A nagy számok törvényének teljesülése végtelen nagy aszimptotikus szórás esetén[1]

STEINER FERENC, HAJAGOS BÉLA[2]

A dolgozat vizsgálatai megmutatják, hogy az általános értelemben felfogott „nagy számok törvénye" akkor is teljesülhet, ha az aszimptotikus szórás végtelen.

F. STEINER, B. HAJAGOS: The fulfilment of the law of large numbers in case of infinite asymptotic scatter

The law of large numbers in general sense can be fulfilled even if the asymptotic scatter is infinite.

1. Bevezetés

A matematikai statisztika legfőbb gyakorlati értéke a geotudományok területén is az, hogy olyan algoritmusokat szolgáltat a felhasználónak, amikor az eredmények egyre pontosabbak, ha az n mért adatszám egyre nagyobb. Legáltalánosabban ezt értjük „a nagy számok törvénye" alatt, s fontossága nyilvánvaló: a mindig véges pontosságú geofizikai műszerekkel (pl. graviméterrel) kapott mérési adataink csak kellően nagy n adatszám esetén képesek elegendő pontossággal szolgáltatni a hatóra (pl. geológiai szerkezetre) vonatkozó adatokat.

A legegyszerűbb feladatkörben, amikor csak egyetlen jellemző kellő pontosságú meghatározását tűzzük ki célul, szűkebb értelemben a nagy számok törvénye alatt azt értjük, hogy meghatározásunk szórását $A/\sqrt{n}$ szerint számíthatjuk $n \to \infty$, azaz gyakorlatilag nagy n-ek esetén, ahol A az ún. aszimptotikus szórás, amely az alkalmazott statisztikai algoritmustól és annak az anyaeloszlásnak a típusától függ, amely megfelel n darab adatunknak. (Triviális ugyan, de megemlíthető, hogy az A aszimptotikus szórás arányos az anyaeloszlás S skálaparaméterével.) Az eloszlástípusok széles skáláját írja le az ún. „$f_a(x)$ szupermodell", amelynek valószínűségfüggvénye standard esetben, azaz $T = 0$ helyparaméter és $S = 1$ skálaparaméter esetén

$$f_a(x) = n(a) \cdot (1 + x^2)^{-a/2} \qquad (1 < a < \infty), \tag{1}$$

ahol az a típustól függő normálási faktor a Γ-függvény segítségével az

$$n(a) = \Gamma(a/2) \cdot \Gamma^{-1}[(a-1)/2] \cdot \pi^{-1/2} \tag{1a}$$

kifejezésből nyerhető. Megemlítendő, hogy egész a értékek esetén az $f_a(x)$-ek azonosak az $(a–1)$ szabadságfokú Student-eloszlástípusokkal.

Ismeretes [pl. STEINER 1990], hogy az $\bar{x}$ számtani átlag $A_{\bar{x}}$-val jelölt aszimptotikus szórása az $f_a(x)$ szupermodell típusainál $A_{\bar{x}} = 1/\sqrt{a-3}$, így már $a = 3$ esetén végtelen az értéke. Az imént idézett könyv azonban azt is bemutatja, hogy általánosan felfogva a nagy számok törvényét, ez utóbbi még a $3 \geq a > 2$ típustartományban is teljesül.

A következő, 2. pontban egy más, több gyakorlati tanulsággal szolgáló esetet vizsgálunk meg közelebbről, ahol szintén végtelen aszimptotikus szórás mellett teljesül a nagy számok törvénye.

2. A σ szórás meghatározásának szintén szórással jellemzett bizonytalanságai

2.1. A Monte Carlo-számítások végrehajtásának módja

Végezzük a σ_j szórás számításait az n = 100; 250; 1000; 2500; 10000 és 40000 mintaelem-számokra. Ha az anyaeloszlás véletlenszámait x_i-vel, ezek számtani átlagát $\bar{x}$-sal jelöljük, a jól ismert

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

képlet szerint kapjuk a σ_j szórást. Ha a σ_j ingadozásait is ezek (σ_σ-val jelölt) szórásával akarjuk jellemezni, miután N-szer megismételtük a (2) szerinti σ_j-meghatározást, nyilván a

$$\sigma_\sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\sigma_j - \bar{\sigma})^2} \tag{3}$$

képlet szerint kell számolnunk, ahol $\bar{\sigma}$ a σ_j-értékek számtani átlaga. Az N értékét célszerű nagyra, mondjuk $N = 10\,000$-re választani, mert a σ_σ nagy pontosságú meghatározására törekszünk ebben az elméleti, de gyakorlati célú vizsgálatainkban, amely több látszólagos ellentmondás valódi tartalmát és jelentését kívánja tisztázni. N-et 10 000-nél kisebbre csak a legnagyobb n mintaméreteknél választottuk.

2.2. Az a = 4-gyel jellemzett $f_a(x)$ eloszlástípus esete

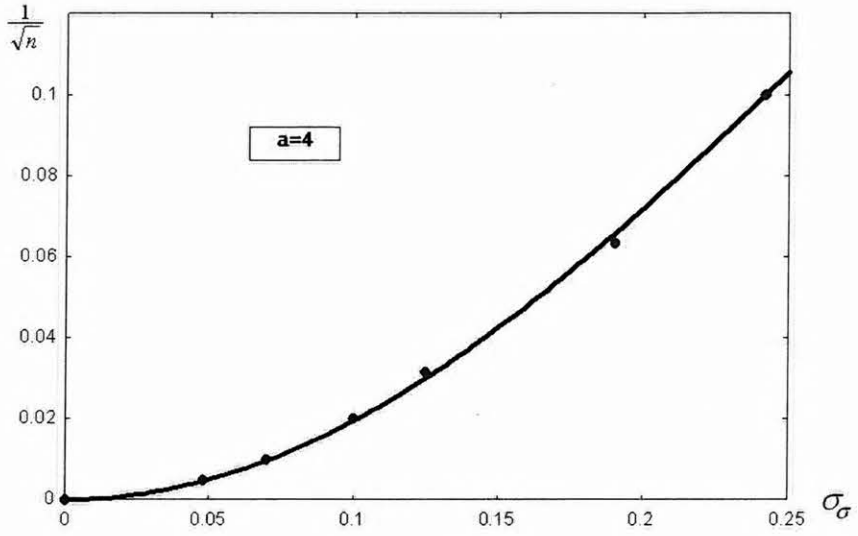
Anyaeloszlásunk legyen először az a = 4 -hez tartozó $f_a(x)$-eloszlás. Ha a hatféle n-hez kapott σ_σ-k értékeit — a szokásokat követve — $1/\sqrt{n}$ függvényében hordjuk fel, az *1. ábrán* látható (nyilván az origóból induló) görbét kapjuk. Célszerűbb azonban most ugyanazon (σ_σ, $1/\sqrt{n}$)

[1] Beérkezett: 2005. július 14-én

[2] Miskolci Egyetem Geofizikai Tanszék, H-3515 Miskolc, Egyetemváros

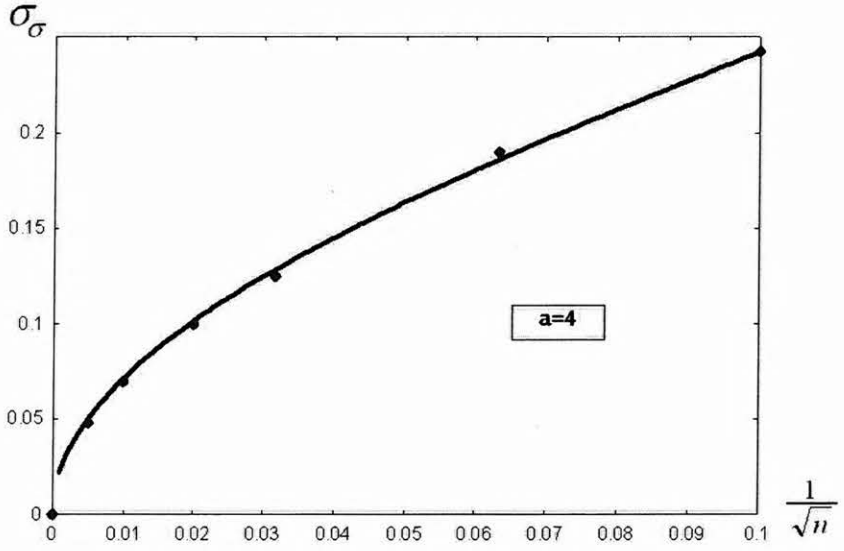
értékpárok pontjait $1/\sqrt{n}$ ordinátájú és σ_σ abszcisszájú koordináta-rendszerben ábrázolni (ld. 2. *ábra*), mert a legnagyobb n-ek tartományában ekkor eredményeinket analitikusan egyszerűen a

$$\sqrt{1/n} = \alpha \cdot \sigma_\sigma^2 + \beta \cdot \sigma_\sigma^4 \qquad (4)$$



1. ábra. A szórások σ_σ-val jelölt szórásának függése $1/\sqrt{n}$ -től (n a mintaméret), ha a minta adatainak eloszlását az $a = 4$-hez tartozó $f_a(x)$ sűrűségfüggvény jellemzi

Fig. 1. The σ_σ scatter of the scatters vs $1/\sqrt{n}$ (n means the number of the data), if the distribution of the data is characterized by $f_a(x)$ in case of $a = 4$



2. ábra. Az $1/\sqrt{n}$ mennyiség görbéje a σ_σ függvényében, $a = 4$ esetén

Fig. 2. The $1/\sqrt{n}$ -curve *vs* σ_σ in case of $a = 4$

képlettel írhatjuk le. Egyszerű számítással az $\alpha = 2$ és a $\beta = -5$ értékek adódnak, de a legfontosabb arra rámutatni, hogy nem létezik lineáris tag, így az origóbeli derivált zérus, ami nyilván egyértelmű azzal, hogy az 1. ábra görbéjének origóbeli deriváltja végtelen. A szokásos megfogalmazással kifejezve eredményünket tehát arra jutottunk, hogy a = 4 esetén a (2) szerinti szórások aszimptotikus szórása végtelen, ugyanezzel egyidőben azonban az 1. ábra pontsora világosan mutatja, hogy a nagy számok törvénye (általános értelemben felfogva) teljesül: *nagyobb n mintaméretekhez a* σ_σ *kisebb volta, azaz a* σ_j*-szórások nagyobb pontossága tartozik.* Végül megemlítjük, hogy α és β egész értékei analitikus vizsgálatok végzésére is csábíthatnának bennünket, ha nem volna eleve bizonyos, hogy ezek

eredményeiből geofizikai szempontból fontos konzekvenciák nem várhatók.

2.3. A geostatisztikus eloszlása esete

Jelenlegi vizsgálatainkban az a = 5-höz tartozó típus fontos határesetnek tekintendő, mivel a szórások relatív aszimptotikus szórását a következő képlet szolgáltatja (ld. [STEINER (*Ed.*) 1997], 96. oldal):

$$\sigma_\sigma / \sigma = \frac{1}{\sqrt{2}} \sqrt{1 + \frac{3}{a-5}} \qquad (\infty > a > 5), \qquad (5)$$

azaz véges értékeket csak $a > 5$ esetén kaphatunk. Az utóbbi típustartományban a növekedésével egyre csökken σ_σ / σ értéke; a Gauss-típusra vezető $a \to \infty$ határesetben érjük el a minimális: $1/\sqrt{2} \approx 0{,}7$ értéket (ld. újra [STEINER (*Ed.*) 1997] imént idézett oldalát).

Az a = 5-re végzett Monte Carlo-számítások meglepetésünkre egyenest közelítő $(1/\sqrt{n}, \sigma_\sigma)$ pontsorra vezettek első lépésben, de szükségesnek tartottuk, hogy az origóhoz közeli tartományt közelebbi vizsgálatok tárgyává tegyük. E célból igen nagy ($n > 40\,000$) mintaméretekre is végeztünk gépidőigényes számításokat, hogy a $(1/\sqrt{n}, \sigma_\sigma)$-görbe esetleges parabolikus indulását legalább nagyon kicsiny σ_σ-intervallumon kimutathassuk. Erőfeszítéseink nem jártak sikerrel, noha a nagy n-ek miatt ezek a Monte Carlo-vizsgálatok *sok száz millió* (!) geostatisztikai véletlenszám generálását igényelték. Arra kell tehát következtetnünk, hogy az $(1/\sqrt{n}, \sigma_\sigma)$-görbe 2. ábrán látható, parabolikusan induló szakasza növekvő a esetén egyre rövidebb lesz, $a \to 5$ esetén pedig már ennek a parabolikus (és ezzel végtelen aszimptotikus szórást eredményező) szakasznak a hossza infinitezimálisan kicsinnyé válik. — Átlépve az $a = 5$ határesetet, tehát $a > 5$ esetén, az (5)-ből láthatóan persze már szó sincs a σ_σ végtelen aszimptótikus szórásáról.

3. Geostatisztikai meggondolások. Gépidő problémák elméleti Monte Carlo-vizsgálatokban és gyakorlati célú számításokban

3.1. A krigeléssel kapcsolatos és egyéb meggondolások

Tegyük fel, hogy egy geológus vagy geofizikus KRIGE professzor nevezetes módszerét óhajtja alkalmazni [ld. pl. STEINER 1990], amelynek kulcsfontosságú függvénye a (fél- vagy szemivariogramnak is nevezett) variogram. Utóbbi azonban szórásnégyzet (azaz variancia) számítása útján határozható meg a mért adatok alapján, s így bizonyos a típusparaméterrel jellemzett eloszlásoknál ugyanakkor találkozunk itt is végtelen aszimptotikus szórásokkal, mint maguknál a szórások számításánál, azaz a variogramra vonatkozóan is az (5) képlet alapján tájékozódhatunk.

A hibák aktuális eloszlástípusa azonban a priori szinte sohasem ismeretes, ezért egy, a matematikai statisztika és a geostatisztika területén egyaránt nagy tapasztalattal rendelkező szaktekintélyhez fordulhatunk tanácsért. JEFFREYS valószínűségelméleti könyvet írt [JEFFREYS 1961], elméleti szaktekintélynek tekinthető tehát, ugyanakkor a geofizikai (elsősorban szeizmológiai) mérések kiértékelésével (ma úgy mondanánk: inverziójával) számos mérési adatrendszer esetében foglalkozott, így komolyan kell vennünk azt a megállapítását [KERÉKFI 1978], hogy Gauss típusú anyaeloszlásból származó hibákkal sohasem találkozott. A XX. század utolsó harmadában rohamosan fejlődő robusztus statisztika könyveiben és dolgozataiban egyre hangsúlyosabb ez a megállapítás: a mérési hibák anyaeloszlásaként Gauss típusú nem fordul elő. Ez ugyan szöges ellentéte annak a klasszikus statisztikában dogmává merevült megállapításnak, hogy a hibák Gauss-típustól való eltérését legfeljebb a mérések helytelen kivitelezésének a számlájára írhatjuk.

Mi tehát az igazság? Forduljunk először újra JEFFREYS tapasztalataihoz, aki mérési hibatípusként a = 6 és a = 10 közötti $f_a(x)$-eloszlásokkal találkozott ugyan, de a Gauss-típushoz ennél közelebb állóval nem, ld. [KERÉKFI 1978] (jogos tehát a $10 > a > 6$ típus-tartományt „Jeffreys-intervallumnak" nevezni).

Hogy a geostatisztikában milyen típus várható a legnagyobb valószínűségsűrűséggel, arra vonatkozóan célszerű [DUTTER 1986/87] ábrájára pillantanunk, amelyet 8.8 ábraként [STEINER 1990] mint valószínűségeloszlási görbepárt minden változtatás nélkül vett át, csak feliratokkal látta el. Mindkét eloszlás egységnyi szórású, az egyik a standard Gauss-eloszlás sűrűségfüggvénye, a másik pedig olyan típust jellemez, amellyel (vagy ahhoz közelállóval) hibaeloszlásként DUTTER a geostatisztikában a leggyakrabban találkozott, ez pedig az a = 5-höz tartozó $f_a(x)$-eloszlás. Úgy gondolom, hogy bízvást hihetünk DUTTER professzornak, hiszen egyrészt közös cikkei jelentek meg HUBERrel (akit akár a robusztus statisztika pápájaként is aposztrofálhatnánk), másrészt a leobeni Montanuniversität geostatisztika előadója. Ezért az a = 5-tel jellemzett típust joggal láthatjuk el a „geostatisztikai" jelzővel, sőt, mivel a Miskolci Egyetem Geofizikai Tanszékének geostatisztikai munkacsoportja egyéb tudományterületek adatrendszereinek típusvizsgálatainál is meglepően sokszor kapott a = 5-höz közeli eredményeket, általánosabban talán a „statisztikai" jelző is jogos lenne. Ha ezt elfogadjuk, a sűrűségfüggvény jele $f_{st}(x)$ lehet, amelyet (az (1) és (1a) alapján, a = 5-tel) az

$$f_{st}(x) = \frac{3}{4}(1 + x^2)^{-5/2} \qquad (6)$$

alakban írhatunk fel standard esetben.

Persze a ritkábban előforduló hibaeloszlásokról sem feledkezhetünk meg. LANDY és LANTOS [1982] Cauchy-eloszlással közelíthető geofizikai adatrendszert mutat be, ami nyilván az a = 2 típus az $f_a(x)$ szupermodellből. Figyelemre méltó [TARANTOLA 1987] megjegyzése is, amely szerint ha ismeretlen elhelyezkedésű outlierek létezését jogos feltételeznünk, a statisztikai procedúrát célszerű Cauchy típusú hibák feltételezésével kialakítanunk. — Bár elenyésző számban a geofizikában $a < 2$ típus is előfordulhat (ld. pl. [STEINER (*Ed.*) 1997] 2.4 alfejezetét, ahol a = 1,6-ra találunk példát), a Cauchy-eloszlás valószínűségsűrűségét kell a nagy szárnyak tartományában még nem elhanyagolhatónak tekintenünk, de ez az érték még a

Jeffreys-intervallum $f_{st}(x)$-nél rövidebb szárnyú eloszlástípusait jellemző értékeknél is valószínűleg kisebb.

Ha a fentieket egyszerű formulába akarjuk sűríteni, akkor — bevezetve a $t=1/(a-1)$ jelölést, — a típuselőfordulások $g(t)$ valószínűségsűrűség függvényét így írhatjuk fel:

$$g(t)=16\cdot t\cdot e^{-4t} \qquad (7)$$

A klasszikus szemlélettel szemben álló $g(0)=0$-t (azaz hogy pontosan Gauss-eloszlást nem várhatunk anyaeloszlásként) némileg feloldja az, hogy a Gauss-típushoz közelálló Jeffreys-intervallum valószínűségsűrűségei még a $g(t)$-görbe maximális értékéhez viszonyítva sem mondhatók kicsinynek, — a Cauchy-eloszlást azonban már e maximális értéknek is csak kb. 20%-a jellemzi. (A (7) képlet $g(t)$-görbéjét [STEINER 1990] a 233. oldalon mutatja be.)

Térjünk vissza a krigelést végrehajtani kívánó barátunkhoz, aki lelkiismeretesen tanulmányozva a geostatisztikai szakirodalmat, rábukkan egyrészt DUTTER professzor állítására, amely szerint mérési eredményeinek hibáit a legvalószínűbben a (6) szerinti $f_{st}(x)$, azaz az a = 5-höz tartozó $f_a(x)$ sűrűségfüggvény írja le, de megtalálja az (5) formulát is, amely szerint — első pillanatra — a variogram számításához nem is érdemes hozzáfognia, hiszen a = 5-höz igen közeli, de a < 5-tel jellemzett esetekben már igen nagy pontatlanságtól félhet, hiszen a szórás, így a variancia aszimptotikus szórása is végtelen. A jelen cikk 2. pontját elolvasva azonban megnyugodhat: a nagy számok törvénye a geostatisztikus hibaeloszlás környékén, pl. a = 4-nél is „működni" fog (ha nem is olyan hatékonyan, mint véges aszimptotikus szórásoknál), de persze csak a $(\infty > a > 3)$ tartományban, hiszen a = 3 esetén már a szórás $1/\sqrt{a-3}$ szerint számítandó elvi értéke is végtelen.

3.2. Gépidő problémák geostatisztikus elméleti vizsgálatok és a geofizikai praxis számításainál

Rövidség kedvéért nem másoljuk át dolgozatunkba azokat a jól ismert klasszikus statisztikai tételeket, amelyek felületes ismerete többnyire úgy marad meg az olvasóban, hogy ha a mérési hiba igen sok igen kicsiny hatás szuperpozíciójaként jön létre, az Gauss típusú lesz. Hangsúlyosan idézzük viszont CRAMÉR szinte sohasem idézett tételét [CRAMÉR 1945], amely szerint *a hatások szuperpozíciójaként előálló hiba akkor és csakis akkor lesz Gauss típusú, ha minden egyes komponens is már eleve Gauss típusú volt.* Ez a tétel akár annak a tapasztalatnak az elméleti alátámasztásaként is felfogható, hogy anyaeloszlásként miért nem kapunk szinte sohasem Gauss típusú mérési hibákat.

A fentiek (és a 3.1. pontban megbeszéltek) után szinte érthetetlen, hogy geofizikai tárgyú, és egyébként magas színvonalú eredményeket felmutató PhD-értekezésekben mindmáig Gauss típusú hibát alkalmaznak szimulációs vizsgálatoknál (a szerzők itt illendőnek találják a hivatkozások mellőzését). Nyilván a (6) sűrűségfüggvényű geostatisztikus hibák szuperpozíciója volna sokkal inkább indokolt. Geostatisztikus véletlenszámok az $F_{st}(x)$ eloszlásfüggvény ismeretében a szokásos módon generálhatók: a gép által szolgáltatott, a (0,1) intervallumban egyenletes eloszlású x_u véletlenszámokat azoknak az értékeknek tekintjük, amilyen valószínűséggel a geostatisztikus adatok $-\infty$-től x_u-ig előfordulnak. [STEINER 1990] 5.2 ábrája ezt mutatja, csak az ott berajzolt eloszlásfüggvény analitikus alakja most

$$F_{st}(x)=\frac{1}{2}+\frac{x}{2\sqrt{1+x^2}}+\frac{x}{4}\left(1+x^2\right)^{-3/2} \qquad (8)$$

(ld. a [STEINER 1990] 50. oldalán látható képleteket).

Nyilvánvaló, hogy a geostatisztikus véletlenszámok mindegyikének fenti előállítása valamilyen iterációs algoritmus alkalmazását igényli, ami hosszabb számítás, mintha rendelkezésünkre állna (8) inverz függvénye, azaz $F_{st}^{-1}(x_u)$. Az iteráció nem nevezhető hosszadalmasnak s így csak akkor okoz problémát, ha pl. elméleti célú Monte Carlo-vizsgálatainkban olyan extrém nagyszámú geostatisztikus véletlenszámra van szükségünk, mint jelen dolgozatunk 2. pontjában. Hasonló lehet a helyzet az inverzió eredményeit pontosabbá tevő, ún. „többlethibamódszer" olyan alkalmazásakor, amikor igen nagyszámú többlethibát szuperponálunk a természetes hibákra (az idézett módszer ismertetésére nézve ld. a [STEINER 2002] dolgozatot). — Mind a többlethiba-módszer esetén, mind nagypontosságú elméleti Monte Carlo-vizsgálatoknál (a 2. pontbeli vizsgálatok sok százmilliós nagyságrendű geostatisztikai véletlenszám generálását igényelték,) nagy segítség az a [HAJAGOS 2005] dolgozatbeli eredmény, amely az $F_{st}(x)$ inverzét szolgáltatja.

A jelen dolgozat kidolgozása a T 049852 számú OTKA-kutatás keretében történt.

HIVATKOZÁSOK

CRAMÉR H. 1945: Mathematical Methods of Statistics. Almquist & Wksells, Uppsala

DUTTER R. 1986/87: Mathematische Methoden in der Montangeologie. Vorlesungsnotizen. Leoben

HAJAGOS 2005: Geostatisztikus eloszlású véletlenszámok gyors előállítása. Magyar Geofizika **46**, 2

JEFFREYS H. 1961: Theory of Probability. Clarendon Press, Oxford

KERÉKFI P. 1978: Robusztus becslések. Alkalmazott Matematikai Lapok **4**

LANDY I., LANTOS M. 1982: Praktisches Beispiel zur Cauchyschen Verteilung. Publications of the Technical University for Heavy Industry, Series A Mining **37**, 1–2. Miskolc

STEINER F. 1990: A geostatisztika alapjai. Tankönyvkiadó, Budapest

STEINER F. (*Ed.*) 1997. Optimum Methods in Statistics. Akadémiai Kiadó, Budapest

STEINER F. 2002: A mérési adatokból nyert információk hibáinak csökkentése általunk ismételten generált többlethibáknak a mérési adatokra történő szuperponálásával. Magyar Geofizika **43**, 2

TARANTOLA A. 1987: Inverse Problem Theory. Elsevier, Amsterdam