Liet. matem. rink., T. 40, spec. nr., 399–407

Namų ūkių klasterinė analizė

Ričardas KRIKŠTOLAITIS (VDU), Gražina BINKAUSKIENĖ (SD)
el. paštas: ricardas@if.vdu.lt, grazina.binkauskiene@mail.std.lt

Įvadas

Lietuvos Respublikos Vyriausybės 1995 m. spalio 5 d. nutarimu Statistikos departamentas atlieka namų ūkių biudžetų tyrimą. Tai vienas iš sudėtingiausių statistinių tyrimų. Tyrimo tikslas – gauti patikimą informaciją apie namų ūkių gyvenimo lygį, t.y. vartojimo išlaidas, pajamas, jų struktūrą, gyvenimo sąlygas, apsirūpinimą ilgalaikio naudojimo daiktais ir pan.

Namų ūkių biudžetų tyrimo objektas yra privatus namų ūkis. Namų ūkis – tai grupė žmonių, kurie susiję giminystės ar kitais asmeniniais ryšiais, turi bendrą biudžetą, kartu maitinasi ir gyvena viename būste. Namų ūkiu gali būti: šeima, susidedanti iš sutuoktinių su vaikais ar be jų, arba vienas iš tėvų su vaikais; kartu gyvenantys ir bendrą biudžetą turintys giminaičiai, pvz., brolis ir sesuo ir pan.; kartu gyvenantys ir bendru biudžetu susiję asmenys, neturintys giminystės ryšio; vieniši asmenys, gyvenantys iš savo pajamų. Per metus tyrime dalyvauja daugiau kaip 8 tūkst. namų ūkių. Namų ūkiai atrenkami atsitiktinių imčių metodu, remiantis gyventojų registru. Tokia atranka užtikrina vienodas galimybes visų visuomenės sluoksnių atstovams būti atrinktiems tyrime. Atrinkti namų ūkiai tyrime dalyvauja tik vieną mėnesį. Po mėnesio jie keičiami naujais. Namų ūkių biudžetų tyrime taikomi du skirtingi informacijos gavimo metodai – apklausos, kurią vykdo apklausėjas, ir savarankiškos registracijos, t.y., kai tam tikrus duomenis respondentai patys surašo specialiuose tyrimo dokumentuose.

Namų ūkių ekonominė padėtis paprastai vertinama pagal pajamas ir išlaidas vienam namų ūkio nariui. Savo ruožtu namų ūkių diferenciacija charakterizuoja ekonominę šalies situaciją. Manoma, kad bendrą namų ūkių padėtį turėtų atspindėti suvartojami produktų kiekiai. Galima patikrinti hipotezę: „vartojamų produktų kiekiai atspindi ekonominę namų ūkių padėtį“. Todėl namų ūkiai buvo klasterizuoti pagal suvartojamus atskirų produktų kiekius, o gauti klasteriai lyginti su įvairiais namų ūkių statusą charakterizuojančiais faktoriais. Faktorių, apibūdinančių namų ūkius, yra nemažai. Tai ir namų ūkių gyvenamoji vieta, namų ūkių demografinė sudėtis, socialinis-ekonominis statusas, išsimokslinimas, amžius, pajamų šaltinis ir pan. Tiriant produktų vartojimo priklausomybę nuo namų ūkių klasterių naudotas Spearman'o ranginės koreliacijos koeficientas. Buvo nustatyta, kad didžiausia koreliacija tarp klasterių ir namų ūkių gyvenamosios vietos ir išlaidų lygio – taip vadinamų decilių. Tarp klasterių ir kitų namų ūkius apibūdinančių veiksnių koreliacija yra labai maža arba jos visai nėra.

Namų ūkių biudžetų tyrime yra išskirti 3 gyvenamosios vietos tipai: 5 didieji miestai (Vilnius, Kaunas, Klaipėda, Šiauliai ir Panevėžys); kiti miestai bei miesteliai; ir kaimas.

Deciliai skaičiuojami padalijus į dešimt lygių dalių eilutę, kurią sudaro tiriamieji namų ūkiai, išdėstyti didėjimo tvarka pagal vartojimo išlaidų lygį vienam namų ūkio nariui. Pirmąjį dešimtadalį, t.y., pirmąjį decilį sudaro namų ūkiai, kurių narių išlaidos vienam nariui yra mažiausios, antrą decilį – namų ūkiai, kurių išlaidos vienam nariui yra didesnės nei pirmojo decilio tiriamųjų, bet mažesnės nei trečiojo ir t.t.

Esame nuoširdžiai dėkingi prof. R. Rudžkiui ir dr. B. Kaminskienei už vertingus patarimus, pateiktus rašant šį straipsnį.

Tyrimui buvo naudota daugiamačių duomenų klasterizavimo programa, sukurta Matematikos ir informatikos instituto Taikomosios statistikos skyriuje, ir statistinė sistema Statistica 4.5 (S/N: SW4064148514D45).

Tyrimo metodika

Trumpai apžvelgsime klasterizavimo metodiką. Stebimą atsitiktinai parinkto namų ūkio vartojimą žymėkime X. Tyrimui buvo pasirinkta 10 maisto produktų grupių: duona ir kruopos; mėsa ir mėsos gaminiai; žuvis ir žuvies produktai; aliejai ir riebalai; vaisiai; daržovės; cukrus, džemas, medus, šokoladas ir konditerijos gaminiai; druska, prieskoniai ir kiti produktai; kava, arbata ir kakava; kiti bealkoholiniai gėrimai. Tuo būdu X yra dešimtmatis vektorius, kurio i-toji komponentė žymi atitinkamos maisto produktų grupės suvartojimo kiekį per mėnesį. Daroma prielaida, kad visi Lietuvos namų ūkiai gali būti sugrupuoti į kelis klasterius, o kiekviename klasteryje minėtas suvartojimo vektorius turi Gauso skirstinį. Skirtingus klasterius atitinka skirtingi skirstinių parametrai. Taigi, taikomas žemiau aprašomas Gauso skirstinių mišinio modelis.

Tarkim, turim q nepriklausomų d-mačių Gauso atsitiktinių dydžių (a.d.) Y_i, kurių skirstinio tankis $\varphi(\cdot; M_i, R_i) \stackrel{def}{=} \varphi_i$, kur vidurkis M_i ir kovariacinė matrica R_i, $i = 1, 2, \ldots, q$, yra nežinomi. Tegul ν yra atsitiktinis dydis, nepriklausomas nuo Y_i, ir įgyjantis reikšmes $1, 2, \ldots, q$ su nežinomomis tikimybėmis $p_i > 0$, $i = 1, 2, \ldots, q$. Pažymėkime d-matį a.d. $X = Y_\nu$. Kiekvienas stebėjimas priklauso vienai iš q klasių, priklausančių nuo a.d. ν. A.d. X skirstinio tankis yra Gauso mišinio tankis

$$f(x) = \sum_{i=1}^{q} p_i \varphi_i(x) \stackrel{def}{=} f(x, \theta), \quad x \in \mathbb{R}^d, \tag{1}$$

kur $\theta = (p_i, M_i, R_i,\ i = 1, 2, \ldots, q)$ yra nežinomas daugiamatis parametras. Tikimybės $p(i, x) = \mathbb{P}\{\nu = i\}$ yra vadinamos apriorinėmis tikimybėmis.

Nagrinėkime bendrą klasifikacijos problemą kaip įvertinti a.d. imties $X^N \stackrel{def}{=} \{X_1, X_2, \ldots, X_N\}$ su skirstinio tankiu (1) aposteriorines tikimybes $\pi(i, x) = \mathbb{P}\{\nu = i | X = x\}$. Pasinaudoję įvestais pažymėjimais, galime užrašyti:

$$\pi(i, X) = \pi_\theta(i, X) \stackrel{def}{=} \frac{p_i \varphi_i(X)}{f_q(x, \theta)}, \quad i = 1, 2, \ldots, q, \tag{2}$$

EM algoritmas. Jeigu klasių skaičius q yra žinomas, tai maksimalaus tikėtinumo (MT) įvertis θ^* yra efektyvus θ įvertis. Bendriausias metodas paskaičiuoti Gauso mišinio MT įvertį yra taip vadinamas EM (*Expectation Maximization*) algoritmas. Tegul $\pi^N = \{\pi(i, X),\ i = 1, 2, \ldots, q,\ X \in X^N\}$ yra imties X^N įvertintos aposteriorinės tikimybės. Duotai π^N, parametras $\theta = (p_i, M_i, R_i, i = 1, 2, \ldots, q)$ yra įvertinamas panaudojus lygybes:

$$
\begin{aligned}
p_i &= \frac{1}{N}\sum_{j=1}^{N}\pi(i, X_j), \quad i = 1, 2, \ldots, q, \\
M_i &= \frac{1}{N}\sum_{j=1}^{N}\frac{\pi(i, X_j)}{p_i}X_j, \quad i = 1, 2, \ldots, q, \\
R_i &= \frac{1}{N}\sum_{j=1}^{N}\frac{\pi(i, X_j)}{p_i}[X_j - M_i][X_j - M_i]^T, \quad i = 1, 2, \ldots, q.
\end{aligned}
\tag{3}
$$

Duotai θ reikšmei tikimybės π^N paskaičiuojamos naudojant (2) formulę. EM algoritmas yra rekurentinė procedūra, kuri pradeda skaičiuoti arba nuo duoto parametro θ, arba nuo duotos tikimybės π^N įverčių, naudojant (2) ir (3) formules. EM algoritmas paprastai nutraukiamas po kažkiek iš anksto užduotų iteracijų skaičiaus. Parametras θ EM algoritme konverguoja į MT įvertį, jeigu tik pradinis įvertis θ^0 pakankamai arti θ^* reikšmės.

Pastebėkime, kad pradinė parametro θ^0 reikšmė gali būti gauta iš pradinių aposteriorinių tikimybių reikšmių.

Schlesinger (1965), Hasselblad (1966) ir Behboodian (1970) nepriklausomai vienas nuo kito pasiūlė EM algoritmą mišinių skirstiniams. Dabar jau yra gerai ištirtos EM algoritmo savybės.

Esant didelės dimensijos duomenims yra problematiška gerai juos klasterizuoti, sunku interpretuoti gautus rezultatus. Vienas iš būdų išvengti šių problemų – projektavimas į mažesnio matavimo erdves. Trumpai aprašysime straipsnyje panaudotą projektavimo metodą.

Diskriminantinė erdvė. Tegul $V = \mathrm{cov}(X, X)$ bus a.d. X kovariacinė matrica ir, paprastumo dėlei, $\mathbb{E}X = 0$. Apibrėžkime vektorių $u, h \in \mathbb{R}^d$ skaliarinę sandaugą $(u, h) = u^T V^{-1} h$ ir įveskime pažymėjimą u_L – vektoriaus $u \in \mathbb{R}^d$ projekcija į tiesinį poerdvį $L \subset \mathbb{R}^d$. Diskriminantinė erdvė H yra apibrėžiama kaip tiesinis poerdvis $H \subset \mathbb{R}^d$, tenkinantis sąlygą $\mathbb{P}\{\nu = i | X = x\} = \mathbb{P}\{\nu = i | X_H = x_H\}, i = 1, 2, \ldots, q$, $x \in \mathbb{R}^d$ ir turintis minimalią dimensiją. Yra žinoma, kad Gauso mišiniams (1) su vienodom kovariacinėm matricom, $\dim H < q$.

Tegul $k = \dim H$ ir vektoriai $u_1, u_2, \ldots, u_k$ sudaro bazę erdvėje H. Pažymėkime $U = (V^{-1}u_1, V^{-1}u_2, \ldots, V^{-1}u_k)^T$. Tada $\pi(i, x) = \mathbb{P}\{\nu = i | UX = Ux\}$, $i = 1, 2, \ldots, q$, $x \in \mathbb{R}^d$. Tai reiškia, kad suprojektuota imtis $\{UX_1, UX_2, \ldots, UX_N\}$ yra pakankama statistika aposteriorinių tikimybių įvertinimui. A.d. UX skirstinys bus Gauso

mišinio tankis

$$f^H(z) = \sum_{i=1}^{q} p_i \varphi_i^H(z) \stackrel{def}{=} f_q^H(x, \theta_H), \quad z \in \mathbb{R}^k, \tag{4}$$

kur $\varphi_i^H = \varphi(\cdot; M_i^H, R_i^H)$, $i = 1, 2, \ldots, q$ yra k-matis Gauso skirstinys su vidurkiais $M_i^H = UM_i$ ir kovariacinėm matricom $R_i^H = U^T R_i U$. $\theta_H = (p_i, M_i^H, R_i^H,\ i = 1, 2, \ldots, q)$ yra daugiamatis parametras.

Tikslinio projektavimo algoritmas. Vienas iš metodų surasti diskriminantinę erdvę – tikslinis projektavimas (*Projection Pursuit (PP)*). Tai nuosekli procedūra, skirta surasti diskriminantinės erdvės bazinius vektorius. *PP* metodas buvo sukurtas Friedman ir Tukey (1974). Šio metodo savybės yra gerai išnagrinėtos, pvz. [2], [7]. Toliau naudosimės [8] staipsnyje įvestais pažymėjimais.

Tegul $\mathbf{F}$ žymi vienamačių Gauso mišinio skirstinių aibę; $\rho = \rho(G, \Psi), G, \Psi \in \mathbf{F}$ – bet kuris funkcionalas, tenkinantis tokias sąlygas: $\rho(G, \Psi) = 0$ ir $\rho(G, \Psi) > 0$, jeigu $G \neq \Psi$. Nenuliniam vektoriui $u \in \mathbb{R}^d$ apibrėžkime projektavimo indeksą $Q(u) = \rho(F_u, \Phi)$, kur F_u yra a.d. $u^T X$ skirstinys, Φ – Gauso skirstinys su nuliniu vidurkiu ir dispersija $\|u\|^2$.

Tegul ortonormuoti vektoriai $u_1, u_2, \ldots, u_k$ bus skaičiuojami pažingsniui, naudojant formules: $U_0 = \{0\}$, visiems $i = 1, 2, \ldots, d$ skaičiuojame $u_i = \arg\max\{Q(u),\ u \in U_{i-1}^{\perp}, \|u\| = 1\}$, $U_i = \text{span}\{u_1, u_2, \ldots, u_i\}$ ir sustojame, kai $Q(u_i) = 0$. Jeigu $Q(u_1) = 0$, tai diskriminantinės erdvės dimensija $k = 1$. Jeigu visiems $i = 1, 2, \ldots, d$ bus $Q(u_i) > 0$, tai diskriminantinės erdvės dimensija $k = d$. Tuo būdu $k = \min\{i; Q(u_{i+1}) = 0\}$. Statistiškai įvertinant $u_1, u_2, \ldots, u_k$, vietoj nežinomos funkcijos F_u panaudojama atitinkama empirinė pasiskirstymo funkcija. Šiame darbe diskriminantinė erdvė buvo vertinta, parinkus funkcionalą

$$\rho(G, \Psi) = n \cdot \int_{-\infty}^{\infty} \frac{G(x) - \Psi(x)}{\Psi(x)(1 - \Psi(x))} \, \mathrm{d}\Psi(x). \tag{5}$$

Šio funkcionalo savybės buvo ištirtos [9] darbe.

Rezultatai

Apžvelgsime namų ūkių klasterizavimo pagal maisto produktų vartojimą rezultatus, tiek grupuojant pirminius duomenis, tiek ir po projektavimo į dvimatę erdvę.

Klasterinė analizė buvo atlikta su atsitiktinai atrinktais 2027 namų ūkiais, kuriuos Statistikos departamentas apklausė 1999 m. antrame ketvirtyje. Gautų duomenų klasterizacija buvo atlikta dviem būdais: 1) pradinai duomenys buvo klasterizuoti, panaudojant aukščiau minėtą daugiamačių duomenų klasterizavimo programą, kuri remiasi EM algoritmo taikymu; 2) pradiniai duomenys buvo projektuoti *PP* metodu į dvimatę diskriminantinę erdvę ir klasterizuotos jų projekcijos.

Pirmu atveju vykdant daugiamačių duomenų klasterizavimo programą automatinės klasterizacijos režimu buvo gauti 4 klasteriai. Į pirmą klasterį pateko apie 63 % namų

ūkių. Į šį klasterį patekę namų ūkiai vienam namų ūkio nariui vartojo visų maisto produktų mažiau nei bendras vidurkis. Papildomai galima būtų paminėti, kad vaisių (P5) buvo suvartota dvigubai, o kitų belakoholinių gėrimų (P10) beveik trigubai mažiau už vidurkį. Į antrą klasterį pateko apie 19 % namų ūkių. Šiems namų ūkiams būdingas taip pat mažesnis nei vidutinis maisto produktų vartojimas, tik išsiskiria vaisių (P5) ir kitų bealkoholinių gėrimų (P10) poreikis, kurių sunaudota beveik dvigubai daugiau nei bendras vartojimo lygis. Pastarųjų dviejų produktų vartojimo skirtumu ir skiriasi pirmas ir antras klasteriai. Į trečią klasterį pateko apie 5 % namų ūkių. Kavos, arbatos (P9) išgerta mažiau nei vidutiniškai, tačiau kitų produktų sunaudota daugiau už vidurkį. Be to, labai išsiskiria duonos ir kruopų (P1), cukraus, džemo ir pan. (P7) bei druskos, prieskonių ir kitų produktų (P8) vartojimas, kurių vidurkis keletą kartų didesnis už bendrą vidurkį. Į ketvirtą klasterį pateko apie 11 % namų ūkių. Šie namų ūkiai vartojo mažiau tik druskos, prieskonių ir kitų produktų (P8), o likusius produktus – daugiau už vidutinį vartojimą. Šiems namų ūkiams būdingas didelis vaisių (P5), kavos ir pan. (P9) bei kitų bealkoholinių gėrimų (P10) vartojimas, kurių vidurkis keletą kartų viršija bendrą vidurkį. Detalesni vidutiniai suvartotų maisto produktų kiekiai vienam namų ūkio nariui per mėnesį yra pateikiami 1 lentelėje.

Čia N pažymėta namų ūkių kiekis, P1 – duona ir kruopos, P2 – mėsa ir mėsos gaminiai, P3 – žuvis ir žuvies produktai, P4 – aliejai ir riebalai, P5 – vaisiai, P6 – daržovės, P7 – cukrus, džemas, medus, šokoladas ir konditerijos gaminiai, P8 – druska, prieskoniai ir kiti produktai, P9 – kava, arbata ir kakava, P10 – kiti bealkoholiniai gėrimai. Ta pačia klasterizavimo programa buvo išskirti 3 ir 5 klasteriai. Tačiau vertinant Spearmen'o ranginės koreliacijos koeficientą tarp namų ūkių maisto vartojimo rodiklių (t.y. klasterių numerių) ir atskirų namų ūkius charakterizuojančių rodiklių buvo gauta, kad geriausiai pradinius duomenis atspindi 4 klasteriai (žr. 4 lentelę), t.y. Spearmen'o ranginės koreliacijos koeficientas buvo gautas didžiausias.

Antru būdu klasterizuojant, t.y. pradžioje pradinius duomenis suprojektavus į dvimatę erdvę ir po to atlikus klasterizavimą daugiamačių duomenų klasterizavimo programa automatinės klasterizacijos režimu, taip pat buvo gauti 4 klasteriai. Į pirmą klasterį patekę apie 48 % namų ūkių visų maisto produktų vartojimas yra analogiškas kaip ir klasterizuojant pirmu būdu. Į antrą klasterį pateko apie 29 % namų ūkių. Jie šiek tiek daugiau už vidurkį vartojo mėsos ir jos gaminių (P2), vaisių (P5) ir kavos ir pan. (P9). Visų kitų produktų vartojimas buvo truputį mažesnis už bendrą vidutinį vartojimą. Trečiam klasteriui

1 lentelė.
Klasterių, gautų nenaudojant projektavimo, vidurkiai

	N	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Bendras	2027	9,51	6,12	1,31	1,67	2,52	16,76	3,27	1,05	0,21	2,11
1 klasteris	1282	8,68	5,30	1,18	1,57	1,30	15,94	2,32	0,84	0,17	0,68
2 klasteris	392	7,17	5,53	0,84	1,12	4,13	12,27	2,86	0,72	0,18	4,96
3 klasteris	92	18,92	6,66	1,60	2,28	2,84	24,30	12,78	2,92	0,16	2,18
4 klasteris	220	12,28	10,71	2,07	2,25	5,79	21,55	4,35	0,99	0,52	5,10

priklauso apie 10 % namų ūkių. Galima būtų išskirti šiek tiek didesnį kavos ir pan. (P9), bei keletą kartų didesnį vaisių (P5) bei bealkoholinių gėrimų (P10) vartojimą. Visų kitų produktų vartojimas truputį mažesnis už vidurkį. Į ketvirtą klasterį pateko daugiausiai maisto produktų vartojantys namų ūkiai. Jų visų maisto produktų vartojimas keletą kartų viršijo vidutinį vartojimą. Tokių namų ūkių iš viso buvo apie 12 %. Detalesni vidutiniai suvartotų maisto produktų kiekiai vienam namų ūkio nariui per mėnesį yra pateikiami 2 lentelėje.

Kadangi atstumas tarp išskirtų klasterių nėra didelis, buvo atliktas klasterizavimas į

2 lentelė.
Klasterių, gautų panaudojant projektavimą, vidurkiai

	N	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Bendras	2027	9,51	6,12	1,31	1,67	2,52	16,76	3,27	1,05	0,21	2,11
1 klasteris	982	8,71	5,02	1,17	1,60	0,86	16,74	2,34	0,82	0,17	0,26
2 klasteris	584	8,74	6,18	1,21	1,47	3,05	13,33	2,99	0,79	0,23	2,05
3 klasteris	206	7,12	5,81	1,04	1,14	5,27	10,07	2,53	0,56	0,25	5,94
4 klasteris	252	16,24	10,44	2,30	2,83	5,47	29,98	7,86	2,65	0,30	6,22

3 lentelė.
Klasterių, gautų panaudojant projektavimą, vidurkiai

	N	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Bendras	2027	9,51	6,12	1,31	1,67	2,52	16,76	3,27	1,05	0,21	2,11
1 klasteris	978	10,26	7,20	1,43	1,76	4,03	17,01	4,18	1,23	0,25	3,88
2 klasteris	1032	8,61	5,02	1,15	1,56	0,96	15,85	2,24	0,76	0,17	0,32

4 lentelė.
Klasterių, gautų nenaudojant projektavimo, ir decilių palyginimas.

	3 klasteriai			4 klasteriai			5 klasteriai		
	χ^2	df	p	χ^2	df	p	χ^2	df	p
Pearson χ^2	397.14	18	.00	530.59	27	.00	584.25	36	.00
M-L χ^2	389.02	18	.00	553.77	27	.00	595.25	36	.00
Spearman R	0.391	t=18.94	.00	0.472	t=23.85	.00	0.259	t=12.05	.00

5 lentelė.
Klasterių, gautų dvimatės projekcijos atveju, ir decilių palyginimas.

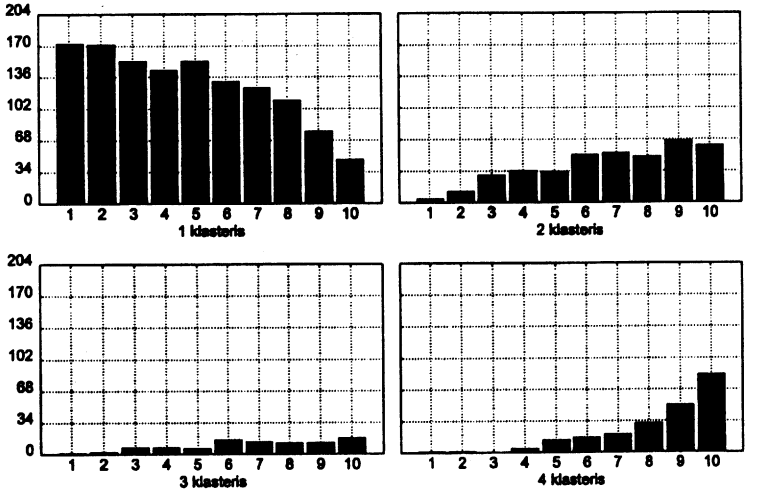
	2 klasteriai			3 klasteriai			4 klasteriai		
	χ^2	df	p	χ^2	df	p	χ^2	df	p
Pearson χ^2	432.67	9	.00	559.87	18	.00	634.09	27	.00
M-L χ^2	475.72	9	.00	588.31	18	.00	655.75	27	.00
Spearman R	-0.459	t=-23.18	.00	-0.031	t=1.40	.16	0.065	t=2.93	.00

mažesnį klasterių skaičių – į du ir į tris. Ekonominių ryšių požiūriu turiningiausi rezultatai gauti, kai namų ūkiai buvo klasterizuoti į du klasterius, panaudojant *PP* metodą (žr. 5 lentelę). Pastaruoju atveju skirtinguose klasteriuose namų ūkių maisto produktų vartojimas labai aiškiai skiriasi (3 lentelė). Į pirmą klasterį patekę 48 % namų ūkiai vartojo visų produktų daugiau nei vidutiniškai. Papildomai galima išskirti vaisių (P5), cukraus ir pan. (P7) bei bealkoholinių gėrimų (P10) vartojimą. Į antrą klasterį pateko apie 51 % namų ūkių. Jų vidutinis vartojimas buvo mažesnis, o ypač vaisių (P5), cukraus ir pan. (P7) bei bealkoholinių gėrimų (P10).

Toliau galima patyrinėti ryšį tarp namų ūkių klasterių numerių ir išlaidų decilių. Švedų mokslininkų nuomone, pagal skirtingų prekių ir paslaugų suvartojimo lygį galima įvertinti skirtumus tarp namų ūkių. Todėl buvo tikrinta hipotezė: „Ar yra ryšys tarp namų ūkių maisto vartojimo rodiklių (klasterių numerių) ir išlaidų decilių?". Kartu buvo lyginta ar vienodai pasiskirstę skirtingų decilių namų ūkiai į minėtus klasterius. Taip pat vertintas Spearmen'o ranginės koreliacijos koeficientas tarp namų ūkių klasterių numerių ir atskirų namų ūkius charakterizuojančių rodiklių. Žemiau pateikiami tik tai klasterizacijos rezultatų palyginimai su deciliniais namų ūkių pasiskirstymais.

Tikrinam hipotezę, kad namų ūkių pajamos ir pasirinktų maisto prekių vartojimas yra nepriklausomi atsitiktiniai dydžiai (4 lentelė). Tyrimui panaudoti Pearson'o ir maksimalaus tikėtinumo χ^2 kriterijai. Lentelėje matome tyrimų rezultatus. p raide pažymėtas ribinis reikšmingumo lygmuo, su kuriuo hipotezė priimama. Matome, kad visi naudoti kriterijai patvirtina išvadą apie stiprią priklausomybę tarp pajamų lygio ir maisto produktų vartojimo.

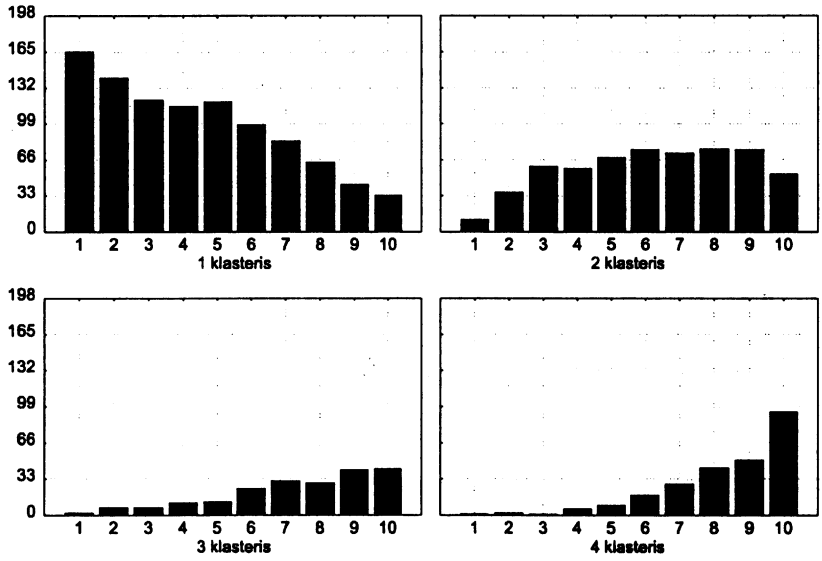
Iš 1 grafiko matyti, kad į pirmą klasterį pateko daugmaž skurdesni namų ūkiai, ketvirtą – turtingesni. Apie trečiam klasteryje esančius namų ūkius nieko konkretaus negalima pasakyti. Jei panagrinėtume, kaip pasiskirstęs maisto produktų vartojimas pagal gyvenamą vietą, tai galima pastebėti, kad pirmame klasteryje miesto žmonės pasiskirstę tolygiai, o kaimo žmonės sudaro daugumą žemesnėse decilėse. Antrame klasteryje kaimo



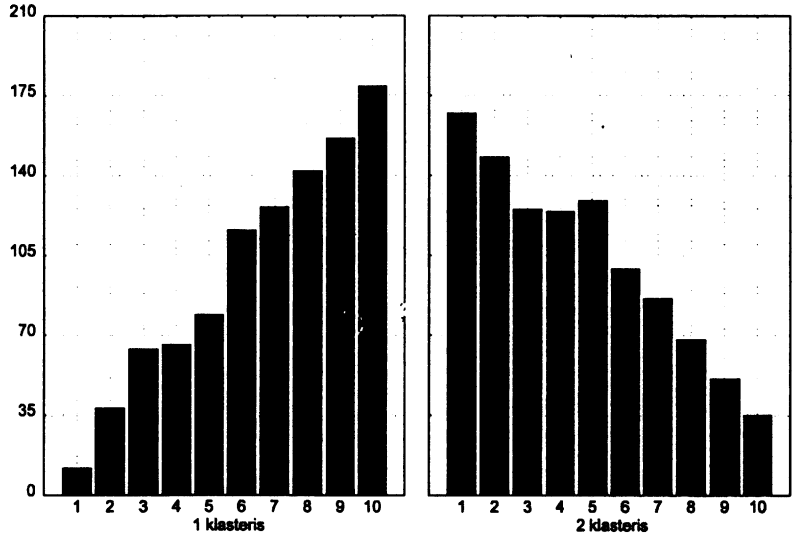
1 grafikas. Dešimtmačio dydžio 4 klasterių ir decilių pasiskirstymas.

žmonės pasiskirstę tolygiai, o miestiečių daugiau aukštesnėse decilėse. Be to, į šį klasterį pateko mažiausiai vartojantys maisto produktų namų ūkiai. Tačiau įdomu tai, kad vaisių ir bealkoholinių gėrimų vartojimas žymiai didesnis už vidurkį. Į ketvirtą klasterį pateko turtingiausi namų ūkiai (tiek kaimiečiai, tiek miestiečiai).

Suprojektavus pradinius duomenis į dvimatę erdvę išlieka visiška analogija rezultatams, gautiems be projektavimo. Tai matosi tiek grafiškai (2 grafikas), tiek panagrinėjus vartojimą pagal namų ūkių gyvenamą vietą.



2 grafikas. Dvimačio dydžio 4 klasterių ir decilių pasiskirstymas.



3 grafikas. Dvimačio dydžio 2 klasterių ir decilių pasiskirstymas.

Gretinant dvimačių duomenų klasterizacijos rezultatus su namų ūkių deciliniu pasiskirstymu matosi taip pat stipri priklausomybė tarp pajamų ir maisto produktų vartojimo lygių (5 lentelė).

Kaip matome iš 3 grafiko į pirmą klasterį pateko daugiausia turtingų namų ūkių, tuo tarpu antrame vyrauja skurdesni. Panagrinėjus maisto produktų vartojimą pagal gyvenamą vietą, pastebėta, kad pirmame klasteryje kaimiečiai pasiskirstę tolygiai, o miestiečiai pateko į aukštesnes deciles. Tuo tarpu antrame klasteryje miestiečiai pasiskirstę tolygiai, o daugumą žemų decilių sudaro kaimo žmonės. Didinant klasterių skaičių mažėja koreliacijos koeficientas ir atsiranda tarpinės grupės, apie kurias nieko negalima pasakyti be papildomo nagrinėjimo.

Literatūra

[1] S.A. Aivazian, V.M. Buchštaber, I.S. Jeniukov and L.D. Mešalkin, Prikladnaja statistika. Klasificacija i snižėnije razmernosti, *Finansy i Statistika*, M. (1989) (rusų k.).

[2] J.H. Friedman, Exploratory projection pursuit, *J. Amer. Statist. Assoc.*, **82**, 249–266 (1987).

[3] Household budget surveys, *Eurostat*, **3–4** (1997).

[4] G. Jakimauskas and R. Krikštolaitis, Influence of projection pursuit on classification errors: computer simulation results, *Informatica*, **11**(2), 115–124 (2000).

[5] G.J. McLachlan and K.E. Basford, *Mixture Models. Inference and Applications to Clustering*, Marcel Dekker, N.Y. (1988).

[6] *Namų ūkių pajamos ir išlaidos 1999 metais*, Statistikos departamentas prie Lietuvos Respublikos Vyriausybės, Vilnius (2000).

[7] R. Rudzkis and M. Radavičius, Statistical estimation of a mixture of Gaussian distributions, *Acta Applicandae Mathematicae*, **38**, 37–54 (1995).

[8] R. Rudzkis and M. Radavičius, Cėlėnapravlienoje projecirovanije v modeliach smesi Gausovskich rasprdėlėnij, sochraniajuščėjė informaciju o klastiernoj struktūre, *Liet. Matem. Rink.*, **37**(4), 550–563 (1997) (Rusų k.).

[9] D. Šimoliūnas, Klasterių skaičiaus nustatymas panaudojant neparametrines savybes, *Magistro tezės*, Matematikos ir statistikos katedra, Vytauto Didžiojo Universitetas, Kaunas (1998).

[10] D.M. Titterington, A.F.M. Smith and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*, Wiley, N.Y. (1985).

Cluster analysis of household income and expenditure

R. Krikštolaitis, G. Binkauskienė

Household income and expenditure under cluster analysis is analyzed. Calculation are made usind data of Lithuanian Department of Statistics, which were collected questioning people at second quarter of 1999 year.