Modern statisztikai módszerek a geofizikai-geológiai törvényszerűségek megbízható felismerésének szolgálatában[1]

STEINER FERENC, HAJAGOS BÉLA, HURSÁN LÁSZLÓ[2]

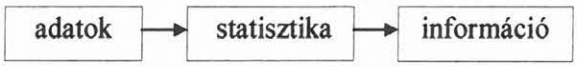
A dolgozat egy karotázs példán szemlélteti, hogy a statisztika modern optimum-módszerei milyen megbízhatósággal képesek törvényszerűségek feltárására is. A dolgozat részét képezi egy rövid, lényegre törő összefoglalás a statisztika optimum-módszereiről.

F. STEINER, B. HAJAGOS, L. HURSÁN: Modern statistical optimum-methods for discovering unambiguously geophysical-geological regularities

It is shown on a practical example that modern statistical optimum-methods are able to discover regularities, too. The paper gives also a short, concise summary about the optimum-methods of statistics.

1. Bevezetés és célkitűzés. Geofizikai-geológiai mérési adatok és információk

A statisztika — teljesen általánosan, a köznapi életben is, — az adatok nyersanyagából származtatja a döntéshez szükséges, közvetlenül felhasználható információt, így szerepét az alábbi egyszerű blokksémával érzékeltethetjük:



(1)

Talán túl vulgáris a példa, de a sarki fűszeres sem egyetlen szombati napon eladott zsemlék darabszáma szerint fogja rendelését feladni, hanem több ilyen adat számtani átlaga, mint információ alapján rendel a következő szombatra. Eközben ugyanúgy nincsen tudatában (de nem is kell, hogy tudatában legyen) annak, hogy az átlagképzéssel az L_2 eltérésnorma minimumhelyét határozta meg, tehát (ösztönösen bár, de) *statisztikát* alkalmazott, — mint ahogyan MOLIÈRE közismert hőse sem tudta, hogy egész életében *prózában* beszélt.

Szakmához közelebb álló példára térünk át: valamely széntelepből származó egyetlen mintán végzett fűtőérték-meghatározás végeredménye egyetlen adat, amelynek alapján senki sem kockáztat semmiféle, pl. külszíni fejtéssel kapcsolatos döntéshozást. Ha azonban nagyszámú x_1, x_2,..., x_i,..., x_n ilyen adatunk van ugyanabból a telepből, akkor átlagképzéssel vagy modernebb módszerekkel egyetlen olyan adatot származtathatunk (pl. T-vel jelölve, a modern statisztikából vett, helyparaméter-jellegre utaló jelöléssel), amely már információ értékű.

Elérkezve végül saját, geofizikai-földtani szakterületünkre: az információ pl. valamely réteghatár-mélység egy adott helyen, de közvetlen mérési adataink (az alkalmazott geofizikai módszer szerint) pl. a látszólagos fajlagos ellenállások, esetleg a nehézségi gyorsulás értékének helyről helyre meghatározott változásai, a rugalmas hullámok beérkezési idői stb. Szakterületünkön tehát a fizikai összefüggések sokféleségének az ismerete, esetleg bonyolult algoritmusok alkalmazása szükséges ahhoz, hogy akár csak a direkt feladat („forward modelling") is megoldható legyen, azazhogy — teljesen általánosan fogalmazva — egy (a realitáshoz minél közelebb álló) modellhez az $\overline{y}_i$ pontra meghatározhassuk a számított $\xi_i \equiv \xi(\overline{p}; \overline{y}_i)$ értéket, még ismert p_1, p_2, ..., p_j, ..., p_J modellparaméterek esetén is (az y_i alatt egy, két vagy három helykoordináta értéket értünk attól függően, hogy mérési pontjaink elhelyezkedése szelvény menti, közel síkbeli — pl. alföldi terepmérések esetén, — vagy térbeli).

A mérési pontokban meghatározott primer x_i értékek még abban az esetben sem lennének azonosak a ξ_i értékekkel, ha modellünk történetesen ideálisan írná le a valóságot, hiszen kisebb vagy nagyobb mérési hiba mindig terhelni fogja az x_i értékeket. Az általános esetben eltérésnek nevezett

[1] Beérkezett: 1997. október 15-én

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

$$X_i = x_i - \xi_i \qquad (2)$$

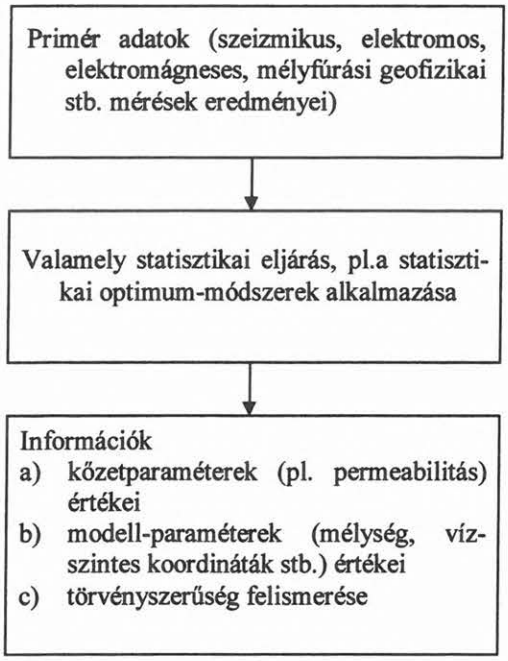
különbségeket tehát irreális volna zérus értékűnek megkövetelni, azt viszont reális célként tűzhetjük ki, hogy azok összességükben minél kisebb értékűek legyenek. Az *1. táblázat*ból tehát egy, *a hibák típusához legjobban megfelelő eltérés-normát választunk, és a* p_1, p_2, ..., p_j,..., p_J *információ jellegű paramétereknek azokat az értékeit fogjuk helyeseknek elfogadni, amelyek minimalizálják a választott eltérés-normát.* Ezt a feladatot egyre több területen már gazdaságos a *„globális optimalizáció"* valamelyik algoritmusával megoldani (ld. pl. SZŰCS [1995], KIS [1996]). Ha röviden *statisztikai optimum-módszer*ről beszélünk, ez alatt valamely *norma-minimalizálási eljárást* értünk, akkor is, ha nem globális optimalizációval érünk célt.

Norma	Képlet	Sajáteloszlás (a hibaeloszlás megadott típusánál vezet a norma optimális statisztikai algoritmusra)	A sajáteloszlás jele és sűrűségfüggvénye (standard esetben)
L_1	$\frac{1}{n}\sum_{i=1}^{n} \lvert X_i \rvert$	Laplace	$f_L(X) = \frac{1}{2} \cdot e^{-\lvert X \rvert}$
L_2	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$	Gauss	$f_G(X) = \frac{1}{\sqrt{2\pi}} \cdot e^{-X^2/2}$
P_J P_J^*	$\varepsilon \cdot \left\{ \prod_{i=1}^{n} \left[1 + (X_i/3\varepsilon)^2 \right] \right\}^{1/2n}$ $20\varepsilon/n \cdot \sum_{i=1}^{n} X_i^2/(27\varepsilon^2 + X_i^2)$	Jeffreys	$f_J(X) = \frac{35}{32\left[\sqrt{1+X^2}\right]^9}$
P P^*	$\varepsilon \cdot \left\{ \prod_{i=1}^{n} \left[1 + (X_i/2\varepsilon)^2 \right] \right\}^{1/2n}$ $10\varepsilon/n \cdot \sum_{i=1}^{n} X_i^2/(12\varepsilon^2 + X_i^2)$	geostatisztikus	$f_{st}(X) = \frac{3}{4\left[\sqrt{1+X^2}\right]^5}$
P_C P_C^*	$\varepsilon \cdot \left\{ \prod_{i=1}^{n} \left[1 + (X_i/\varepsilon)^2 \right] \right\}^{1/2n}$ $4\varepsilon/n \cdot \sum_{i=1}^{n} X_i^2/(3\varepsilon^2 + X_i^2)$	Cauchy	$f_C(X) = \frac{1}{\pi} \frac{1}{1+X^2}$
P_{lt} P_{lt}^*	$\varepsilon \cdot \left\{ \prod_{i=1}^{n} \left[1 + (2X_i/\varepsilon)^2 \right] \right\}^{1/2n}$ $5\varepsilon/2n \cdot \sum_{i=1}^{n} X_i^2/(3\varepsilon^2/4 + X_i^2)$	Az összes P_k és P^*_k normánál az ε-nal jelölt dihéziónak a következő egyenletet kell teljesítenie: $\varepsilon^2 = 3 \cdot \frac{\sum_{i=1}^{n} X_i^2/(X_i^2+\varepsilon^2)^2}{\sum_{i=1}^{n} 1/(X_i^2+\varepsilon^2)^2}$	

1. táblázat. Normák és sajáteloszlásaik

Table 1. Norms and eigen-distributions

A jelen dolgozat azt szeretné hangsúlyozni és mint leglényegesebbnek tartott mondanivalóját példával is bemutatni, hogy *modern statisztikai eljárások, információként alkalmazva, törvényszerűségek világos felismeréséhez is vezethetnek.* Így az általános, de túl szűkszavú (1) blokkséma szakterületünkre vonatkozóan a (3) blokkséma szerint részletezendő.



2. Statisztikai normák és optimum-elvek

2.1. Eltérés-normák

A (3) blokkséma lényegi részét képező optimum-módszerek kellő alaposságú ismeretében a jelen cikk olvasója a 2. és 3. pont elhagyásával azonnal a 4. pontra térhet át. Ellenkező esetben viszont a szerzők farizeuskodásnak tartanák a maguk részéről, ha puszta utalás történnék három könyvre [STEINER 1990, STEINER (*Ed.*) 1991, STEINER (*Ed.*) 1997], amelyek összterjedelme meghaladja az ezer oldalt (pláne az lenne, ha az utolsónak felsorolt könyv bibliográfiájára történnék hivatkozás, ahol a cikkek terjedelme összesen kb. 2000 oldal). Elhagyhatatlannak tűnik tehát egy minél rövidebbre szabott, de érthető és egy-egy részletet kiemelő összefoglalás.

Az 1. táblázat 10 féle eltérés-normát definiál, megadva mindegyikhez azt a hibaeloszlás-típust is, amelynél a norma minimumhelyének a meghatározása optimális (100%-os hatásfokú) statisztikai algoritmust definiál. Mind a tíz esetben X_i $(1 \le i \le n)$ jelentheti a (2) egyenlet szerinti eltérést (általában $J > 1$ db meghatározandó paraméterérték esetén, amely J darabszámra lehetőleg $J \ll n$ teljesül), vagy egyszerűen az $x_i - T$ különbséget, ha direkt méréseket végzünk egyetlen ismeretlenre. A P_J-, P-, P_C- és P_{lr}-normák közül az utolsó a Cauchy-félénél is súlyosabb szárnyakkal bíró (long tailed) hibaeloszlásoknál működik 100%-hoz közeli hatásfokkal, az index nélküli P-norma pedig éppen azért nincs indexszel ellátva, mivel ennek alkalmazása javasolható abban az esetben, ha nincs semmiféle előzetes információnk a hibaeloszlás típusáról. A felső *-gal jelölt variánsok nagyon hasonlóan viselkednek, mint * nélküli párjaik, de speciális esetekben előnyösebb lehet outlierekkel (durva hibájú adatokkal) szembeni nagyobb érzéketlenségük, azaz nagyobb rezisztenciájuk. Az 1. táblázat tartalmazza az ε-nal jelölt dihéziót definiáló formulát is (az X_i eltérések nagy valószínűséggel esnek 2ε hosszúságú intervallumba).

2.2. Optimum-elvek

Néhány olyan optimum-elvet írunk fel, amelyek a 2.1.-ben ismertetett normákhoz szorosan kapcsolódnak.

A statisztikai elv neve:	A teljesítendő követelés:	
„a legkisebb négyzetek elve" (1795)	$\sum_{i=1}^{n} X_i^2 = \min.$	(4)
„a legnagyobb reciprokok elve" (1965)	$\sum_{i=1}^{n} \frac{1}{S^2+X_i^2} = \max.$	(5)
„a legkisebb szorzatok elve" (1988)	$\prod_{i=1}^{n} \left[S^2 + X_i^2\right] = \min.$	(6)

Az S értékeket célszerű 2ε-nak választani (az ε dihézió meghatározási formulája látható az 1. táblázat jobb alsó sarkában). Ha elvégezzük ezt a behelyettesítést a (6) kifejezésben, nyilvánvaló, hogy az 1. táblázatbeli P minimalizálásával a legkisebb szorzatok elvének teszünk eleget, hiszen P-nek ε-nal való osztása, $2n$-ik hatványra emelése, valamint mindegyik szorzótényezőnek $(2\varepsilon)^2$-tel való szorzása után valóban a

$$\prod_{i=1}^{n} \left[(2\varepsilon)^2 + X_i^2\right] = \min. \qquad (7)$$

kifejezést nyerjük (figyeljük meg, hogy egyetlen felsorolt átalakítás sincs hatással a minimumhelyre).

2.3. A leggyakoribb érték

Legyen adva a legegyszerűbb esetünk: az x_i -k direkt mérési eredményeket jelentenek valamely mennyiségre, így általános esetben $X_i=x_i-T$ írandó. Ha azonban a (7) követelés minimumhelyét keressük, legyen szabad T helyett azonnal M-et (vagy *MFV*-t) írni, anticipálva annak az ismeretét, hogy így a legnagyobb adatsűrűségre jellemző, ezért „leggyakoribb értéknek" nevezhető érték fog eredményül adódni; a *most frequent value* betűszava pedig *MFV*, ezt tovább egyszerűsítve M-et írhatunk.

Kiindulásunk tehát a

$$\prod_{i=1}^{n}\left[(2\varepsilon)^2+(x_i-M)^2\right]=\min. \qquad (8)$$

követelés; célunk a követelést teljesítő M meghatározása. Logaritmizálás után (amellyel természetesen nem változik a minimumhely,) összeg alakú feltételt kapunk, amelynek extrémumhelyén nyilván teljesülnie kell a

$$\frac{\partial}{\partial M}\sum_{i=1}^{n}\ln\left[(2\varepsilon)^2+(x_i-M)^2\right]=0 \qquad (9)$$

egyenlőségnek. Elvégezve a differenciálást, egyszerűsítés és átrendezés után M-et a következőképpen fejezhetjük ki:

$$M=\frac{\sum_{i=1}^{n}\dfrac{x_i}{(2\varepsilon)^2+(x_i-M)^2}}{\sum_{i=1}^{n}\dfrac{1}{(2\varepsilon)^2+(x_i-M)^2}} \qquad (10)$$

Mivel M a jobboldalon is szerepel, a (10)-et iterációs utasításnak fogjuk fel, amit pl. $M_{ind}=\overline{x}$ -nál indíthatunk, ahol $\overline{x}$ az x_i-k számtani középértékét jelenti. Az ε ugyan eleve ismert is lehet, de ellenkező esetben az 1. táblázat ε formuláját egyszerűen átírhatjuk az $X_i = x_i - M$ esetre:

$$\varepsilon^2=3\cdot\frac{\sum_{i=1}^{n}\dfrac{(x_i-M)^2}{\left[(2\varepsilon)^2+(x_i-M)^2\right]^2}}{\sum_{i=1}^{n}\dfrac{1}{\left[(2\varepsilon)^2+(x_i-M)^2\right]^2}} \qquad (11)$$

és ez szintén tekinthető iterációs eljárás definiálásának, amit pl. a jobboldalon $\varepsilon_{ind}=\max(x_i)-\min(x_i)$ értékkel (és az $M_{ind}=\overline{x}$ helyettesítéssel) indíthatunk. Az így kapott ε-nal (és $M=\overline{x}$-sal) számíthatjuk a következő lépésben a (10) kifejezést. A kapott M (és az első lépésben nyert ε) kerül a (11) jobboldalára, ami új ε-ra vezet, és így tovább. Ez az első pillanatban bonyolultnak tűnő, de még PC-szinten is észlelhetetlenül rövid idő alatt befejeződő „ping-pong iteráció" a keresett M mellett az ε dihéziót is szolgáltatja, ha ez a vizsgált esetben nem volna eleve ismert.

3. Néhány norma tesztelése általános szupermodellen

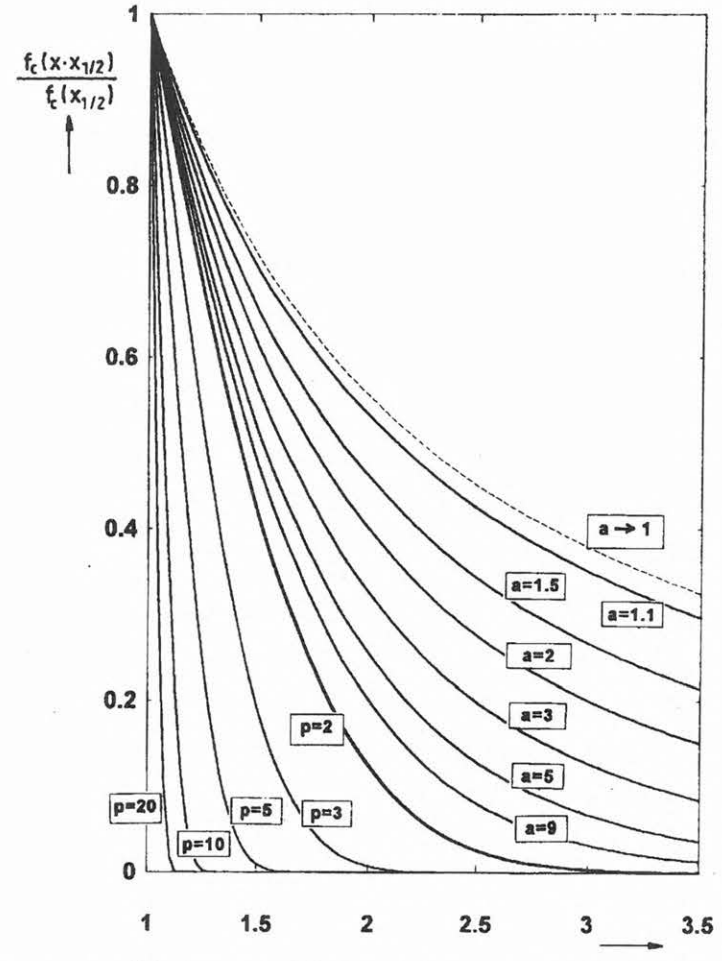
Az 1. táblázat ugyan közli, hogy milyen hibatípusnál 100%-os az illető norma hatásfoka, de semmilyen tájékoztatást nem nyújt arról, hogy egyéb hibatípusoknál mennyire működik hatékonyan az illető norma minimumhely-kereséseként definiált statisztikai algoritmus. A táblázatból az L_1-, L_2-, P_J- és P-normákat teszteltük, amihez persze célszerű minél általánosabb szupermodellen végezni a vizsgálatokat. Választásunk természetszerűleg esett a STEINER, HAJAGOS [1998] dolgozatban definiált $f_c(x)$ szupermodellre: az indexben szereplő c a szupermodell *complete* voltára utal.

3.1. A harang alakú hibaeloszlástípusok $f_c(x)$ szupermodellje

Az $f_c(x)$ típuscsalád sűrűségfüggvényeit standard alakban, azaz a $T=0$ és $S=1$ esetre a következő képletekkel definiáljuk:

$$f_c(x)=\begin{cases}\dfrac{p^{1-1/p}}{2\cdot\Gamma(1/p)}e^{-|x|^p/p} & (\infty>p\geq 2)\\[2ex] \dfrac{\Gamma(a/2)}{\sqrt{\pi}\cdot\Gamma((a-1)/2)}\cdot\dfrac{1}{(1+x^2)^{a/2}} & (\infty>a>1)\end{cases} \qquad (12)$$

A p=2 esetre nyilván a Gauss-típus adódik az első formulából, de jól ismert [pl. STEINER 1990], hogy $a \to \infty$ esetén a második formula is (amely önmagában az $f_a(x)$ szupermodell képlete,) a Gauss-típus sűrűségfüggvénye felé tart. Hogy az $f_c(x)$ típusai folyamatos átmenettel realizálják az összes szárnyhosszúságot zérustól a hibaeloszlás szárnyainak extrém nagy súly értékeiig, arra nézve lássuk az *1. ábrát.*



1. ábra. Az $f_c(x)$ típuscsalád által modellezett hibaeloszlások szárnytartományai

Fig. 1. Flanks of the error types $f_c(x)$ defined by Eq.(12)

Itt az $f_c(x \cdot x_{1/2}) / f_c(x_{1/2})$ értékeket hordtuk fel az x függvényeként az $1 \le x \le 3{,}5$ intervallumban, a típusparaméterek p = 20, 10, 5, 3, 2, ill. a = 9, 5, 3, 2, 1,5 és 1,1 értékeire (szaggatott vonallal a $\sqrt{2/(1+x^2)}$ görbét is feltüntettük, amelyhez $a \to 1$ esetén tart az $f_c(x \cdot x_{1/2}) / f_c(x_{1/2})$ hányados görbéje). Az $x_{1/2}$ értéket az $f(x_{1/2}) = f_{max}/2$ definiálja, azaz — önkényesen — a sűrűségfüggvényeknek (balról és jobbról egyaránt) azokat a végtelenbe nyúló szakaszait tekinti az 1. ábra szárnyaknak, ahol a sűrűségfüggvény már a modusznak (a szimmetriapontbeli maximális $f_c(0)$ értéknek) a felét sem éri el. A szárnyaknak ettől eltérő definíciója sem szolgáltatna más következtetéssel, mint első ábránk: a statisztikai módszerek effektivitására nagy hatással levő szárnyakat az $f_c(x)$ szupermodell nullától a lehetőségek maximumáig jól modellezi.

3.2. A Gauss-eloszlástól mért típustávolság definíciója

Ha Φ-vel jelöljük a Gauss-féle eloszlásfüggvényt, akkor az origóra szimmetrikus, egyébként tetszőleges, valamely F eloszlásfüggvénnyel jellemzett valószínűségi változó típusának a távolságát a Gauss-típustól a

$$D[\Phi, F] = \min_S \left\{ \max_x \left[\left| \Phi(0,1;x) - F(0,S;x) \right| \right] \right\} \qquad (13)$$

kifejezés adja meg (ld. a HAJAGOS, STEINER [1994] szerinti általános definíciót).

A $D(\Phi,F)$ — távolság lévén — mindig pozitív érték, márpedig $F=F_c$ esetén ez a távolság ugyanúgy jellemezhet a Gauss-félénél rövidebb vagy súlyosabb szárnyakat: az utóbbira példa az a=5-tel jellemzett geostatisztikus eloszlás, amelyre a (13) kifejezés a 0,016 értéket szolgáltatja, de ugyanez a $D(\Phi,F_c)$ típustávolság adódik a rövid szárnyakkal jellemzett p=10/3 hibatípusra is. Megkülönböztetés céljából ezért az alábbi jelöléseket vezetjük be:

$D = D(\Phi,F_c)$, ha az $f_c(x)$ szárnyai súlyosabbak a Gauss-eloszlás szárnyainál (14)

$D^- = D(\Phi,F_c)$, ha az $f_c(x)$ szárnyai rövidebbek a Gauss-eloszlás szárnyainál (15)

Mivel $f_c(x)$ $p \to \infty$ esetén az egyenletes eloszláshoz tart, könnyen meggyőződhetünk arról, hogy D^- ugyanakkor a

$$D^-_{max} \approx 0{,}048 \qquad (16)$$

értékhez tart. Ami D-t illeti, ez a

$$D_{max} = 1/4 \qquad 17)$$

értékhez konvergál, ha $a \to 1$ [CSERNYÁK 1995]; ugyanez a cikk bizonyítja be, hogy ez egyben a maximális lehetséges távolság két szimmetrikus eloszlástípus között).

1 Az aktuális eloszlás modellként szolgáló $f_c(x)$ típusparamétere p és a	2 Az $f_c(x)$-típus távolsága a Gauss-féle eloszlástípustól D^- vagy D		3 Az $f(x_{1/2}) = f_{max}/2$ egyenlettel definiált $x_{1/2}$ ($2x_{1/2}$ az ún. félterjedelem) $x_{1/2}$	4 Egyetlen $f_c(x)$ -szárny W_{fl} súlya a $P(x > x_{1/2})$ valószínűséggel mérve W_{fl}
$p \to \infty$	0,04804	▲	1,00000	0,00000
$p = 200$	0,04745		1,02497	0,00095
$p = 100$	0,04686		1,04330	0,00191
$p = 50$	0,04569		1,07348	0,00384
$p = 20$	0,04222		1,14049	0,00981
$p = 15$	0,04031		1,16900	0,01323
$p = 10$	0,03656	D^-	1,21363	0,02028
$p = 20/3$	0,03119		1,25809	0,03132
$p = 5$	0,02584		1,28221	0,04289
$p = 4$	0,02088		1,29039	0,05490
$p = 10/3$	0,01619		1,28562	0,06730
$p = 3$	0,01322		1,27639	0,07575
$p = 5/2$	0,00762		1,24597	0,09301
$p = 20/9$	0,00370		1,21458	0,10619
$p = 2$	0,00000	▼	1,17741	0,11952
$a = 9$	0,00803	▲	0,40808	0,14086
$a = 5$	0,01601		0,56525	0,16073
$a = 3$	0,03129		0,76642	0,19585
$a = 2{,}5$	0,04090		0,86087	0,21605
$a = 2$	0,05807		1,00000	0,25000
$a = 1{,}75$	0,07241	D	1,09917	0,27713
$a = 1{,}5$	0,09742		1,23282	0,31733
$a = 1{,}25$	0,14210		1,42528	0,38205
$a = 1{,}15$	0,17276		1,52916	0,42007
$a = 1{,}1$	0,19150		1,58945	0,44407
$a = 1{,}05$	0,21536		1,65664	0,46946
$a = 1{,}01$	0,24077		1,71622	0,49352
$a \to 1$	1/4 = 0,25000	▼	$\sqrt{3}$ = 1,73205	1/2 = 0,50000

2. táblázat A (12) egyenlettel definiált $f_c(x)$ eloszláscsalád jellemző adatai

Table 2. Characteristic data of the supermodel $f_c(x)$ defined by Eq.(12)

A *2. táblázat* az $f_c(x)$ szupermodell p ill. a típusparamétereihez adja meg a D^-, ill. D távolságokat, az $x_{1/2}$ értékeit, valamint egyetlen szárnynak a $P(x > x_{1/2})$ valószínűséggel definiált W_{fl} súlyát. A táblázat 2. és 4. oszlopainak összehasonlításából kiderül, hogy ha abszcisszaként a $D^-_{max} \approx 0{,}048$ értéktől jobbra haladva lineárisan csökkenő D^- értékeket szerepeltetünk nulláig, onnan pedig (persze azonos skálabeosztással) növekvő D-ket 0,25-ig, akkor ez az abszcissza a szárnyak monoton növekvő súlyainak felel meg. Célszerű tehát ezt az abszcisszát alkalmazni minden általános, az $f_c(x)$-re támaszkodó vizsgálatnál; ezt fog-

juk a következőkben tenni a statisztikai hatásfokok analízisekor is.

3.3. Hatásgörbék és robusztussági mérőszámok

Az *e* statisztikai hatásfok definíciója jól ismerten

$$e = \frac{A^2_{\min}}{A^2} \cdot 100\% \qquad (18)$$

ahol A^2 az éppen alkalmazott statisztikai (pl. helyparamétert meghatározó) algoritmus aszimptotikus szórásnégyzete. A leggyakoribb értékek számítása esetén pl. ez a következő formulából határozható meg:

$$A^2 = \frac{\int_{-\infty}^{\infty} \frac{x^2}{\left[(k\varepsilon)^2 + x^2\right]^2} f_c(x)\mathrm{d}x}{\left[\int_{-\infty}^{\infty} \frac{(k\varepsilon)^2 - x^2}{\left[(k\varepsilon)^2 + x^2\right]^2} f_c(x)\mathrm{d}x\right]^2} \qquad (19)$$

(k=2 esetén a P-norma minimalizálásának megfelelő M-meghatározás, k=3 esetén a P_J szerinti leggyakoribb érték számítás történik, így k=2 esetén A^2 helyébe $A^2(M)$-et, k=3-nál pedig $A^2(M_J)$-t írhatunk). Amennyiben a fentiekben említett aszimptotikus szórásnégyzet, vagy az alábbiakban előforduló fogalmak nem lennének kellő mélységben ismertek az olvasó előtt, azokat részletesebben ismerteti a STEINER-monográfia [1990], sőt abban az alább összefoglalt képletek egy része is megtalálható, pl. a (20) egyenlet az idézett könyv 5.5. táblázatából azonnal következik.

A (18)-ból láthatóan a hatásfok számításához az $A^2_{\min}$ képletekre is szükségünk van:

$$A^2_{\min} = \begin{cases} \dfrac{\Gamma(1/p)}{\Gamma(2-1/p)} \cdot p^{\frac{2}{p}-2} & \text{a } D^- \text{ tartományban} \\ \dfrac{a+2}{a\cdot(a-1)} & \text{a } D \text{ tartományban} \end{cases} \qquad (20)$$

A *2. ábrán* a P és P_J jelű hatásfokgörbék a (19) és a (20) egyenletek felhasználásával voltak felrajzolhatók; az L_2-vel jelölt görbénél az A^2-eket természetesen az

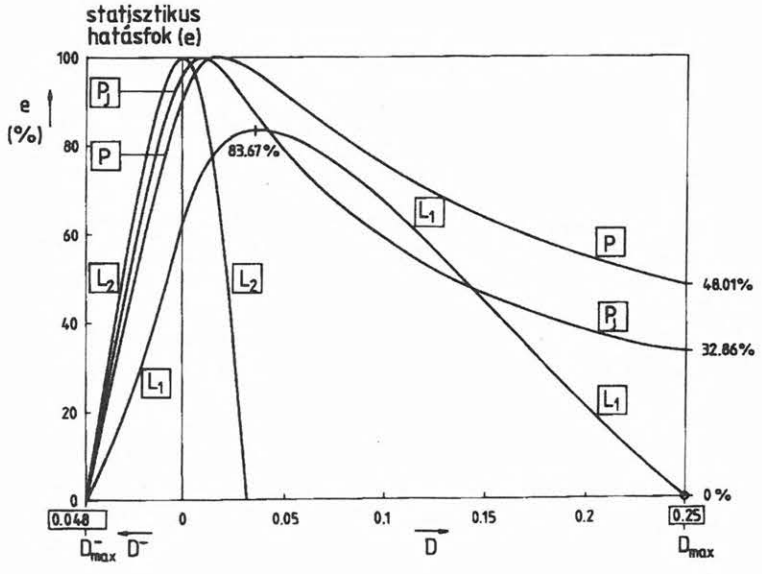
$$A^2_{L_2} = \int_{-\infty}^{\infty} x^2 f_c(x)\mathrm{d}x \qquad (21)$$

formula szerint számítottuk.

Végül pedig, ami az L_1 görbét illeti, itt célszerű magával az e_{L_1} hatásfoknak a képletét közvetlenül megadni:

$$e_{L_1} = \begin{cases} \dfrac{1}{\Gamma(1/p)\cdot\Gamma(2-1/p)} & \text{a } D^- \text{ tartományban,} \\ \dfrac{4\cdot\Gamma^2(a/2)\cdot(a+2)}{\pi\cdot\Gamma^2[(a-1)/2]\cdot a\cdot(a-1)} & \text{a } D \text{ tartományban.} \end{cases} \qquad (22)$$

A hagyományos statisztikának megfelelő L_2-görbe csak D=0,032-ig vesz fel zérusnál nagyobb értékeket, de a maximumtól jobbra és balra egyaránt gyorsan csökken a hatásfok, márpedig a robusztusság követelménye éppen az ellenkezője volna: t.i. hogy ez a csökkenés a típustávolság növekedésével lassú legyen. Geo-



2. ábra. Hatásfokgörbék a teljes $f_c(x)$ szupermodellre, valamint négyféle statisztikai normára

Fig. 2. Curves of the statistical efficiencies for the norms L_1, L_2, P and P_J vs all type-distances belonging to the whole supermodel $f_c(x)$

fizikai-geológiai esetekben a hibatípusok döntően a D tartományban (leggyakrabban a D=0,016 távolságnak megfelelő típus környezetében) várhatók, így — ha nem általános statisztikai meggondolásokat követünk, — a görbéknek D tartománybeli viselkedése alapján ítélhetjük meg az egyes normák előnyös vagy kevésbé előnyös sajátságait.

A fentieket figyelembe véve a P-norma alkalmazása mutatkozik a vizsgált négy norma közül a legelőnyösebbnek, ezt követi a P_J-norma. Ez utóbbit még a hibák Gauss-típusa esetén is célszerű az L_2-norma helyett alkalmazni, mert a D=0-nál tapasztalható néhány %-nyi hatásfokcsökkenés igazán csekély „biztosítási díjnak" tekinthető (ANSCOMBE [1960] találó kifejezésével élve), hogy ezáltal védve legyünk a durva hibák el nem hanyagolható (sőt esetleg katasztrofális mértékű) torzításai ellen. Az L_1-norma közbülső helyet foglal el (azaz az összehasonlításban a harmadik a négy közül), a maga maximálisan 84% körüli (és a Gauss-típusnál 64%-os) hatásfokával, valamint azzal a sajátságával, hogy $e_{L_1} \to 0$, ha $D \to 0{,}25$, azaz az extrém nagy súlyú szárnyakkal rendelkező hibaeloszlásoknál gyakorlatilag nem használható.

A robusztusságot mind ez ideig a 2. ábrán látható hatásfokgörbék egy-két jellemzőjének kiemelésével hasonlítottuk össze. Nyilván gyakran jelentene előnyt, ha ezt a sajátságot egyetlen számmal jellemezhetnénk. Ha egy pillanatra eltekintünk szakterületünk speciális szempontjaitól, akkor az volna az ideális (de gyakorlatilag természetesen megvalósíthatatlan), ha e=100% lenne a 2. ábra egész abszcisszája mentén, azaz minden $f_c(x)$ típusra. Az R-rel jelölt *általános robusztusság*

$$R = \frac{1}{0{,}298} \cdot \left[\int_0^{0{,}048} e(D^-)\,\mathrm{d}D^- + \int_0^{0{,}25} e(D)\,\mathrm{d}D \right] \quad (23)$$

definíciója ehhez az ideális esethez viszonyít, szakterületünkre vonatkozóan viszont STEINER és HAJAGOS [1993] adott adekvát definíciót arra az r-rel jelölt robusztusságra, amely az egyes hibatípusok geofizikai-földtani előfordulási valószínűségsűrűségére is tekintettel van. A *3. táblázat* három normára adja meg mind R, mind r értékeit; a robusztusság táblázatbeli mérőszámai önmagukért beszélnek.

	A robusztusság mérőszáma	
	általános	földtudományi hibatípusokra
	(R)	*(r)*
L_2	16%	36%
L_1	47%	80%
P	68%	96%

3. táblázat. A robusztusság mérőszámai

Table 3. Indices of robustness

4. Optimum-módszerek a geofizikai-geológiai törvényszerűségek felismerésének szolgálatában

Már az 1.2. alpontban megadtuk a jelen dolgozat célkitűzését, nevezetesen annak bemutatását egy jól definiált gyakorlati példán, hogy a modern optimum-módszerek képesek információként nemcsak kőzet- vagy modellparamétereket pontosan szolgáltatni, hanem valamely törvényszerűségre is egyértelműen rávilágítani. A bemutatott példa a T 014027 számú mélyfúrási geofizikai témakörű OTKA-kutatás egy részeredménye, tehát szükségképpen karotázs jellegű. A lényeg megértése azonban távolról sem igényli azt, hogy az olvasó specialista legyen ezen a területen.

4.1. Az összehasonlítandó karotázs szelvények és a mérési terület kiválasztása

Válasszunk példaként egy karotázs adatrendszert, s egy viszonylag egyszerű, agyagos, homokos, kavicsos rétegösszletben keressünk kapcsolatot két önkényesen választott szelvény, a természetes gamma (*TG*) és egy közel azonos térrészt érzékelő, L=10 cm hosszúságú potenciálszondával felvett ellenállásszelvény (R_a) között. Ilyen rétegösszlettel csaknem minden hazai fúrásban találkozunk: a vízkutató fúrások több mint 95%-a agyagos, laza homokos, kavicsos rétegeket harántol, a széntelepek fedőösszlete a legtöbb területen hasonló, a szénhidrogén-kutató fúrások zöme szintén agyagos, gyengén vagy közepesen cementált homokkövekben mélyül.

A vizsgált fúrások a kelet-borsodi szénmedence részét képező dubicsányi területen találhatók. A karotázs méréseket az OFKFV Északmagyarországi Üzemvezetősége végezte 1986–88 között.

4.2. A terület rövid földtani jellemzése, a vizsgálandó rétegösszlet kiválasztása

A dubicsányi szénterületen a *pleisztocént* és a *pliocént* változó vastagságú, helyenként kivékonyodó agyagos, kőzetlisztes, homokos, kavicsos rétegek képviselik. A *szarmata kori* rétegek csaknem mindegyike andezit vulkanizmushoz köthető. Az andezittufák különböző mállott formái (agyagos, lapillis, törmelékes andezittufa, tufás agyag, homok) mind megtalálhatók. A *badeni* összletben az agyagok, az agyagos riolittufák, tufás homokok az uralkodók. Ami az *ottnangi* széntelepes összletet illeti, a széntelepek feküjében riolittufás homokok, agyagok fordulnak elő, a fedőben pedig 70–200 m vastag agyagos, homokos rétegeket harántoltak a fúrások.

E rétegsorból agyagos, uralkodóan kvarchomokos rétegösszletet kívántunk vizsgálat tárgyává tenni, ezért a szarmata és bádeni, valamint a széntelep feküjében lévő tufás homokok, agyagok nem vehetők számításba. A kőzetek radioaktivitását ugyanis a mállás csak akkor befolyásolja lényegesen, ha a mállást követő üledékképződés radioaktív elemtranszporttal jár együtt. Területünkön ez nem következett be, így a tufás agyagok, homokok aktivitása csaknem azonos. A pleisztocén és pliocén összletek helyenként annyira kivékonyodtak, hogy az őket harántoló fúrásokban elhelyezett vezércső lehetetlenné tette az R_a mérését.

A fentieket figyelembe véve a vizsgált összlet a széntelep feletti vastag, a statisztikai analízis céljára kiválasztott területünk minden fúrásában meglévő agyagos-homokos rétegcsoport volt. Ez tufát nem tartalmaz, és ebből a szempontból hasonló különösen a fiatalabb (pleisztocén, pliocén), más hazai területeken is mindenütt előforduló rétegösszletekhez.

4.3. A természetes gamma aktivitás (TG) és a látszólagos fajlagos ellenállás (R_a) közötti kapcsolat vizsgálata

4.3.1. Néhány ismert összefüggés az agyagtartalomnak a természetes gamma aktivitásra, ill. elektromos fajlagos ellenállásra való hatására

Közismert, hogy az agyagos, homokos üledékes rétegek fajlagos ellenállása adott R_w rétegvíz-ellenállás mellett a kőzetalkotó szemcsék átmérőjével azonos irányban változik. Ellenállás-csökkentő komponens az agyag és a kőzetliszt. *TG*-növelő hatása elsősorban az agyagnak van, a kőzetliszteké minimális, mivel zömükben kvarckeveréket tartalmaznak, s aktivitásuk csak az abszorbeált sugárzó ionoktól függ.

A *TG*-aktivitás és az agyagtartalom kapcsolatát többféle empirikus összefüggés írja le. Durva becslésnek számít az agyagtartalomra a következő összefüggés:

$$V_{sh} = i_\gamma \,, \tag{24}$$

ahol

$$i_\gamma = \frac{TG - TG_{homok}}{TG_{agyag} - TG_{homok}}$$

a természetes gamma index.

Pretercier kőzetekre a

$$V_{sh} = 0{,}33 \cdot \left(2^{2i_\gamma} - 1\right) \tag{25}$$

a tercier korú és fiatalabb kőzetekre a

$$V_{sh} = 0{,}083 \cdot \left(2^{3{,}7 i_\gamma} - 1\right) \tag{26}$$

összefüggéseket szokás elfogadni.

Amennyiben kellő laboratóriumi vizsgálati eredmény is rendelkezésre áll, célszerű az empirikus kapcsolatokat területenként és rétegösszletenként külön megvizsgálni, illetve megállapítani.

Az agyagtartalom ellenállás-csökkentő (vezetőképesség-növelő) hatását kvantitatíve számos összefüggéssel közelítik. Ezek mindegyike azt fejezi ki, hogy az agyagos homok, homokkő σ_0 vezetőképességét a rétegvíz és az agyag vezetőképessége, azaz σ_w és σ_{sh} befolyásolja:

$$\sigma_0 = A\sigma_w + B\sigma_{sh} \tag{27}$$

azaz

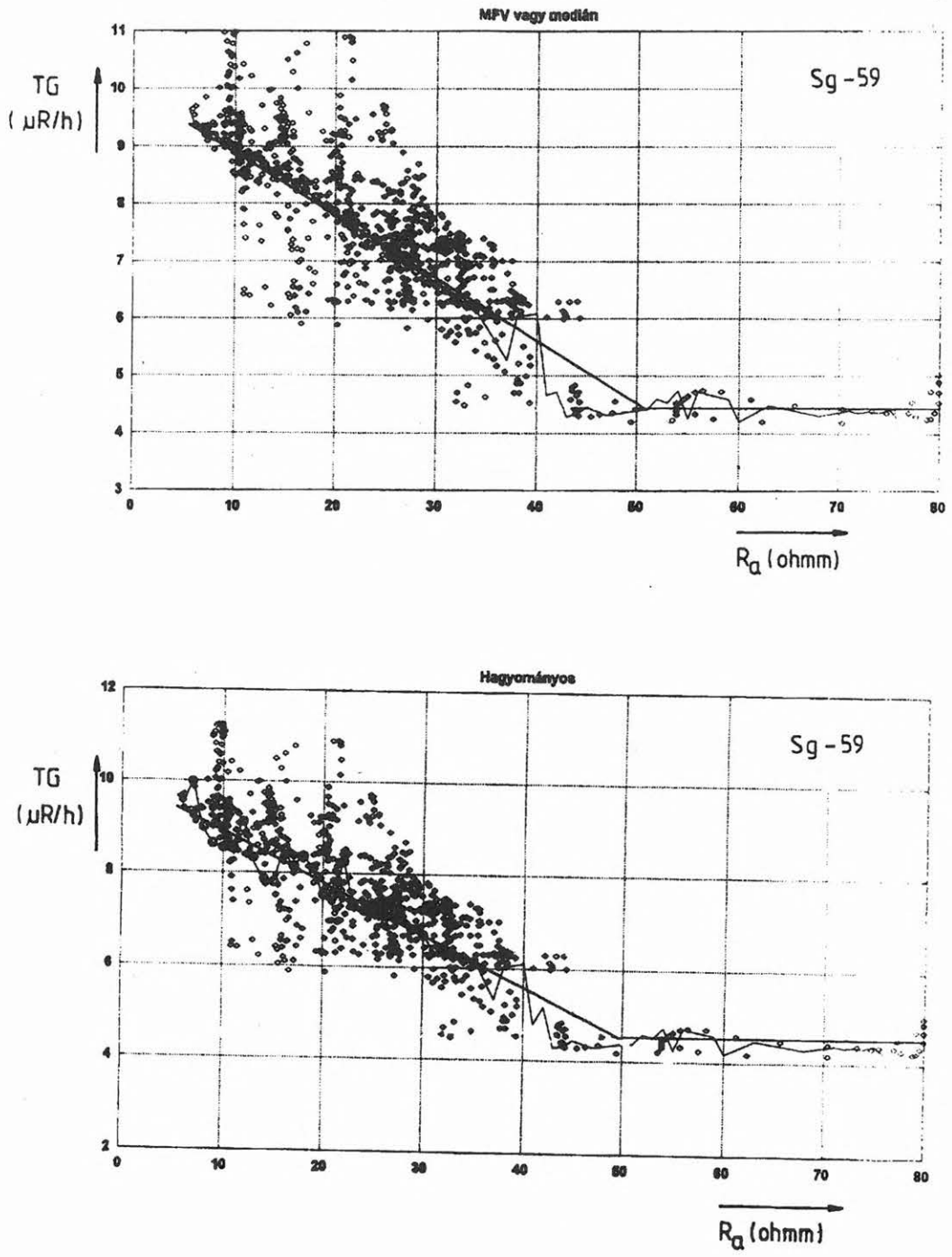
$$\frac{1}{R_0} = A\frac{1}{R_w} + B\frac{1}{R_{sh}} \tag{28}$$

Különféle kőzetmodellekre más és más összefüggés érvényes. A kapcsolatok rendkívül bonyolultak, és semmi sem garantálja, hogy megfelelnek az általunk vizsgált rétegösszleteknek.

4.3.2. A *TG* és az R_a kapcsolatának statisztikai vizsgálata

A 4.3.1. alpont meggyőzhetett bennünket arról, hogy a *TG* és R_a kapcsolatának a meghatározása analitikus úton reménytelen, feltétlenül statisztikai módszerekre vagyunk utalva.

Eredményeinket a dubicsányi terület Salgógalgóc-59 (Sg-59) jelű fúrásában mért adatokon végzett vizsgálatain keresztül mutatjuk be (*3. ábra*), mivel itt állt rendelkezésre az adatpárok legnagyobb száma, így itt várhattuk a statisztikai szempontból leginkább demonstratív, legmegbízhatóbb eredményeket.



3. ábra. A felső ábra *P*-normával (MFV-számítással) kapott nullkörei egy lineáris kapcsolatra egyértelműen mutatnak rá a vizsgált mérési adatrendszeren belül a természetes gamma aktivitás (*TG*) és a látszólagos fajlagos elektromos ellenállás (R_a) között, a 7–33 ohmm-es R_a tartományban. Ez a törvényszerűség a hagyományos (L_2 szerinti) optimum-módszerrel lényegesen bizonytalanabbul mutatkozik csak meg, ld. az alsó ábrát

Fig. 3. The circles on the upper figure (got by using the *P*-norm) show unambiguously a linear regularity between the natural γ-activity (*TG*) and the apparent electrical resistivity (R_a) for the investigated pairs of data in the R_a-domain of 7–33 ohmm. The same regularity appears also in the lower figure where conventional (L_2-based) statistics were used but disturbed by a considerable statistical fluctuation

Az analóg szelvények digitalizálásakor a mintavételi köz 10 cm volt A mikroröntgen/órában (μR/h) adott *TG*-szelvényt a statisztikus ingadozás eliminálása céljából 5-pontos átlagszűréssel szűrtük. Az így nyert adatokat az azonos mélységben, ohmm-ben mért R_a értékekkel párosítva nyertük a TG–R_a plotot.

Hogy ne növeljük feleslegesen a jelen dolgozat ábráinak a számát, kérjük az olvasót, hogy a 3. ábrán (ábrapáron) először csak a rombuszokkal jelölt pontokra legyen tekintettel, mégpedig az alsó ábrán lévőkre. (A felső ábrán ugyanazokat a pontokat látjuk ugyan, csak a TG>11μR/h természetes gamma-indikációkkal jellemzett néhány pont kivételével.) E pontok képezik a TG–R_a plotot a kiválasztott (Sg-59) fúrásra.

Rutinszerű matematikai approximációs gondolatmenettel egy $TG = A + B/R_a$ összefüggés szerinti kiegyenlítés jut először eszünkbe. Azonban bármelyik norma minimálásával végeztük is a kiegyenlítést, a kapott hiperbolák nem voltak elfogadhatóak a TG–R_a kapcsolat leírására (ezért ábrán felesleges is lenne ezeket bemutatni), azt viszont mindegyik hiperbola egyértelműen jelezte, hogy az egész adatrendszerre egyetlen, egyszerű analitikus alak rákényszerítése megakadályozza azt, hogy a statisztikai törvényszerűség mintegy maga mutassa meg magát.

Az utóbbit úgy sikerült lehetővé tennünk, hogy 5 ohmm hosszúságú R_a részintervallumokra határoztuk meg a *TG* leggyakoribb értékét (ha az intervallumra 10-nél kevesebb pont esett, a mediánját), valamint a számtani átlagát. A részintervallumokkal 1 ohmm-enként haladva, az utóbbi esetben (azaz a hagyományos statisztikával) az alsó ábra vékony vonallal rajzolt törtvonalú görbéjét nyerjük eredményül, és hasonló görbe mutatja a felső ábrán a modern statisztikával nyert eredményeket.

Az R_a értékek 7–33 ohmm intervallumán markáns nullkörökkel jeleztük a vékony vonalú görbék „töréspontjait". Az idézőjelet a felső ábra indokolja: a nullkörök elhanyagolható (első pillantásra alig látható) ingadozással esnek egyetlen egyenesre, ebben az R_a intervallumban *egyértelműen mutatva a TG és R_a értékek lineáris összefüggésének törvényszerűségét.* Miután a felső ábráról leolvastuk ezt a törvényszerűséget, felismerjük, hogy a hagyományos statisztikával nyert alsó ábra is ugyanezt a törvényszerűséget tükrözi ugyan, de olyan mérvű statisztikus ingadozással, hogy pusztán ennek alapján csak óvatosan lehetett volna megkockáztatni azt a következtetést, amelyet a modern statisztikával nyert felső ábra alapján egyértelműen lehetett kimondani.

A TG–R_a plot alapján puszta ránézéssel is megállapítható (amit a vékony törtvonalú görbe különösen a felső ábrán még csak megerősít), hogy az R_a értékek 50–80 ohmm-es tartományára a *TG* érték gyakorlatilag konstans: az agyagtartalom hiánya miatt az R_a-változások ebben a tartományban a szemcsenagyság változásainak tulajdoníthatók. A *TG* *R*-től való függését tehát két egyenes írja le: az elsőt az $R_a \leq 40$ ohmm-es tartomány pontjainak lineáris kiegyenlítésével, a vízszintes egyenes szakaszt pedig konstans kiegyenlítéssel nyertük az $R_a > 40$ ohmm tartománybeli pontok alapján. Ugyanezzel a módszerrel határoztuk meg a többi fúrásra is az egyenespárokat; nem meglepő, hogy ezek paraméterei nagyon hasonlóaknak adódtak.

Gyakorlati alkalmazása a fentieknek, hogy a 35 ohmm-nél kisebb R_a tartományban pusztán az R_a értékekből tudunk a *TG* értékére következtetni, abból pedig a V_{sh} agyagtartalom becsülhető.

Bár a fentiekben gyakorlati alkalmazásra is hivatkoztunk, nyilván azt az elvi eredményt kell befejezésül újra hangsúlyoznunk, hogy a modern statisztika optimum-módszerei képesek törvényszerűségek feltárására.

HIVATKOZÁSOK

ANSCOMBE F. J. 1960: Rejection of outliers. Technometrics **16,** 147-185

CSERNYÁK L. 1995: Distance of probability distribution types. (Study of the distance of the normal and the generalized Student-distributions). Acta Geodaet. et Geoph. Acad. Sci. Hung. **30,** 2-4, 281-284

HAJAGOS B., STEINER F. 1994: Definition der Entfernung von Wahrscheinlichkeitsverteilungstypen. Eine mögliche Charakterisierung der Robustivität für die Praxis. Publ. Univ. of Miskolc, Series D. Natural Sciences **35**, 97-108

KIS M. 1996: Globális optimalizáció a geofizikában a Simulted Annealing algoritmus alkalmazásával. Magyar Geofizika, **37**, 170-181

STEINER F. 1990: A geostatisztika alapjai. Tankönyvkiadó, Budapest, 357 p.

STEINER F. (*Ed*) 1991: The Most Frequent Value. Akadémiai Kiadó, Budapest, 315 p.

STEINER F. (*Ed*) 1997: Optimum Methods in Statistics. Akadémiai Kiadó, Budapest, 370 p.

STEINER F., HAJAGOS B. 1993: Practical definition of robustness. Geophysical Transactions **38**, 4, 193–210

STEINER F., HAJAGOS B. 1998: Error-types characterized by arbitrary short or heavy flanks. Acta Geodaet. et Geoph. Acad. Sci. Hung. (Sajtó alatt; a kéziratleadás 1997 szeptemberében történt)

SZŰCS P. 1995: Theoretical and practical consequences of the global optimization methods. Acta Geodaet. et Geoph. Acad. Sci. Hung. **30**, 2–4, 301–312