# Hozzászólás Bán István: A kiválasztási matematikai modell felhasználása a trágyázási szaktanácsadáshoz c. dolgozatához

KABOS SÁNDOR és VARGHA MÁRTON

MTA Talajtani és Agrokémiai Kutató Intézete
és KSH Államigazgatási Számítógépes Szolgálat, Budapest

Dr. BÁN ISTVÁN 1979-ben e hasábokon ismertette talajerő-utánpótlási módszerét [1], amely a Szerző jellemzése szerint az agronómiai gyakorlaton alapul, felhasználja a matematikát és a számítógép-technikát. A módszer agronómiai és számítástechnikai bírálatára nem vállalkozunk, figyelmünket a matematikai modellre összpontosítjuk. Előbb összefoglaljuk a Szerző eljárását, lépéseit futólag diszkutáljuk, majd kissé részletesebben foglalkozunk azzal a kérdéssel, vajon indokolt-e esetünkben paraméteres regressziót használni lokális becslésre.

Ahol lehet, ragaszkodunk [1] jelölésrendszeréhez, a statisztikában szokásos elnevezéseket zárójelben adjuk. Az elérni kívánt Q termésszintet (a függő változót) a

$$Q = f(K1, K2, \ldots, Kn, N, P, K) + \varepsilon \qquad (*)$$

alakban állítjuk elő, ahol f az ismeretlen regressziós függvény, K1, . . ., Kn a környezetet leíró „kiválasztási állapotjellemzők" (a csoportosító változók), N, P, K a kiszórt hatóanyag-mennyiségek (a független változók), $\varepsilon$ a meg nem figyelt állapotjellemzők hatása (a véletlen hiba).

A szaktanácsadás azt jelenti, hogy a Q, K1, . . ., Kn, N, P, K változókra vett nagy mintából becslést adunk rögzített $(Q_*, K1_*, \ldots, Kn_*)$-hoz tartozó N, P, K értékekre. A rendszer működését tehát úgy képzeljük, hogy a felhasználó megadja táblája lényeges állapotjellemzőit és az elérni kívánt termésszintet, és a számítógép az adatbázisból kikeresett múltbeli termelési adatok alapján megmondja, mennyi N-, P-, K-hatóanyag szükséges. Lehetne vitatni, hogy a gyűjtött adatok mennyiben használhatók közvetlenül tervezésre, és mennyiben a meglevő műtrágyázási szokások leírására. Cikkünk célja szerint az ilyen és hasonló meggondolásokat félretesszük, és megjegyzés nélkül követjük a Szerző gondolatmenetét, hogy az alkalmazott matematikai apparátusra vonatkozó mondanivalónkat kifejthessük.

A „kiválasztási matematikai modell" lényegében a következő:

(A) Válasszunk egy alkalmas $\delta$ számot, és szorítkozzunk azokra a táblákra, ahol $(1 - \delta) \cdot Kj_* \leq Kj \leq (1 + \delta) \cdot Kj_*$ $(j = 1, \ldots n)$ vagyis hagyjuk el azokat a táblákat, amelyeknek van olyan kiválasztási állapotjellemzője, amely „nagyon" eltér attól, amire a tanácsot adjuk.

(B) Az előző lépésben megmaradt mintából külön-külön becsüljük a N — Q, P — Q, ill. K — Q összefüggéseket, tehát 3 db egyváltozós regressziót számolunk.

(C) A regressziós függvény első-, másod- vagy magasabb fokú polinom. $Q_*$ termésszint elérésére azokat a N, P, K értékeket javasoljuk, ahol a megfelelő regressziós függvény értéke $Q_*$.

(D) Elvárjuk, hogy a regressziós függvénynek a megfigyelési értékek által definiált értelmezési tartományán legfeljebb egy szélső értéke legyen. Elvárjuk, hogy a (C)-beli helyettesítési értékek a megfelelő polinomok monoton növő szakaszán legyenek.

Az első észrevételünk az, hogy a (*) egyenlet független változóit az algoritmus (A) és (B) lépései különbözőképpen kezelik. Érthető, hogy a kategória (nominális) változók csak mint csoportosító tényezők jönnek tekintetbe (pl. talajtípus), bár ekkora minta esetén a csoportosítást pl. variancia-analízis előzhetné meg, kijelölve a ténylegesen eltérő csoportokat. Nem tartjuk indokoltnak viszont, hogy pl. a talaj N-ellátottsága kiválasztási állapotjellemző (tehát a regressziós egyenletből eliminálják), a kiszórt N-hatóanyag viszont cél állapotjellemző (tehát bevonjuk a regresszióba). Kézenfekvő lenne sokváltozós regresszió alkalmazása: a várható termést, mint több, egymással kölcsönhatásban levő tényező függvényét becsüljük.

Még világosabb, hogy sokváltozós módszerek szükségesek, ha az algoritmus (B) lépését vizsgáljuk, ahol a N-, P-, K-hatóanyagokra külön egyváltozós regressziók szerepelnek. Jogos lenne ez pl. akkor, ha mindig arányosnak feltételezhetnénk a kiszórt N-, P-, K-adagokat (ilyenkor persze fölösleges három regresszió), egyébként belátható, hogy az átlagosnál magasabb termésszint esetén a módszer várhatóan túlbecsüli a hatóanyagigényt. Sokváltozós megközelítés esetén a (C) lépés természetesen nem egyetlen (N, P, K) javaslatot ad, hanem egy kétdimenziós sokaságot, amelyen egy alkalmas célfüggvényt (pl. a műtrágyázás költségét) lehet optimalizálni.

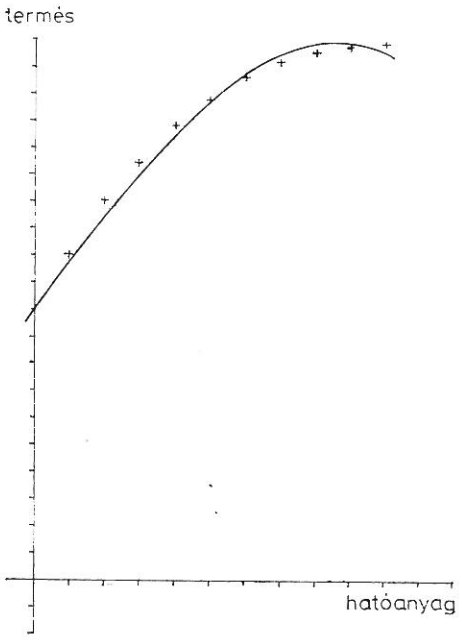
Az algoritmus (D) lépése az előzőeknél mélyebb kérdést vet fel, és rámutat a konstrukció alapproblémájára, ezért ezt részletesebben tárgyaljuk. Feladatunk inverz becslés: hatóanyag → termés regressziót számolunk, és ebből kívánunk visszakövetkeztetni az egy adott termésszinthez tartozó hatóanyagra. Világos, hogy ez a visszakövetkeztetés csak ott lehetséges, ahol a regressziós görbe elég meredek, azaz a hatóanyag változtatására a termés elég határozottan reagál. Megjegyezzük, hogy ez az oka, hogy a visszakövetkeztetéses becslés nem rendelkezik a statisztikában szokott tulajdonságokkal, pl. nem torzítatlan (ezt [3]-ban részletesen kifejtettük).

Tekintsük az 1. ábrán feltüntetett következő (az egyszerűség kedvéért egyváltozós) példát, ahol a mintapontokra harmadfokú regressziós polinomot illesztettünk. Az illeszkedés statisztikusan jó ($R^2 = 0{,}98$). Vegyük észre azonban, hogy a regressziós polinom nem szándéka szerint a véletlen hibák kiszűrésével javítja, hanem éppen elrontja a becslést. A minta végig emelkedő tendenciát mutat, tehát még lehetőséget adna a kívánt visszakövetkeztetésre, miközben a regressziós polinom növekedése megáll, sőt csökkenni kezd. Adatainkat szemléltető céllal konstruáltuk, de belátható, hogy a jelenség tipikus, amikor egy várhatóan telítődési görbe jellegű hatóanyag → termés regressziót kívánnánk polinommal közelíteni. A paraméteres regresszió e jól ismert hibája ellen a reziduumok vizsgálatával szokás védekezni; esetünkben ez jellegzetes U alakú görbe, és világosan mutatja, hogy a hibák eloszlása nem

véletlenszerű, tehát más függvénnyel kell próbálkozni. Nem ajánlunk ilyen és hasonló (nemparaméteres) módszereket a polinomiális regresszió finomítására, egyrészt azért, mert úgy látjuk, hiába tennénk; ezek az egyedi goodness-of-fit vizsgálatok a szaktanácsadás menetébe nem lennének jól beilleszthetők. A másik ok súlyosabb: ebben az esetben hibásnak tartjuk a feladat interpretálását paraméteres regressziós modellben, véleményünk szerint nemparaméteres, robusztus becslésekre van szükség. Röviden elmondjuk az érveinket, képletekkel a Függelékben illusztráljuk gondolatmenetünket.

A Függelék (2) képletéből azt látjuk, hogy a lineáris regresszió a mért értékek súlyozott összegeként állítja elő a becslést. Nem meglepő a (4) képlet, amely azt mondja: a ténylegesen mért és a regresszióval becsült érték egészen gyengén korrelál. Ez a tulajdonság általában jellemzi a paraméteres regressziós becsléseket, és önmagában véve természetesen nem hiba, hanem erény. A becslés minden pontban az (1) hipotézis helyességén alapul, és a teljes mintán; az egy-egy pontbeli esetlegességek alig befolyásolják. Így érhető el a becslés (3)-beli szórása, amelyről jól ismert, hogy a lineáris torzítatlan becslések között optimális. Megismételjük, hogy ezek a kiváló tulajdonságok csak az (1) feltételek szigorú teljesülése esetén garantálhatók. Ha a modell csak közelítőleg igaz, mondjuk az elméleti regressziós egyenes egy kicsit lehajlik, akkor itt teljesen félrevezető becslést kaphatunk.

Másként viselkednek a nemparaméteres becslések. A Függelék (5) képlete mutatja, hogy itt is a mintapontok súlyozott összegét kell kiszámítani. Döntő különbség, hogy a távolabbi pontokban mért értékeket kevésbé vesszük figyelembe — nem úgy, mint (2)-ben, ahol a súlyok aránya lényegében állandó. Világos, hogy a mért értékkel jól korreláló becsléshez jutunk, amint ezt a (7)




*1. ábra*
A harmadfokú regressziós polinom telítődési görbe jellegű hatóanyag → termés adatokon

képlet példája mutatja. A becslés szórása természetesen nem lesz olyan jó, mint a paraméteres esetben (l. a (6) képletet).

Példáinkkal a következő szemléletes képet kívántuk kialakítani: a nemparaméteres becslés a paraméteres modell szigorú teljesülése esetén kevésbé pontos (nagyobb szórású) ugyan, viszont lokálisan jobban korrelál a mintával, jobban figyelembe veszi a modelltől való eltérést. A kiválasztási matematikai modell esetében a polinomiális regresszió feltételezése eleve csak közelítés, és mivel végül is csak lokális becslésre használjuk, időnként nagy hibákra kell számítanunk. Azt pedig semmiképp sem várhatjuk, hogy e közelítő regressziós polinom becslésének lefutása kövesse a valódi regresszióét.

## Függelék

Számoljunk utána a fenti általános megállapításoknak néhány egyszerű esetre. Legyen

$$y_i = f(x_i) + \varepsilon_i \tag{1}$$

ahol $y_i$ az $x_i$ helyen mért érték, $f$ ismeretlen, becsülendő lineáris függvény, $\{\varepsilon_i\}$ független, 0 várható értékű, $\sigma$ szórású normális eloszlású valószínűségi változók, $i = 1, 2, \ldots, N$. Az $f$ függvény jól ismert legkisebb négyzetes becslése

$$\hat{f}(x_j) = \bar{y} + \frac{\sum_{i=1}^{N} (x_j - \bar{x}) \cdot (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{2}$$

E becslés szórásnégyzete

$$\operatorname{var}\left(\hat{f}(x_j)\right) = \sigma^2 \cdot \left( \frac{1}{N} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right) \tag{3}$$

A szórás a mintaátlagban a legkisebb: $\frac{\sigma}{\sqrt{N}}$, egy az átlagtól „átlagosan messze" eső pontra: $\sqrt{2}\,\frac{\sigma}{\sqrt{N}}$.

Az $x_i$-beli mintapont $(y_i)$ és a rá adott $\hat{f}(x_i)$ becslés korrelációja

$$\operatorname{corr}\left(y_i, \hat{f}(x_i)\right) = \sqrt{\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}} \tag{4}$$

Ezt a mennyiséget ritkán számítják ki, ne tévesszük össze az R többszörös korrelációs együtthatóval, ami itt az $(y_1, \ldots, y_N)$ és az $\left(\hat{f}(x_1), \ldots, \hat{f}(x_N)\right)$ vektorok korrelációja. A (4) korreláció a mintaátlagban: $\frac{1}{\sqrt{N}}$, és máshol is

ilyen (kicsi) nagyságrendű az átlagtól átlagosan messze eső pontra: $\frac{\sqrt{2}}{\sqrt{N}}$.

A legismertebb nemparaméteres regressziós becslés a Rosenblatt—Parzen sorfejtés (pl. [2])

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} y_i \cdot h_{x_i}(x)}{\sum_{i=1}^{N} h_{x_i}(x)} \tag{5}$$

ahol $h_{x_i}$ $x_i$ várható értékű, alkalmasan választott szórású normális sűrűségfüggvény. Pontosabb eredményeket tartalmaz a [4] cikk.

Tekintettel arra, hogy változóink diszkrétek, pontosabban szólva a mintanagysághoz képest kevés különböző értéket vesznek fel, kézenfekvő az ún. k-Nearest Neighbour (legközelebbi szomszéd) módszerek alkalmazása. Ábrázoljuk változóinkat (esetünkben: K1, . . ., Kn, N, P, K) sokdimenziós koordinátarendszerben, hozzárendelve az egyes rácspontokhoz a megfelelő $y_i$ (esetünkben: az elért termés) értékeket. Kijelöljük a rácspont első, második, . . . . . ., $k$-adik szomszédságát, és a pontban a $\sum_{i=0}^{k} w_i \cdot y_i / \sum_{i=0}^{k} w_i$ becslést adjuk, ahol $w_i$ az $i$-ik szomszédság befolyásának súlya, a 0-ik szomszéd önmaga. Megjegyzendő, hogy a sokdimenziós szomszédság (közelség) definiálása, a $k$ és a $w_i$ számok megválasztása nem triviális feladat. A szóbajövő statisztikai módszereket jól összegzi [5], emellett persze célszerű szakmai (agronómiai) tapasztalatokra is támaszkodni. Amikor a súlyokat konstansnak választjuk, a jól ismert mozgóátlag becsléseket kapjuk. M. Kendall (pl. [2]) nyomán rámutatunk, hogy ez a becslés azonos a mintára $k$-anként illesztett súlyozott polinomiális regresszióból kapottal.

Legyen szemléltetésül a következő egyszerű példa:

$$\hat{f}(x_i) = \frac{1}{4} y_{i-1} + \frac{1}{2} y_i + \frac{1}{4} y_{i+1}$$

(ezt a simítást csak a minta belső pontjain értelmezzük). A becslés szórása

$$\mathrm{var}\left(\hat{f}(x_i)\right) = 0{,}375 \cdot \sigma^2 \tag{6}$$

Megjegyezzük, hogy a szórás nem függ $i$-től, azaz a görbén állandó. A számított és mért értékek korrelációja

$$\mathrm{corr}\left(y_i, \hat{f}(x_i)\right) = 0{,}82 \tag{7}$$

szintén állandó. A fenti számítások természetesen úgy értendők, hogy közben fennáll az (1) paraméteres modell.

A nemparaméteres regresszió fontos tulajdonsága, amit a főszövegben igyekeztünk érzékeltetni: a robusztusság. Ez azt jelenti, hogy a modell „csekély" változására a becslés keveset változik. A nemparaméteres regresszióra vonatkozó ilyen irányú eredmények összefoglalását adja Stone [6].

Végezetül említésre méltónak tartjuk azt a tényt, hogy a kiválasztási matematikai modell és újabb változatai, valamint az adatbázis számítógépes megvalósítását 1977 óta a KSH Államigazgatási Számítógépes Szolgálat munkatársai végzik.

## Összefoglalás

Nem állítottuk, hogy elvileg lehetetlen kidolgozni a növény életfeltételei és termése közti összefüggés paraméteres regressziós modelljét. Úgy tudjuk, ez a modell jelenleg nem létezik, láthatólag osztja e nézetet dr. BÁN is. Ebben a helyzetben módszertani tévedésnek tartjuk termésbecslésre vállalkozni polinomiális regressziós görbék alapján. Egyetértünk viszont a Szerzővel, amikor az egyedi táblákra a teljes mintából lokális becslést keres; úgy találtuk, az alapgondolat következetes végigvitele a vázolt robusztus becslésekre vezet.

## Irodalom

[1] BÁN, I.: A kiválasztási matematikai modell felhasználása a trágyázási szaktanácsadáshoz. Agrokémia és Talajtan. **28.** 243–248. 1979.
[2] FRITZ, J. & RÉVÉSZ, P.: Az alakfelismerés statisztikus módszerei. BJMT kiadása. Budapest. 1974.
[3] KABOS, S. & REICZIGEL, J.: Sokváltozós regressziós becslések. KSH ÁSzSz Tanulmány. Budapest. 1980.
[4] MAJOR, P.: On non-parametric estimation of the regression function. Studia Sci. Math. Hung. **8.** 347–361. 1973.
[5] Smoothing techniques for curve estimation. (Ed.: GASSER, TH. & ROSENBLATT, M.) Springer. Berlin. 1979.
[6] STONE, CH.: Consistent nonparametric regression. Ann. Statistics. **5.** 595–622. 1977.

*Érkezett: 1982. április 5.*