EESTI NSV TEADUSTE AKADEEMIA TOIMETISED. XV KÖIDE
BIOLOOGILINE SEERIA. 1966, Nr. 4

ИЗВЕСТИЯ АКАДЕМИИ НАУК ЭСТОНСКОЙ ССР. ТОМ XV
СЕРИЯ БИОЛОГИЧЕСКАЯ. 1966, № 4

T. FREY, L. VÕHANDU

UUS MEETOD KLASSIFIKATSIOONIÜHIKUTE PÜSTITAMISEKS

Kaasaegset bioloogiat iseloomustavad nii detailidesse tungimine kui ka laienenud huvi paljutunnuseliste objektide vastu. Eriti köidavad need probleemid, mis on seotud «kõrgemat järku ühikute», nagu koosluste, ökosüsteemide jne., uurimisega ja püstitamisega.

Paljutunnuselise, keerulise loomuga objektidesüsteemi objektiivne jaotamine klassidesse on üks probleeme, mille lahendamisega on juba aastaid vaeva nähtud. Matemaatilises kirjanduses leidub mitmeid meetodeid indiviidi klassifitseerimiseks ühesse olemasolevatest klassidest (näit. Anderson, 1958), kuid objektiivsed meetodid klasside eneste püstitamiseks praktiliselt puuduvad. Viimase kümne aasta jooksul esitatud meetoditest pole ükski leidnud täielikku tunnustust. Nende ühiseks puuduseks on tohutu ajakulu andmete töötlemisel, mida ei korva ka elektronarvutid, sest sooritatavate tehete arv on väga suur.

Alljärgnevalt esitatakse kommentaare bioloogiliste objektide klassifitseerimise eri meetoditele, viimaseid endid lähemalt tutvustamata, ja kirjeldatakse üht uut klassifikatsiooniühikute püstitamise meetodit.

Geomeetriline mudel

Et bioloogiliste objektide üldine klassifikatsiooniteooria saab tugineda ainult ruumilisele mudelile, on tarvis hankida informatsiooni kõigi võrreldavate objektide omavahelise sarnasuse või erinevuse kohta. Teiste sõnadega: kogu vajalik informatsioon sisaldub kõikide objektide kui paljumõõtmelise ruumi punktide vastastikuses asetuses. Niisuguse ruumimudeli mitmekülgse analüüsi esitab D. W. Goodall (1963) oma fütotsönoloogilises töös. Toetudes Goodalli mudelile, täpsustame kolme seal käsitletud erijuhtu.

1) Regulaarne jaotus. Mingi objekti (Q_m) ja selle lähima naaberobjekti (Q_n) vahekaugus (d_{mn}) on konstantne uuritava ruumi kõigis osades (sarnaselt kristallvõrele).

2) Juhuslik jaotus. Objektid on ruumis jaotatud suhteliselt ühtlaselt, kuid moodustavad hõredamaid ja tihedamaid piirkondi. Suuruse d_{mn} jaotus vastab normaaljaotuskõverale.

3) Agregeerunud jaotus. Objektid koonduvad ruumis kogumikesse, mis on üksteisest eraldatud objektidevaese või -tühja piirkonnaga. Suurus d_{mn} annab mitmetipulise või asümmeetrilise jaotuskõvera, olenevalt erinevate kogumike tihedusest.

On ilmne, et kolmandal juhul on objektide klassifitseerimine kergesti teostatav, sest tihedad kogumikud on ülejäänud objektidest hästi eraldatavad. Tegemist on looduses reaalselt eksisteerivate ühikutega, millesse kuuluvad objektid on omavahel väga sarnased, kuid naaberkogumikesse kuuluvatest objektidest selgesti erinevad.

Esimesel juhul on võimatu objekte loomulikeks ühikuteks klassifitseerida, sest puudub igasugune kirjeldatav ruumiline muster (objektid moodustavad homogeense massiivi).

Kõige keerukam on olukord teisel juhul. Siin on võimalik kindlaks teha vaid tihedamate kogumike keskosi, kuid kogumike tihedus langeb nende perifeeria suunas ühtlaselt. Piirjooni kahe tihedama kogumiku vahele ei saa tõmmata, sest üks läheb sujuvalt teiseks üle. Rühmitada saab üksnes mingi kogumiku tihedamasse keskossa kuuluvaid objekte. Valdav osa objektidest erineb nendest «keskosade» esindajatest aga sedavõrd, et neid pole võimalik lugeda ühesse neist kuuluvaks.

Nullhüpotees

Niisiis on klassifitseerimine võimalik vaid objektide heterogeense paigutuse puhul, millal teatud ruumiosad on asustatud märksa tihedamalt kui ülejäänud. Missugune neist kolmest juhust vastaks aga nullhüpoteesile? Regulaarse paigutuse puhul on ilmne, et ühe punkti olemasolu korral on tõenäosus teise punkti leidmiseks esimese naabruses (piirkonnas raadiusega d_{mn}) piiratud. Kolmandal juhul on tõenäosus naaberpunkti leidmiseks antud punkti läheduses suurem, kui juhuslikkuse alusel võiks oodata. Järelikult vastab kogumanalüüsi nullhüpoteesile kõige enam teine juht, kus tõenäosus objekti esinemiseks ükskõik millises uuritava ruumi punktis ei sõltu ülejäänud objektide asukohast. Niisiis väidab nullhüpotees, et vaadeldavad objektid paiknevad uuritavas ruumiosas juhuslikult. Ruumimustrite statistilist hindamist on käsitlenud mitmed uurijad (Barton, David, 1962; Schneiderman, Smith, 1962 jt.), tuginedes dispersioonanalüüsile. Siiski tuleb arvestada, et reaalsete bioloogiliste objektide puhul ei saa kaks või mitu neist asuda täpselt samas punktis, sest looduses ei leidu absoluutselt sarnaseid indiviide. Seetõttu tekib ruumimustrite summaarsel dispersioonanalüüsil teatav viga (Edwards, Cavalli-Sforza, 1965), mille kõrvaldamine jääb edasise uurimistöö ülesandeks.

Kogumanalüüsi meetoditest

N objekti, mida iseloomustavad M tunnust, moodustavad M-dimensioonilises ruumis teatava N punktist koosneva ruumimustri. Objektidena tulevad kõne alla kas kromosoomid mitoosifotol, tähed galaktikates, mingit liiki organismi üksikesindajad, patsiendid, biotsönoosid jne., kusjuures vastavateks tunnusteks võiksid olla vahekaugus fototasapinnal, ruumiline distants, morfoloogilised tunnused, haigussümptoomid ja kooslusi (tsönoose) moodustavad liigid. Algandmete teisendamine võimaldab uuritavat mustrit väljendada kas objektidevaheliste erinevuste või sarnasuste kaudu maatriksina. Käesolevas töös ei pöörata tähelepanu probleemidele, millega uurija puutub kokku enne maatriksi valmimist. Oletame, et maatriksis esitatud arvud peegeldavad objektide poolt moodustatud ruumilist mustrit küllaldase täpsusega. Järgmiseks probleemiks on maatriksi analüüs, mille eesmärgiks ongi objektide poolt moodustatud ruumimustri kindlakstegemine, seega kogumanalüüs. Teiste sõnadega: uurija püüab objekte klassifitseerida, kui see on võimalik.

D. W. Goodall (1953) ning W. T. Williams ja J. M. Lambert (1959, 1960) lähenesid klassifitseerimisprobleemile faktoranalüüsi *R*-tehnika alusel, püüdes objektide algmassiivi jaotada alaklassideks (jaotust korrati seni, kuni saadi homogeensed grupid). Viimaseid defineeritakse kui ühikuid, kus tunnustevaheline korrelatsioon puudub. Teine võimalus on üksikobjekte vaadelda faktoranalüüsi *Q*-tehnika eeskujul, tuginedes mitte üksiktunnustele, vaid objektide sarnasusele või erinevusele (Edwards, Cavalli-Sforza, 1965; Mattson, Dammann, 1965). Viimasega sarnanevad ka mitmed intuitiivsed meetodid (Rao, 1952; Florek jt., 1951), mis lähtuvad aga induktiivsest printsiibist (objektide ühendamine rühmadesse). Töömahukuse seisukohalt on viimane suund märksa vastuvõetavam (vt. Harman, 1960).

Rühma eraldamine peaks üheaegselt tuginema nii objektide sarnasusele kui ka erinevusele. Antud massiivis saavad objektid olla omavahel sarnased sellega, et nad ülejäänutest erinevad. Seega, kui mingi grupisisene keskmine distants $\bar{d}_S$ on küllalt väike, peab kogu massiivi keskmise distantsi $\bar{d}_N$ puhul ülejäänud objektide omavaheline keskmine distants $\bar{d}_{N-S}$ olema sellest suurem. Vastasel juhul on tegemist regulaarse jaotumusega. Niisiis püüdleme rühmade moodustamisel rühmasisese varieeruvuse ΣS_i minimiseerimise ja rühmadevahelise varieeruvuse ΣRV_i maksimiseerimise poole. Nagu näitavad A. W. Edwards ja L. L. Cavalli-Sforza (1965), on niisugune printsiip kooskõlas R. A. Fisheri (1936) diskriminantanalüüsi põhimõtetega, kusjuures paljumõõtmelise ruumi analüüsimisel tuleb vahekaugused ruutida. Lähtudes asjaolust, et kogu objektidemassiivi varieeruvus massiivi keskmise suhtes võrdub vahekauguste ruutude summa $\sum_{i=1}^{N} d_{ij}^2$ ja objektide arvu N jagatisega, arendavad nad oma meetodi kõikvõimalike kombinatsioonide võrdlemiseni, mille eesmärgiks on leida võimalus massiivi jagamiseks kahte rühma nii, et rühmadesisese varieeruvuse summa oleks minimaalne. Iga saadud rühma sees toimetatakse uuesti sama protseduuri, kuni saadakse kaheliikmelised lõppgrupid. Meetodi põhiliseks puuduseks on jällegi töömahukus: $(n-1)^2 2^{n-1}$ sekundit elektronarvutil «Olivetti Elea 6001» ajakarakteristikuga 5 miljondikku sekundit tehtele (Edwards, Cavalli-Sforza, 1965). Teiselt poolt pole hierarhiline kaheks klassifitseerimine kaugeltki alati vastuvõetav.

Rühmitamise protseduur

Otsides võimalusi kogumanalüüsi uue meetodi väljatöötamiseks, torkab silma mitme uurija (Goodall, 1953; Lambert, Williams, 1962; Williams, Lambert, 1959, 1960; Mattson, Dammann, 1965; Edwards, Cavalli-Sforza, 1965) püüe objektide algkogumit hierarhiliselt j a o t a d a. Seevastu tunduvad Harmani (1960), Floreki (Florek jt., 1951) jt. meetodid, mis seisnevad objektide ü h e n d a m i s e s homogeensetesse gruppidesse, olevat loogiliselt enam põhjendatud, sest nad peavad võimalikuks ka rühmadesse mittekuuluvaid, üleminekutunnustega objekte.

Võttes arvesse, et rühm on seda tihedam, mida väiksema reaalse ruumiosa iga üksik tema liikmeist hõlmab, jõuame rühmitamisele sarnasuse alusel. Ideaalne oleks üheaegselt arvestada nii sarnasust kui ka erinevust. Sarnasuse mõõduks sobib antud n objektist koosneva i-nda kogumi rühmasisene varieeruvus:

$$S_i = \sum_{i=1}^{n} d_{ij}^2.$$

Erinevus ülejäänud objektidest on iseloomustatav aga rühmavälise varieeruvusega:

$$V_i = \sum_{n+1}^{N} d_{ij}^2 .$$

Mõlema väljendusvormiks sobib keskmine ruutdistants (keskmise sideme suurus)

$$\bar{d}_S = \frac{2\sum_{i=1}^{n} d_{ij}^2}{n(n-1)}$$

ja

$$\bar{d}_V = \frac{\sum_{n+1}^{N} d_{ij}^2}{n(N-n)},$$

kusjuures

$$2\sum_{i=1}^{n} d_{ij}^2 + \sum_{n+1}^{N} d_{ij}^2 = \sum_{i=1}^{N} d_{ij}^2 .$$

Järelikult defineerime rühmavälist varieeruvust nende distantside ruutude summana, mis eraldavad antud rühma kuuluvaid objekte sinna mittekuuluvatest objektidest. Seega oleks kogumanalüüsimeetodi ülesandeks anda arvutuseeskiri rühmasisese varieeruvuse minimiseerimiseks ja rühmavälise varieeruvuse maksimiseerimiseks. Kui kõnesolev meetod sellele nõudele vastab, ei tarvitse maatriksis esitada ruutdistantse, vaid kas lihtsalt vahekaugusi d_{ij} või koguni mõnesuguseid sarnasuse hinnanguid. On ilmne, et rühmitamine algab objektide paarist Q_a ja Q_b, millede vahekaugus (erinevus) on vähim (ehk sarnasus suurim). Leidnud maatriksist vähima arvu d_{ab}, vaatleme kolmanda liikme Q_c lisandamise võimalusi. Ilmselt on kahele eelnevale liikmele üheaegselt lähim see objekt, mille sidemete summa $d_{ac} + d_{bc}$ on minimaalne. Neljas liige Q_k peab andma

$$\sum_{j \neq k}^{j} d_{jk} = d_{ak} + d_{bk} + d_{ck} = \min \quad \text{jne.}$$

Sellist protseduuri kogu maatriksile rakendades ühendame lõpuks kõik rühma kuuluvad objektid. Edasi aga lisanduvad üleminevate tunnustega objektid ja seejärel varjutab analüüs ülejäänud tegelikult eksisteerivad rühmad, kuna objektid lisanduvad sfääriliselt lähterühma keskpunktile. Arvutades iga objekti lisandumise järel $\bar{d}_S$, saame rühmasisese keskmise vahekauguse ehk rühma poolt hõivatud ruumi iseloomustava raadiuse. Kui rühma keskosa on ammendatud, hakkab $\bar{d}_S$ kiiremini suurenema. Seejärel, kui üleminevad objektid on kaasa haaratud, on suurenemine väike või $\bar{d}_S$ hakkab koguni kahanema, sest naaberrühmade tuumikud on samuti keskmisest tihedamad. Suuruse $\bar{d}_S$ selline käitumine võimaldab otsustada rühma piiride üle. Et saada õiget pilti ka naaberrühmade kohta, tuleb sama protseduuri alustada kõigist rühmadest. Selleks leiame kõikidele objektidele lähimad naaberobjektid. Kui esimene rühm on ammendatud, alustame kasutamata arvudest vähimaga. Lõpuks jõuame sinnamaale, et rühmitamata objekte enam pole.

Kirjeldatud meetod on seotud korduvate liitmistega, mistõttu suure objektide massiivi korral tuleks arvutusi võrdlemisi palju. Selle töölõigu hõlbustamiseks koostati Tartu Riikliku Ülikooli biofüüsika laboratooriumis vastav programm elektronarvutile «Ural-4».

Rühmitamise kriteerium

Eespool märgiti, et objektiivne kogumanalüüs peaks üheaegselt tuginema nii sarnasusele kui ka erinevusele.

H. H. Harman (1960) lähtuski sellisest käsitlusest, tuletades rühma karakteristiku

$$B = \frac{200\,\bar{d}_{S_i}}{\bar{d}_{V_i}} = \frac{200(N-n)S_i}{(n-1)V_i}.$$

Märganud loogilise kaalutluse põhjal, et nii S_i kui ka V_i suurenevad rühma kasvatamisel pidevalt, vaatleb Harman eeskätt nende suhte kasvamise hüppelisust. Et muud kriteeriumi pole esitatud, ei saa Harmani B-kordaja kasutamisest erilist efekti oodata, sest samaks otstarbeks sobib ka lihtsamini arvutatav $\bar{d}_{S_i}$. Viimane kasvab rühma kasvatamisel, sooritades samuti «hüppe», kui rühm on ammendatud. Seega on nii B kui ka $\bar{d}_{S_i}$ kasutatavad üksnes selgesti eraldunud gruppide puhul. Et aga grupid on enamasti üleminevad, pole kerge välja selgitada «olulist hüpet», mistõttu võib rühma sageli kasvatada kogumassiivini välja erilist hüpet leidmata, kuigi massiivis esinevad objektiivselt defineeritavad rühmad.

Objektiivse kogumanalüüsi muudab keerukaks just rühmade ebaselgus. Nii väidab C. R. Rao (1952), et rühmade leidmiseks pole võimalik anda formaalseid eeskirju, sest rühm on vaevu defineeritav mõiste. See on suurel määral tõsi. Siiski peaks mõtlema kriteeriumi leidmisele vähemalt niisuguseks juhuks, kui rühmad avalduvad enam-vähem selgesti.

Loogilised kaalutlused näitavad, et selgesti avalduv heterogeensus vastab minimaalse rühmadesisese ΣS_i ning maksimaalse rühmadevahelise ΣRV_i ja ka maksimaalse rühmadevälise ΣV_i varieeruvuse nõudele. Et nii S_i kui ka V_i koosnevad üksikute vahekauguste summast, oleneb nende suurus summeeritud liikmete arvust. Seetõttu on $\bar{d}_S$ ja $\bar{d}_V$ võrdlemiseks sobivamad. Esimese suurenemine näitab rühmasisese tiheduse langust ehk rühma poolt hõivatud kerakujulise ruumi raadiuse kasvu. Rühmaväline keskmine distants näitab, kuivõrd eraldi seisab vaadeldav rühm. Seni kuni rühmale liidetakse objekte, mis sellesse rühma tõepoolest kuuluvad, suurenevad nii $\bar{d}_S$ kui ka $\bar{d}_V$. Kui rühma hõlmatakse teise rühma kuuluv objekt, suureneb $\bar{d}_S$ tugevasti, $\bar{d}_V$ aga kas üldse mitte või ta koguni väheneb, sest rühma keskpunkt nihkub tublisti uue objekti suunas, rühma mittekuuluvate objektide keskpunkti aga ühe objekti kõrvaldamine nii oluliselt ei muuda. Järelikult oli kahe keskpunkti vahekaugus enne «vale» objekti lisamist maksimaalne. Väljapääs peitub selles, et mõlemad keskpunktid on vastavalt $\bar{d}_S$ ja $\bar{d}_V$ kaudu kaudselt kirjeldatud. Nimelt näitab kummagi suurenemine ühtlasi vastava keskpunkti nihkumise astet. Kuni $\bar{d}_S$ kasvab aeglaselt, nihkub ka rühma keskpunkt aeglaselt. Et samal ajal kasvab $\bar{d}_V$ kiiremini, leiab aset keskpunktide vahekauguse suurenemine. Et meie eesmärgiks on seda vahekaugust maksimiseerida, tuleb rühmitamine lõpetada siis, kui $\bar{d}_S$ hakkab kasvama sama kiiresti kui $\bar{d}_V$. Järelikult sobib rühmitamise kriteeriumiks

$$K = \frac{\bar{d}_{V_{j+1}} - \bar{d}_{V_j}}{\bar{d}_{S_{j+1}} - \bar{d}_{S_j}} > 1.$$

K väärtus 1.0 vastab rühma kuuluvate objektide keskpunkti ja ülejäänud objektide omavahelise kauguse maksimumile. Seega kestab rühma kasvatamine seni, kuni $K > 1.0$.

Numbriline näide

Hea võrreldavuse huvides kasutame C. R. Rao (1952) andmeid A. W. Edwardsi ja L. L. Cavalli-Sforza (1965) poolt ruutdistantsideks teisendatud kujul (tabel 1), käsitades neid lihtsuse mõttes lihtdistantsidena.

Tabel 1

Kaheteistkümne India antropoloogilise rühma erinevuste maatriks
(erinevused ruutdistantsidena Edwards, Cavalli-Sforza, 1965 järgi)

Q_{ij}	B_1	B_2	C_1	C_2	D	Bh	Ch	M	A_1	A_2	A_3	A_4	Rass	
1	0	5	65	42	54	84	57	54	22	28	40	62	B_1	Brahmin, Basti
2	5	0	68	31	53	72	54	49	15	19	28	51	B_2	Other Brahmin
3	65	68	0	25	85	96	99	84	50	56	63	79	C_1	Bhatu
4	42	31	25	0	40	46	88	70	24	29	31	54	C_2	Habru
5	54	53	85	40	0	22	72	46	55	45	43	50	D	Dom
6	84	72	96	46	22	0	94	59	48	42	33	42	Bh	Bhil
7	57	54	99	88	72	94	0	8	64	40	51	42	Ch	Chattri
8	54	49	84	70	46	59	8	0	46	25	27	17	M	Muslim
9	22	15	50	24	55	48	64	46	0	6	9	29	A_1	Ahir
10	28	19	56	29	45	42	40	25	6	0	2	11	A_2	Kurmi
11	40	28	63	31	43	33	51	27	9	2	0	8	A_3	Other Artisan
12	62	51	79	54	50	42	42	17	29	11	8	0	A_4	Kahar
$\sum_{i=1}^{N} d_{ij}$	513	445	770	480	565	638	669	485	368	303	335	445	$\Sigma R_k = 6016$	

Tähistame rea summa

$$\sum_{i=1}^{N} d_{ij} = R_k,$$

kus N — objektide (ridade) arv;
d_{ij} — maatriksi element.

Rühmitamisel on vajalikud veel järgmised suurused:

n — objektide arv rühmas;

s — sidemete (vabadusastmete) arv rühmas; $s = \frac{n(n-1)}{2}$;

S_i — rühmasiseste distantside (sidemete) summa; $S_i = \sum_{i=1}^{n} d_{in}$;

$\bar{d}_S$ — rühmasisene keskmine distants; $\bar{d}_S = \frac{S_i}{s}$;

$(N-n)$ — antud rühma momendil mittekuuluvate objektide arv;

v — rühmaväliste sidemete arv (rühma kuuluvate ja mittekuuluvate objektide vahel); $v = n(N-n)$;

V_i — rühmaväliste sidemete summa; $V_i = \sum_{i=1}^{n} R_k - 2S_i$;

$\bar{d}_V$ — rühmaväline keskmine distants; $\bar{d}_V = \frac{V_i}{v}$;

ΔS — uue objekti liitmisest rühmaga tulenev $\bar{d}_S$ juurdekasv; $\Delta S = \bar{d}_{S_{(j+1)}} - \bar{d}_{S_j}$;

ΔV — samasugune $\bar{d}_V$ juurdekasv; $\Delta V = \bar{d}_{V_{(j+1)}} - \bar{d}_{V_j}$; $(\bar{d}_{V_1} = \bar{d}_{S_1} = 0)$;

K — rühmitamise kriteerium; $K = \frac{\Delta V}{\Delta S}$.

Tabel 2

Q_{in}	Q_{ij}												ΣR_k
	1	2	3	4	5	6	7	8	9	10	11	12	
10	28	19	56	29	45	42	40	25	6	—	**2**	11	303
10, 11	68	47	119	60	88	75	91	52	**15**	(2)	(2)	19	638
10, 11, 9	90	62	169	84	143	123	155	98	(15)	(8)	(8)	**48**	1006
10, 11, 9, 12	152	**113**	248	138	193	165	197	115	(44)	(19)	(16)	(48)	1451

Vähim distants maatriksis on $2 (= d_{10, 11})$; järelikult liigitame ühte rühma kõigepealt objektid 10 ja 11. Arvutused teeme kahes paralleelses skeemis. Esimese (tabel 2) põhjal leiame rühmaväliste distantside summa (rühma kuuluvate objektide summaarsed kaugused rühma mittekuuluvatest objektidest) objektide kaupa, et otsustada, millise objekti hõivamine rühma annab rühmasisesele vahekauguste summale vähima lisa.

Teises skeemis (tabel 3) leiame, kas uus objekt, mis eelmistele rühma kuulunud objektidele summaarselt on kõige lähemal, tõepoolest rahuldab kuuluvuskriteeriumi. Selleks koostame kõigepealt rühma kahe esimese objekti jaoks rea 10, 11 (üksikobjekti puhul $\bar{d}_{V_1} = \bar{d}_{S_1} = 0$) (vt. tabel 3):

Tabel 3

Q_{in}	n	ΣR_k	S_i	s	$\bar{d}_S$	V_i	v	$\bar{d}_V$	ΔV	ΔS	K
10, 11	2	638	2	1	2.00	634	20	31.70	31.70	2.00	15.85
9	3	1006	17	3	5.67	972	27	36.00	4.30	3.67	1.17
12	4	1451	65	6	10.83	1321	32	41.28	5.28	5.16	1.02
2	5	1896	178	10	17.80	1540	35	44.00	2.72	6.97	0.39

Esimesest skeemist näeme, et objektidele 10 ja 11 on summaarselt kõige lähemal objekt 9 (summa on 15). Leiame nüüd, kas objekt 9 sobib rühma. Täitnud tabelis 3 teise rea, näeme, et objekt 9 tõepoolest sobib. Liidame nüüd esimeses skeemis kõigile teise rea (10, 11) elementidele vastavate objektide kaugused objektist 9. Vähima summaga (rida 10, 11, 9) on objekt 12. Kontrollime jälle teises skeemis objekti 12 sobivust. Et $K > 1$, siis objekt 12 sobib. Esimesest skeemist näeme, et järgmisena annab vähima lisa S_i-le objekti 2 hõivamine. Teine skeem aga ütleb, et objekt 2 enam ei sobi. Sellega on üks rühm, kuhu kuuluvad objektid 10, 11, 9, 12, lõppenud. Esimene skeem annab vastuse ka rühma keskse elemendi kohta. Sulgudes esinevatest summadest viimases reas on vähim objekti 11 summa. See näitab, et objekti 11 kaugus teistest rühma kuu-

luvatest objektidest on kõige väiksem. Järelikult ongi objekt 11 rühma keskseks objektiks.

Teise rühma saamiseks otsime kauguste maatriksist vähima elemendi, mis paikneb väljaspool 9., 10., 11., 12. rida ja veergu. Sellise vähima kaugusena leiame $d_{1,2} = 5$. Rakendame jälle kahte arvutusskeemi (tabelid 4 ja 5).

Tabel 4

Q_{in}	Q_{ij}												ΣR_k
	1	2	3	4	5	6	7	8	9	10	11	12	
1, 2	(5)	(5)	133	73	107	156	111	103	37	47	68	113	958

Tabel 5

Q_{in}	n	ΣR_k	S_i	s	$\bar{d}_S$	V_i	v	$\bar{d}_V$	ΔV	ΔS	K
1, 2	2	952	5	1	5.00	948	20	47.40	47.40	5.00	9.48
9	3	1326	42	3	14.00	1242	27	46.00	—1.40	7.00	—0.20

Näeme, et objektid 1 ja 2 jäävad omaette rühma.

Kolmanda rühma alustamiseks leiame objektide 3, 4, 5, 6, 7 ja 8 omavahelistest vahekaugustest vähima: $d_{min} = d_{7,8} = 8$. Analoogiliste arvutuste põhjal näeme, et omaette rühmadeks osutuvad veel objektid 7, 8 ($K = 7.1$), 5, 6 ($K = 2.5$) ja 3, 4 ($K = 2.4$).

Edasi vaatame, millised rühmad on omavahel paremini ühendatavad. Selleks koostame kõigepealt rühmadevaheliste kauguste tabeli (tabel 6).

Tabel 6

Rühm	I	II	III	IV	V	S_i	Q_{in}
$A_2A_3A_1A_4$	—	265	312	358	386	65	9, 10, 11, 12
B_1B_2	265	—	214	263	206	5	1, 2
MCh	312	214	—	271	341	8	7, 8
DBh	358	263	271	—	267	22	5, 6
C_1C_2	386	206	341	267	—	25	3, 4

Selles maatriksis kujutab iga arv vastavate rühmade vahel eksisteerivate seoste kogupikkust ΣRV_i. Rühmade ühendamiseks tuleb võtta nendevaheliste kauguste summa ja liita sellele rühmasisesed kaugused S_i kummastki rühmast. Jagades saadud summa seoste arvuga, saame liitrühmasisese objektidevahelise keskmise kauguse:

Rühm

I + II $265 + 65 + 5 = 335$; $\bar{d}_{I,II} = \frac{335}{15} = 22.33$;

II + III $214 + 5 + 8 = 227$; $\bar{d}_{II,III} = \frac{227}{6} = 37.83$ jne.

Vastavad andmed esitatakse tabelis 7, kust ilmneb, et esimese liitrühma annavad rühmad I ja II.

Järgmise liitrühma leiame tabeli 7, *a* põhjal, kust selgub, et kõige vastuvõetavamaks on III rühma liitmine liitrühmaga I, II.

Tabel 7

Rühm	I	II	III	IV	V	Q_{in}
$A_2A_3A_1A_4$		**22.3**	25.7	29.7	31.7	9, 10, 11, 12
B_1B_2	22.3	—	37.8	48.3	39.3	1, 2
MCh	25.7	37.8	—	50.2	62.3	7, 8
DBh	29.7	48.3	50.2	—	52.3	5, 6
C_1C_2	31.7	39.3	62.3	52.3	—	3, 4

Tabel 7a

Rühm	III	IV	V
I, II	**31.0**	34.9	34.0
III	—	50.2	62.3
IV	—	—	52.3

Tabel 7b

Rühm	IV	V
I, II, III	**39.6**	40.6
IV	—	52.3

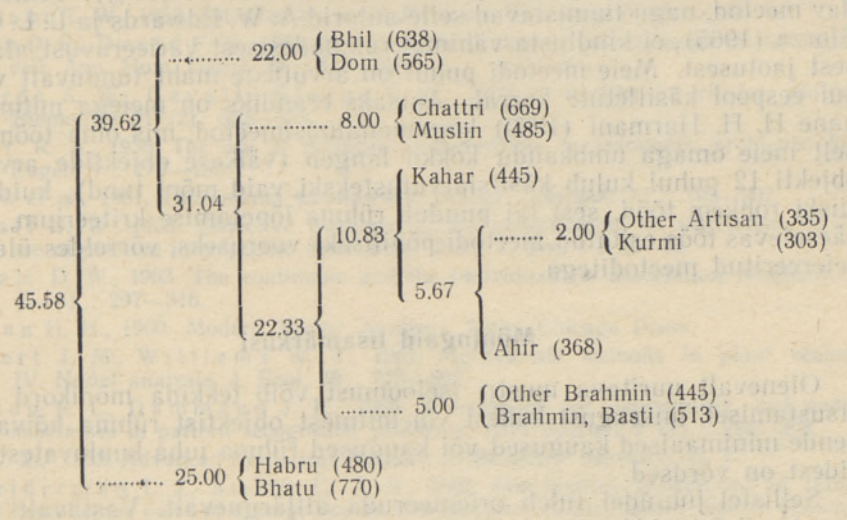
Tabel 7, *b* näitab, et rühmad IV ja V on ikka veel liiga erinevad ja kõige sobivam oleks liita rühm IV olemasolevale liitrühmale I, II, III, millele viimasena liitub rühm V. Et viimase rühma hõivamisega kujuneb kõigi algtabeli arvude summaks

$$\sum_{i=1}^{N} d_{ij},$$

on selle jagatis objektidevaheliste sidemete arvuga $\frac{n(n-1)}{2}$ ehk $\bar{d}_N = 45.58$ keskmiseks objektidevaheliseks kauguseks. Viimane on orientiiriks tabeli 7 hindamisel.

Tulemuste võrdlus

Kordame analüüsi käiku graafiliselt (vt. skeemi), kus arvud tähistavad rühmasisest keskmist sidet, mis üheaegselt iseloomustab grupeeringu tihe-



Klassifikatsiooni skeem Frey-Võhandu meetodil

dust ja annab võrdluse võimaluse eri rühmade täpsustaseme kohta. (Sulgudes on toodud vastava objekti kauguste summa.)

Loomulikult pakub huvi võrrelda tulemusi sama lähtematerjali eelmiste grupeeringutega (tabel 8: Edwards, Cavalli-Sforza, 1965 järgi; *RS* — rühmasisene varieeruvus, *SS* — koguvarieeruvus).

Tabel 8

Kaheteistkümne India hõimu ja rassi grupid

Rao, 1952		Edwards, Cavalli-Sforza, 1965	
Rühm	$\frac{RS \cdot 100}{SS}$ %	Rühm	$\frac{RS \cdot 100}{SS}$ %
Ahir Kurmi Other Artisan Kahar	6.48	Kurmi Other Artisan Kahar	2.79
Chattri Muslim	1.60	Chattri Muslim	1.60
Dom Bhil	4.39	Dom Bhil	4.39
Brahmin, Basti Other Brahmin	1.00	Brahmin, Basti Other Brahmin Ahir	5.59
Bhatu Habru	4.99	Bhatu Habru	4.99
Kokku	18.46	Kokku	19.36

Tabeli 8 andmete võrdlusest selgub, et käesolevas töös kasutatud meetodiga loodud klassifikatsioon langeb täielikult kokku bioloogi skeemiga ning annab vähima rühmadesisese varieeruvuse (18,46%). Teine võrreldav meetod, nagu tunnistavad selle autorid A. W. Edwards ja L. L. Cavalli-Sforza (1965), ei kindlusta vähimat rühmadesisest varieeruvust alates teisest jaotusest. Meie meetodi puhul on arvutuste maht tunduvalt väiksem kui eespool käsitletute korral. Ainsaks erandiks on meiega mitmeti sarnane H. H. Harmani (1960) kogumanalüüsimeetod, mis oma töömahukuselt meie omaga umbkaudu kokku langeb (väikese objektide arvu, näit. objekti 12 puhul kulub käsitsiarvutustekski vaid mõni tund), kuid nõuab siiski rohkem tööd, sest tal puudub rühma lõpetamise kriteerium, mis on käesolevas töös esitatud meetodi põhiliseks vooruseks, võrreldes ülejäänud refereeritud meetoditega.

Mõningaid lisamärkusi

Olenevalt uuritava mustri iseloomust võib tekkida mõnikord raskusi otsustamisel, missugust kahest või mitmest objektist rühma hõivata, kui nende minimaalsed kaugused või kaugused rühma juba kuuluvatest objektidest on võrdsed.

Sellistel juhtudel tuleb orienteeruda alljärgnevalt. Vastavalt rühmasiseste distantside summa S_i minimiseerimise ja rühmaväliste distant-

side summa V_i maksimiseerimise nõudele on objektidest, mille hõivamine annab S_i-le võrdse juurdekasvu, eelistatavam see, mis samal ajal annab suurima lisa V_i-le. Järelikult tuleb otsustada, millise objekti j lisamisel $R_{kj} - 2S_{ij} = max$. Samuti võib kaalutleda ka väga sarnast S_{ij} omavate objektide puhul. Näiteks tabeli 2 neljandas (10, 11, 9, 12) reas annavad praktiliselt võrdse S_{ij} objektid Q_2 (= 113) ja Q_8 (= 115). Lisa V_i-le aga on küllaltki erinev:

$$S_{i2} = R_{k2} - 2S_{i2} = 445 - 226 = 219$$

ja

$$S_{i8} = R_{k8} - 2S_{i8} = 485 - 230 = 255.$$

Viimasest võrdlusest järeldub, et eelistatavam peaks olema objekt 8. Ja tõepoolest, asendades tabelis 3 neljanda rea objektile 8 vastavate arvudega, saame järgmised tulemused: 8; 5; 1936; 180; 10; 18.00; 1576; 45.03; 3.75; 7.17; 0.52. Kuuluvuskriteerium kasvas $K_2 = 0.39$ pealt $K_8 = 0.52$ peale, mis kinnitab objekti 8 paremat sobivust. Seetõttu on sarnaste objektide sobivust otstarbekas hinnata sobimiskriteeriumi

$$g_{ij} = \frac{R_{kj} - 2S_{ij}}{S_{ij}}$$

alusel, rühmitades eelisjärjekorras kõige suurema g_{ij} väärtusega objektid. Niisiis $g_{i2} = 1.94$ ja $g_{i8} = 2.21$.

Samasugusele kriteeriumile saab tugineda ka rühma alustamiseks sobivate objektide valikul (kui minimaalsed vahekaugused on võrdsed või väga vähe erinevad),

$$G_{ij} = \frac{R_{ki} + R_{kj} - 2d_{ij}}{d_{ij}},$$

kuigi enamasti pole selleks ilmset tarvidust.

KIRJANDUS

Anderson T. W., 1958. Introduction to Multivariate Statistical Analysis. N. Y.

Barton D. E., David F. N., 1962. The analysis of chromosome patterns in the normal cell. Ann. Hum. Genet. **25** : 323—329.

Edwards A. W., Cavalli-Sforza L. L., 1965. A method for cluster analysis. Biometrics **21** (2): 362—375.

Fisher R. A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. **7** : 179—188.

Florek et al., 1951. Taksonomia wroclawska. Przegl. Antropol. **17** : 193—207.

Goodall D. W., 1953. Objective methods for the classification of vegetation. I. The use of positive interspecific correlation. Aust. J. Bot. **1** : 39—63.

Goodall D. W., 1963. The continuum and the individualistic association. Vegetatio **11** (5—6) : 297—316.

Harman H. H., 1960. Modern Factor Analysis. Univ. Chicago Press.

Lambert J. M., Williams W. T., 1962. Multivariate methods in plant ecology. IV. Nodal analysis. J. Ecol. **50** : 775—802.

Mattson R. L., Dammann J. E., 1965. A technique for determining and coding subclasses in pattern recognition problems. IBM Journal **9** (4) : 294—302.

Rao C. R., 1952. Advanced Statistical Methods in Biometric Research. N. Y.

Schneiderman L. J., Smith C. A. B., 1962. Non-random distribution of certain homologous pairs of normal human chromosomes in metaphase. Nature [London] **195** : 1229—1230.

Williams W. T., Lambert J. M., 1959. Multivariate methods in plant ecology. I. Association-analysis in plant communities. J. Ecol. 47 : 83—101.

Williams W. T., Lambert J. M., 1960. Multivariate methods in plant ecology. II. The use of an electronic digital computer for association-analysis. J. Ecol. 48 : 689—710.

Eesti NSV Teaduste Akadeemia
Zooloogia ja Botaanika Instituut
Tartu Riiklik Ülikool

Saabus toimetusse
23. V 1966

Т. ФРЕЙ, Л. ВЫХАНДУ

НОВЫЙ МЕТОД ВЫДЕЛЕНИЯ КЛАССИФИКАЦИОННЫХ ЕДИНИЦ

Резюме

В работе определяется нулевая гипотеза при изучении пространственного распределения объектов на фоне изучаемых признаков, а затем — понятие классификационных единиц как подмножеств, для которых дисперсия внутри — их минимальное при максимальности межгрупповых дисперсий.

В работе дается алгоритм для соединения объектов в такие подмножества данного множества, а также найден объективный критерий для окончательного образования данного подмножества. Подробно рассмотрен пример.

Институт зоологии и ботаники
Академии наук Эстонской ССР
Тартуский государственный университет

Поступила в редакцию
23/V 1966

T. FREY, L. VÕHANDU

A NEW METHOD FOR THE ESTABLISHING OF CLASSIFICATIONAL UNITS

Summary

The authors present a definition of the null hypothesis for pattern recognition in a multidimensional space. A classificational unit is defined as a subset which has the minimum within-cluster variation provided the between-clusters variation is maximized. An algorithm is described for building up such subsets of a given set, and an objective criterion for completing the given cluster is also developed. An example is considered in detail.

Acaaemy of Sciences of the Estonian SSR,
Institute of Zoology and Botany
Tartu State University

Received
May 23, 1966