Kutatói tájékoztató

Útmutató a KSH kutatószobai környezetében folyó kutatómunkához

Központi Statisztikai Hivatal

2015. december

Kutatói tájékoztató

Tartalomjegyzék

Kutatói tájékoztató célja ... 3

1. A kutatás folyamata és az adatvédelem jogi háttere ... 3

1.1. A kutatás folyamata ... 4

1.2. A kutatási eredmények adatvédelmi ellenőrzését végző szakértő feladata ... 5

1.3. A kutató feladata ... 6

1.4. Jogi háttér ... 7

2. A kutatás dokumentációja ... 7

2.1 A fájlok tárolására szolgáló könyvtárstruktúra ... 7

2.2 Dokumentáció programfájlokban, változók felcímkézése ... 11

2.3 A kutatás dokumentációjának egyéb lehetőségei ... 15

3. A kutatási eredmények utólagos adatvédelmi szempontú ellenőrzésére vonatkozó módszertani eljárások és szabályok ... 16

3.1. A KSH gyakorlatában alkalmazott adatvédelmi szabályok ... 16

3.2. Az adatvédelmi módszerek csoportosítása és a KSH gyakorlata ... 17

3.3. Táblázatos adatok védelme ... 20

3.3.1. Gyakorisági táblák adatvédelme ... 22

3.3.2. Értékösszegtáblák adatvédelme ... 23

4. Egyéb adatvédelmi eljárások ... 26

4.1. Teljeskörű felvételekre és mintabeli sokaságokra vonatkozó eljárások ... 27

5. Tanácsok kutatók számára biztonságosnak ítélhető kutatási eredmények előállításához ... 28

Melléklet ... 32

Kutatói tájékoztató célja

Jelen dokumentum célja, hogy eligazítást nyújtson a kutatók számára a Központi Statisztikai Hivatal (KSH) kutatószobai hozzáférésével kapcsolatban felmerülő jogi, adminisztratív, technikai, és módszertani adatvédelmi kérdésekben. A kutatói tájékoztató kitér a biztonságos környezetbeli kutatási szabályokra, kiemelt figyelemmel a biztonságos környezetben létrehozandó kutatási eredmények szabályaira és azok utólagos adatvédelmi szempontú ellenőrzésére.

A tájékoztató, ezekre az igényekre válaszként, elsősorban a kutatók és adatvédelmi szakértők munkáját kívánja segíteni, a kutatási eredmények adatvédelmi szempontú ellenőrzésének gördülékenyebbé tételéhez. Jelen kutatószobai tájékoztató ezen felül hasznos információkkal szolgálhat mindazoknak, akik a KSH adathozzáférési csatornáival kapcsolatban felmerülő adatvédelmi kérdésekről kívánnak tájékozódni, illetve a KSH biztonságos környezetében létrejövő kutatási eredmények adatvédelmi szempontú ellenőrzésének folyamatában általános betekintést nyerni.

1. A kutatás folyamata és az adatvédelem jogi háttere

A Központi Statisztikai Hivatal statisztikai célú adatkezelést végez, egyedi adatokat harmadik fél részére – törvény által nevesített kivételektől eltekintve – nem ad át. Az egyedi adatok védelmére vonatkozó törvényi kötelezettség teljesítése és adatszolgáltatóink bizalmának megőrzése érdekében fokozott figyelmet fordítunk az egyedi adatok védelmére. A Hivatal statisztikai adatszolgáltatási kötelezettségével összhangban feladatunknak tekintjük, hogy a KSH által kezelt statisztikai adatokat tájékoztatásra, továbbá – a tudományos előrehaladás érdekében – tudományos célú kutatások támogatására használjuk fel.

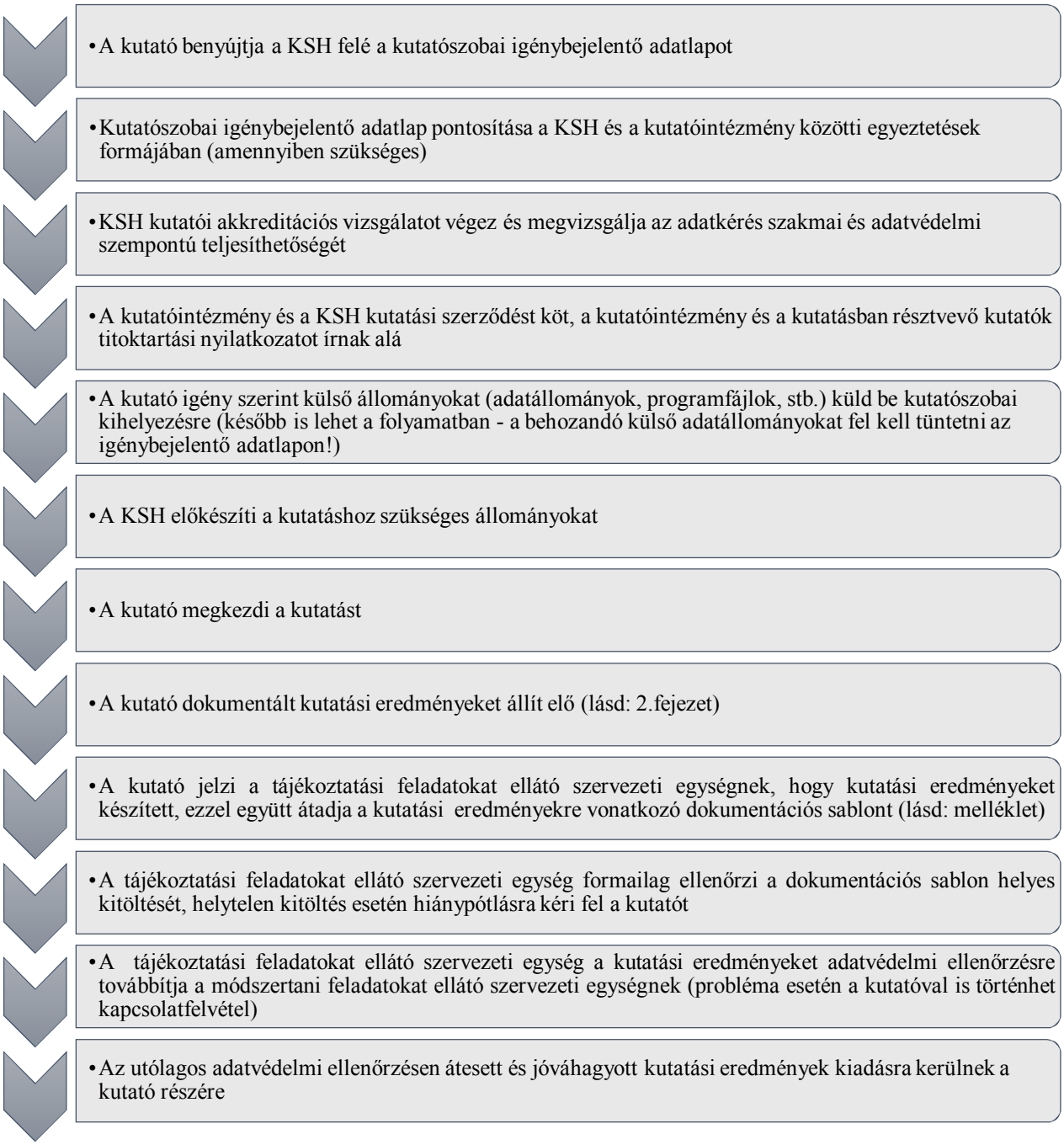
A KSH többféle csatornán biztosít hozzáférést statisztikai adatokhoz az érdeklődők számára. Ezek közül a tudományos célú adathozzáférési csatornák kifejezetten kutatók számára teszik lehetővé kutatásra előkészített statisztikai adatállományokhoz való hozzáférést. Az egyedi adatok védelme érdekében tett jogi garanciák, és módszertani intézkedések mellett ebben az esetben a fizikai adatvédelem is különös hangsúlyt kap.

A tudományos célú adathozzáférési csatornák között a kutatószobai hozzáférést, a távoli hozzáférést, a távoli végrehajtást és az anonimizált mikroadat-kiadást különböztetjük meg. Jelen kutatószobai tájékoztató a kutatószobai hozzáférésre összpontosít. A Kutatószobában szigorúan ellenőrzött körülmények között, a KSH belső hálózatától, az ott tárolt adatoktól és a külső hálózati kapcsolatoktól (internet) technikailag elkülönített környezetben történik az adatokhoz

való hozzáférés. Az egyedi adatok védelmének törvényi kötelezettsége értelemszerűen a Kutatószobából kivitt kutatási eredményekre is kiterjed.

1.1. A kutatás folyamata

Az alábbi ábra szemlélteti a kutatószobai kutatási folyamat legfontosabb lépéseit a kutatószobai igénybejelentő adatlap benyújtásától a kutatási eredmények adatvédelmi ellenőrzését követő adatkiadásig.



1. ábra Mikroadatokhoz való hozzáférés folyamata biztonságos környezetben

A **1. ábra** jelzett munkamenet gördülékeny haladása érdekében:

- ✓ a kutató egyértelműen jelzi a tájékoztatási feladatokat ellátó szervezeti egységnek, amikor kutatási eredményeket kíván kivinni, és egyúttal arról is információt ad, hogy mely mappába helyezte el a kutatási eredményekre vonatkozó dokumentációs sablont.
- ✓ a kutató a KSH által előírt könyvtárstruktúrát alkalmazza az eredményei és munkafájljai tárolására (lásd 2.1 A fájlok tárolására szolgáló könyvtárstruktúra.)
- ✓ a kutató a kutatói tájékoztató alapján minden segédtáblát elkészít.

Ha olyan kutatási eredményeket készít a kutató, melyek csak nagymértékű információveszteség árán lennének kiadhatók, akkor a KSH értesíti a kutatót a problémáról, illetve javasolja, hogy az eredményeket valamilyen alternatív módszerrel, esetleg aggregáltabb formában állítsa elő.

1.2. A kutatási eredmények adatvédelmi ellenőrzését végző szakértő[1] feladata

Az adatvédelmi szakértő fő feladata a *kutatási eredmények adatvédelmi ellenőrzése* (*output checking*), a felfedés elleni védelem biztosítása. Hangsúlyozzuk, hogy a vizsgálat célja *nem* a kutatási módszertan *helyességének ellenőrzése*, hanem annak megállapítása, hogy a kiadott adatok *adatvédelmi szempontból* megfelelnek-e a jogszabályi [2] és módszertani követelményeknek és alapelveknek.

Az adatvédő tehát a kutatási eredményeket *kizárólag adatvédelmi szempontból* ellenőrzi. Sem az adatvédelmi szakértőnek, sem a KSH-nak nem feladata, és nem is kötelessége a kutatási eredmények helyességét igazolni vagy cáfolni.

Az *adatvédelmi szakértő* feladata az alábbiakban foglalható össze:

- minden, a kutatók által beküldött külső állományt módszertani adatvédelmi szempontból ellenőriz;
- minden egyes kutatási eredményt átvizsgál módszertani adatvédelmi szempontból, és írásbeli véleményt alkot arról, hogy a kutatási eredmény eredeti formájában kiadható-e;
- módszertani szempontból védetté teszi a kutatási eredményt, amennyiben a felfedés kockázata magas;

[1] Továbbiakban adatvédelmi szakértő vagy adatvédő.

[2] Lásd 1.4-es fejezetet

- indokolt esetben felfedés elleni védelmi módszereket alkalmaz, majd elérhetővé teszi a kiadhatónak ítélt kutatási eredményeket a tájékoztatási feladatokat ellátó szervezeti egységnek;
- a kutatási eredményekkel kapcsolatos adatvédelmi kérdésekben szükség esetén egyeztet a kutatóval

1.3. A kutató feladata

A kutatás megtervezése és az alkalmazott elemzési módszerek helyes megválasztása a kutató feladata, így a kivitt és publikált eredmények szakmai tartalmáért kizárólagosan a kutató felel.[3]
A kutató köteles betartani a kutatószobai hozzáférés szabályait, melyek a KSH honlapján érhetők el. Ezen felül köteles a következő módon elősegíteni az adatvédelmi ellenőrzés folyamatát:

- Minden behozni kívánt külső állományhoz megadja a változók metainformációit is (külső adatállományban szereplő változók neve, pontos tartalma és az ezekhez kapcsolódó, az állomány tartalmának megértését szolgáló leírások).
- Tudomásul veszi, hogy a Kutatószobában a KSH adatvédelmi okokból nem tesz elérhetővé olyan állományt, amely közvetlen azonosítót[4] tartalmaz.
- Minden kivitelre szánt értékösszegtáblához egy külön állományba elkészíti a hozzá tartozó *abszolút gyakorisági táblát* (lásd 3. táblázat).
- Minden értékösszegtáblában szereplő értékhez megadja *a legnagyobb hozzájáruló értékösszegben való részesedését* (százalékos arányát) cellánként, szintén egy külön fájlban (lásd 4. táblázatt).
- A kutatási eredményben szereplő *újonnan képzett* vagy *átnevezett* változókat *definiálja* és *felcímkézi* (lásd 2.2 Dokumentáció programfájlokban, változók felcímkézése).
- Ésszerű mértékű *dokumentációt, magyarázatot biztosít* a kutatási eredményhez (lásd 2.2 Dokumentáció programfájlokban, változók felcímkézése).

[3] A kutatások szakmai kérdéseibe tehát semmilyen módon nem szólunk bele, a kutatók ebből a szempontból *teljes szabadságot* élveznek.

[4] **Közvetlen azonosító**: a statisztikai egységhez rendelt egyedi azonosítási kód, illetve a statisztikai egység megnevezése/neve, a statisztikai egységhez tartozó pontos címadat (pl. lakcím, székhelycím, telephelycím, stb.), valamint a statisztikai egység elérhetőségi adatai (pl. e-mail cím, telefonszám, stb.).

Amennyiben a kutatási eredmények adatvédelmi szempontú ellenőrzése indokolja, akkor – az adatvédelmi szempontú ellenőrzés elvégzése céljából – az adatvédelmi szakértő az *eredmények további értelmezését* (szóban és/vagy írásban), a kutatási eredményekre vonatkozó *dokumentáció kiegészítését* és az alkalmazott *módszer részletesebb leírását* kérheti a kutatótól.

1.4. Jogi háttér

A KSH–ban a statisztikai adatgyűjtés az *adatszolgáltatók bizalmán* alapul. Ezért kiemelten fontosnak tartjuk az egyedi adatok védelmét, amit a kutatószobai munkafolyamat minden fázisában is biztosítunk. A jogi háttérre vonatkozó információk a KSH honlapon elérhetők.

2. A kutatás dokumentációja

A kutatási eredmények megértése elengedhetetlen az adatvédelmi ellenőrzés elvégzéséhez, ezért a kutató köteles az ebben a fejezetben leírtaknak megfelelően elvégezni a kutatás dokumentációját. Amennyiben a dokumentáció nem a leírtaknak megfelelően történik, vagy hiányos, úgy a kutatási eredmény ellenőrzését nem tudjuk elvégezni. Ebben az esetben a kutatót hiánypótlásra kérjük fel.

2.1 A fájlok tárolására szolgáló könyvtárstruktúra

A kutatáshoz dedikált tárhelyen (*X:\Kimeno\<Felhasználónév>*) az alábbiak szerint kell tárolni az adatvédelmi ellenőrzéshez szükséges állományokat.

- Minden kutatáshoz, még a kutatási folyamat megkezdése előtt a KSH hozzárendel egy kutatási felületet, annak sorszámával együtt (pl. az adott kutatáshoz hozzárendeljük a „kutato17” felületet).
- Minden kutató, aki részt vesz ebben a kutatásban, ugyanazzal a felhasználónévvel (a kutatási felület neve, pl. „kutato17”) és jelszóval tud belépni az adott kutatás keretében. Egy kutatási felületet egyszerre csak egy kutató használhat.
- Egy kutatási felületen a kutatásban részt vevő kutatók tehát láthatják, módosíthatják, sőt, törölhetik is egymás munkáit.
- A kutató köteles a kutatáshoz rendelt felületen egy saját nevével elnevezett mappát létrehozni (pl. Kovacs_Istvan).
- Minden alkalommal, amikor a kutató kutatási eredményt szeretne kivinni, ebben a mappában egy alkönyvtárat hoz létre *’<Felhasználónév>_<Vezetéknév>_<Keresztnév>_<Dátum>’* névvel. Ennek az

alkönyvtárnak a nevében a <*Felhasználónév*> helyén értelemszerűen az adott kutatási felület neve, a <Vezetéknév>_<Keresztnév> helyén a kutatási eredményt beadó kutató vezeték- és keresztneve áll ékezetek nélkül, a <*Dátum*> helyén pedig a kutatási eredmény beadásának dátuma szerepel *év, hónap, nap* sorrendben, elválasztó karakterek nélkül.

- Ebben a könyvtárban a kutató további három mappát hoz létre a kutatási eredmények és a hozzájuk kapcsolódó fájlok tárolására a következő elnevezésekkel: *'Kutatasi_eredmenyek'*; *'Munkafajlok'*; *'Adv_ellenorzeshez'*.

Példa:

Tegyük fel, hogy a *kutato17* felületen lévő kutatásban résztvevő *Kovács István* nevű kutató kutatási eredményt állított elő, amelyet 2014. október 30-án adatvédelmi ellenőrzésre ad át. A kutató a kutato*17* mappában hozzon létre egy 'kutato17_Kovacs_Istvan_20141030' elnevezésű alkönyvtárat. Ezen belül, egymással egy szinten hozza létre a '*Kutatasi_eredmenyek', 'Munkafajlok' és 'Adv_ellenorzeshez'* mappákat.

A fent leírtak az alábbi *hierarchia* szerint néznek ki:

- *kutato17*
    - *Kovacs_Istvan*
        - *kutato17_Kovacs_Istvan_20141030*
            - *Adv_ellenorzeshez*
            - *Kutatasi_eredmenyek*
            - *Munkafajlok*
            - *kut_eredm_dok_Kovacs_Istvan_20141030.docx*

Felhívjuk a figyelmet arra, hogy az adatvédelmi ellenőrzést a fenti módon definiált könyvtárakba helyezett állományok alapján végezzük. A kutató ettől eltérő struktúrát is létrehozhat az állományainak tárolására, de mivel azok adatvédelmi szempontú ellenőrzésen nem esnek át, így kiadásra sem kerülhetnek.

- A '<**Felhasználónév> <Vezetéknév> <Keresztnév> <Dátum>**' mappa tartalmaz minden olyan fájlt, amit a kutató a megadott dátum szerinti kutatási eredményekhez felhasznált, illetve eredményül kapott és ki szeretne vinni (programkódok,

segédállományok, adatvédelmi ellenőrzéshez szükséges állományok, valamint maguk a kivitelre készített kutatási eredmények és a kitöltött dokumentációs sablon)

- A '**Kutatasi_eredmenyek**' mappa tartalmaz minden olyan fájlt, amit a kutató *ki szeretne vinni* (táblázatok, regressziós eredmények, programkódok, szöveges dokumentumok, stb.).
- A '**Munkafajlok**' nevű könyvtár minden olyan adatállomány (munkaállomány) tárolására szolgál, amely szükséges volt ahhoz, hogy a '*Kutatasi_eredmenyek'*-ben szereplő állományok előálljanak, de a kutatónak *nem áll szándékában kivinni* őket (azt, hogy melyik kutatási eredmény közvetlenül mely adatállományból származik, a dokumentációs sablonon kell megadni.) Ebbe a mappába kerülnek azok a programfájlok is, amelyek a KSH által biztosított nyers állományokból és/vagy a kutató által behozott külső állományokból előállítják a kutatási eredmények inputjául szolgáló munkaállományokat.
- Az '**Adv_ellenorzeshez'** mappa tartalmazza az alábbiakat:
  - minden értékösszegtáblához a hozzá tartozó *abszolút gyakorisági táblát*
    Ezek elnevezése annyiban különbözik az eredeti értékösszegtábla elnevezésétől, hogy a fájlnév a „**gyak_**" előtagot kötelezően tartalmazza
  - minden értékösszegtáblához a hozzá tartozó *legnagyobb hozzájárulások arányának százalékos értékét*[5]
    Ezek elnevezése annyiban különbözik az eredeti értékösszegtábla elnevezésétől, hogy a fájlnév a „**leg_hozz_**" előtagot kötelezően tartalmazza
  - a kutatási eredmények előállításához közvetlenül szükséges programkódokat, és egyéb kapcsolódó programkódokat
  - a kutató által képzett, a kutatási eredményekben szereplő változók definícióját könnyen előkereshető formában (pl. a változókat lehet definiálni a programfájl kommentjében, szöveges dokumentumban, Excel táblázatban, stb., az adatállományokban pedig fel kell címkézni őket. A cél, hogy az adatvédelmi

[5]Azaz minden cella esetén fel kell tüntetni a cellában szereplő értékösszeghez legnagyobb mértékben hozzájáruló egyed részesedését százalékban kifejezve.

ellenőrzéshez minden szükséges információ elérhető legyen, és könnyen meg lehessen találni egy-egy változó jelentését, képzésének módját)

- minden egyéb olyan dokumentumot, táblázatot, amely az adatvédelmi ellenőrzéshez segítséget nyújthat (pl. szöveges leírások a kutatás céljáról, menetéről, felépítéséről, milyen sorrendben keletkeztek a kutatási eredmények, naplófájl, stb.)

A kutatási eredmények adatvédelmi ellenőrzésének megkezdéséhez szükséges a dokumentációs sablon (lásd melléklet) kitöltése. A kitöltött dokumentációs sablont *kut_eredm_dok_<Vezetéknév>_<Keresztnév>_<Dátum>.docx* módon kell elnevezni, ahol a *<Vezetéknév>*, illetve *<Keresztnév>* helyén értelemszerűen a kutató vezetékneve, illetve keresztneve szerepel, a *<Dátum>* helyén pedig a kutatási eredmény leadásának dátuma a korábban leírtak szerint.

Példa:

➔ *kutato17*

- *Kovacs_Istvan*
  - *kutato17_Kovacs_Istvan_20141030*
    - *Kutatasi_eredmenyek*
      - adatelokeszites.dta
      - merleg_input.sav
      - jegyzetek.docx
      - valtozoleirasok.xlsx
      - kutatas.do
      - eredmenyek_1.smcl
      - …
    - *Munkafajlok*
      - adattisztitas.do
      - leiro_statisztikak.sps
      - export_panel.dta
      - ipari_termeles2.sav
      - …
    - *Adv_ellenorzeshez*
      - gyak_adatelokeszites.dta

— gyak_merleg_input.sav
— …
— leg_hozz_adatelokeszites.dta
— leg_hozz_merleg_input.sav
— …
— program_merge.do
— program_regressions.sps
— program_merge.smcl
— ...
— *magyarazat.docx*

- kut_eredm_dok_Kovacs_Istvan_20141030.docx

2.2 Dokumentáció programfájlokban, változók felcímkézése

A kutatói munkafolyamat és a kutatási eredmények megfelelő dokumentációja az adatvédelmi szakértő számára könnyebben követhetővé, érthetőbbé teszi a kutatási eredményeket, ami jelentősen felgyorsítja az adatvédelmi szempontú ellenőrzés folyamatát. A kutatás dokumentálását a programfájlokon belül az alábbiakban felsorolt pontok figyelembe vételével kérjük elvégezni:

1. Ésszerű mértékben magyarázó megjegyzésekkel (kommentek) legyen ellátva a programfájl. Az *elemzés logikája*, felépítése legyen *követhető* (pl. adatállományok összeállítása, leíró jellegű elemzések, analitikus eredmények).

   A programfájl azon részeinek dokumentációjára kérjük a hangsúlyt tenni, amelyek a *kutatási eredményekhez közvetlenül*, vagy azok előállítása révén közvetett módon, de még szorosan kapcsolódnak.

   A programfájl elején kérjük feltüntetni, hogy a kutató a KSH által a kutatási felületen elérhetővé tett állományok közül melyeket használta fel (közvetlen vagy közvetett módon) a kutatási eredmények előállításához.

   A programfájl elején kérjük tömören azt is leírni, hogy mi az *elemzés célja*, milyen állományokon dolgozott, milyen eredményekre jutott, (lásd: Példa 1) és ehhez milyen főbb lépéseken keresztül jutott el a kutató.

Mindez már jelzi a programfájl tagolását, felépítését és számunkra is kontextusba helyezi, könnyebben érthetőbbé teszi az eredményeket.

Példa 1:

„A kutatás során az 5 fő feletti Magyarországon bejegyzett *feldolgozóipari vállalatok termelékenységi viszonyait* elemeztem 2002-re. Ehhez vállalati *mérlegadatokat* (merleg.dta) és *külkereskedelmi adatokat* (kulker.dta) használtam fel. A programfájl elején *adattisztítást* és az állományok előkészítését, majd összekapcsolását végeztem el. Ezek után TEÁOR kódok alapján leszűrtem a feldolgozóipari vállalatokat, melyre *lineáris regressziókat* futtattam. Ezekben azt vizsgáltam, hogy *X, Y, Z* magyarázó változók szignifikánsan magyarázzák-e W változót. A modellt különböző (*a, b, c*) almintákra is lefuttattam, végül *u, v, h* és az *e* változókra a szokásos leíró statisztikákat is lekértem…"

2. A programfájl egyes szakaszait jól láthatóan el kell különíteni, az egyes szakaszokhoz rövid leírást, címet kell adni, hogy világossá váljon, mely szakaszban mi történik, illetve mely kutatási eredmények állnak elő az adott programrészletből.

Példa 2:

```
**************************

** (1) Az adatbázis konzisztenciájának ellenőrzése**

**************************



/* Az „id", a „szuldat" és a „nem" változók konzisztenciájának ellenőrzése */

cd X:\Kimeno\kutato17\Kovacs_Istvan\
use munkaero_adatb1, clear

replace szuldat=szuldat[_n-1] if (id==id[_n-1])
replace nem=nem[_n-1] if (id==id[_n-1])
…



**********************************************************
```

** (5) Az eredmények grafikus ábrázolása **

****

* „Régi” tagországok *

****

/ Az egy főre jutó GDP növekedési üteme és az exportált autóalkatrészek növekedési üteme közötti kapcsolat grafikus ábrázolása */*

twoway (scatter GDP_ann_growth valcomp_ann_growth, mysymbol(square) msize(tiny) mcolor(black) mlabel(partnercode) mlabsize(tiny) mlabcolor(black)) ///

(lfit GDP_ann_growth valcomp_ann_growth) if country_dummy==0, ytitle(Az egy főre jutó GDP éves átlagos növekedési üteme (%))ytitle(, size(small)) ///

ylabel(, labsize(small)) xtitle (Az exportált autóalkatrészek éves átlagos növekedési üteme (%)) xtitle(, size(small)) title(Az egy főre jutó GDP és az exportált autóalkatrészek növekedési üteme között fennálló kapcsolat (Régi tagállamok), size(small)) ///

note((Az egy főre jutó GDP konstans 2005-ös árakon; Az exportált alkatrészek konstans 2006-os árakon), size (small) position(6)) legend(size(small))

...

3. Az *átnevezett* és az *újonnan képzett* változók definiálása a programfájlban szövegesen is, mellyel a kutató minden olyan információt megad, amely a változók tartalmára vonatkozik.

Példa 3:

****

** (1) Változók átnevezése, új változók képzése**

```
**********************************************************


/* Mérleg változók átnevezése */

ren Rövidlejáratúköt s_term_debt
ren Szállítók payables

/* Rövid távú hitelállomány változó képzése */

gen scredit=s_term_debt-payables

...

**********************************************************
** (2) Változók címkézése**
**********************************************************


/* Alkatrészek és kész autók definiálása, termékkód elnevezése e szerint*/

label define vehicletype 0 „components” 1 „final vehicles”

label values comm_code vehicletype

label variable comm_code „Category var.: Components and final vehicles”
```

4. Ha viszonylag sok programfájlt használt a kutató a kutatás, illetve a kutatási eredmények előállítása során, érdemes lehet készíteni egy „vezérlő” programfájlt, amely megfelelő sorrendben hívja meg a többi programfájlt. Egy ilyen programfájl megjegyzésekkel ellátva bonyolultabb esetekben nagyban segítheti a kutatás menetének megértését.

   Példa 4:

```
**************************************************************************

/*Ez a programfájl más programfájlok meghívásával, az eredeti állományokból több lépésben
állítja elő a kutatási eredményeket és a hozzájuk tartozó segédtáblákat.*/

**************************************************************************

cd ”X:\Kimeno\kutato17\Kovacs_Istvan\”
```

```
/*Panel adatállomány elkészítése adattisztítással, valamint külkereskedelmi- és mérlegadatok összekapcsolásával*/

do prepare_paneldataset.do

...

cd "X:\Kimeno\kutato17\Kovacs_Istvan\Adv_ellenorzeshez"

/*Aggregálás évek szerint, a programfájl a kutatási eredményeket és a hozzájuk tartozó segédtáblákat állítja elő*/

do aggreg_year.do
```

2.3 A kutatás dokumentációjának egyéb lehetőségei

A kutató akkor is köteles megfelelő dokumentációval ellátni a kutatási eredményt, ha azt nem programfájllal állította elő. A dokumentációra ebben az esetben tetszőleges formát választhat, azzal a céllal, hogy átláthatóvá tegye a kutatás célját, menetét, továbbá biztosítson minden olyan egyéb információt az adatvédelmi szakértő számára, amely a kutatási eredmény adatvédelmi ellenőrzéséhez szükséges. A kutatási eredmény előállításának folyamatát olyan mélységben kell leírnia, hogy annak alapján az adatvédelmi szakértő akár újra elő tudja állítani a kutatási eredményt (mint ahogy egy programfájl alapján is meg tudná tenni). Minden általa használt, átnevezett, illetve képzett változót megfelelően, egyértelműen definiálnia kell. A változókat a használt szoftver lehetőségei szerint címkéznie kell.

Természetesen ilyen segédanyagok biztosítása minden esetben, a programfájlokban lévő dokumentáción túl is hasznos lehet, azonban ha a kutatási eredményt nem programfájllal állította elő a kutató, úgy ilyen segédanyagok biztosítása nemcsak hasznos, hanem kötelező is.

Példák az adatvédelmi ellenőrzést segítő segédanyagokra:

szöveges dokumentumok: leírások a kutatás céljáról, menetéről, arról a folyamatról, melynek a végén a kutatási eredmények születtek.

táblázatok: pl. változók definícióinak összegyűjtésére, különböző változóátnevezések nyomon követésére alkalmas forma lehet.

programkódok kiíratása: pl. SPSS menüből való használata esetén általában lehetőség van arra, hogy végrehajtást jelentő „Ok" helyett a „Paste" gombra kattintással kiíratjuk a szoftverrel egy programfájlba a parancsnak megfelelő programkódot.

naplózás bekapcsolása, naplófájl elmentése: a naplófájlok segíthetik az adatvédelmi szakértőt a kutatás menetének megértésében.

3. A kutatási eredmények utólagos adatvédelmi szempontú ellenőrzésére vonatkozó módszertani eljárások és szabályok

3.1. A KSH gyakorlatában alkalmazott adatvédelmi szabályok

A kutatási eredmények adatvédelmi ellenőrzésekor különböző adatvédelmi szabályok [6] alkalmazására van lehetőség[7]. A továbbiakban a KSH jelenlegi gyakorlatában alkalmazott szabályokat ismertetjük.

küszöbszabály: egy táblázatban a küszöbszabály szerint érzékenynek tekinthető egy cella, ha az adatszolgáltatók száma egy meghatározott küszöbértéknél kevesebb.

> Megjegyzés: Például: m=5-öt választva küszöbértéknek, a cella érzékenynek tekinthető, ha az értékéhez hozzájáruló válaszadók száma 5-nél kevesebb.
>
> ***hármas szabály***: A Statisztikai törvény végrehajtási rendelete által nevesített küszöbérték, mely szerint összesítve sem lehet nyilvánosságra hozni olyan adatot, melynél az adatszolgáltatók száma háromnál kevesebb (Statisztikai törvény végrehajtási rendelete, 19. §).
>
> Megjegyzés: Ennél megengedőbb küszöbérték (1-es vagy 2-es szabály) nem alkalmazható, erősebb azonban igen. (például 4-es vagy 5-ös szabály). Megjegyezzük továbbá, hogy legalább hármas szabály ($m \geq 3$) alkalmazását nem csak a törvényi kötelezettség indokolja, hanem az is, hogy ez az a minimális küszöbszám, ami biztosítja, hogy az elnyomott cellához tartozó hozzájárulók száma ne legyen *egyértelműen* meghatározható. Ezáltal csökkentjük az adatszolgáltató *beazonosításának* és/vagy az adatszolgáltatóhoz kapcsolódó új információ *felfedésének* valószínűségét. *Gyakorisági*

[6] Például küszöbszabály, dominancia szabály, p%-szabály, p/q-szabály

[7] *Handbook on Statistical Disclosure Control* (http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf).

táblákra és *értékösszeg*táblákra is megvizsgáljuk, hogy legalább a hármas szabálynak megfelelnek-e a kivinni kívánt adatok. Megjegyezzük, hogy a küszöbszabályt olyan értékösszegtáblák esetén is alkalmazzuk, amelyekhez a kutató nem kívánja a hozzá tartozó gyakorisági táblát kivinni.

(n,k)-dominancia szabály: egy táblázatban az *(n,k)-dominancia szabály* szerint érzékenynek tekintendő a cella, ha az értékhez hozzájáruló adatszolgáltatók közül a legnagyobb *n* darab részesedésének összege meghaladja a cella értékének *k* %-át.

Megjegyzés: Az „*n*" és a „*k*" értékét mindig az illetékes adatgazda szervezeti egység javaslatát figyelembe véve határozzuk meg, ami kutatási eredményenként is változhat a felfedési kockázat függvényében. Mint bármely, a kutatási eredmények védettségének biztosítására alkalmazott módszertani szabály esetén, a dominanciaszabály és a küszöbszabály esetén is érvényesül az az előírás, hogy az alkalmazott adatvédelmi szabályokról tájékoztatjuk a kutatót, ugyanakkor az egyes módszereknél alkalmazandó pontos paramétereket nem adhatjuk ki.

A dominanciaszabályt *értékösszeg*táblák esetében alkalmazzuk. A szabály alkalmazása például abban az esetben lehet indokolt, ha egy kutatás erősen *koncentrált iparágakra* terjed ki, ahol kevés számú, ún. domináns vállalat fedi le a piaci részesedés döntő részét. Ebben az esetben valamely iparági jellemző nyilvánosságra hozatala egy viszonylag pontos becslést adna az iparágban *domináns vállalat*(*ok*) jellemzőjére is. Sőt, ha egy iparágban kevés számú vállalat tevékenykedik, akkor felmerülhet annak a lehetősége is, hogy (*n-1*) vállalat *koalíciót* alkotva pontosan meghatározza a domináns vállalat jellemzőjét az iparági összesen és a saját értékeik ismeretében.

3.2. Az adatvédelmi módszerek csoportosítása és a KSH gyakorlata

Az adatvédelmi módszerek alapvetően két csoportba sorolhatók: (1) *perturbatív* és (2) *nem-perturbatív* eljárások.

Az első csoportba tartoznak például a kerekítési eljárások, véletlen zaj hozzáadása, rekordok/értékek megkeverése, míg a másodikba a globális átkódolás, alsó/felső kódolás, mikroaggregálás és a cellaelnyomás. A KSH jelenlegi gyakorlatában a táblázatos kutatási eredmények adatvédelmét egységesen a minimálisan szükséges *cella elnyomásával* biztosítjuk.

Ez ugyan mennyiségi szempontból információveszteséget jelent a perturbatív eljárásokhoz képest, viszont a kivihető kutatási eredmények nem lesznek torzítottak.[8]

Amennyiben túl sok érték kerülne elnyomásra, akkor előnyösebb lehet a kutató számára az eredményeket más bontásban (például bizonyos kategóriák összevonásával) elkészíteni. Általában több, adatvédelmi szempontból elfogadható megoldás is létezik (nem-perturbatív eljárások különböző kombinációi) az eredmények módosítására, ezért a kutatóra bízzuk, hogy amennyiben szükségesnek tartja, akkor az adatvédelmi szabályok figyelembe vételéve válassza ki azt, ami számára a leginkább használható eredmények kivitelét teszi lehetővé (a tapasztalat azt mutatja, hogy nem érdemes alacsony területi bontás mellett részletes tevékenységkódokat vizsgálni. Például: megye-szakágazat kereszttábla esetén célszerűbb régióra aggregálni vagy a TEÁOR-t ágazatonként vagy áganként elemezni.).

Amennyiben az adatvédelmi szakértő a kutatási eredmények adatvédelmi ellenőrzésekor úgy ítéli meg, hogy a kutatási eredmények csak jelentős információveszteség árán adhatók ki, úgy ezt jelzi a kutatónak. Mivel ebben a fázisban a kutatási eredmények kiadása még nem történt meg, a kutatónak lehetősége van az eredményeit módosítania. Hangsúlyozzuk, hogy amíg az eredmények előállítása, a kutatási módszertan megválasztása/módosítása a kutató feladata, addig az adatvédelmi eljárások kiválasztása és a kiadhatóságra vonatkozó vélemény kialakítása kizárólag az adatvédelmi szakértő feladata.

A fenti eljárásrend alkalmazása abból a szempontból is indokolt, hogy azonos mikroadat(ok)ból többféle változatban (bontásban) készített kutatási eredmények (például kereszttábla) esetén előfordulhat, hogy a kutatási eredménytáblák bár külön-külön kiadhatóak lennének, de együttesen már lehetővé válna felfedés vagy akár az azonosítás is.

> Például vizsgáljuk meg a következőt: Legyen három táblánk: *nem-város*, *nem-büntetett előélet*, *város-büntetett előélet*. A cellaértékek minden esetben a könyvárusok számai. Tulajdonképpen ez egy 3-dimenziós tábla, annak 3 db 2-dimenziós, a részösszeseneket tartalmazó tábláiról van szó.

Könyvárusok száma

[8]A kutató biztos lehet abban, hogy a kivitt eredmények nem tartalmaznak például kerekítési hibát vagy véletlen zajt, vagyis a kivitt kutatási eredmények ún. „valós” adatok.

1. Tábla

	Miskolc	**Debrecen**	*Összesen*
Férfi	21	12	33
Nő	16	19	35
Összesen	37	31	68

2. Tábla

(Büntetett előélet)	**Igen**	**Nem**	*Összesen*
Férfi	23	10	33
Nő	8	27	35
Összesen	31	37	68

3. Tábla

(Büntetett előélet)	**Igen**	**Nem**	*Összesen*
Miskolc	11	26	37
Debrecen	20	11	31
Összesen	31	37	68

Vegyük a következő jelöléseket:

- Nem: *f = férfi*
  n = nő
- Város: *M = Miskolc*
  D = Debrecen
- Büntetett előélet: *i = igen*
  n = nem

Tehát például: X_{fDn} jelölje azoknak a férfi könyvárusoknak a számát, akik Debrecenben dolgoznak és nem büntetett előéletűek. Ez egy cella értéke a 3-dimenziós táblázatban.

Ezek, és a 3-dimenziós táblázat alapján az alábbi egyenletek írhatóak fel:

- 1. Tábla alapján

$$X_{fMi} + X_{fMn} = 21$$

$$X_{fDi} + X_{fDn} = 12$$

$$X_{nMi} + X_{nMn} = 16$$

$$X_{nDi} + X_{nDn} = 19$$

- 2. Tábla alapján

$$X_{fMi} + X_{fDi} = 23$$
$$X_{nMi} + X_{nDi} = 8$$
$$X_{fMn} + X_{fDn} = 10$$
$$X_{nMn} + X_{nDn} = 27$$

- 3. Tábla alapján

$$X_{fMi} + X_{nMi} = 11$$
$$X_{fMn} + X_{nMn} = 26$$
$$X_{fDi} + X_{nDi} = 20$$
$$X_{fDn} + X_{nDn} = 11$$

Ezekből azt kapjuk, hogy:

$X_{fMi} = 11$	$X_{fMn} = 10$
$\mathbf{X_{fDi} = 12}$	$\mathbf{X_{fDn} = 0}$
$X_{nMi} = 0$	$X_{nMn} = 16$
$X_{nDi} = 8$	$X_{nDn} = 11$

Azaz kiderül, hogy **Debrecenben minden férfi könyvárus büntetett előéletű**!

3.3.Táblázatos adatok[9] védelme

Ahogy már a 3.2-es pontban is jeleztük, táblázatos kutatási eredmények adatvédelmi ellenőrzése során az adatok védelmét *cellaelnyomással* biztosítjuk, ami az egyik leggyakrabban alkalmazott adatvédelmi módszer a *táblázatos adatok* esetében. Az alábbiakban a cellaelnyomás fogalmát tisztázzuk, majd a *gyakoriságí- és értékösszegtábla* példáján keresztül a gyakorlatban is illusztráljuk a módszer működését.

Cellaelnyomás: a táblázatos adatvédelem egyik módszere, melynek eredménye, hogy adott cellák értékei nem közölhetőek, azok egyezményes jellel helyettesítendőek. Amennyiben cellaelnyomást alkalmazunk, egyértelműen jelezzük, hogy mely cellák kerültek elnyomásra

[9] **Táblázatos adat:** olyan, táblázatos formába rendezett adatállomány, amely aggregált adatokat tartalmaz. Alapvetően kétféle táblázatos adatot különböztetünk meg: gyakorisági táblát és értékösszeg táblát.

(például ne lehessen összekeverni a hiányzó értékekkel). Az alábbiakban ismertetett elsődlegesen és másodlagosan elnyomott cellákat azonban egységes jellel látjuk el, nem különböztetjük meg őket egymástól.

A cellaelnyomás folyamata két egymásra épülő részből áll, az *elsődleges* és a *másodlagos* cellaelnyomásból, melyek alkalmazása már védetté teszi a táblázatot. Elsődleges cellaelnyomás után mindig meg kell vizsgálni, hogy szükséges-e másodlagos cellaelnyomást is alkalmazni, mert az elsődleges cellaelnyomás önmagában nem elégséges a táblázat védettségének biztosításához, amennyiben az elnyomott cella értéke valamilyen módon visszaszámolható.

- **Elsődleges cellaelnyomás:** táblázatos adatok védelmére alkalmazott eljárás, melynek lényege, hogy valamilyen adatvédelmi szabály(ok) eredményeként *érzékenynek* ítélt cella értéke nem közölhető, hanem egyezményes jellel helyettesítendő.
- **Másodlagos cellaelnyomás:** táblázatos adatok védelmére alkalmazott eljárás, melynek során az elsődleges cellalenyomás során elnyomott cellákon felül további cellákat nyomunk el a táblázat védettségének biztosítása érdekében. Ennek célja kizárólag az, hogy az elsődlegesen elnyomott, érzékenynek ítélt cellák védettek maradjanak. A másodlagos cellaelnyomásra kijelölt cella értékét önmagában ugyanis nem kellene védeni.

A másodlagos cellaelnyomás alapgondolata, hogy amikor adatokat és azokból képzett aggregátumokat együtt szerepeltetünk egy állományban, akkor az elsődlegesen elnyomott adat egyértelműen ne legyen meghatározható a belőle képzett aggregátumból és a többi adatból. Például egy táblázatban az elsődlegesen elnyomott cella értéke a sorában vagy oszlopában szereplő többi adatból, illetve a sor- vagy oszlop összesen értékekből alapvető matematikai műveletekkel egyértelműen meghatározható lehet.

A másodlagos cellaelnyomás egy optimalizálási feladat, melynek megoldására különféle algoritmusok léteznek. Nincs egységes álláspont, hogy melyik cellaelnyomási algoritmust kell alkalmazni, de *két szempontot* mindenképpen figyelembe veszünk:

1. A táblákban az *érzékenynek ítélt cellák védve* legyenek (az elsődlegesen védendő cellák értékeit ne lehessen kiszámítani).
2. A lehető *legkisebb* legyen az *információveszteség*. Ez a gyakorlatban azt jelenti, hogy:

    - A lehető legkevesebb cellát próbáljuk elnyomni.

- Amennyiben több lehetőség is adódik a minimális számú cella elnyomására, akkor azt a megoldást választjuk, ahol összességében a legkevesebb a hozzájáruló adata kerül elnyomásra.
- Az összesenek elnyomását kerüljük.

A táblázatos adatoknak két alapvető típusa létezik: *gyakorisági táblák* (1) és *értékösszegtáblák* (2). A két tábla a felfedés két különböző típusához kapcsolódik. Míg a gyakorisági táblák csupán azonosítást tesznek lehetővé, addig az értékösszegtáblák esetében egy új információ (például árbevétel) is kiderülhet egy adott egyedről. Az egyed azonosítása ugyanúgy nem megengedett, mint egy új információ nyilvánosságra hozatala.

3.3.1. Gyakorisági táblák adatvédelme

A táblázatos adatok védelmének egy gyakran alkalmazott alternatív módszere a kerekítés. A cellaelnyomás előnye ezzel szemben az, hogy mind a táblán belüli el nem nyomott értékek, mind a sor és oszlop végösszegek értékei *változatlanok* maradnak.

Régiónkénti vállalatok darabszáma Nemzetgazdasági áganként 2010-ben (fiktív adatok)							
Régiók	**Nemzetgazdasági ágak**						
	K	**C**	**D**	**E**	**J**	**M**	**Total**
Közép-Magyarország	18	500	20	11	326	1281	**2156**
Közép-Dunántúl	2	145	2	15	22	140	**326**
Nyugat-Dunántúl	4	105	5	7	10	105	**236**
Dél-Dunántúl	4	58	4	6	21	136	**229**
Észak - Magyarország	1	100	*3*	5	21	119	**249**
Észak-Alföld	1	99	*8*	8	16	158	**290**
Dél-Alföld	8	107	9	8	31	163	**326**
Total	**38**	**1114**	**51**	**60**	**447**	**2102**	**3812**

1. táblázat Cellaelnyomás mintapélda

A cellaelnyomás két alaplépése:

Az *elsődlegesen védendő cellákat* kijelöljük (itt a küszöbszabály alapján, általában a háromnál kevesebb elemet tartalmazó cellaértékek törlését javasoljuk, jogszabály szerint is). Sok esetben nem elegendő az elsődlegesen kijelölt cellák elnyomása (*világos*

szürkével jelölt cellák, *aláhúzott* cellaértékek), mert a sor- és oszlop összesenekből kiszámíthatóak az elsődlegesen elnyomott cellák értékei.

Az elsődlegesen elnyomott cellák visszaszámolhatósága miatt néhány újabb cellát is el kell nyomni, ez a *másodlagos cellaelnyomás* (*sötét szürkével* jelölt cellák, *dőlt* cellaértékek).

Például: Az elsődlegesen védendő cellák:
K – Közép–Dunántúl
K – Észak– Magyarország
K – Észak–Alföld
D – Közép–Dunántúl

A „*K – Észak–Magyarország*", a „*K – Észak–Alföld*" és a „*D – Közép–Dunántúl*" cellaértéke is egyértelműen kiszámítható a sor- és oszlopösszesenekből. Figyelembe véve, hogy a *D* oszlopban mindenképpen másodlagos cellaelnyomást kellene végrehajtani a „D – *Közép–Dunántúl*" cellaértéke miatt, ezért a *D* oszlopból választjuk ki a megfelelő cellákat a „K – Észak– Magyarország" és „*K – Észak-Alföld*" értékeinek levédésére is. A „*D – Észak– Magyarország*" és „*D – Észak–Alföld*" cellába tartozó értékek kerülnek tehát másodlagosan elnyomásra, így már egyik elsődlegesen elnyomott cellaérték sem számolható vissza egyértelműen.

3.3.2. Értékösszegtáblák adatvédelme

Értékösszegtáblák esetén nemcsak a szabályt, hanem a *dominanciaszabályt* is figyelembe vesszük. Az előbbi indokolja az értékösszegtáblához tartozó *abszolút gyakorisági* tábla (*3. táblázat*), míg az utóbbi az *értékösszeghez legnagyobb mértékben hozzájáruló vállalat részesedését* (%) mutató táblázat (*4. táblázat*) kiszámítását. Ezeket a táblákat a 2. fejezetben leírtaknak megfelelően a kutató az **„Adv_ellenorzeshez"** mappába gyűjti.

A kutató által adatvédelmi ellenőrzésre beadott táblázat:

Vállalatok árbevételének összege régiónként és a kiemelt nemzetgazdasági ágak szerint 2010-ben (fiktív adatok)					
	Vizsgált nemzetgazdasági ágak				
Régiók	A	B	C	D	Total
1	7 767 971328.0	211 091 899 472.0	9 943 678 279.0	303 314 418 304.0	**532 117 967 383.0**

2	293 999 322 944.0	482 392 636 132.0	195 788 826 974.0	132 201 948 800.0	**1 104 382 734 850.0**
3	109 496 225 408.0	8 703 476 032.0	125 583 012 384.0	251 129 981 347.5	**494 912 695 171.5**
4	199 566 570 752.0	327 763 841 000.0	97 802 072 160.0	12 144 741 376.0	**637 277 225 288.0**
5	275 043 249 600.0	70 936 308 800.0	161 725 637 616.0	430 395 294 040.0	**938 100 490 056.0**
6	165 900 728 448.0	126 406 154 216.0	334 304 238 378.0	212 415 489 584.0	**839 026 610 626.0**
7	233 459 129 720.0	156 755 016 340.0	8 416 344 064.0	70 604 533 024.0	**469 235 023 148.0**
Total	**1 285 233 198 200.0**	**1 384 049 331 992.0**	**933 563 809 855.0**	**1 412 206 406 475.5**	**5 015 052 746 522.5**

2. táblázat [10] Mintapélda értékösszeg táblára

1. **A kutató elkészíti a 2. táblázathoz tartozó <u>abszolút gyakorisági</u> táblázatot:** Felhívjuk a figyelmet arra, hogy a gyakoriságok kiszámításakor az adott értékösszeghez hozzájáruló különböző adatszolgáltatók számát kell figyelembe venni. Tegyük fel, hogy a hármas szabályt alkalmazzuk. Emiatt a három alatti cellaértékek törlésre kerülnek (*világos szürkével* jelölt cellák, *aláhúzott* cellaértékek - *elsődleges cellaelnyomás*)

Vállalatok száma régiónként és a kiemelt nemzetgazdasági ágak szerint 2010-ben (fiktív adatok)					
Régiók	**Vizsgált nemzetgazdasági ágak**				
	A	**B**	**C**	**D**	**Total**
1	<u>1</u>	42	10	54	107
2	51	97	40	26	214
3	19	<u>2</u>	30	57	108
4	35	69	19	<u>2</u>	125
5	51	15	34	79	179
6	33	28	70	43	174
7	43	35	<u>1</u>	14	93
Total	233	288	204	275	1000

3. táblázat [11] A 2. táblázathoz tartozó abszolút gyakorisági tábla

2. **A kutató elkészíti a 2. táblázathoz a <u>legnagyobb hozzájárulónak az értékösszegből való százalékos részesedésére</u> vonatkozó táblázatot.** Tegyük fel, hogy *k*=90%, ezért a *sötét szürkével* jelölt cellák (félkövérrel jelölt cellaértékek) *elsődleges* elnyomásra kerülnek.

[10] A kutató ezt a '*Kutatasi_eredmenyek*' mappába helyezi el.
[11] A kutató ezt az '*Adv_ellenorzeshez*' mappába helyezi el.

Az értékösszeghez legnagyobb mértékben hozzájáruló vállalat részesedése (%)					
Régiók	Vizsgált nemzetgazdasági ágak				
	A	B	C	D	Total
1	**100.00**	4.73	**95.30**	3.07	1.88
2	3.38	2.06	4.51	7.36	0.90
3	9.05	**91.52**	7.19	3.88	2.00
4	4.97	3.01	10.00	54.61	1.56
5	3.61	12.78	6.09	2.32	1.06
6	5.17	7.75	2.94	4.63	1.17
7	4.26	5.87	**100.00**	13.11	2.12
Total	0.77	0.72	1.05	0.71	0.20

4. táblázat [12] A 2. táblázathoz tartozó, az értékösszeghez tartozó legnagyobb mértékben hozzájáruló részesedésére (%) vonatkozó táblázat

Az elsődleges cellaelnyomásokat az *5. táblázat* mutatja (*hármas szabály* és *dominanciaszabály* szerint):

Vállalatok árbevételének összege régiónként és a kiemelt nemzetgazdasági ágak szerint 2010-ben (fiktív adatok)					
Régiók	Vizsgált nemzetgazdasági ágak				
	A	B	C	D	Total
1	…	211 091 899 472.0	…	303 314 418 304.0	**532 117 967 383.0**
2	293 999 322 944.0	482 392 636 132.0	195 788 826 974.0	132 201 948 800.0	**1 104 382 734 850.0**
3	109 496 225 408.0	…	125 583 012 384.0	251 129 981 347.5	**494 912 695 171.5**
4	199 566 570 752.0	327 763 841 000.0	97 802 072 160.0	…	**637 277 225 288.0**
5	275 043 249 600.0	70 936 308 800.0	161 725 637 616.0	430 395 294 040.0	**938 100 490 056.0**
6	165 900 728 448.0	126 406 154 216.0	334 304 238 378.0	212 415 489 584.0	**839 026 610 626.0**
7	233 459 129 720.0	156 755 016 340.0	…	70 604 533 024.0	**469 235 023 148.0**
Total	**1 285 233 198 200.0**	**1 384 049 331 992.0**	**933 563 809 855.0**	**1 412 206 406 475.5**	**5 015 052 746 522.5**

5. táblázat Elsődleges cellaelnyomást tartalmazó táblázat

Látható, hogy a fenti táblázatnak *még nem védett* minden elnyomott cellája, mert például az *A1*-es cella értéke visszaszámolható. Az elsődlegesen elnyomott cellák védelme érdekében *másodlagos cellaelnyomást* is alkalmaznunk kell.

[12] A kutató ezt az '*Adv_ellenorzeshez*' mappába helyezi el.

A kiadható eredmény a *6. táblázatban* látható (*elsődleges* és *másodlagos* cellaelnyomásokat alkalmazva):

Vállalatok árbevételének összege régiónként és a kiemelt nemzetgazdasági ágak szerint 2010-ben (fiktív adatok)					
Régiók	Vizsgált nemzetgazdasági ágak				
	A	B	C	D	Total
1	**…**	211 091 899 472.0	**…**	303 314 418 304.0	**532 117 967 383.0**
2	293 999 322 944.0	482 392 636 132.0	195 788 826 974.0	132 201 948 800.0	**1 104 382 734 850.0**
3	109 496 225 408.0	**…**	125 583 012 384.0	…	**494 912 695 171.5**
4	199 566 570 752.0	…	97 802 072 160.0	**…**	**637 277 225 288.0**
5	275 043 249 600.0	70 936 308 800.0	161 725 637 616.0	430 395 294 040.0	**938 100 490 056.0**
6	165 900 728 448.0	126 406 154 216.0	334 304 238 378.0	212 415 489 584.0	**839 026 610 626.0**
7	…	156 755 016 340.0	**…**	70 604 533 024.0	**469 235 023 148.0**
Total	**1 285 233 198 200.0**	**1 384 049 331 992.0**	**933 563 809 855.0**	**1 412 206 406 475.5**	**5 015 052 746 522.5**

6. táblázat Elsődleges és másodlagos cellaelnyomást tartalmazó táblázat

Megjegyzés: Amennyiben nem ismertek a sor/oszlopösszesenek és nem is áll fenn annak reális lehetősége, hogy egyéb adatforrásokból hozzá lehet jutni, vagy ki lehet számítani, akkor a másodlagos cellaelnyomás alkalmazása nem indokolt.

Felhívjuk a figyelmet arra, hogy amennyiben egy értékösszegtáblában több változóra is szerepel értékösszeg (pl. árbevétel, profit, hozzáadott érték stb.), úgy mindegyikhez külön-külön el kell készíteni a gyakoriságokat és a legnagyobb hozzájárulók részesedéseit, kivéve, ha ezek nem különböznek az egyes változókra.

A gyakoriságok kiszámításakor mindig az adott értékösszegtábla struktúráját kell figyelembe venni: Pl. ha az értékösszeg vállalatoknak 2 jegyű TEÁOR szerinti bontásban kiszámított árbevétele, akkor a gyakoriság ehhez az árbevételhez hozzájáruló különböző vállalatok száma lesz.

4. Egyéb adatvédelmi eljárások

Eddig a táblázatos adatok védelmét tárgyaltuk, a következőkben az egyéb, gyakran előforduló kutatási eredményeknél alkalmazott szabályokról, elvekről ejtünk néhány szót. Hangsúlyozzuk, hogy a KSH adatvédelmi gyakorlata az ún. *principle based* modellt követi, amely a konkrét, számszerű hüvelykujjszabályok és „*best practice*”-ek mechanikus alkalmazása helyett minden egyes kutatási eredmény esetében egyedi mérlegelést ír elő, illetve tesz lehetővé.

Ez alól csak a már korábban említett hármas szabály számít kivételnek, amit lényegében automatikusan alkalmazunk minden sokaságból származó eredményre [13] . Ennek a megközelítésnek az előnye, hogy a kutató várhatóan kisebb információveszteséggel számolhat, mivel a kutatási eredményeket rugalmasan, a reálisan szóba jöhető felfedési forgatókönyvek figyelembevételével bíráljuk el. Ez a megközelítés megköveteli a kutató által előállított eredmények pontos megértését is, így kérdéses esetekben a kutató és az adatvédelmi szakértő közötti szakmai konzultációk elengedhetetlenek. Az eljárás hátránya, hogy a kutatási eredmények utólagos adatvédelmi szempontú ellenőrzése valamelyest időigényesebb, illetve a kutatók számára kevésbé adhatóak jól definiált, számszerű szabályok, küszöbértékek.

A Hivatalban a kutatási eredmények adatvédelmi ellenőrzése során a *Guidelines for the checking of output based on microdata research* [14] *(ESSNet SDC, 2009)* útmutatóban szereplő ún. 'principle based' szabályok, ajánlások tekinthetőek irányadónak. Amennyiben az általunk leírtak és az útmutatóban foglaltak között ellentmondás tapasztalható (pl. 3-as szabály vs. 10-es szabály), akkor természetesen az általunk leírtak a mérvadóak.

4.1. Teljeskörű felvételekre és mintabeli sokaságokra vonatkozó eljárások

A *gazdaságstatisztikai* állományokból készített kutatási eredményekre automatikusan a hármas szabályt alkalmazzuk, függetlenül attól, hogy teljes körű a felvétel vagy minta. Ennek indoka, hogy a vállalatokról sokkal több publikus információ (mérleg, éves beszámoló, stb.) áll rendelkezésre, mint az egyénekről vagy háztartásokról, így általánosságban a felfedés kockázata is magasabb, mint lakossági felvételek esetén.

A lakossági mintás felvételeknél és a népszámlálási mintaállományoknál nem alkalmazzuk automatikusan a hármas szabályt, mert ezzel túl nagy információveszteséget okoznánk. Ezekben az esetekben egyedileg, a szakfőosztály javaslatait figyelembe véve mérlegelünk a kiadhatóságról.

[13] A mintákról ld. 4.1-es fejezet

[14] Letölthető: http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf

A kiadhatóság esélyét csökkenti, ha a felvétel több *érzékenynek* tekinthető változót (pl. nemzetiség, felekezeti hovatartozás, szexuális orientáció, egészségi állapot, jövedelem, stb.) tartalmaz.

A kutatási eredmények vizsgálata során külön figyelmet szentelünk az ún. *azonosítást lehetővé tevő változókra* is, úgy, mint:

- földrajzi bontás (pl. megye, település, számlálókörzet, stb.);
- osztályozások (pl. FEOR, TESZOR, TEÁOR, ÖVTJ, stb.);
- életkor;
- nem;
- családi állapot;
- legmagasabb iskolai végzettség;
- gyerekek száma a háztartásban;
- egy háztartásban élők; száma.

5. Tanácsok kutatók számára biztonságosnak ítélhető kutatási eredmények előállításához

A fent leírtak alapján az alábbi hasznos tanácsokat ajánlunk kutatóink figyelmébe a kutatási eredmények előállításával kapcsolatban.

- Törekedjen arra, hogy a lehető leginkább véglegesnek tekinthető állapotban állítsa elő a kutatási eredményeket.
- Gondolja végig, hogy melyek azok, a számára leglényegesebb eredmények, amelyeket mindenképpen ki szeretne vinni. Erre a következő okokból van szükség:
    - Ha a kutatási eredmények sok fájlt tartalmaznak, vagy azok nagy méretűek, akkor ez lassíthatja az adatvédelmi szempontú ellenőrzés folyamatát. Ugyanis az adatvédelmi ellenőrzés elvégzéséhez szükséges, hogy az ellenőrzést végző személy bizonyos mértékig megértse a kutatás célját, menetét, a szükséges mértékben átlássa az összefüggéseket.
    - Ha sok fájlt szeretne kivinni a kutató, illetve pl. sok változó van egy-egy táblázatban, akkor nagyobb a valószínűsége annak, hogy a táblázatok közötti összefüggések miatt tekinthető adatvédelmi szempontból aggályosnak egy kutatási eredményfájl kiadása.

- Értékösszegtáblák, illetve gyakorisági táblák esetén mennyire jellemző a nagyon alacsony (1 vagy 2) számú hozzájárulás a cellákhoz? A hármas szabály alapján minden esetben megvizsgáljuk a táblákat.
- Ha nagy arányban vannak kevés hozzájáruló adatát tartalmazó cellák a táblában, lehet, hogy érdemes átstrukturálni, kevésbé részletes bontásban előállítani a táblát, aggregálást végezni, stb. A kutatási eredmények adatvédelmi ellenőrzése során érzékenynek talált cellákra minden esetben cellaelnyomást végzünk.
- Van-e valamilyen sor/oszlopösszeg a táblázatban? Ekkor az érzékeny cellák elnyomásán túl másodlagos cellaelnyomás is szükséges lehet, ami nagymértékben megnövelheti az elnyomott cellák számát.
- Amennyiben lehetséges, a kutatási eredményeket programfájllal állítsa elő, mivel így visszakövethető a kutatás menete és reprodukálhatóak a kutatási eredmények is.
- Programfájllal előállított kutatási eredmény estén javasoljuk, hogy a teljes futásról készüljön logfájl, amit a kutató az 'Adv_ellenorzeshez' könyvtárba kell elhelyezni. Ez segítséget nyújt nekünk a kutatás gyorsabb és jobb megértéséhez.
- Javasoljuk, hogy egy kutatási eredményt egy programfájllal állítson elő, mivel abból egyértelműen kiolvasható a kutatás menete. Amennyiben kényelmesebb, ha több programfájl állítja elő a kutatási eredményt, úgy azt javasoljuk, hogy:
    - Legyen egyetlen ún. „vezérlő" programfájl és az egyes részszámításokat végző egyéb programfájl ebben legyen meghivatkozva, vagy
    - Egy külön szöveges dokumentumban jelezze a programfájlok lefuttatásának logikai sorrendjét. Ezt az állományt az 'Adv_ellenorzeshez' mappába helyezze el.
- A reprodukálhatóság érdekében kérjük, hogy mindig új néven mentse el azokat az állományokat, amelyeken több alkalommal is végez valamilyen műveletet. Amennyiben az állományt felülírja és csak az utolsó mentés szerinti állapot áll rendelkezésre, akkor nem tudunk olyan résszámításokat elvégezni, amihez az adott állomány egy korábbi verziójára lenne szükség.
- Mielőtt lefuttatja a programfájlt, mindenképpen mentse el azt. Ne futtasson a 'Temp' könyvtárban található .tmp kiterjesztésű programfájlt, különben a hozzá tartozó esetleges futási logból nem tudjuk visszafejteni, hogy konkrétan melyik programfájl futott le.

- Gyakran előfordul, hogy az adatvédelmileg kívánatos eredmény elérése érdekében nem elegendő a végeredményeket módosítani, hanem az alkalmazott modelleket is újra kell kalibrálni és futtatni. A kutatói szabadság elvéből következik, hogy az előállított eredmények szakmai tartalmáért kizárólag a kutató tartozik felelősséggel, így a módosítási eljárások helyes megválasztása is a kutató joga és feladata.

1.melléklet

Kutatási eredmények dokumentációs űrlapja[15]

1. Kutatási alapinformációk

1.1 Kutató bejelentkezési azonosítója:	
1.2 Kutatási projekt megnevezése:	
1.3 Kutatási intézmény neve (szervezet neve):	
1.4 Kutatási eredményt előállító kutató neve:	
1.5 Kutatási eredményt előállító kutató e-mail címe:	
1.6 Kutatási eredmény leadásának dátuma (éééé/hh/nn):	

2. Kutatási eredményre vonatkozó dokumentáció

2.1 Programkóddal előállított[16] kutatási eredmények

Kutatási eredmény fájl neve(i)[17]	Kutatási eredmény fájl típusa[18]	Adja meg a programfájl nevét és sorát/sorait, mely(ek) a kutatási eredményt előállítja/előállítják	Nevezze meg az eredeti szakstatisztiká(ka)t/inputállomány(oka)t[19], ami(k)ből a kutatási eredmény származik

[15] A kutatási eredmény utólagos adatvédelmi szempontú ellenőrzés megkezdésének feltétele, hogy a kutató a nyomtatványt hiánytalanul kitöltse, illetve azt a tájékoztatási feladatokat ellátó szervezeti egység részére leadja.

[16] A kutatási eredmény előállításához programkód is rendelkezésre áll (pl. STATA-ban, SAS-ban végzett kutatómunka).

[17] Ugyanazon cellában minden olyan fájl felsorolható, amely ugyanabból a mikroadat állományból származik és az adott programfájl ugyanazon sora generálja le.

[18] Pl. gyakorisági tábla, értékösszeg tábla, regresszió analízis, log fájl. Ha a kutatási eredmény egy log fájl, akkor nem kell az abban szereplő minden táblát, regressziós outputot stb. egyenként felsorolni, de a megfelelő programfájlban kérjük ezeket rövid kommentekkel ellátni.

[19] Pl. KSH-inputok: Népszámlálás, Halálozás, K+F adatok stb.; Kutató által behozott inputok esetén azok pontos megnevezése.

2.2 Programkód nélkül[20] előállított kutatási eredmények

2.2.1 Kérjük, hogy röviden ismertesse a kutatási eredmény előállításának **célját**, a kutatási eredmény előállításának **logikai menetét**, előállításának **főbb lépéseit**!

2.2.2 Kérjük, adja meg az **újonnan képzett vagy átnevezett változók definícióját**, melyek a kutatási eredményekben, illetve a kutatási eredmények előállításához közvetlenül felhasznált állományokban szerepelnek[21]!

Változó neve	Változó jelentése	Melyik állományban szerepel

[20] A kutatási eredmény előállításához programkód nem áll rendelkezésre (pl. Excel-ben, SPSS-ben végzett kutatómunka).

[21] Ha egy külön dokumentumban (pl. Word, Excel fájlban) már definiálásra kerültek a változók, akkor a táblázatot nem kell kitölteni, de ezt a dokumentumot kérjük a 'Dokumentacios_urlap' mappába helyezni.

2.2.3 Kérjük, ismertesse a **kutatási eredmények**et[22]!

Kutatási eredmény fájl neve	Kutatási eredmény fájl típusa[23]	Mely(ik) állomány(ok)ból származik/származnak	Az eredmény rövid értelmezése

A KSH biztonságos környezetéből a kutatási eredmények kivitelének szükséges feltétele a kutatási eredmények dokumentációjának helyes kitöltése, melyet kitöltöttség szempontjából a tájékoztatási feladatokat ellátó szervezeti egység munkatársai vizsgálnak át. Csak helyesen és teljes mértékben kitöltésre került űrlap megléte esetén kerülnek továbbításra a kutatási eredmények utólagos adatvédelmi szempontú ellenőrzésre, az azt végző szakértők számára. A tájékoztatási feladatokat ellátó szervezeti egység 2-3 munkanapon belül visszajelez a kutató felé a dokumentációs űrlap kitöltésének helyességéről.

Amennyiben a dokumentációs űrlap kitöltése helytelen vagy hiányos, úgy a tájékoztatási feladatokat ellátó szervezeti egység a kutatót hiánypótlásra kéri fel.

Minden mező esetén kötelező a válaszadás, az alábbi megjegyzés figyelembevételével:

Amennyiben minden kutatási eredményt programkód generál le, akkor értelemszerűen csak a 2.1-es pont kitöltendő, míg amennyiben egyáltalán nem áll rendelkezésre programkód, akkor elégséges csak a 2.2-es pontot kitölteni.

[22] Ha a kutatási eredmények értelmezését már tartalmazzák a kivinni szándékozott kutatási eredményeket tartalmazó állományok, akkor itt elég utalni rá (pl. lásd az <x_y>.<xyz> fájlban).

[23] Pl. gyakorisági tábla, értékösszeg tábla, regresszió analízis, logfájl. Ha a kutatási eredmény egy logfájl (pl. spv output), akkor az abban szereplő összes eredményt (kereszttábla, regressziós output stb.) egyenként fel kell tüntetni és értelmezni.

2.melléklet

A honlapunkon feltüntetett, **kutatásra előkészített és ingyenesen kutatható mikroadat-állományok** kutatási eredményeinek utólagos adatvédelmi szempontú ellenőrzése térítés nélkül történik.

Az **egyedi igények szerint összeállított kutatási adatállományok** esetében viszont a kutatási eredmények utólagos adatvédelmi szempontú ellenőrzése során szakértői díj kerül felszámításra, melynek összege 60.000 Ft/fő/nap + ÁFA.

A díj számítása a kutatási eredmények (output) típusától függően az alábbiak szerint történik:

Azonos típusú kutatási eredmények			
Output típus	**Megnevezés**	**Kutatási eredmények darabszáma**	**Ráfordítás embernap**
1. típus	ábrák, grafikonok	max. 25 db ábra	¼ nap
2. típus	statisztikai modell eredmények	max. 50 db futási eredmény	¼ nap
3. típus	átlag, szórás és esetszámokat tartalmazó egydimenziós táblák	max. 25 db táblázat	¼ nap
Vegyes típusú kutatási eredmények			
Output méret típusok	**Kutatási eredmények darabszáma**	**Output irányadó összmérete***	**Ráfordítás embernap**
4. típus	max. 5 db fájl	max. 1 MB	¼ nap
5. típus	max.10 db fájl	max. 3 MB	½ nap
6. típus	max. 25 db fájl	max. 5 MB	$^2/_3$ nap
7. típus	max. 50 db fájl	max. 8 MB	1 nap
8. típus	max 100 db fájl	max. 10 MB	2 nap
9. típus	101 db fájl és fölötte**	-	output darabszáma alapján kerül meghatározásra

*Indokolt esetben eltérhet az output mérete.

**A kutatási projekt kivitelezéséhez elengedhetetlen eredmények, melyet szakmai indokok is alátámasztanak.

Kutatási eredményfájlokra vonatkozó **általános** kikötések:

- Táblázatos adatok (definíciót lásd. Kutatói tájékoztató 20. oldal) esetén az output csak .xlsx, .dta formátumban adható be;
- .dta kiterjesztés esetén 1 db fájl 1 táblázat;
- Excel kiterjesztés esetén egy táblázat egy fájlnak felel meg,

- A specifikációt a típusok részletesebb leírásánál adjuk meg.
- Statisztikai modell eredmények előállítása esetén 1 db fájl legfeljebb 50 db modell eredményt tartalmazhat;
- Ábrák előállítása esetén 1db fájl 1 db ábrával egyenlő;
- Tömörített mappát és fájlt (.zip, .7z,. rar stb.) nem tartalmazhat az output.
- A kutatási eredményeket előállító programkódok nem növelik a kutatási eredmények darabszámát.

Kutatási eredményfájlok egyes típusaira vonatkozó **részletes/speciális** kikötések:

4. típus:
   - A kutatási eredmények csak egyféle KSH-s mikroadatállományból származnak (pl. Munkaerő-felmérés 2010,2011,2012-es évekre).
   - Táblázatos adatok előállítása esetén a fájlok összesen, max. 3000 vizsgálandó cellát tartalmazhatnak.
   - Példa a 4. típusra: 3 db 100x10-es táblázat, 1 db becslési eredmény, 1 db ábra.
   - Az output tartalma, változók átnevezései, a változók képzései könnyen átláthatóak, részletesen dokumentáltak.

5.,6.,7.,8.,9. típus:

- Táblázatos adatok esetén 1 táblázat nem tartalmazhat 12000 cellánál több vizsgálandó értéket, ha mégis tartalmaz, akkor az a fájlok darabszám méretének növekedését eredményezi. (Pl. 1 fájl tartalmaz 144000 vizsgálandó cellát, az 12 db fájlnak feleltethető meg.);
- logfájl (.txt, log, .smcl., .xlsx, .docx) mérete legfeljebb 15.000 sor lehet, nem tartalmazhat táblázatokat; egy programfájlhoz egy logfájl tartozhat.