lekérhető adatokat, táblázatokat. Az adatközlés egy harmadik típusa a mikroadat-állomány közzététele, amely rekordsorosan tartalmazhat egy adatfelvételt vagy annak egy részét.

A tájékoztatás kötelezettségét, annak alapelveit az uniós statisztikai jogszabályok is részletesen rögzítik. A tájékoztatási tevékenységgel kapcsolatban megfigyelhető az a tendencia, hogy egyre nyitottabbá válnak a statisztikai szervezetek, egyre több adat kerül nyilvánosságra. Egyre több formában válik lehetségessé a tájékoztatás, egy-egy adat, szám közlése mellett egyre részletesebb összesítések, táblázatok jelennek meg, és akár teljes adatállományok is hozzáférhetővé válnak. Ennek konkrét bizonyítéka, hogy euró-

[1] 1993. évi XLVI. Törvény a statisztikáról 17.§ (1)
[2] 1993. évi XLVI. Törvény a statisztikáról 23.§ (2)

pai uniós szinten jogilag is megnyílt a lehetőség a kutatók, tudományos élet képviselői előtt, hogy bizalmas, egyedi adatokhoz férjenek hozzá. (Az Európai Unió 1997-ben született statisztikai törvénye már megfogalmazta ezt a lehetőséget (17. cikk), 2002-ben azonban rendelet is született, amely részletezi azt.)

A 831/2002/EK rendelet a bizalmas adatokhoz való tudományos célú hozzáférésről[3] lehetővé teszi a közösségi hatóság (az Európai unió statisztikai hivatala, más néven Eurostat) hivatali helyiségeiben a bizalmas adatokhoz való hozzáférést, és anonimizált mikroadatok kibocsátását is. Egyetemek, felsőoktatási intézmények, tudományos kutatással foglalkozó szervezetek, intézmények, hivatalok, számára nyitott ez a lehetőség (részletesen lásd 831/2002/EK rendelet 3. cikk). A rendelet az adatokhoz való hozzáférés módjáról, engedélyezéséről szól, annak érdekében, hogy pontosan tudható legyen – az adatok bizalmas volta miatt –, hogy az adathozzáférés folyamán ki mikor jut hozzá valamihez és mi alapján, mit tehet, mik a kötelezettségei stb.

Adatvédelmi intézkedések természetesen itt is vannak, a mikroadatok kiadásakor eltávolítják a közvetlen azonosítókat, és a rendelkezésre álló legjobb eljárás alkalmazásával minimálisra csökkentik az érintett statisztikai egységek közvetett azonosításának veszélyét. Az Eurostat hivatali helyiségeiben (a gyakorlatban kutatószoba) engedélyezhető hozzáférés pedig mindig csak hivatalos személy felügyelete mellett történhet, és a kutatás eredményeit – mielőtt kikerülnek az intézményből – ellenőrzik, biztosítva, hogy azok nem tartalmaznak bizalmas adatokat.

Nevesítve a hozzáférés négy felmérésből, illetve statisztikai adatforrásból lehetséges: a közösségi háztartási panelből, a munkaerő-felmérésből, a közösségi innovációs felmérésből és a szakmai továbbképzési felmérésből. (Az adatszolgáltató nemzeti statisztikai hivatalok megtagadhatják az adataikhoz történő hozzáférést, de engedélyezhetik is a felsoroltaktól eltérő bizalmas adatokhoz való hozzáférést.)

A rekordsoros adatokhoz történő hozzáférés nagyon nagy nyitottságot jelent az adatgazda statisztikai hivataloktól a tájékoztatásban, ezért is követeli meg a legszigorúbb adatvédelmet.

DEFINÍCIÓK

A cikkben tárgyalt statisztikai információk bizonyos jogszabályokon alapulnak, így azokra támaszkodunk mi is. Az előbbiekben használtuk a *bizalmas adatok* kifejezést. A következőkben ismertetjük a *személyes adat*, a *bizalmas adat* és az *azonosíthatóság* fogalmát, amelyek témánk szempontjából meghatározóak. A magyar adatvédelmi, statisztikai törvények, így a fogalmak is számos európai jogszabály és ajánlás alapján születtek. Először az európai uniós megfogalmazások lényegét ismertetjük, majd röviden a hazairól szólunk.

Személyes adat: A személyhez kapcsolódó adat definíciója alapvetően fontos, hiszen a statisztikai felmérések nagy része emberekre vonatkozik. A fogalom igen jól körülírható. „Személyes adat bármely, azonosított vagy azonosítható természetes személyre (’adatalany’) vonatkozó információ; a személy különösen akkor tekinthető azonosíthatónak, ha őt – közvetlenül vagy közvetve – azonosítószám vagy egy vagy több fizikai, fizi-

[3] A Bizottság 831/2002/EK Rendelete (2002. május 17.) a bizalmas adatokhoz való tudományos célú hozzáférés tekintetében a közösségi statisztikáról szóló 322/97/EK tanácsi rendelet végrehajtásáról.

ológiai, mentális, gazdasági, kulturális vagy szociális azonosságára jellemző tényező alapján azonosítani lehet."[4]

Bizalmas adat: Ez ugyancsak kulcsfontosságú fogalom, hiszen ez alapján definiálhatjuk majd a statisztikai titkosságot. A bizalmas adat a személyes adatnál bővebb kategória. A személyes adaton túl egyéb adatok is beletartoznak, pl. a gazdasági szervezetek adatai. A bizalmas adat lényeges tulajdonsága, hogy az a megfigyelési egységekre – személyekre, cégekre stb. – vonatkozó adat, információ. Az „adatok bizalmasnak tekintendők, amennyiben segítségükkel a statisztikai egységek akár közvetlenül, akár közvetve azonosíthatók és így egyedi információt fednek fel."[5]

A bizalmas – egyes szövegekben védettnek nevezett – adat alapján a *statisztikai titkosság* magának a tevékenységnek, az egyes statisztikai egységekkel kapcsolatos adatoknak a védelme.

A bizalmas adat tehát megköveteli, hogy ne lehessen sem közvetlenül, sem közvetve azonosítani a vonatkoztatási, statisztikai tárggyal. (A bizalmas adatokhoz való tudományos célú hozzáférésről szóló rendeletben bizalmas adatok alatt már csak a közvetett azonosíthatóságot értik, hiszen a statisztikai munkában a közvetlen azonosítást a feldolgozási folyamat elején lehetetlenné teszik, illetve az idősoros elemzéseknél külön kezelik az azonosítókat.)

A nemzetközi joganyagok fogalmai egységesek a tekintetben, hogy megkövetelik a közvetlen azonosítók leválasztását, illetve, hogy az egyértelmű azonosíthatóságot és a lehetséges azonosíthatóságot is a fogalom részévé teszik. Az igazán lényegi információt azonban azok a meghatározások adják, amelyek magáról a kikövetkeztethetőségről, azonosíthatóságról szólnak.

Azonosíthatóság: A közvetlen azonosíthatóság egyértelműen definiálható az egyedi azonosítók leválasztásával (személyeknél: név, lakcím; gazdasági szervezeteknél: név, telephely vagy azonosítószám).

A közvetett azonosíthatóságról vagy felfedésről már csak durva körülhatárolás lehetséges:

– „A statisztikai egység azonosíthatóságának megállapításakor figyelembe kell venni mindazokat az eszközöket, amelyeket egy harmadik fél ésszerűen *(reasonably)* igénybe vehet az említett statisztikai egység azonosításához."[6] (A harmadik fél úgy értendő, hogy az első két fél az adatszolgáltató és a statisztikai hivatal, hiszen ők jogosultak az adatot ismerni.)

– „A személy nem tekinthető azonosíthatónak, ha az azonosítása ésszerűtlenül hosszú időt és munkabefektetést igényel."[7]

Magyarországon két alaptörvény szabályozza a kérdéskört, a statisztikai törvény (1993. évi XLVI. Törvény), valamint az adatvédelmi törvény (1992. évi LXIII. Törvény), hivatalos nevén Törvény a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról.

Az adatvédelmi törvény határozza meg a személyes adatot.

[4] Az Európai Parlament és a Tanács 95/46/EC Irányelve az egyénnek a személyes adatok feldolgozásával kapcsolatos védelméről és ezeknek az adatoknak a szabad áramlásáról 2. cikk (a.)

[5] A Tanács 1997. február 17-i. 322/97. (EK) számú rendelete a közösségi statisztikákról 13. cikk (1)

[6] A Tanács 1997. február 17-i. 322/97. (EK) számú rendelete a közösségi statisztikákról V. fejezet 13. cikk

[7] A tagállamok minisztereinek bizottsága által 1997. szeptember 30.-án elfogadott 97/18 sz. ajánlás a statisztikai célból gyűjtött és feldolgozott személyes adatok védelméről Fogalmak 1. bekezdés

Személyes adat: bármely meghatározott (azonosított vagy azonosítható) természetes személlyel kapcsolatba hozható adat, az adatból levonható, az érintettre vonatkozó következtetés. A személyes adat az adatkezelés során mindaddig megőrzi e minőségét, amíg kapcsolata az érintettel helyreállítható. A személy különösen akkor tekinthető azonosíthatónak, ha őt – közvetlenül vagy közvetve – név, azonosító jel, illetőleg egy vagy több, fizikai, fiziológiai, mentális, gazdasági, kulturális vagy szociális azonosságára jellemző tényező alapján azonosítani lehet.[8]

A statisztika és a statisztikai törvény azonban a nemzetközi gyakorlat alapján védi a többi adattípust is, például a gazdasági szervezetek adatait. Ennek érdekében bevezeti az egyedi adat fogalmát és azt védi.

Egyedi adat: a statisztikai célt szolgáló, a természetes és a jogi személy, valamint a jogi személyiséggel nem rendelkező adatszolgáltatóval kapcsolatba hozható adat.[9] Egyedi adat tehát az, ami a nemzetközi joganyagokban bizalmas vagy védett adat. (A jelenleg folyó uniós jogszabályok fordításában elképzelhető, hogy a bizalmas adatok helyett egyedi adat szerepel majd.) Egyedi adat csak statisztikai célra használható.

Azonosíthatóság: A hazai gyakorlat, jog is elsődleges védelmi kritériumként az egyedi azonosítók leválasztását követeli meg. Az azonosítók leválasztása a közvetlen azonosítás megakadályozását szolgálja: „A természetes személy személyére vonatkozó adatgyűjtésnél az érintett nevét és a lakcímét (személyazonosító adat) – kivéve azt, amelynek adathordozóját a levéltári anyag védelmére vonatkozó jogszabály értelmében levéltári őrizetbe kell adni – a statisztikai feldolgozás befejezésekor, az adatok teljességének és összefüggésének ellenőrzését követően, de legkésőbb a tárgyidőszakot követő egy éven belül kell törölni, adatátadás esetén ezt megelőzően is."[10]

(„Az egy évnél hosszabb időszakra vonatkozó idősoros vizsgálatok esetében az adatállományt belső azonosítóval kell ellátni, amelyből az érintett személyazonossága nem állapítható meg. Az érintett személyazonosító adatait az adatállománytól elkülönítetten kell kezelni."[11])

A gazdálkodó szervezet akkor tekinthető anonimnak, ha elnevezése és telephelye nincs feltüntetve (*Statisztikai igazgatás* [2000]).

Egyetlen kritérium van a statisztikai törvény végrehajtási rendeletében, amely a közvetett azonosítást kívánja megakadályozni. Azt mondja a szabály, hogy összesítve sem lehet nyilvánosságra hozni olyan adatot, amelynél az adatszolgáltatók száma háromnál kevesebb.[12]

A jogszabályok definíciói után szeretnénk tisztázni egy, a gyakorlatban elterjedt félreértést. Az adatvédelem során gyakori, hogy megkülönböztetik a jogi védelmet a technikai védelemtől, mondván, hogy amikor például egy szerződést ír alá valaki egy adathozzáférésről, akkor az jogi védelem, míg amikor beavatkozást végzünk egy táblázaton, vagy adatbázison, akkor az technikai. A valóságban ez a két dolog nem különíthető így el, hanem egyik a másikon alapul. A jogszabályok megfogalmazzák a kereteket, fogalmakat, teendőket, s ennek alapján készülnek a gyakorlatban technikák, módszerek azok megvalósítására.

[8] 1992. évi LXIII. Törvény a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról 2.§ 1.
[9] 1993. évi XLVI. Törvény a statisztikáról 17.§ (2)
[10] 1993. évi XLVI. Törvény a statisztikáról 19.§ (1)
[11] 1993. évi XLVI. Törvény a statisztikáról 19.§ (2)
[12] 1993. évi XLVI. Törvény végrehajtásáról szóló 170/1993. (XII. 3.) Kormány rendelet 19.§

AZ ADATKÖZLÉS PROBLEMATIKÁJA

Az adatközlés egyik, nagy problémát jelentő kérdése a *közvetett azonosíthatóság*, azaz az adatfelfedés lehetősége. Maga az *adatvédelem* az a technika vagy módszer, amely alkalmazásával minimálisra csökkenthető a statisztikai egységek azonosításának veszélye.

Az adatközlés során az adatvédelmet készítők maguk döntik el, hogy a jogszabályban megfogalmazott „nagy időbefektetés során lehetővé válható kikövetkeztethetőség" mikor válhat lehetségessé. A közvetett felfedés elleni védekezés bonyolult, komoly munkát igényel, hiszen egy külső, harmadik fél technikai és tudásbeli hátterével szemben kell eszközöket találni. A külső fél, a lehetséges adatfelfedő jó- és rosszindulatú is lehet, különféle motivációkkal és eszközökkel. Az adatközlés számos publikációs formában ölthet testet, a papír alapútól az internetes közlésen át, és ezek eltérő védelmi technikákat, stratégiákat igényelnek.

Az adatfelfedés teljes mértékű megakadályozása által tökéletesen lehetetlenné válna az adatközlés, az adatokhoz való hozzájutás. Az egyre biztonságosabb adatközlés, az egyre nagyobb védelem mindig együtt jár azzal, hogy egyre több és több adatot kell elzárni a felhasználók elöl, és végül az elrejtett információknak köszönhetően használhatatlanná válhatnak adatbázisok.

A cél és egyben a legnagyobb kihívás a felfedés elleni védekezésben az, hogy megtaláljuk azt az optimális arányt az elrejtett, védett és a tájékoztatás révén közölt adatok közt, amivel már biztonságosnak tekinthetőek az adatok, és a felhasználók is hozzájuthatnak a megfelelő részletezettségű információkhoz. Ehhez ismernünk kell, hogy milyen kockázat rejlik a különféle adatközlésekben, és kik lehetnek a felhasználók *(Eurostat* [1999]).

(Azonosításon, azonosíthatóságon azt értjük, hogy egy anonim információhoz valamilyen módon hozzárendelhető, hozzákapcsolható egy egyedi azonosító (azonosítószám vagy kulcs). E mellett az adatfelfedés azt jelenti, hogy egy személyre vagy egy intézményre vonatkozóan új, plusz információ birtokába jutunk az azonosítás által. A két kategória tehát egymásból következik, hiszen plusz információ birtokába akkor jutunk, ha azonosítjuk a személyt. Tanulmányunkban mi e két kategóriával, s a kialakítandó védelemmel együtt foglalkozunk.)

Felfedési lehetőségek és kockázatok

Az informatika nagyfokú elterjedtségének és technikai fejlődésének következtében a közölt adatok analizálásával, kombinálásával olyan új információ birtokába juthat egy külső, harmadik személy, amelyet az adatközlőnek nem állt szándékában közölni. Az adatok felfedése, kikövetkeztethetősége az adatok egyedisége, bizalmassága miatt kockázatossá válhat.

A területi szintű tájékoztatásban kiemelten jelentkezik a probléma: a terület nagysága, az alacsony lélekszám, vagy az adattartalom miatt válik nem közölhetővé az adat. Például:

– Ritka foglalkozások közlése (például: a budapesti agglomeráció egyik kis településén élő operaénekesnő közölt adatai név nélkül egyértelmű felfedést jelentenek).

– Egy átlagos foglalkozású (például bolti eladó) ember is azonosíthatóvá válik, ha csak egy emberről van szó a területen.

– Ugyancsak védendőek bizonyos egyedi, ritka családi vagy egyéb körülmények kis területre vonatkozó adatközlésben (például: 8 gyermekes család; magas jövedelmű személy).

Gazdasági szervezetek adatközlésénél számos probléma merülhet fel. A legkiemelkedőbb a dominancia problémája, vagy a monopol pozíciójú szervezetek, cégek adatai. Azonos jellemzőkkel rendelkező, azonos adatszolgáltatói csoportba tartozó, azonos terméket gyártó, azonos szolgáltatást nyújtó gazdálkodó szervezetek adatai statisztikai összesítés formájában bármikor közölhetők, ám amint valamelyik szervezet egyik mutatója kiugró, domináns értékkel bír (például legmagasabb foglalkoztatotti szám, legnagyobb bevétel, előállított egyedi termék stb.), akkor érzékennyé válik az adat.

Közérdekű és védendő adat együttes közlése során is felmerülhetnek adatvédelmi aggályok. (A Központi Statisztikai Hivatal pontosan felsorolja a közérdekű adatok körét.[13]) Közérdekű adat például a központi vagy helyi önkormányzati költségvetésből finanszírozott bölcsődei ellátásra vonatkozóan az ellátók száma, az ellátottak száma, a forgalom, a befogadóképesség és az ellátottak által fizetett hozzájárulás összesen. Amennyiben egy településen három bölcsőde működik, amelyből kettő állami és egy magán, akkor nagyon megfontoltnak kell lenni a felsorolt adattípusok együttes közlésekor, hiszen az egy magán bölcsőde adata így felfedhetővé, azaz nyilvánossá válik.

A mintavételes felvételek védelmét gyakran feleslegesnek tartják, holott a tájékoztatás módjától függően bizonyos esetekben védendő adattá válnak:

– Ha a megszerzett adatokból, tehát a mintából becslünk egy tulajdonságot egy legalább három fős sokaságra, akkor ezek az adatok nem lesznek védendőek, még akkor sem, ha becsült (és egyben a tájékoztatott) adatok egybeesnek valamely mintaelemmel.

– Abban az esetben viszont, ha a mintaelemeket „nyersen", mikroadat formájában szeretnénk közreadni, védelemmel kell ellátni őket. Gondoljuk csak el, hogy a szomszédunk elmeséli, hogy egy kérdezőbiztos a napi időbeosztásáról és tevékenységéről érdeklődött. Amennyiben birtokában vagyunk egy-két alapinformációnak szomszédunkról, az illető könnyen beazonosíthatóvá válik az adatbázis segítségével.

Előfordulhat, hogy az adatszolgáltatók magas száma ellenére is védenünk kell a cellát, például a kategóriák alacsony száma miatt. Olyan kérdésnél, amire igennel illetve nemmel lehet válaszolni, vagy kevés számú válaszlehetőség van – különösen, ha az adott kérdés valamely kényes, különleges dologra kérdez rá (például betegség, vallási hovatartozás, politikai vélemény) –, fokozottan figyelnünk kell. Ha ugyanis minden válaszadó azonosan, mondjuk igennel válaszol, akkor, amennyiben a többi, nem érzékeny kérdésre adott válaszból felfedünk valakit, akkor arról az egyedről olyan plusz, és érzékeny információ birtokába jutunk, aminek nem kellene tudtunkra jutnia: például, hogy milyen betegsége van, droghasználó-e vagy sem, milyen vallási közösség tagja stb. (Ez a fajta adatvédelmi probléma egyébként meglehetősen ritkán merül fel a nagy esetszámok miatt.)

Problémát okozhat az ugyanazon kiadványban vagy ugyanazon adatbázison alapuló különböző adatközlésekben a különböző táblák összeolvasásából azonosítható adatszolgáltató. Annak megoldása, hogy a keresztinformációkból ne váljon kikövetkeztethetővé az adatszolgáltató, a nagyon pontosan megtervezett és nyomon követett adatközléseken múlik.

FELHASZNÁLÓK ÉS MOTIVÁCIÓK

Az adatok felhasználói az igényelt adatok és az igénylés módja szerint jól definiálható csoportokra oszlanak. Fontos ismernünk a különböző felhasználókat, hogy tudjuk, kitől, mikor és miért várható az adatfelfedés, és milyen következményekkel számolhatunk.

[13] IV/1997. (SK 3.) KSH szabályzat a statisztikáról szóló jogszabályokból adódó feladatok végrehajtásáról X.

A felhasználók alábbi osztályozását, amelyet az OECD a legjobbnak minősített, a dán statisztikai hivatal állította össze:

1. „Farmerek" (szerződéses ügyfelek): Mindig ugyanarra a statisztikai adatra, szolgáltatásra van szükségük, általában ciklikusan. Az adatok közt nem válogatnak, csak olvassák azokat. Az igényeik kielégítése védett lekérdezés biztosításával, gyors adatátadással (például e-mailen keresztül), speciális információk előjegyzésével, valamint speciális formák (általuk elkészített egyéni táblázatok) alkalmazásával történik. A felhasználóknak ebbe a körébe tartoznak a pénzügyi szektor, a gazdálkodó és egyéb szervezetek képviselői, akik a statisztikán keresztül általában saját szakterületük alakulására kíváncsiak.

2. „Turisták" (alkalmi böngészők): Ezek a felhasználók általános statisztikai adatokat igényelnek különböző területekről, témákról. Az adatok, dokumentumok könnyű, gyors elérésében érdekeltek. Mindenképpen szükséges számukra köznapi fogalmak és nyelvezet alkalmazásával magyarázatot fűzni a számokhoz. Ők azok, aki csak „felnéznek" az Internetre, és ők teszik ki a felhasználók mintegy 15 százalékát. Közéjük tartoznak a köznapi emberek, a diákok, a sajtó mindenre kíváncsi munkatársai.

3. „Bányászok" (szakértők): Mélyre ásnak az adatokban, igénylik a minél részletesebb metszeteket. Vizsgálataikat sok és részletes információval, bizonyítékkal szeretnék alátámasztani. Egyedi adatokra is szükségük van, amelyekhez külön szerződés keretében hozzá is juthatnak. Elsősorban a kutatók és a tervezők tartoznak közéjük.

Az adatfelfedés szempontjából a farmerek felől érkező támadás a legvalószínűtlenebb. A *Bányászok* csoportba sorolható egyre több és több részletes adatkérőt viszont már potenciális támadónak kell tekintenünk. A legnagyobb – akár jó- vagy rosszindulatú – támadás a turisták, alkalmi böngészők csoportja felől érkezhet. Ugyanis a diákok, egyetemisták akár véletlenül is felfedhetnek bizalmas adatot, a média embere pedig szándékosan is kereshet bizalmas, védendő adatot.

Az adatfelfedést motiválhatják társadalmi, gazdasági, pszichológiai, politikai tényezők. Az adatvédelmi stratégia kialakításához fontos ezen tényezők ismerete.

Az azonosítási kísérletek elkövetőit két nagy csoportba sorolhatjuk a szerint, hogy elsősorban információszerzés céljából követik-e el a támadást, vagy pedig az adatgyűjtések, a statisztikai hivatal lejáratása, ilyen módon a közbizalom rontása a céljuk.

A támadás eszközei is széles skálán mozognak, a statisztikai és informatikai ismeretek alkalmazásától a nagy fokú számítógépes támogatásig. Ehhez járulnak a meglévő, mindenki számára elérhető információk, a köztudomású tények, a nyilvános adatbázisok, és az egyéb ismeretek egy témáról.

Természetesen minden egyes statisztikai felvétel esetén maghatározhatók támadási okok, motivációs célok, attól függően, hogy milyen érzékeny vagy érzékenynek vélhető kérdések szerepelnek a kérdőívben.

Egy angol tanulmány – elsősorban a népszámlálási adatok feltörése kapcsán – a következő lehetséges forgatókönyveket állapítja meg az adatfelfedési motivációkra, célokra nézve (*Elliot* [1996]):

1. Adatbázis gazdagítása népszámlálási adatokkal;

2. Adatbázis adatainak összevetése és megerősítése (lopott) népszámlálási adatokkal;

3. Egy jó újságíró sztori megírása annak bizonyítására, hogy adataink mennyire nem védettek;
4. Az állami adatgyűjtéseket és a mindenkori kormányzat hitelét rontó támadások;
5. Személyazonosító (szám vagy kulcs) ellopása;
6. Gazdasági versenytárs adatainak azonosítása.

(A példa az angol népszámlálásra vonatkozik.)

Az adatfelfedés hatása természetesen függ a behatoló céljától, a kísérlet sikerétől vagy kudarcától. Legyen azonban szó akárcsak egy adatszolgáltató személy azonosításáról, vagy egy gazdasági versenytárs adatainak megszerzéséről, az adatbiztonság mindenképpen sérül. Egy adatfelfedési kísérlet megtörténtének nyilvánosságra hozatala mindenképpen rontja a közhangulatot – akár sikeres volt a kísérlet, akár nem –, hiszen maga a tény, hogy egy behatolás (azonosítás) véghezvihető, önmagában rontja a statisztikai szolgálat hitelét. Ennek eredménye végső soron az lehet, hogy megszűnik maga a minőségi adatszolgáltatás.

ADATVÉDELMI TECHNIKÁK, TÁBLÁZATOS ÉS MIKROADAT VÉDELEM

A természetes jóindulatú állampolgári adatigény és az ezzel szemben jelentkező rosszindulatú támadások, lejáratások miatt az adatközlőknek szükségük van egy olyan stratégiára, amellyel biztonsággal közölhetnek adatokat, és minimálisra csökkentik a támadható felületet.

A magyar jogszabály által nevesített szabály, azaz a minimum 3 elem egy cellában jó és szükséges védelmi szabály, de nem minden esetben nyújt elégséges védelmet a felfedés ellen. A világ országaiban számos jól bevált technika létezik az adatfelfedés megakadályozására. Ezek alapjait mutatjuk be a következőkben.

Az adatfelfedhetőség szempontjából alapvetően kétféle tájékoztatási formát különböztetünk meg. Az első, hagyományosnak mondható forma az, amikor táblázatos formában, azaz bizonyos dimenziók (tulajdonságok) kereszthivatkozásaiban hozzuk nyilvánosságra az adatokat. A másik esetben az adatokat rekordsorosan tesszük közzé. Ez utóbbit nevezzük mikroadatoknak. Ennek megfelelően beszélünk *táblázatos*-, illetve *mikro*-adatvédelemről. A kétféle védelem alapjait tekintve hasonlít egymásra, mégis különböző technikák alkalmazását igénylik.

Kockázat

Az adatvédelem kialakításának első lépéseként felmérjük és megbecsüljük az adatközlésben felmerülő felfedési kockázatot. Ehhez ismernünk kell az adott statisztikai felvétel összes jellemzőjét: a sokaságot, az elemszámot, az esetleges mintát, a mintakiválasztás módját, a változókat, az adatokat, a főbb megoszlásokat stb. Ennek ismeretében tudunk dönteni valamelyik védelmi technika mellett, és határozhatjuk meg az adatvédelmi stratégiát.

Adatfelfedés több tényező együttes jelenléte esetén történhet meg, így a felfedési kockázat becslésére is több mód kínálkozik.

Táblázatos adatokban rejlő kockázat

A felfedés kockázata a táblázatos adatoknál párhuzamban áll a cella érzékenységének kritériumával. Az egyes cellák érzékenységének megléte és mennyisége határozza meg a kockázat mértékét. A gyakorlatban négy alapvető módszer terjedt el arra, hogy egy celláról kiderítsük, szükséges-e védeni vagy sem (*Carlson* [2002], *Merola* [2003]).

Jelölések:

n – az adott cellába tartozó adatszolgáltatók száma,
$z_1 \geq z_2 \geq ... \geq z_n \geq 0$ – az adatszolgáltatóktól származó adatok nemnövekvő rendszere,
T – a cella értéke, azaz $T = \sum_{j=1}^{n} z_j$.

Ehhez kapcsolódóan definiálunk még három értéket:

$$t_m = \sum_{j=1}^{m} z_j, \qquad r_m = \sum_{j=m+1}^{n} z_j, \qquad R_{l,m} = \sum_{j=m+1}^{m+l} z_j ,$$

ahol $1 \leq m \leq n$.

Küszöb szabály: Ha az adatszolgáltatók száma egy meghatározott M küszöbértéknél ($M \geq 1$) kevesebb, akkor a cella érzékeny. A cella biztonságosnak tekinthető, ha $n > M$.

Dominancia szabály: Érzékenynek tekinthető a cella, ha az értékét adó z-k közül m db legnagyobb összegének a T-hez viszonyított aránya meghalad egy k értéket (azaz dominánsak a cellában), ahol $0 < k < 1$. Az m és a k változtatásával alakíthatjuk a rendszerünk biztonságát: nagy m-mel és kicsi k-val nagy biztonság érhető el.

Választott m és k mellett a cella biztonságosnak tekinthető, ha

$$\frac{t_m}{T} < k .$$

Ez a szabály tulajdonképpen azt méri, hogy a legnagyobb elem vagy elemek mekkora arányban szerepelnek a teljes összegben. Ha egy elem 99 százalékát adja a cellaértéknek, akkor ezt nyílván nem szabad közölni, mivel nagyon kis hibával lehet következtetni erre az értékre. A nemzetközi gyakorlatban a két legnagyobb elem 80-85 százalékos részesedésénél már veszélyesnek tekintik a cellát. A paraméterekre nézve ez azt jelenti hogy: m=2, k=0,8 – 0,85.

p-szabály: Ez a szabály közvetlenül vizsgálja az egyes adatszolgáltatók adatainak a részvételét a teljes értékösszegben és feltételezi, hogy a támadó személy a cellát alkotó válaszadók közül kerül ki (z_i). Függetlenül az n nagyságától, $T - z_i$ –t tekinthetjük úgy is, hogy egy becslés minden egyes z_h-ra ($1 \leq h \leq n$), azaz $\hat{z}_h = T - z_i$. A felfedési kockázat mértéke ennek a becslésnek a relatív hibája: $(\hat{z}_h - z_h)/ z_h$. Minél kisebb a z_h, annál roszszabb ez a becslés. Nyílván a T-hez legközelebb álló értékek (z_1, z_2) adják a legjobb becslést, és ebben az esetben a legvalószínűbb is az adatfelfedés. Tehát ezt alapul véve kell megállapítani a kockázatot:(h=1, i=2): $(T - z_1 - z_2)/ z_1$. A szabály megköveteli, hogy ez a

relatív hiba nagyobb legyen, mint egy előre megadott $p > 0$ érték (*Cox* [1981]). Így biztonságosnak tekinthető egy cella, ha

$$\frac{r_2}{z_1} > p\,.$$

pq-szabály: Ez tulajdonképpen a p-szabály általánosítása, ahol a p-t alulról korlátozzuk egy $q \geq 0$ számmal. Tehát a $0 \leq q < p$ figyelembevételével biztonságosnak mondható a cella, ha

$$\frac{r_2}{z_1} > \frac{q}{p}\,.$$

A mikroadatokban rejlő kockázat

Számos módszer kínálkozik adataink ellenőrzésére. A szakirodalom (*Skinner–Elliot* [2002], *Carlson, M.* [2002]) alaposan tárgyalja ezeket a számítási módokat, ezek közül a legáltalánosabbat részletezzük. A módszer alapjául az egyednek a sokaságban való előfordulási gyakorisága szolgál. Első lépésként megvizsgáljuk minden egyes egyed gyakoriságát a sokaságban, kiszámítjuk az egyes egyedek, rekordok kockázatát, majd ez után kiszámítható a teljes adatstruktúra kockázata. Fontos betartani ezt a két lépcsős számítást, mivel az egyes rekordokban rejlő esetleges alacsony kockázat nem jelenti automatikusan a teljes adatstruktúra biztonságosságát. Ennek az az oka, hogy a rekordszintű kockázati valószínűségek összeadódnak.

Példa: U jelöli a teljes (véges) sokaságot, X az azonosító változók lehetséges kombinációinak összességét, J pedig a kombinációk számát. Ekkor az X például olyan elemekből fog állni, hogy „Férfi–50éves–Fogorvos”. Minden egyes ilyen (rész)sokaságnak meg kell határozni a gyakoriságát, azaz hogy hány egyed tartozik ebbe a tulajdonságkörbe. F_j a j-edik sokaság gyakorisága. I az indikátor függvényt jelöli.

$$F_j = \sum_{i \in U} I(X_i = j), \qquad j = 1,\ldots,J$$

Ebből már látható, hogy milyen gyakoriságúak az egyes sokaságok. Fontos tudni azt is, hogy az egyes gyakoriságokból hány darab van, mivel ha kétszer több egyelemű sokaság van, akkor kétszer nagyobb a felfedés kockázata is. Ha

$$N_r = \sum_{j=1}^{J} I(F_j = r), \qquad r = 1,2,\ldots\,,$$

ami a gyakoriságok gyakoriságát jelöli, akkor ez alapján fel tudjuk írni a felfedési kockázatot:

$$P = \frac{N_1}{N} = \frac{\sum_j I(F_j = 1)}{N}.$$

N a sokaság méretét jelöli. N_1 került a számlálóba, mivel a felfedés legnagyobb kockázata az egy elemű sokaságokban (N_1) rejlik. Ha N_1=0 akkor a következő legkisebb nemnulla részsokaság gyakoriságát kell tekintenünk. A P meghatározza, milyen valószínűséggel fedhetőek fel az adataink. A rendszerünktől megkövetelt biztonságától függ, hogy mikor tekintjük ezt elfogadhatónak és mikor nem. Az alacsony elemszámú sokaságok megszűntetésével természetszerűen csökkenthető a P értéke. Ennek módjáról a következőkben mutatunk be módszereket.

TÁBLÁZATOS ADATVÉDELEM

A tájékoztatás szempontjából kétféle táblázatot különböztetünk meg. Az egyik a gyakorisági (frequency), másik a értékösszeg (magnitude) tábla. A gyakorisági tábla tartalmazza az adatszolgáltatók számát, az értékösszeg tábla pedig az ezen adatszolgáltatók által szolgáltatott adatok összességét. A védendő adatok feltérképezéséhez minden egyes tájékoztatásra kerülő táblához el kell készíteni annak gyakorisági tábláját is. A következő példa ezt szemlélteti.

1. tábla

Értékösszeg tábla
Árbevétel (millió forint)

	Ipar	Mező-gazdaság	…	Összesen
1. város	124	0	…	…
2. város	236	377	…	…
Összesen	*360*	*377*	…	…

Gyakorisági tábla
Vállalatok száma

	Ipar	Mező-gazdaság	…	Összesen
1. város		0	…	…
2. város	6	1	…	…
Összesen	*7*	*1*	…	…

Természetesen a két tábla megegyezik abban az esetben, ha a tájékoztatásra kerülő adataink pont az adatszolgáltatók számát jelöli.

Ha a két táblázat nem egyezik meg, és csak az értékösszeg táblát jelentetjük meg, akkor az adatokból nem derül ki, hogy mely cellák rejtenek mindössze 1 vagy 2 adatszolgáltatót. Ez is jelent önmagában egy minimális védelmet, de a védelem kialakításánál fel kell tételeznünk, hogy ezen információ megszerzéséhez nem kell különösebb detektív képességgel rendelkezni, hiszen például a gazdasági életben a cégek tudják, hány hozzájuk hasonló van a piacon.

A példákban a sötéttel jelzett cellák értékei jelentik azokat az adatokat, amelyeket nem közölhetünk. Az egyszerűség kedvéért az értékösszeg- és gyakorisági tábla adatai egyértelműen megfeleltethetők egymásnak, és az érzékenység kritériuma az 1 vagy 2 adatszolgáltató ténye. A cél tehát az, hogy „megszüntessük” ezeket a cellákat.

Aggregálás

A módszer lényege, hogy oszlopok illetve sorok összevonásával cellákat egyesítünk, növelve ezzel az egy cellában lévő adatszolgáltatók számát (*Eurostat* [1996]).

Az összevonás alapja a következő két ismérv:

– *minőségi ismérv*: Két hasonló, vagy minimális számú hasonló dimenzióértékeket vonunk össze;
– *mennyiségi ismérv*: A skálázás alapjául vett mennyiségértékeknek állapítunk meg új határokat.

Példák a módszer szemléltetésére:

a) *A tábla méretének kicsinyítése* (minőségi ismérv)

2. tábla

Eredeti tábla

	Kék szemű	Zöld szemű	Barna szemű	Albínó (piros) szemű	*Összesen*
Férfi	12	10	**2**	6	*30*
Nő	24	**2**	6	8	*40*
Összesen	*36*	*12*	*8*	*14*	*70*

Az „Zöld szemű” és „Barna szemű” oszlopokban kis értékű cellákat találhatunk. Összevonjuk őket, feltételezve azt, hogy ezek a dimenzióértékek egy meghatározott szempont szerint összetartozónak tekinthetők. Egy érzékeny cellát tartalmazó oszlopot természetesen összevonhatunk olyan oszloppal is, amelyben nem szerepel érzékeny adat.

3. tábla

Védett tábla

	Kék szemű	Zöld és barna szemű	Albínó (piros) szemű	*Összesen*
Férfi	12	12	6	*30*
Nő	24	8	8	*40*
Összesen	*36*	*20*	*14*	*70*

b) *A karakterisztika újrakódolása* (mennyiségi ismérv)

4. tábla

Eredeti tábla

Kor	<12	12	13	14	15	16	17	18	19	20	>20	*Összesen*
Férfi	23	3	3	7	7	3	4	4	7	4	15	*80*
Nő	**2**	**2**	**1**	**1**	**1**	**2**	**2**	**2**	**1**	**1**	5	*20*
Összesen	*25*	*5*	*4*	*8*	*8*	*5*	*6*	*6*	*8*	*5*	*20*	*100*

Ennél a módszernél egy meghatározott skálázási tulajdonság alapján új intervallumokat állapítunk meg.

5. tábla

Védett tábla

Kor	<13	13-15	16-19	20 vagy <	*Összesen*
Férfi	26	20	19	15	*80*
Nő	4	5	6	5	*20*
Összesen	*30*	*25*	*25*	*20*	*100*

Ezt az adatvédelmi technikát gyakran alkalmazzák a statisztikai munkában. (Papír alapú kiadványokban előfordul, hogy nem pusztán a védelem miatt használják, hanem a táblázatok kisebb mérete miatt az összevont kategóriák áttekinthetőbbek. Az összevonás során például minden egyes korév helyett öt évenként összevont korcsoportok jelennek meg.)

Az aggregálás előnye:

– az adatok nem torzulnak, azaz a táblázat adatai a valóságot tükrözik, ami a a *legfontosabb* szempont felhasználók számára;
– könnyen megvalósítható;
– az adatbázisban léteznek olyan ismérvek (régió-megye-város stb.), amelyek a hierarchikus felépítés miatt közvetlen alapjául szolgálhatnak az eljárásnak.

Az aggregálás hátránya:

– az összevonások során a háttérbe kerülhetnek részletes tulajdonságok, vagyis *információveszteséggel* kell számolni. A fenti példában minden egyes korév helyett például csupán négy összevont korcsoport kategória jelenik meg.

Igen szemléletes példáját láthatjuk itt az adatvédelmi mérlegelésnek. Dönteni kell, hogy megengedhető-e az összes korév megjelenése, vagy 5, esetleg 10 éves korcsoportokat kell közölni, netán csak 3-4 korcsoport kategória jelenhet csak meg

Cellaelnyomás

Amennyiben adatok nagyfokú részletességgel történő közlése a cél, az egyik megoldás a cellák elnyomásának módszere.

A cellaelnyomás lényege, hogy az érzékenynek ítélt cellák tartalmát egyszerűen kitöröljük, ezt nevezzük *elsődleges elnyomásnak*. Mivel a kitörölt adat sorában lévő többi adatból, illetve az „összesen" mezőből ezt követően is egyértelműen meghatározható lenne a cella értéke, ezért a biztonság növelése érdekében további cellákat kell „elnyomni", ezt nevezzük *másodlagos elnyomásnak*. Különféle algoritmusok léteznek annak meghatározására, hogy mely cellákat kell még járulékosan kitörölni a védelem biztosításához (*Hundepool* [1999]).

Kétféle cellaelnyomás létezik:

– a cella tartalmának teljes kitörlése; illetve
– olyan intervallum megadása a cellában, amelybe a cella értéke belelesik.

Példa a módszer szemléltetésére:

6. tábla

Eredeti tábla

	Barna szemű	Kék szemű	*Összesen*
Fekete hajú	Védendő cella	3	*7*
Barna hajú	2	1	*3*
Szőke	3	3	*6*
Összesen	*9*	*7*	*16*

7. tábla

Elsődlegesen és másodlagosan elnyomott cellák

	Barna szemű	Kék szemű	*Összesen*
Fekete hajú	X	X	*7*
Barna hajú	X	X	*3*
Szőke	3	3	*6*
Összesen	*9*	*7*	*16*

	Barna szemű	Kék szemű	*Összesen*
Fekete hajú	3-6	1-4	*7*
Barna hajú	0-3	0-3	*3*
Szőke	3	3	*6*
Összesen	*9*	*7*	*16*

Amennyiben nem kívánjuk teljesen elrejteni a számokat, egyetlen tartományt adunk meg az elsődlegesen és másodlagosan elnyomott cellákra.

A cellaelnyomás előnye:

– a látható adatokat részletes felosztásban kapjuk meg, ami több információt jelent;
– a látható adatok a valóságot tükrözik, vagyis nem torzítottak;
– léteznek szoftverek, melyek optimalizálják a másodlagos cellaelnyomást.

A cellaelnyomás hátránya:

– a másodlagos cellaelnyomásokkal olyan cellák is rejtve maradnak, melyek egyébként közölhetőek lennének. Egy érzékeny cellához további 2-3 cellára kell alkalmazni a másodlagos cellaelnyomást, ennek következtében jelentősen megritkulhat a táblázat;

– bonyolult és hosszadalmas algoritmus végrehajtása szükséges ahhoz, hogy meghatározzuk azt a minimális számú törlendő cellát, amellyel a védelem még fennáll.

Kerekítés

Ennél az adatfelfedés elleni módszernél nem követeljük meg a cellaadatoktól, hogy pontosan tükrözzék a valóságot. A felfedési valószínűséget úgy is lehet csökkenteni, ha

nem szolgáltatunk a felhasználónak pontos értékeket, hanem az összes cella értékét – beleértve az „összesen" cellákat is – kerekítjük egy hozzá közel eső szintre.

A módszer legnagyobb előnye hihetetlenül egyszerű megvalósításában rejlik, a felhasználók körében mégsem arat osztatlan sikert, mivel meglehetősen bizalmatlanul kezelik ezeket az adatokkal. A felhasználóknak azonban nem szabad elfeledkezniük arról, hogy az általunk „elrejtett" értékek is hordozhatnak magukban hibákat (például mintavételi hibát).

A kerekítési folyamat:

Elsődlegesen megválasztunk egy *b* értékét, amit az egészszámú kerekítés *alapjának* nevezünk. N_{ij} jelöli az *i*-edik sor *j*-edik oszlopának cellaértékét.

Egy cellában lévő érték kerekítésének lépései (*Eurostat* [1996]):

1. Meghatározzuk azt a legnagyobb *h*-t (*szorzóérték*) amelyre teljesül, hogy $N_{ij} \geq h*b$

2. Így adódik a r_{ij} maradék: $N_{ij} = h*b + r_{ij}$, ahol $0 \leq r_{ij} < b$.

3. A maradékot kerekítjük, 0-ra vagy *b*-re. Így adódik a cella új értéke (N_{ij}'): ha $r_{ij}=0$ vagy *b*, akkor nyilvánvalóan: $N_{ij} = N_{ij}'$.

A különböző megoldási módozatok sajátosságai a *b* értékének megválasztásában rejlenek.

8. tábla

Eredeti tábla

	Kék szemű	Zöld szemű	Barna szemű	*Összesen*
Fekete hajú	1	4	0	*5*
Barna hajú	15	10	10	*35*
Vörös hajú	2	10	8	*20*
Szőke	2	6	15	*23*
Összesen	*20*	*30*	*33*	*83*

a) *Rögzített kerekítés*

A *b* értéket rögzítjük, és az előbb leírtak alapján alkalmazzuk a kerekítést.

A tábla minden egyes cellájára elvégezzük a kerekítést (példánkban legyen *b*=5) és akkor a táblázat az alábbi módon alakul.

9. tábla

Rögzített kerekítéssel védett tábla

	Kék szemű	Zöld szemű	Barna szemű	*Összesen*
Fekete hajú	0	5	0	*5*
Barna hajú	15	10	10	*35*
Vörös hajú	0	10	10	*20*
Szőke	0	5	15	*25*
Összesen	*20*	*30*	*35*	*85*

A módszer előnye, hogy egyszerűen kiszámolható, és minimalizálja a tényleges értéktől való eltérést.

b) *Véletlen kerekítés*

A *b* értéke itt is rögzített, viszont a kerekítés már nem a hagyományos módon történik. Az N_{ij} értéket *p* valószínűséggel kerekítjük lefelé, és 1–*p* valószínűséggel kerekítjük felfelé. Ez a következőt jelenti:

Ha *b*=5, akkor a kerekítés valószínűségei a maradék függvényében a következőképpen alakulnak:

N_{ij} b-vel való osztásának a maradéka	0	1	2	3	4	5
0-ra való kerekítés valószínűségei, *p*=	1	4/5	3/5	2/5	1/5	0
1-re való kerekítés valószínűségei, 1–*p*=	0	1/5	2/5	3/5	4/5	1

A *p*-t tehát egyenletesen kell megválasztani, a maradék és a *b* hányadosaként. Ily módon a valószínűséggel történő kerekítés biztosítja a módszer torzítatlanságát, azaz $E(N_{ij}')=N_{ij}$ (*Eurostat* [1996]). (Ha például a maradék 1, akkor $E(N_{ij}')=4/5(N_{ij}-1)+1/5(N_{ij}+4)= N_{ij}$)

A módszer sem biztosítja, hogy az oszlop és sorösszegek kiadják az egyes elemek összegét, mivel minden egyes elemre (beleértve az összegeket is) külön-külön végzzük el a kerekítéseket, és nem vesszük figyelembe az elem és az összegértékek viszonyát.

10. tábla

Véletlen kerekítéssel védett tábla (b=5)

	Kék szemű	Zöld szemű	Barna szemű	*Összesen*
Fekete hajú	0	0	0	*5*
Barna hajú	15	10	10	*35*
Vörös hajú	0	10	10	*20*
Szőke	0	10	15	*20*
Összesen	*20*	*30*	*35*	*85*

c) *Ellenőrzött kerekítés*

A kerekítésnek ez a fajtája annyiban különbözik a véletlen kerekítéstől, hogy járulékos ellenőrzéssel megpróbálunk eleget tenni az additivitásnak is, vagy annak, hogy a sorok és oszlopok kiadják az „összesen” mező értékeit. Ennek megvalósítására a leggyakrabban a Cox & Ernst algoritmust használják, melynek során a fel-le kerekítéseket úgy határozzák meg, hogy az kiadja a sor illetve oszlopösszegeket. (*Fischetti– Salazar-González* [1998], *Eurostat* [1996], *Ernst* [1989]).

11. tábla

Ellenőrzött kerekítéssel védett tábla

	Kék szemű	Zöld szemű	Barna szemű	*Összesen*
Fekete hajú	0	5	0	*5*
Barna hajú	15	10	10	*35*
Vörös hajú	0	10	10	*20*
Szőke	5	5	15	*25*
Összesen	*20*	*30*	*35*	*85*

Dimenziókorlátozás

Ez a védelmi módszer csak az elektronikus tájékoztatási formánál alkalmazható. Egyes tájékoztatási rendszereknél olyan formában érhetőek el az adatok, hogy a felhasználó által kiválasztott tulajdonságoknak (dimenziók) megfelelő adatokat kapja meg az adatigénylő táblázatos formában. Az adatkérő a kiválasztott tulajdonságok növelésével egyre részletesebb adatokhoz jut, és egyben növeli az azonosíthatóságot és ezzel együtt a felfedési kockázatot is. Ilyen esetben célravezető védelmi megoldás, hogy maximalizáljuk a választható tulajdonságok számát (legyen ez a szám n). Ezt az értéket úgy kell megválasztani, hogy a tulajdonságokból bármely n darabot választva sem juthassunk olyan információhoz, ami védendőnek tekinthető.

A módszer egyszerű és könnyen megvalósítható. A probléma csak az, hogy sok információ maradhat rejtve, ha az adatstruktúra egy részében kevés tulajdonság választása esetén is sok védendő adatot kapunk, és emiatt kicsire kell választanunk az n-t.

Ebből kifolyólag a gyakorlatban ezt a módszert csak „elővédelemnek” szokták alkalmazni, olyan formában, hogy egy alkalmas n választásával levágják az adathalmaz peremét (mivel itt a legvalószínűbbek az egyedi adatok), a továbbiakban felmerülő eseteket pedig lokális védelemmel látják el.

A dimenziókorlátozás klasszikus módszertanának vannak változatai, amelyekkel átfogóbb védelmet alakíthatunk ki:

Szelektív dimenziókorlátozás: Meg kell vizsgálni, hogy mi az a maximális n, amely mellett nem érhető el védendő cella. Az n-t 3-4-nél kevesebbre nincs értelme választani, még akkor sem, ha a vizsgálatok azt bizonyítják, hogy kevesebbnél kellene meghúzni a határt, mivel az elérhető adatok aránya vészesen lecsökken. A gyakorlatban megfigyelhető, hogy sokszor csak egy-két dimenziópárosítás választásával érhetőek el védendő cellák. Ezeknek a párosításoknak a letiltásával növelhetjük az n értékét.

Differenciált dimenziókorlátozás: A lekérdezett dimenziók számához különböző részletességű adatbázist párosítunk. A választott dimenziók számának növelésével csökkentjük a megjelenítendő adatok részletességét. A felhasználó természetesen egy dimenzió választása esetén kapja a legbővebb adatbázist.

MIKROADAT-VÉDELEM

Mikroadatok alatt az egy statisztikai egységről birtokunkban lévő legrészletesebb adatokat értjük. Ezek az adatok a gyakorlatban rekordsoros állományokban vannak eltárolva**,** az *egy sor egy adatszolgáltató* elv alapján.

A jogszabályok alapján anonimizált mikroadatok olyan egyedi statisztikai adatok, amelyeket annak érdekében módosítottak, hogy a mindenkori legjobb eljárással összhangban minimálisra csökkenjen az érintett statisztikai egységek azonosításának veszélye.

Ilyen adatállományok teljes vagy részleges publikálása is csak az megfelelő anonimizálás után tehető meg. A jogi részben megismertek alapján az egyedi azonosítók esetében nincs mérlegelési jogkörünk azok megtartására, egyszerűen *ki kell* törölni őket. A további vizsgálataink tárgyát a fennmaradó oszlopok képezik.

12. tábla

Személyek adataiból álló mikroadatbázis

Név	Lakhely	Születési hely	Születési idő	Foglalkozás	Vallás	…
Kala Pál	Iszapszentmotoros	Iszapszentmotoros	1881.01.02.	Tűzkő árus	–	…
Hó Virág	Tápiórettentő	Tápiórettentő	2031.02.12.	Tűzoltó	Szombatista	…
…	…	…	…	…	…	…

13. tábla

Gazdasági szervezetek adataiból álló mikroadatbázis

Cégnév	Telephely	Alapítás dátuma	Tevékenység	Alaptőke (millió forint)	…
Gépolaj Rt.	Markotabödöge	1844.06.12.	Szállítmányozás	234	…
Sikattyu Kft.	Nagybajom	2021.12.22.	Költöztetés	133	…
…	…	…	…	…	…

Látható, hogy a mikroadatok esetében sokkal szorosabb kapcsolat van az *azonosítás* és a *felfedés* között, mint a táblázatos adatoknál. Ezért beszélünk itt anonimizálásról, nem pedig felfedésről: a speciális rekordsorban szereplő adatok miatt az azonosítás itt önmagában felfedést is jelent. Tehát itt nem az érzékeny adatok elrejtésén van a hangsúly, hanem a rekordnak az egyénhez való társításának megakadályozásán. Ahhoz, hogy kicsi legyen a kockázat, csökkenteni kell a legkisebb gyakoriságú részsokaságok számát.

Az anonimizálás itt is tartalmaz bizonyos fokú információveszteséget, de a védelmi technikáknál éppen az a célunk, hogy megtaláljuk azokat az adatokat amelyek elrejtésével a legkevesebb az információveszteség, és közben az anonimitásnak is eleget teszünk.

Csonkolás

Ez a technika a legnyilvánvalóbb és egyben első helyen alkalmazott. Csonkolásnál egy teljes oszlopot kitörlünk az adatbázisból. Ezt a módszert alkalmazzuk akkor is, amikor az egyedi azonosítókat leválasztjuk az adatbázisról.

14. tábla

Védelem kialakítása csonkolással

Születési hely	Szül.idő	Foglalkozás	Vallás	…
XXX	1881.01.02.	Tűzkő árus	–	…
XXX	2031.02.12.	Tűzoltó	Szombatista	…
XXX	…	…	…	…

A fő probléma annak eldöntésében rejlik, hogy mely oszlop kitörlésével érhetjük el a kellő anonimitást. Ennek eldöntésére meg kell vizsgálnunk, hogy vannak-e olyan tu-

lajdonság- kombinációk (például: „Születési hely” és „Születési idő”) amelyek egyedivé teszik az egyes vagy akár az összes rekordokat. Ezek azok az oszlopok potenciális jelöltjei a csonkolásnak. A csonkolási technika alkalmazása egy ciklikus folyamat. Minden egyes lépésnél csakis egyetlen oszlopot szabad kitörölni, és ezután újra meg kell vizsgálni, mely rekordok maradtak még továbbra is kritikusak a felfedés szempontjából.

A csonkolási technika adatbázisok védelménél igen durva beavatkozásnak számít, hiszen hatására dimenziók tűnnek el. Mivel a tájékoztatás célja, hogy minél több információt biztosítsunk a felhasználóknak, így a mikroadatbázis egészénél a csonkolás mellett gyakran más módszereket is alkalmaznak.

Cellaelnyomás

A cellaelnyomás során egyes tulajdonságok „vészesen kevés számú” előfordulásait kell kitörölni. A tulajdonságok vészesen kevés előfordulásai során arra kell gondolni, hogy az adatbázis információi egyediek, így közvetlenül beazonosítható az adatszolgáltató. A minőségileg egyedi és a mennyiségileg kevés vagy kiugróan sok elemszám teszi kritikussá, azonosíthatóvá a rekordot, és így az adatszolgáltatót is.

Példa: Ha Iszapszentmotoroson csak egy tűzkőárus van, akkor ez egyértelmű azonosítást, közvetett adatfelfedést tesz lehetővé. Tehát itt a lakhely és a foglalkozás kombinációja kritikus a védelem szempontjából. Bármelyik cella rejtetté tétele megoldja a problémát. A döntés a védelmet kialakító egyéntől függ, illetve attól, hogy lakhelynek vagy a foglalkozásnak a kombinációja fontos-e a többi adatéval összevetve.

15. tábla

Védelem kialakítása cellaelnyomással

Születési hely	Születési idő	Foglalkozás	Vallás	…
Iszapszentmotoros	1881.01.02.	XXX	–	…
Tápiórettentő	2031.02.12.	Tűzoltó	Szombatista	…
…	…	…	…	…

Ez a módszer enyhébb, mintha csonkolással eltávolítottuk volna az egész foglalkozási vagy vallási oszlopot, de tény, hogy ez is adatvesztéssel jár.

Átkódolás

Az adatszolgáltatók kilétével kapcsolatos bizonytalanság kialakítható úgy is, ha nem az általunk ismert legpontosabb adatot írjuk a cellába. Ez nem az jelenti, hogy a cellák nem valós értékeket tartalmaznak, hanem csak annyit, hogy egy bővebb tartományba helyezzük át a tulajdonságot, tulajdonságokat.

Példa: Ha egy városban csak egy balettcipő-készítő van, akkor érdemes összevonni a cipőkészítővel. Ennek megfelelően a cipőkészítőt és a balettcipő-készítőt át kell írni „Cipő- és balettcipő-készítő”-re. A következő két példa talán még szemléletesebbé teszi az elvet.

16. tábla

Védelem kialakítása átkódolással

Születési hely	Születési idő	Foglalkozás	Vallás	…
Iszapszentmotoros	1881.01.02.	Tűzoltó-Tűzkőárus	–	…
Tápiórettentő	2031.02.12.	Tűzoltó-Tűzkőárus	Szombatista	…
…	…	…	…	…

Alapítás dátuma	Tevékenység	Alaptőke (millió forint)	…
1844.06.12.	Szállítmányozás-Költöztetés	234	…
2021.12.22.	Szállítmányozás-Költöztetés	133	…
…	…	…	…

Kerekítés

A táblázatos adatok védelménél bemutatott kerekítés, teljesen azonos módon alkalmazható számértékű mikroadatoknál is. A *b* érték növelésével növekszik a bizonytalanság a tényleges értékre, viszont ezzel arányosan sajnos növekszik az adat használhatatlansága is. A legnagyobb feladat az optimális *b* választásában rejlik, amely mindig függ a kerekítendő szám nagyságrendjétől, illetve attól, hogy az egyes értékek milyen tartományban mozognak. Nyílván lényegesen eltérő *b*-t kell választani milliós, illetve ezres nagyságrendű értékeknél. Ügyelni kell arra is, hogy ne forduljon elő az, hogy a kerekítést követően olyan számot kapjunk, amelyet az egyébként jellemzett tulajdonság fel sem vehet.

17. tábla

Védelem kialakítása kerekítéssel (b=50 millió forint)

Alapítás dátuma	Tevékenység	Alaptőke (millió forint)	…
1844.06.12.	Szállítmányozás	250	…
2021.12.22.	Költöztetés	150	…
…	…	…	…

Összekeverés

A felfedési kockázatot azzal is tudjuk csökkenteni, hogy az egyes emberekhez tartozó érzékeny adatokat véletlenszerűen összekeverjük. Így az egyes emberek adatain nincs mit felfedni, viszont a sokaságok egészére nézve nem változott semmi.

18. tábla

Védelem kialakítása összekeveréssel

Születési idő	Lakhely	Foglalkozás	Vallás	…
…	…	…	…	…
1881.01.02.	Iszapszentmotoros	Tűzkő árus	–	
2031.02.12.	Tápiórettentő	Tűzoltó	Szombatista	…
…	…	…	…	…

*

Tanulmányunkban bemutattuk az adatfelfedés elleni védelem környezetét, valamint a a védelem különböző eszközeit. Mi ezen eszközök „tiszta" bemutatására törekedtünk, és olyan példákat hoztunk, amelyek megfelelően reprezentálták az eszközök bemutatását. A gyakorlati statisztikai munkában azonban komoly háttérmunkát jelent annak eldöntése, hogy mely módszer vagy módszerek alkalmazása a legcélravezetőbb egy adott statisztikai felvétel közlésekor; a megvalósítás pedig összehangolt statisztikai-matematikai-informatikai eszköztárat igényel.

Egyes módszerek információ-veszteséggel vagy információ-torzulással járnak. Mint korábban jeleztük, az adatvédelmi technikák alkalmazása egy finom mérleghez hasonlít, ahol azt kalkuláljuk, hogy a tájékoztatás formájának célja, módja milyen védelmi technikát igényel, azaz mit nyerünk az egyik oldalon és mit veszítünk a másikon, s hogyan kerül a kettő egyensúlyba.

A tájékoztatás során a felhasználót mindig tájékoztatják az adatvédelemről. Ez történhet úgy, hogy egy kerekítésnél megadják a kerekítés mértékét, lábjegyzetben vagy mellékletben jelezik, hogy milyen eljárással módosították az adatokat, mikroadat-állomány közlése során pedig csatolnak az állományhoz egy leírást az adatvédelem módjáról és hatásáról. Egyes országok nem közlik pontosan az alkalmazott adatvédelmi technikát, csak a pontosság és megbízhatóság mértékét, illetve, hogy az adatok milyen korlátok és feltételek között alkalmazhatóak.

A védelmi eljárás alkalmazásának tényéről azonban mindig történik tájékoztatás. Ez a nemzetközi gyakorlat is, hiszen ez az etikai lépés biztosítja a kölcsönös bizalmat, és a korrekt statisztikai munkát az adatközlő és a felhasználó oldalán is.

A célhoz, a minél szélesebb körű tájékoztatáshoz kezünkben vannak tehát az eszközök, amelyek alkalmazásával elégedett lehet mind a tájékoztató statisztikai intézmény, mind az adatigénylő felhasználó.

IRODALOM

BÁNSZEGI K. [1997]: Felfedést akadályozó módszerek a statisztikai tájékoztatásban. *Statisztikai Szemle*. 75. évf. 12. sz. 1039–1046. old.

BÁNSZEGI K. – LAKATOS M. [1994]: Információszabadság – adatvédelem – statisztika (III.). *Statisztikai Szemle*. 72. évf. 10. sz. 761–777. old.

CARLSON, M. [2002]: Assessing microdata disclosure risk using the poisson inverse gaussian distribution. Stockholm. Kézirat. (http://www.matstat.umu.se/banocoss/papers/carlson.pdf)

COX, L. H. [1981]: Linear sensitivity measure in statistical disclosure control. *Journal of Statistical Planning and Inference*. 5. évf. 2. sz. 153–164. p.

DUNCAN, G. T. – KELLER-MCNULTY, S. A. – STOKES, S. L. [2001]: *Disclosure risk vs. data utility: The R-U confidentiality map*. Kézirat. (http://www.niss.org/technicalreports/tr121.pdf)

ELLIOT, M. [1996]: Attacks on census confidentiality using the sample of anonymised records: an analysis. 3rd International Seminar on statistical confidentiality. Bled 1996.

ERDEI V. – SÁNTA J. [2000]: *A statisztikai adatok védelmének nemzetközi szabályozása, módszertani kérdései.* Népszámlálások az ezredfordulón 3. *(Tanulmányok)* Központi Statisztikai Hivatal. Budapest.

ERNST, L. R. [1989]: Further application on linear programming to sampling problems. Kézirat. (http://www.census.gov/srd/papers/pdf/rr89-05.pdf)

Eurostat [1996]: *Manual on disclosure control methods*. Luxemburg.

Eurostat [1999]: Statistical data confidentiality.

FAGAN, J. T. – GREENBERG, B. V. – HEMMING, B. [1988]: *Controlled rounding of three dimensional tables.* Kézirat. (http://www.census.gov/srd/papers/pdf/rr88-02.pdf)

FISCHETTI, M. – SALAZAR-GONZÁLEZ, J. J. – Caprar, A. [1998]: *Computational experience with the controlled rounding problem in statistical disclosure control*. Padova. Kézirat. (http://neon.vb.cbs.nl/casc/ISIBerlin/Salazar.pdf)

FISCHETTI, M. – SALAZAR-GONZÁLEZ, J. J. [1998]: Experiments with controlled rounding for statistical disclosure control in tabular data with linear constraints. *Journal of Official Statistics*. 4. évf. 4. sz.

HUNDEPOOL, A. [1999]: Statistical disclosre limitation in practice. Kézirat. (http://europa.eu.int/en/comm/eurostat/research/conferences/etk-99/papers/hundepool.pdf)

LAKATOS M. [1994]: Információszabadság – adatvédelem – statisztika (I.). *Statisztikai Szemle*. 72. évf. 7. sz. 547–559. old.
MEROLA, G. [2003]: *Generalized risk measure for tabular data*. Roma. Kézirat. (http://neon.vb.cbs.nl/casc/ISIBerlin/merola.pdf)
SKINNER, C.J. – ELLIOT, M. J. [2002]: *A measure of disclosure risk for microdata*. Kézirat. (http://www.ccsr.ac.uk/publications/occasion/occ23.pdf)
Statistical Journal of the United Nations ECE [2001]. Data confidentilaity. 285–407. old.
Statisztikai igazgatás [2000]. (Közigazgatási szakvizsga tankönyv). Budapest.

SUMMARY

This study is about the statistical tools on data confidentiality and the background of the confidentiality issue. It gives an overview on the European and Hungarian legislation, the different and new forms of dissemination, and the confidentiality problems. It shows the different risks on micro and tabular data dissemination and the possible statistical tools of protection with examples.