GEOGRAFICKÝ ČASOPIS

ROČNÍK 36 1984 ČÍSLO 3

JAROSLAVA HORVÁTHOVÁ

TEORETICKO-METODOLOGICKÉ PRINCÍPY NUMERICKEJ TAXONÓMIE PÔD

II. HODNOTENIE PODOBNOSTI OBJEKTOV A ZÁKLADNÉ AGLOMERÁCIE

Jaroslava Horváthová: Theoretical-methodological Principles of Numerical Soil Taxonomy. II. Estimation of Objects Similarity and Basic Agglomerations. Geogr. Čas., *36*, 1984, 3; 2 figs, 4 tabs, 22 refs.

Estimation of multidimensional objects similarity (in our case pedons and polypedons) originates from calculations of the similarity coefficients, or distance. More coefficients, which were used for the soil profiles and pedo-ecological units comparation are considered. The basic characteristics of individual agglomerative processes are presenting.

1. ÚVOD

V článku *Teoreticko-metodologické princípy numerickej taxonómie pôd I.* sme v krátkosti spomenuli klasický model klastrovej analýzy, ktorý sa zakladá na troch predpokladoch:

1. Pre každú množinu objektov $o_1, \ldots, o_N$ a znakov $z_1, \ldots, z_n$ vytvoríme údajovú maticu $X_{(o)}$ typu $N \times n$ s prvkami x_{ik}.
2. Pre každú dvojicu $\{o_i, o_j\}$ vypočítame koeficient podobnosti P, resp. koeficient vzdialenosti V, čím redukujeme maticu dát na maticu podobnosti $P_{(o)}$, resp. maticu vzdialenosti $V_{(o)}$ typu $N \times N$ s prvkami P_{ij}, resp. V_{ij}.
3. Vyberieme určitú stratégiu, ktorá každej matici podobnosti $P_{(o)}$, resp. matici vzdialenosti $V_{(o)}$ jednoznačne priradí klastery $R_1, \ldots, R_q$.

2 Koeficienty vzdialenosti a podobnosti

Pri obidva koeficienty platí, že hodnoty ich funkcií sú určované pomocou všetkých znakov $z_1, \ldots, z_n$. Na vzdialenosť kladieme požiadavku, že koeficient V musí vyhovovať axiómam metriky:

— nezápornosti $V\{h_1, h_2\} \geqq 0$,
— identity $V\{h_1, h_2\} = 0$ práve vtedy, ak $h_1 = h_2$,
— symetrie $V\{h_1, h_2\} = V\{h_2, h_1\}$,
— trojuholníka $V\{h_1, h_2\} + V\{h_2, h_3\} \geqq V\{h_1, h_3\}$.

Každú metrickú funkciu $V\{h_1, h_2\}$ možno previesť na funkciu podobnosti podľa univerzálneho vzťahu:

$$P = \frac{c}{1 + V}$$, kde c je kladná konštanta.

Všeobecne platí, že čím je menšia vzdialenosť, tým je väčšia podobnosť a naopak. Hodnota funkcie podobnosti $P(h_1, h_2)$ môže byť ľubovoľné reálne číslo. Maximálnej hodnote P_{max} odpovedá maximálna podobnosť a naopak.

V pedológii sa použilo niekoľko mier vzdialeností. Základnou mierou je Euklidova (Pytagorova) vzdialenosť, ktorá je použiteľná pre kvantitatívne hodnoty. Ak máme dva objekty s ich znakovými hodnotami (vektory) uvažovať súčasne, musíme ich usporiadať v znakovom priestore a ich podobnosť merať ako vzdialenosť medzi nimi. Ak súradnice dvoch objektov i a j sú x_{i1}, x_{i2} a x_{j1}, x_{j2} (v dvojrozmernom znakovom priestore), potom vzdialenosť V_{ij} medzi objektami i a j bude

$$V_{ij} = \{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2\}^{1/2} .$$

Princípy vzdialenosti platia pre akýkoľvek počet rozmerov. Ak máme n-rozmerný priestor (čo je v praxi obvyklý prípad), potom vzorec pre výpočet vzdialenosti bude

$$V_{ij} = \left\{ \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \right\}^{1/2} .$$

Priemerná vzdialenosť bude

$$\bar{V}_{ij} = \left(\frac{V_{ij}^2}{n} \right)^{1/2} .$$

Z priemernej vzdialenosti $\bar{V}_{ij}$ alebo z obyčajnej euklidovskej vzdialenosti V_{ij} môžeme vypočítať koeficient podobnosti P_{ij} podľa vzťahu

$$P_{ij} = 1 - \tilde{V}_{ij} ,$$

kde $\tilde{V}_{ij}$ označuje buď priemernú vzdialenosť $\bar{V}_{ij}$ zobrazenú do intervalu (0, 1), alebo euklidovskú vzdialenosť V_{ij} s tou istou vlastnosťou. Euklidova vzdialenosť sa ukázala ako doteraz najvhodnejšia miera pre výpočet podobností medzi pôdnymi jednotkami (Gruijter [7]).

Niekedy býva Euklidova vzdialenosť neuspokojivá a používa sa absolútna vzdialenosť (Cain, Harrison [1]), ktorá predstavuje priemernú sumu absolútnych rozdielov:

$$V_{ij} = \frac{1}{n} \sum_{k=1}^{n} |x_{ik} - x_{jk}| .$$

Canberrova miera sa použila v niekoľkých štúdiách (Moore, Russell [13], Campbell a kol. [2], Webster, Burrough [21]). Mieru vytvorili Lance, Williams [11] a definovali ako

$$V_{cij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}| / (x_{ik} + x_{jk}) .$$

Táto miera je výhodná z hľadiska výpočtových možností, pretože údaje nemusia byť štandardizované a zvlášť je vhodná pre škálované kvantitatívne premenné.

Korelačný koeficient použili pre porovnanie pôdnych profilov Moore, Russell [13], Cuanalo, Webster [3]. Vzorec pre porovnanie dvoch jedincov i a j je

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{n} (x_{jk} - \bar{x}_j)^2\}^{1/2}},$$

kde $\bar{x}_i$ a $\bar{x}_j$ predstavujú priemerné hodnoty jedincov i a j. r_{ij} nadobúda hodnoty od -1 do $+1$, pričom záporné hodnoty nemožno dosť dobre interpretovať. Znaky musia byť štandardizované a centrované tak, aby ich priemery boli nula. Celkove sa korelačný koeficient nedoporučuje.

Gowerov všeobecný koeficient podobnosti [5] má prednosť v tom, že môže súčasne porovnávať rôzne typy údajov (napr. kvantitatívne s nominálnymi a dichotomickými), čo je pri pôdnych jednotkách zvlášť výhodné. Možnosť porovnávania dvoch jedincov i a j je znázornená veličinou δ_{ijk}, ktorá sa rovná 1, ak znak môže byť porovnaný a 0, ak znak nemôže byť porovnaný (napr. ak x_{ik} alebo x_{jk} sú neznáme alebo neaplikovateľné hodnoty). Podobnosť je definovaná ako

$$P_{ij} = \sum_{k=1}^{n} s_{ijk}\,\delta_{ijk} \Big/ \sum_{k=1}^{n} \delta_{ijk},$$

pričom s_{ijk} predstavuje hodnotu pre platné porovnanie k-teho znaku.
Podmienky pre porovnanie rôznych údajov navzájom sú určené takto:

a) pre dichotomické znaky prítomnosť hodnoty znaku je označená (+) a jej neprítomnosť (−). Ak sú známe všetky hodnoty, potom pre objekty i a j existujú štyri kombinácie hodnôt,

b) pre kvalitatívne znaky $s_{ijk} = 1$, ak sa dva objekty zhodujú v k-tom znaku a $s_{ijk} = 0$, ak sa rozlišujú,

c) pre kvantitatívne znaky s hodnotami $x_1, \ldots, x_v$ bude

$$s_{ijk} = 1 - |x_i - x_j| / R_k,$$

kde R_k označuje rozsah znaku v populácii alebo vo vzorke. Ak $x_i = x_j$, potom $s_{ijk} = 1$ a ak x_i a x_j sa nachádzajú na opačných koncoch svojho rozsahu R_k, hodnota $s_{ijk} = 0$.

Gower odporúča udať každému znaku váhu w_k a ak je možné porovnanie $\delta_{ijk} = w_k$, potom všeobecný koeficient podobnosti má vzorec

$$P_{ij} = \sum_{k=1}^{n} s_{ijk}\, w_k \Big/ \sum_{k=1}^{n} w_k.$$

Rayner [14] bol prvý, ktorý použil Gowerov všeobecný koeficient podobnosti pre výpočet podobností medzi pôdnymi horizontmi.

Informačný rádius Jardina, Sibsona [8] sa vypočítava pre kvantitatívne údaje za predpokladu, že všetky x_{ik} sú kladné čísla.

$$P_{\mathrm{inf}\,ij} = \sum_{k=1}^{n} x_{ik} \log \frac{2x_{ik}}{x_{ik} + x_{jk}} + \sum_{k=1}^{n} x_{jk} \log \frac{2x_{jk}}{x_{ik} + x_{jk}},$$

Tab. 1. Platnosť dichotomického porovnania

	Hodnoty znaku			
i	+	+	—	—
j	+	—	+	—
s_{ijk}	1	0	0	0
δ_{ijk}	1	1	1	0

Tab. 2. Počty binárnych údajov

		Objekt j	
		+	—
Objekt i	+	a	b
	—	c	d

kde log je pri základe 2. Informačný rádius poskytuje najlepšie výsledky vtedy, ak sú všetky znaky štatisticky nezávislé.

Asociatívne koeficienty sa vypočítavajú pre dvojhodnotové diskrétne znaky. V podstate možno pre tieto znaky použiť každú údajovú maticu, pretože každý znak za určitých podmienok pripúšťa dichotomizáciu. Avšak treba vziať do úvahy výber logicky nezávislých znakov a do určitej miery počítať so stratou informácie. V tab. 2 sú znázornené počty vyskytujúcich sa binárnych údajov (tzv. kontingenčná tabuľka).

a — počet znakov, ktoré sú prítomné pri obidvoch objektoch,

b, c — počet znakov, prítomných len pri jednom z dvoch znakov,

d — počet znakov, ktoré chýbajú pri obidvoch objektoch,

$$a + d = g \text{ (súhlasné znaky)},$$

$$b + c = u \text{ (nesúhlasné znaky)},$$

$$g + u = n \text{ (úplný počet znakov)}.$$

Koeficient podľa Jaccarda a Sneatha

$$P_{jS} = \frac{a}{a + u} .$$

Koeficient podľa Sokala a Michenera

$$P_{SM} = \frac{g}{n} .$$

Koeficient podľa Rogersa a Tanimota

$$P_{RT} = \frac{g}{g + 2u} .$$

Koeficient podľa Hamanna

$$P_H = \frac{g - u}{n} .$$

Phi-koeficient

$$P_{Phi} = \frac{ad - bc}{f[(a + b)(c + d)(a + c)(b + d)]} ,$$

kde $f_{(x)} = \sqrt{x}$.

Existuje veľké množstvo asociatívnych koeficientov, prehľadne sú uvedené v Sokal, Sneath [18]. Ich použiteľnosť sa overuje prepočítavaním jednotlivých koeficientov na zvolených modeloch.

2.1 Faktorová analýza (extrakcia znakov)

Doteraz sme redukciu demonštrovali takým spôsobom, že sme všetky znaky pri porovnávaní dvoch objektov redukovali na taxometrické merné číslo. Pri extrakcii znakov, ktorá vychádza z faktorovej analýzy, sa súčasne sledujú všetky znaky a objekty. Princíp tejto metódy načrtneme iba stručne, pretože exaktné pojednanie o faktorovej analýze vyžaduje rozsah celej knihy.

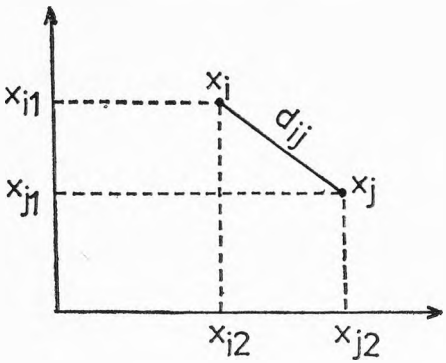
Základom faktorovej analýzy je redukcia n-počtu znakov na prehľadný počet faktorov $F_1, \ldots, F_p$, čím možno s takto získanými veličinami lepšie zaobchádzať. Na začiatku sa hodnoty údajovej matice štandardizujú. Zo štandardizovaných znakových hodnôt sa vypočíta korelačný koeficient medzi znakom k a l. Koreláciu medzi n-počtom znakov si môžeme predstaviť ako účinok p-počtu faktorov. Za určitých matematických predpokladov možno nájsť vzťahy medzi korelačnými koeficientmi a faktorovými nábojmi, ktoré určíme metódou hlavného faktora. Faktorové náboje (matica $n \times p$), ktoré sprostredkujú súvislosť lineárnych vzťahov medzi štandardizovanými znakmi a faktormi, nanesieme na súradnicový systém. Podľa určitých matematických pravidiel môže súradnicový systém rotovať takým spôsobom, aby vytvoril jednoduchú štruktúru faktorových nábojov. Matematickými výpočtami možno zistiť hodnoty faktorov pre každú operačnú taxonomickú jednotku, ktorá už nie je charakterizovaná n-počtom znakov, ale p-počtom faktorov (matica $N \times p$).

Metóda extrakcie znakov má jedinečnú výhodu v kompletnom zohľadnení znakov, jej nevýhodou býva vysoká časová náročnosť v uskutočňovaní výpočtových operácií.

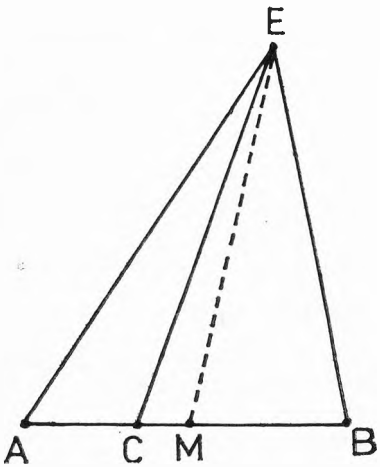
3. AGLOMERÁCIA — HIERARCHICKÉ SYSTÉMY

Veličiny rozpoznávania, ktoré sa získali v procese redukcie, musia sa spracovať takým spôsobom, aby sa dosiahlo usporiadanie taxonomických jednotiek do skupín. Rozoznávame dva základné druhy metód: hierarchické a nehierar-

chické. V hierarchických metódach sa používa buď zoskupovací, alebo rozdeľovací postup. Pri zoskupovacom procese sa vyberajú najpodobnejšie fázy jedincov a spájajú sa do podtried na určitej úrovni podobnosti. Ďalší krok je v prepočítavaní podobnosti medzi novovzniknutým podsúborom a ostatnými jedincami. Rozdeľovacie postupy sú z výpočtového hľadiska prácnejšie, pri týchto postupoch sa pravidelne získa menší počet tried (Edwards, Cavalli-Sforza [4]). Princíp nehierarchických metód sa zakladá na získaní prehľadu o vzá-



Obr. 1. Dvojrozmerný znakový priestor



Obr. 2. Geometrické vzťahy centroidnej stratégie

jomnej podobnosti alebo rozdielnych klasifikovaných objektov (Schnell [15]). Hlavnou prednosťou hierarchických metód pred nehierarchickými je, že úprava a uskladnenie informácií je ľahšie, pretože akýkoľvek jedinec sa identifikuje postupným umiestnením do určitej skupiny.

Keďže existuje veľký počet aglomeračných stratégií, sústredíme sa len na tie, ktoré sa použili pri zoskupovaní pôdnych taxonomických jednotiek.

Metóda jednoduchej väzby (Single-linkage)

Metódu prvýkrát opísal Sneath [16] a pre pôdnu klasifikáciu ju použili Rayner [14] a Moore, Russell [13]. V ekológii je táto metóda známa ako stratégia najbližšieho suseda (Lance, Williams [11]). Stratégia „single-linkage" vychádza z matice vzdialenosti a zakladá sa len na vzdialenosti medzi pármi jedincov. Vzdialenosť medzi jedincom a skupinou je vzdialenosťou medzi týmto jedincom a najbližším členom skupiny. Vzdialenosť medzi skupinami je vzdialenosťou medzi ich najbližšími členmi. Metóda má niekoľko slabostí: neberie do úvahy usporiadanie jedincov (je nehierarchická) a často sa vyskytuje tendencia k tzv. reťazovaniu jedincov.

Wishart [23] modifikoval svoj postup takým spôsobom, ktorý predchádza

reťazovaniu jedincov. Svoju metódu nazval módová analýza. V tejto analýze uvažuje počet susedov a rádius v nich, na základe čoho sa identifikuje určitý počet ohraničených klastrov.

Centroidná metóda

Z geometrického aspektu sa najvýhodnejšou zoskupovacou stratégiou javí centroidná metóda Gowera [6]. Pozícia novej skupiny — syntetických jedincov — je v euklidovskom znakovom priestore definovaná centroidom skupiny. Situácia je znázornená na obr. 2. Skupina A obsahuje n_A jedincov a spája sa so skupinou B o n_B jedincoch. Ich zoskupením sa vytvorí nová skupina, ktorá obsahuje $n_A + n_B$ jedincov. Skupina bude umiestnená na jej centroide C na priamke AB tak, že pomer $AC{:}CB = n_B{:}n_A$. Vzdialenosť napr. bodu E bude

$$d_{CE} = \left\{ \frac{n_A}{n_A + n_B} \; d_{AE}{}^2 + \frac{n_B}{n_A + n_B} \; d_{BE}{}^2 - \frac{n_A n_B}{(n_A + n_B)^2} \, d_{AB}{}^2 \right\}^{\frac{1}{2}}.$$

Centroidná metóda oddeľuje dva hlavné klastre zreteľnejšie ako metóda jednoduchej väzby.

Vážená centroidná metóda

Vážená centroidná aglomerácia je známa ako mediánové zoskupenie [Lance, Williams [9]]. Metóda je použiteľná hlavne vtedy, ak dve skupiny obsahujú rozdielne množiny jedincov. Ak napr. n_A je počtom jedincov oveľa väčšia ako n_B, potom vzdialenosť novej skupiny M ku skupine E je nevyvážená. Ak umiestnime novú skupinu M na stred bodov priamky AB, potom jej vzdialenosť ku bodu E bude

$$d_{EM} = \left\{ \frac{1}{2} \; d_{AE}{}^2 + \frac{1}{2} \; d_{BE}{}^2 - \frac{1}{4} \, d_{AB}{}^2 \right\}^{\frac{1}{2}}.$$

Metóda skupinového priemeru

Metódu skupinového priemeru vyvinuli Sokal, Michener [17]. Ak máme spojiť dva jedince alebo skupiny, vzdialenosť alebo podobnosť medzi nimi a novou skupinou vypočítame ako priemer vzdialenosti alebo podobnosti medzi nimi a všetkými členmi novej skupiny. Priemerné vzdialenosti sú zvyčajne o trochu väčšie ako vzdialenosti medzi centroidmi.

Metóda úplnej väzby (Complete linkage)

Táto metóda, ktorej iné pomenovanie je stratégia najvzdialenejšieho suseda [Lance, Williams [9]] je presným opakom metódy jednoduchej väzby. Kritériom metódy je najväčšia vzdialenosť medzi jedincom a ďalšími členmi zosku-

povania. Výsledkom rastu skupiny od jej susedov býva skutočnosť, že skupiny sa v znakovom priestore rozširujú a majú hypersférický smer. Porovnávaním tejto metódy s ostatnými sa zistilo, že v pedológii nemá žiadne použitie.

Kombinatorické znázornenie

Lance, Williams [9, 10] ukázali, že všetky doteraz uvedené metódy sú varianty jednoduchého lineárneho systému. Znovu uvažujeme obr. 2. Dve skupiny na bodoch A a B o n_A a n_B jedincoch sa spoja napr. v bode K v blízkosti bodov C a M. Ak sú známe hodnoty vzdialeností d_{AB}, d_{AE}, d_{BE} a veľkosti skupín n_A a n_B, potom môžeme vypočítať d_{EK} podľa vzorca

$$d_{EK} = \alpha_A d_{AE} + \alpha_B d_{BE} + \beta d_{AB} + \gamma |d_{AE} - d_{BE}|.$$

Tento vzťah nazývame kombinatorický a pre každú stratégiu sú odvodené hodnoty parametrov α_A, α_B, β a γ. Platí, že ak $\gamma = 0$, aglomerácia bude monotónna za predpokladu, že $\alpha_A + \alpha_B + \beta \geq 1$. Hodnoty parametrov pre každú stratégiu sú:

Nevážený centroid

$$\alpha_A = \frac{n_A}{n_A + n_B}, \qquad \alpha_B = \frac{n_B}{n_A + n_B},$$

$$\beta = -\frac{n_A n_B}{(n_A + n_B)^2} = -\alpha_A \alpha_B, \quad \gamma = 0.$$

Vážený centroid

$$\alpha_A = \frac{1}{2}, \qquad \alpha_B = \frac{1}{2}, \qquad \beta = -\frac{1}{4}, \quad \gamma = 0.$$

Nevážený skupinový priemer

Predpokladáme, že skupina na bode E obsahuje n_E jedincov a že nepodobnosť medzi j-tym jedincom v n_E a i-tym jedincom v n_A je $1 - S_{ij}$. Priemer nepodobnosti medzi dvoma skupinami na bodoch A a E je

$$d_{AE} = \frac{1}{n_A} \frac{1}{n_E} \sum_{i=1}^{n_A} \sum_{j=1}^{n_E} (1 - S_{ij}),$$

$$d_{BE} = \frac{1}{n_B} \frac{1}{n_E} \sum_{i=1}^{n_B} \sum_{j=1}^{n_E} (1 - S_{ij}).$$

Priemer nepodobnosti d_{EK} bude

$$d_{EK} = \frac{1}{n_E} \frac{1}{n_A + n_B} \sum_{i=1}^{n_A + n_B} \sum_{j=1}^{n_E} (1 - S_{ij}) =$$

$$\frac{n_A}{n_A + n_B} d_{AE} + \frac{n_B}{n_A + n_B} d_{BE}.$$

Parametre pre kombinatorickú rovnicu sú

$$\alpha_A = \frac{n_A}{n_A + n_B}, \quad \alpha_B = \frac{n_B}{n_A + n_B}, \quad \beta = 0, \quad \gamma = 0.$$

Vážený skupinový priemer

$$\alpha_A = \alpha_B = \frac{1}{2}, \quad \beta = 0, \quad \gamma = 0.$$

Jednoduchá väzba

$$\alpha_A = \alpha_B = \frac{1}{2}, \quad \beta = 0, \quad \gamma = -\frac{1}{2}.$$

Úplná väzba

$$\alpha_A = \alpha_B = \frac{1}{2}, \quad \beta = 0, \quad \gamma = \frac{1}{2}.$$

Flexibilná stratégia

V postupe, ako napr. stratégia najvzdialenejšieho suseda sa rastom skupín vzdialenosti medzi nimi zväčšujú tak, že sa zdá, že priestor okolo nich dilatuje. Ak je táto tendencia zvlášť silná, môžeme uplatniť stratégiu v moderovanou priestorovou dilatáciou. Lance, Williams [9] navrhli flexibilnú kombinatorickú stratégiu s týmito parametrami:

$$\alpha_A + \alpha_B + \beta = 1, \quad \alpha_A = \alpha_B, \quad \beta > 1, \quad \gamma = 0.$$

Parametre β možno prispôsobiť a meniť tak stupeň priestorového zakrivenia okolo narastajúcich skupín. Ak sa hodnoty β približujú k 1, vytvára sa veľmi silné sťahovanie priestoru a následné reťazovanie. Všeobecne sa odporúča hodnota $\beta = -0{,}25$. Flexibilnú stratégiu s touto hodnotou použili Moore, Russell [13], Campbell a kol. [2] a Moore a kol. [12].

Wardova metóda

Cieľom tejto metódy [19] je minimalizovať chybu súčtu štvorcov, kde suma vzdialeností medzi jedincami a ich skupinovými centroidmi je Σd^2. Východis-

kovým bodom je matica vzdialeností. Na každom stupni spojenia je také spojenie, ktoré spôsobuje zvýšenie aspoň o Σd^2. Pre spojenie dvoch skupín A a B o jedincoch n_A a n_B je toto zvýšenie ekvivalentné

$$\frac{n_A - n_B}{n_A + n_B} d_{AB}^2,$$

kde d_{AB} je vzdialenosť medzi skupinovými centroidmi. Metóda uprednostňuje spájanie jedincov do malých skupín a silne zväčšuje priestor.

Asociačná analýza

Metódu vyvinuli Williams, Lambert [22] a najviac je známa v ekologických klasifikáciách. Východiskovým bodom je množina binárnych údajov, kde 1 označuje prítomnosť znaku a 0 jeho neprítomnosť. Ak niektoré vlastnosti predstavujú škálované hodnoty, potom každú škálu môžeme dichotomizovať. Tab. 3 ukazuje jednotlivé počty jedincov. Miera $\varkappa^2$ sa vypočíta z počtov jedincov podľa vzorca

$$\varkappa^2_{kl} = \frac{(n_A n_D - n_B n_C)^2 (n_A + n_B + n_C + n_D)}{(n_A + n_B)(n_C + n_D)(n_A + n_C)(n_B + n_D)}.$$

Pre každú vlastnosť k je hodnota $\varkappa^2$ sumovaná cez zostávajúcich $n - 1$ vlastností. Vyberie sa vlastnosť, pre ktorú je hodnota $\varkappa^2$ najväčšia a súbor jedincov sa rozdelí do dvoch skupín. Jedna skupina obsahuje jedince, ktoré túto vlastnosť majú, a druhá tie, ktoré túto vlastnosť nemajú. Postup sa opakuje, pričom sa použije vlastnosť s druhou najväčšou hodnotou $\varkappa^2$. Konečné rozdelenie sa dosiahne vytvorením skôr určeného počtu tried. Treba poznamenať, že táto metóda nie je vhodná pre úplné kvantitatívne údaje.

Faktorová analýza

Metódu možno použiť len pre maticu korelačných koeficientov a zakladá sa na rovnakom matematickom princípe ako analýza hlavných komponentov pri extrakcii znakov. Avšak korelačné koeficienty sa nevypočítavajú medzi znakmi, ale medzi jedincami. Výpočet prebieha po štádium získania faktorových nábojov, pretože hodnoty faktorov nepripúšťajú interpretáciu. Výsledné hodnoty možno znázorniť v tab. 4. Počet taxónov treba zvoliť tak, aby $p \ll N$. Metóda faktorovej analýzy neposkytuje hierarchické usporiadanie objektov, nanajvýš môžeme analogicky odvodiť hierarchické poradie.

ZÁVER

Problematika numerickej taxonómie je pomerne rozsiahla. V obidvoch článkoch sme sa usilovali o načrtnutie základných princípov riešenia uvedenej problematiky pre praktického používateľa, hlavne pôdoznalca. Nezahrnuli sme

Tab. 3. Počty jedincov: n_A — nemajú znaky k a l,
n_B — nemajú znak k, majú znak l,
n_C — nemajú znak l, majú znak k,
n_D — majú znaky k a l

	Znak 0	l	Spolu
Znak o	n_A	n_B	$n_A + n_B$
Znak k	n_C	n_D	$n_C + n_D$
Spolu	$n_A + n_C$	$n_B + n_D$	$n_A + n_B + n_C + n_D$

Tab. 4. Výsledné hodnoty získané faktorovou analýzou

	Faktor 1 (= taxón 1)	Faktor 2 (= taxón 2)	. . .	Faktor p (= taxón p)
o_1	a_{11}	a_{12}	. . .	a_{1p}
o_2	a_{21}	a_{22}	. . .	a_{2p}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
o_N	a_{N1}	a_{N2}	. . .	a_{Np}

do nich matematicko-štatistické spracovanie pôvodných údajov a programovanie. Používateľ sa môže s týmito problémami obrátiť na matematického štatistika a programátora. Nemusí byť presne oboznámený s priebehom programovania, ale musí ovládať dve skutočnosti:
— čo a akým spôsobom zadáva počítaču,
— čo mu počítač vydá na základe spracovania programu.

Výsledné hodnoty musí vedieť interpretovať a overiť novovytvorenú klasifikáciu v praxi.

LITERATÚRA

1. CAIN, A. J.: An analysis of the taxonomist's judgement of affinity. J. Soil Res., *1*, 1970. — 2. CAMPBELL, N. A. a kol.: Numerical classification of soil profiles on the basis of field morphological properties. Austr. J. Soil Res., *8*, 1970, ss. 43—58. — 3. CUANALO, H. E., WEBSTER, R.: A comparative study of numerical classification and ordination of soil profiles in a locality near Oxford. Soil Sci., *21*, 1970, ss. 340—352. — 4. EDWARDS, A. W. F., CAVALLI-SFORZA, L. L.: A method for cluster analysis. Biometrics, *21*, 1965, ss. 362—375. — 5. GOWER, J. C.: A general coefficient of similarity and some of its properties. Biometrics, *27*, 1971, ss. 857—871. — 6. GOWER, J. C.: A comparison of some methods of cluster analysis. Biometrics, *23*, 1967, ss. 623—637. — 7. GRUIJTER, J. J.: Numerical classification of soils and its application in survey. Neth. Soil Sur. Inst., Wageningen 1977. — 8. JARDINE, N., SIBSON, R.: Mathematical taxonomy. John Wiley nad Sons New York, 1977. — 9. LANCE, G. N. WILLIAMS, W. T.: A general theory of classificatory strategies. I. Hierarchical sys-

tems. Comput. J., *9,* 1967, ss. 373—380. — 10. LANCE, G. N., WILLIAMS, W. T.: A generalized sorting strategy for computer classifications. Nature Lond., 212, 1966.

11. LANCE, G. N., WILLIAMS, W. T.: Mixed-data classificatory programs. I. Agglomerative systems. Austr. Comp. J., *1,* 1967, ss. 15—20. — 12. MOORE, A. W., a kol.: Numerical analysis of soils: a comparison of three soil profile models with field classification. Soil Sci., *23,* 1972, ss. 193—209. — 13. MOORE, A. W., RUSSELL, J. S.: Comparison of coefficients and grouping procedures in numerical analysis of soil trace element data. Geoderma, 1, 1967, ss. 139—158. — 14. RAYNER, J. H.: Classification of soils by numerical methods. Soil Sci., *17,* 1966, ss.79—92. — 15. SCHNELL, P.: Eine Methode zur Auffindung von Gruppen. Biom. Zeitsch. *6,* 1964, s. 47—78. — 16. SNEATH, P. H. A.: The application of computers to taxonomy. J. Gen. Microbiol., *17,* 1957, ss. 201—226. — 17. SOKAL, R. R., MICHENER, C. D.: A statistical method for evaluating systematic relationships. Kans. Univ. Sci. Bull., *38,* 1958, ss. 1409—1438. — 18. SOKAL, R. R. SNEATH, P. H. A.: Principles of numerical taxonomy. Freeman San Francisco, 1963. — 19. WARD, J. H.: Hierarchical grouping to optimize an objective function. J. Am. Statist. Ass., *58,* 1963, ss. 236—244. — 20. WEBSTER, R.: Qunatitative and numerical methaods in soil classification and survey. Claredon Press Oxford, 1977.

21. WEBSTER, R., BURROUGH, P. A.: Computer-based soil mapping of small areas from sample data. I. Multivariate classification and ordination. Soil Sci., *23,* 1972, ss. 210—222. — 22. WILLIAMS, W. T., LAMBERT, J. M.: Multivariate methods in plant ecology. I. Association-analysis in plant communinies. J. Ecol., 47, 1958, ss. 83—101.

Ярослава Горватова

ТЕОРЕТИЧЕСКО-МЕТОДОЛОГИЧЕСКИЕ ПРИНЦИПЫ НУМЕРИЧЕСКОЙ ТАКСОНОМИИ ПОЧВ

2. ОЦЕНКА СХОДСТВА ОБЪЕКТОВ И ОСНОВНЫЕ АГЛОМЕРАЦИИ

Классическая модель кластер-анализа обоснована на:

— создании матрици данных типа Nxn с элементами x_{ik},

— расчете коэффициентов сходства или расстояния и создании матрици сходства $P(o)$, или матрици расстояния $V(o)$ типа NxN с элементами P_{ij}, или V_{ij},

— подборе стратегии сепарации, которая для матрици сходства или расстояния определит кластеры $R_1, \ldots, R_q$.

В почвоведении было для расчета сходства почвенных единиц применено несколько мер расстояния и коэффициентов сходства. Чаще всего применяется расстояние Евклида, абсолютное расстояние и мера Кэнбера. Коэффициент корреляции не оказался подходящим. Коэффициент Говера обращает внимание на разные типы данных и позволяет дать каждому отдельному признаку свой удельный вес. Для расчета сходства между почвенно-экологическими единицами самими подходящими оказались коэффициенты ассоциации с применением бинарных данных. Метод экстракции признаков можно применить в случае, если объект характеризован большим качеством признаков.

Для классификации можно применить несколько стратегий, которые позволяют соединение или рассоединение. Приводятся основные характеристики метода простого соединения, центроидного метода, метода групповой средней, целостного соединения, комбинаторического изображения, анализа ассоциации, флексибильной стратегии, метода Варда и факторного анализа.

Рис. 1. Двухразмерное признаковое пространство.

Рис. 2. Геометрические связи центроидной стратегии.

Таб. 1. Справедливость дихотомического сравнения.

Таб. 2. Количество бинарных данных.
a — количество признаков свойственных обоим объектам, b, c — количество признаков свойственных только одному из двух объектов, d — количество признаков отсутствующих обоим объектам.

Таб. 3. Количество особей.
n_A — неимеющих признак k и признак l, n_B — неимеющих признак k, имеющих признак l, n_C — имеющих признак k, неимеющих признак l, n_D — имеющих признак k и признак l.

Таб. 4. Результаты полученные факторным анализом.

Перевод: Ю. Грашко

Jaroslava Horváthová

THEORETICAL-METHODOLOGICAL PRINCIPLES OF NUMERICAL SOIL TAXONOMY / II. ESTIMATION OF OBJECTS SIMILARITY AND BASIC AGGLOMERATIONS

Classical model of cluster analysis is based on:
— formation of $N \times n$ type data matrix with x_{ik} elements,
— computation of similarity and/or distance coefficients and formation of $N \times N$ type similarity $\{P_{(o)}\}$ and/or distance $\{V_{(o)}\}$ matrices with P_{ij} and/or V_{ij} elements,
— selection of classification strategy assigning clusters $R_1, \ldots, R_q$ to similarity and/or distance matrices.

In pedology, there were several distance metric and similarity coefficients used for the computation of similarity between soil units. Euclidean distance, absolute distance and Canberra metric are most commonly used. Correlation coefficient is considered to be not suitable for soils. Gower's general similarity coefficient considers different data types and allows to assign weight to each character. Associative coefficients computed from binary data are most reasonable for similarity computation between soil-ecological units. Method of character extraction can be usefully used for objects with large number of characters.

When clustering, research worker has several strategies at disposal, where either distributing or grouping process in the matter. We have briefly outlined basic characteristics of these agglomerations: single-linkage grouping, centroid method, group-average method, complete linkage grouping, combinatorial representation, association analysis, flexifile grouping strategy, Ward's method and factor analysis.

Fig. .1 Two-dimensional space.

Fig. 2. Geometrical relations of centroid strategy.

Tab. 1. Validity of dichotomous comparison.

Tab. 2. Numbers of binary data.
a — number of character present in both objects, b, c — number of characters present in one out of two objects, only, d — number of characters absent in both objects.

Tab. 3. Numbers of individuals
n_A — with no k and l characters, n_B — with no k character, with l character, n_C — with k character, with no l character, n_D — with k nad l characters.

Tab. 4. Resulting values obtained by factor analysis.

Translated by T. Antalová