LR Seimo narių elgsenos tyrimas, naudojant klasterinę analizę ir daugiamačių skalių metodą

Vytautas Mickevičius
Vytauto Didžiojo universitetas,
Informatikos fakultetas
Kaunas, Lietuva
El. paštas: vytautas.mickevicius@fc.vdu.lt

Tomas Krilavičius
Vytauto Didžiojo universitetas
Baltijos pažangių
technologijų institutas
Kaunas ir Vilnius, Lietuva
El. paštas: t.krilavicius@bpti.lt

Vaidas Morkevičius
Kauno technologijos universitetas
Politikos ir viešojo
administravimo institutas
Kaunas, Lietuva
El. paštas: vaidas.morkevicius@ktu.lt

***Santrauka*—Racionalūs rinkiminės elgsenos modeliai akcentuoja pakankamos informacijos poreikį rinkėjams priimant sprendimus, už ką balsuoti. Stebėti net pavienių politikų elgseną nėra paprasta, o bandyti fiksuoti ir suprasti viso parlamento narių veiklą yra dar sudėtingesnis uždavinys, nes dideliam kiekiui informacijos apdoroti būtina taikyti statistinius metodus. Šiame darbe siekiama pasiūlyti tinkamas metodikas LR Seimo balsavimų stebėsenai, leidžiančias aiškiau identifikuoti parlamentarų balsavimo tendencijas. Statistiškai analizuojami LR Seimo balsavimai 2008-2012 metų kadencijos pabaigoje (priešrinkiminiu laikotarpiu). Naudojamos klasterizavimo procedūros, o gauti rezultatai daugiamačių skalių metodo pagalba vaizdžiai pateikiami grafiškai. Lyginami skirtingi balsavimų kodavimo ir klasterizavimo metodai.**

Raktiniai žodžiai – klasterizavimas, daugiamatės skalės, politinių duomenų analizė, duomenų gavyba

I. ĮVADAS

Vienas pagrindinių demokratinės valstybės požymių – jos piliečių laisvai renkama valdžia [1]. Lietuvoje vyksta Prezidento, Seimo, savivaldybių bei Europarlamento rinkimai. Svarbiausiais iš jų laikomi Prezidento ir parlamento rinkimai. Tiek Prezidento, tiek Seimo veikla kadencijos metu būna atidžiai sekama. Tai natūralu – šiuos valdžios organus daugumos balsų principu suformuoja patys Lietuvos piliečiai, todėl iš išrinktųjų pagrįstai tikimasi, kad jie tinkamai atstovaus rinkėjų interesus. Vienas iš būdų įvertinti, ar rinkėjų lūkesčiai yra išpildomi – stebėti ir analizuoti valdžios atstovų veiksmus.

Prezidento veiklą stebėti galima tiesiogiai nagrinėjant jo padarytus sprendimus ar išreiškiamą nuomonę tam tikrais klausimais. Tačiau Seimo narių veiklos analizė yra gerokai sudėtingesnė – mat stebime jau nebe vieno, o 141 žmogaus nuomones, sprendimus ir veiksmus. Todėl labai svarbu pasirinkti tinkamus tyrimo metodus Seimo narių veiklos tendencijoms nustatyti. Šiame darbe pagrindinis dėmesys yra skiriamas Lietuvos parlamentarų balsavimams, siekiant nustatyti balsavimo vieningumą frakcijų viduje, o taip pat bendras frakcijų pozicijas viena kitos atžvilgiu.

Straipsnyje daroma prielaida, kad vieninga (disciplinuota) frakcija yra įtakingesnė už tokio pat dydžio nedrausmingą frakciją, todėl jos pozicija labiau įtakoja balsavimų baigtį. Seimo frakcijos yra formuojamos išrinktų partijų atstovų pagrindu. Taip pat jos gali jungtis į koalicinius darinius, kuriose skirtingos frakcijos susitaria veikti išvien. Paprastai, valdančiąją daugumą sudarančios frakcijos yra vadinamos pozicija, o joms nepriklausančios (nepritariančios vyriausybės programai) frakcijos – opozicija. Dažniausiai pozicijos ir opozicijos nuomonės įvairiais klausimais skiriasi, ir tai atsispindi balsavimuose [3], [6].

Kita vertus, kartais girdimi svarstymai, kad nepriklausomai nuo atstovaujamų frakcijų kai kurie (ar net dauguma) parlamentarų siekdami asmeninės naudos (ar atstovaudami tam tikroms interesų grupėms) balsuoja skirtingai nei jų frakcijos. Nors ir mažai tikėtina, kad tokia prielaida pasirodytų esanti teisinga, bet jos patikrinimas yra įdomus ir gana svarbus (politiškai) uždavinys.

Sekiant pasiūlyti tinkamą metodiką analizuoti LR Seimo narių balsavimų tendencijas šiame darbe sprendžiami tokie uždaviniai:

1) Tikrinamas LR Seimo narių balsavimo lojalumas savo frakcijoms, o taip pat – pozicijai arba opozicijai;
2) Tikrinamas atitikimas tarp LR Seimo narių priklausymo frakcijoms ir statistiniais metodais pagal jų balsavimus išskirtoms grupėms;
3) Ieškoma LR Seimo narių reguliariai balsuojančių kitaip nei nuosava frakcija ir panašiai kaip kitoms frakcijoms priklausantys LR Seimo nariai.

Parlamento balsavimų analizė pasaulyje yra gerai žinomas dalykas [2], [3], [4]. Ne išimtis ir Lietuva. Seimo veikla yra tiriama įvairiais metodais, tiek iš politinės [5], tiek iš statistinės pusės. Įdomių rezultatų gaunama politologams ir statistikams dirbant drauge [6]. Taikomi įvairūs metodai, tokie kaip homogeniškumo analizė [7], socialiniai tinklai [8], daugiamatės skalės [9].

Klasterizavimas taip pat nėra naujiena Lietuvos parlamento balsavimų analizėje [9]. Pagrindiniai šiuose darbuose išskirti iššūkiai: tinkamai parinkti duomenų kodavimą, panašumo matus, analizės bei rezultatų vizualizavimo metodus.

II. DUOMENYS

A. LR Seimas nagrinėjamu laikotarpiu

Tyrime naudojami LR Seimo balsavimų duomenys paimti iš projekto *atviras-seimas.info* duomenų bazės [10].

Nagrinėjamas laiko periodas yra 2008-2012 m. LR Seimo kadencijos pabaiga, sutampanti su priešrinkiminiu laikotarpiu. Jis truko nuo 2012-03-10 iki 2012-11-16, bei apėmė LR Seimo 8 eilinę, 9 neeilinę ir 9 eilinę sesijas. Šiame laikotarpyje Seime egzistavo 8 frakcijos (1 lentelė).

1 LENTELĖ. LRS FRAKCIJOS PRIEŠRINKIMINIU LAIKOTARPIU

Frakcijos santrumpa	Frakcijos pilnas pavadinimas	Atstovaujama kryptis
DPF	Darbo partijos frakcija	Opozicija
FTT	Frakcija „Tvarka ir teisingumas“	Opozicija
KPF	Krikščionių partijos frakcija	Opozicija
LCSF	Liberalų ir centro sąjungos frakcija	Pozicija
LSDPF	Lietuvos socialdemokratų partijos frakcija	Opozicija
LSF	Liberalų sąjūdžio frakcija	Pozicija
TS-LKDF	Tėvynės sąjungos – Lietuvos krikščionių demokratų frakcija	Pozicija
Kiti	Seimo nariai, kurie nagrinėjamu laiko periodu priklausė kelioms frakcijoms („perbėgėliai“) arba Mišriai Seimo narių grupei	Kiti

Nagrinėjamame laiko periode įvyko 1489 balsavimai. Apie 2 Seimo narių didžiąją daugumą balsavimų trūksta informacijos, todėl nagrinėjami tik likusių 139 parlamentarų balsavimai.

B. Kodavimas

Kiekvienas LR Seimo nario dalyvavimas balsavime gali turėti 6 baigmes:

- **Už**: balsuoja už įstatymo projektą ar pasiūlymą;
- **Prieš**: balsuoja prieš įstatymo projektą ar pasiūlymą;
- **Susilaikė**: balsavimo metu susilaiko;
- **Nebalsavo**: užsiregistruoja balsavimui, bet nebalsuoja;
- **Neatvyko**: neatvyksta į posėdį, kurio metu balsuojama;
- **Nėra informacijos**: apie balsavimo baigmę nėra informacijos (nagrinėjamuose duomenyse šios baigmės nėra).

Kad galėtume analizuoti balsavimus statistiniais metodais, jų baigmes turime sukoduoti skaitinėmis reikšmėmis. Tai galima padaryti daugybe skirtingų būdų, o šiame darbe naudojami 2 skirtingi kodavimo būdai (2 lentelė).

2 LENTELĖ. LRS BALSAVIMŲ KODAVIMAI

	Standartinis	Alternatyvus
Už	2	1
Neatvyko	1	0
Nebalsavo	0	–1
Susilaikė	–1	–1
Ne	–2	–1
Nėra info	–10	0

Standartinis kodavimas skirtinga reikšme apibrėžia kiekvieną skirtingą balsavimo baigtį, todėl šiuo atveju yra skirtumas, kokiu būtent būdu parlamentaras nepritarė įstatymo projektui ar pasiūlymui (susilaikė, nebalsavo, ar balsavo prieš). Šio kodavimo tikslas yra maksimaliai sugretinti Seimo nario poziciją su jo frakcijos pozicija (tarkime, jei frakcijos dauguma linkusi susilaikyti, o kažkuris jos narys balsuoja prieš, jis yra šiek tiek kitos pozicijos, nei "turėtų"). Kodavimo reikšmės buvo pasiūlytos straipsnyje [6]. Reikšmė, atitinkanti informacijos nebuvimą, naudojama gerokai besiskirianti nuo kitų, kad parlamentarai, apie kurių balsavimus trūksta per daug informacijos (išskirtys), būtų geriau matomi bendrame vaizde.

Alternatyvaus kodavimo tikslas – pateikti supaprastintą bendrų tendencijų vaizdą. Kadangi daugumoje balsavimų galioja principas *ne už = prieš* (įstatymas ar pasiūlymas priimamas skaičiuojant balsų *už* skaičių), tai balsavimo baigtys *prieš*, *susilaikė*, *nebalsavo* yra laikomos lygiavertėmis, kuomet parlamentaras išreiškia neigiamą nuomonę balsavimo klausimu. Nedalyvauti posėdyje Seimo narys gali dėl įvairių priežasčių, todėl tai laikoma neutralia balsavimo baigtimi (kaip ir informacijos apie balsavimą nebuvimas).

III. ĮRANKIAI

Tyrimas (tiek klasterizavimas, tiek duomenų vizualizavimas) buvo atliktas naudojant matematinės statistikos paketą R [11]. Jis plačiai naudojamas įvairiems statistiniams tyrimams daugelyje sričių [12]. Tai yra nemokama programinė įranga, kurios vartotojai turi galimybę kurti programos papildymus (angl. *packages*) [13].

IV. METODAI

A. Klasterizavimas

Klasterizavimas, arba klasterinė analizė – tai objektų suskirstymas į grupes pagal panašumą [14]. Savo ruožtu klasteris – tai panašių objektų grupė. Klasterizavimo tikslas yra suskirstyti objektus taip, kad skirtumai klasterių viduje būtų kuo mažesni, o tarp klasterių – kuo didesni.

B. Panašumo matai

Objektų panašumui nustatyti yra naudojami įvairūs panašumo matai. Dažniausiai sutinkami panašumo matai yra metriniai, dar vadinami metrikomis [14].

Populiariausios metrikos yra šios:

- Minkovskio:

$$\left(\sum_{i=1}^{m}|x_i - y_i|^l\right)^{1/l}, \quad l > 0. \tag{1}$$

Atskiri Minkovskio atstumų atvejai:

- Manheteno (kai $l = 1$):

$$d(X,Y) = \sum_{i=1}^{m}|x_i - y_i|; \tag{2}$$

- Euklido (kai $l = 2$):

$$d(X,Y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}; \tag{3}$$

- Čebyševo:

$$d(X,Y) = \max_i |x_i - y_i|. \quad (4)$$

C. K-means metodas

Vienas iš populiariausių klasterizavimo metodų yra *k*-means [15]. Jis puikiai tinka dideliems duomenų masyvams. Metode naudojamas toks algoritmas:

1) Objektai suskirstomi į *k* pradinių klasterių;
2) Paeiliui apskaičiuojamas kiekvieno objekto atstumas iki klasterių centrų;
3) Objektas skiriamas į artimiausią klasterį;
4) Klasterių centrai perskaičiuojami;
5) 2-4 žingsniai kartojami tol, kol perskirstymų daugiau nėra.

Naudojant *k*-means metodą gali iškilti nepatogumas – klasterių skaičių reikia nustatyti iš anksto. Tam padaryti egzistuoja daug įvairaus sudėtingumo metodų [15], tačiau šiame straipsnyje klasterių kiekis pasirinktas lygus esamų frakcijų kiekiui – 8.

D. Klasterizavimo kokybės įvertinimas

Darbe naudojami 2 vidiniai ir 2 išoriniai kokybės vertinimo kriterijai.

Dunn indeksas. Šis kriterijus yra vidinis (angl. *internal*), t.y. nusako pačio klasterizavimo proceso tikslumą, neatsižvelgiant į a priori objektams priskirtas klases [17]. Jo reikšmė gaunama pagal formulę:

$$DUNN = \min_{1\le i\le c}\left\{\min\left\{\frac{d(c_i,c_j)}{\max_{1\le i\le c}(d(X_k))}\right\}\right\}, \quad (5)$$

čia $d(c_i,c_j)$ žymi atstumą tarp klasterių X_i ir X_j, $d(X_k)$ apibrėžia atstumus tarp objektų klasterio X_k viduje, o c yra klasterių kiekis.

Paprasčiau kalbant, *Dunn* indeksas yra mažiausio atstumo tarp klasterių centrų, ir didžiausio atstumo tarp dviejų objektų viename klasteryje, santykis. Šiuo atveju skaičiuojamas atstumas yra trumpiausias kelias tarp dviejų taškų – Euklido atstumas.

Davies-Bouldin (DB) indeksas. Šis indeksas taip pat yra vidinis kokybės vertinimo kriterijus [17]. Jo vertė randama pagal formulę:

$$DB = \frac{1}{c}\sum_{i=1}^{c}\max_{i\ne j}\left\{\frac{d(X_i)+d(X_j)}{d(c_i,c_j)}\right\}, \quad (6)$$

čia c yra klasterių kiekis, $d(X_i)$ ir $d(X_j)$ yra atstumai nuo objektų iki klasterių X_i ir X_j centrų, o $d(c_i,c_j)$ – atstumai tarp šių klasterių centrų.

Davies-Bouldin indeksas yra panašus matas į *Dunn* indeksą, tačiau pastarajame naudojami dydžiai randami bendrai visiems klasteriams, o *DB* – dydžiai apskaičiuojami kiekvienam klasteriui atskirai ir dalinama iš klasterių kiekio (gaunamas vidurkis).

Rand indeksas. Šis kriterijus išreiškia statistiškai suformuotų klasterių ir pradinių klasių (šiuo atveju – frakcijų) panašumą [16]. Tai išorinis (angl. *external*) kokybės vertinimo kriterijus. Indeksas apskaičiuojamas pagal formulę:

$$RI = \frac{TP+TN}{TP+TN+FP+FN}, \quad (7)$$

čia dydžiai *TP*, *TN*, *FP* ir *FN* apibrėžti 3 lentelėje.

3 LENTELĖ. *TP*, *TN*, *FP* IR *FN* REIKŠMĖS

	Priklauso klasei	Nepriklauso klasei
Priskirtas klasteriui	Teisingai priskirtas (*TP* – true positive)	Neteisingai priskirtas (*FP* – false positive)
Nepriskirtas klasteriui	Neteisingai nepriskirtas (*FN* – false negative)	Teisingai nepriskirtas (*TN* – true negative)

Rand indeksas – tai teisingų klasterizavimo algoritmo padarytų sprendimų dalis.

Purity (grynumo) indeksas. Tai taip pat išorinis kokybės vertinimo kriterijus. *Purity* yra paprastas ir populiarus kokybės vertinimo matas [16]. Jo reikšmė randama pagal formulę:

$$PURITY(Q,Z) = \frac{1}{n}\sum_i \max_j(q_i \cap z_j), \quad (8)$$

čia n yra objektų skaičius, $Q = (q_1, q_2, \dots, q_i)$ yra klasterių aibė (i – klasterių kiekis), $Z = (z_1, z_2, \dots, z_j)$ yra pradinių klasių aibė (j – klasių kiekis).

Paprasčiau tariant, *Purity* mato įvertis gaunamas susumavus populiariausios klasės kiekviename klasteryje objektų skaičių, ir padalinus šią sumą iš viso objektų skaičiaus.

Matuojant klasterizavimo kokybę pageidaujami rezultatai: kuo mažesnė *Davies-Bouldin* indekso reikšmė, ir kuo didesnė kitų 3 kriterijų reikšmė.

E. Daugiamačių duomenų vizualizavimas

Daugiamačiai duomenys vizualizuojami tiesioginiais ir projekciniais metodais. Vizualizavimo tiesioginiais metodais privalumas yra tas, kad neprarandama pradinė informacija, tačiau esant dideliam kintamųjų kiekiui, šių metodų rezultatus yra labai sunku, ar net neįmanoma interpretuoti [18]. Projekcinių metodų tikslas yra didelį kintamųjų (matavimų) kiekį sumažinti iki 2 arba 3 dimensijų (matavimų), kurių reprezentacijas galima būtų pateikti grafine forma (daugelio kintamųjų projekcija į 2 ar 3 dimensijas). Žinoma, atliekant tokius pertvarkymus išsaugoma tik dalis turimos pradinės informacijos.

Vizualizuojant daugiamačius duomenis projekciniais metodais yra svarbu pasirinkti tinkamą dimensijų kiekį. Keturmačiame grafike tilptų daugiau pradinės informacijos, tačiau trimatį grafiką daug lengviau suprasti ir interpretuoti. Čia galioja taisyklė, kad didelės dalies pradinės informacijos negalima aukoti vien tik vardan sprendinio paprastumo, ir atvirkščiai – aiškumo neverta aukoti dėl per mažos dalies išsaugomos pradinės informacijos.

F. Daugiamatės skalės (MDS)

Tarkime, turime I duomenų (stebinių) rinkinį. Atstumą tarp i-ojo ir j-ojo stebinio pažymėkime $\delta_{i,j}$. Visi šie atstumai kartu sudaro kvadratinę atstumų matricą:

$$\Delta = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \dots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \dots & \delta_{2,I} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{I,1} & \delta_{I,2} & \dots & \delta_{I,I} \end{pmatrix} \quad (9)$$

Daugiamačių skalių metodo tikslas [20] yra turint matricą Δ rasti I vektorius $x_1, \dots, x_I \in \mathbb{R}^N$ tokius, kad būtų tenkinama sąlyga:

$$\|x_i - x_j\| \approx \delta_{i,j}, \quad (10)$$

visiems $i, j \in I$. Čia $\|\cdot\|$ yra vektoriaus norma. Dažniausiai ši norma būna Euklido atstumas (taip skaičiuojama ir šiame darbe).

Paaiškintos pradinės informacijos kiekis (angl. *variance explained*) naudojant MDS metodą randamas iš atstumų matricos tikrinių reikšmių vektoriaus [19]:

$$PI = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{n} \lambda_j}, \quad (11)$$

čia m yra pasirinktas matavimų skaičius, n yra pilnas matavimų skaičius, λ_i ir λ_i – atitinkamai i-oji ir j-oji tikrinės reikšmės. Paprasčiau tariant, $\sum_{i=1}^{m} \lambda_i$ yra m pirmųjų tikrinių reikšmių suma, o $\sum_{j=1}^{n} \lambda_j$ – visų tikrinių reikšmių suma.

V. EKSPERIMENTŲ EIGA

A. Duomenų paruošimas

Pirminiai duomenys buvo įrašyti į CSV tipo failus..

Pradinėje duomenų matricoje skirtingų balsavimų rezultatai surašyti eilutėse, parlamentarų ID – stulpeliuose. Ši matrica tyrimo pradžioje transponuota, Seimo narius paverčiant stebiniais (eilutėmis), o balsavimus – kintamaisiais (stulpeliais).

Iš viso paruošti 2 duomenų failai – dviems skirtingiems balsavimų kodavimams (standartiniam ir alternatyviam)..

B. Seimo narių paskirstymas į klasterius

Pasirinktas klasterių kiekis – 8. Tai optimalus variantas, kadangi 8 klasterių elementus galima palyginti su 8 frakcijų nariais.

Abiems nagrinėjamiems kodavimams taikytas klasterizavimas, naudojant 3 skirtingas metrikas, taip iš viso gauti 6 skirtingi klasterizavimo rezultatų variantai. Atrenkant geriausius variantus remtasi klasterizavimo kokybės įverčiais (4 lentelė).

Naudojant Euklido ir Čebyševo metrikas gautas identiškas suskirstymas į klasterius (tiek su standartiniu, tiek su alternatyviu kodavimu), todėl ir kokybės įverčių reikšmės identiškos.

Atsižvelgiant į išorinius kokybės vertinimo kriterijus (4 lentelė, *Rand*, *Purity*) ir darant prielaidą, jog klasteriai turi kuo tiksliau atitikti realų pasiskirstymą į frakcijas, galima teigti, jog standartinis kodavimas ir Manheteno metrika yra nežymiai geresni už alternatyvų kodavimą ir kitas metrikas.

4 LENTELĖ. KLASTERIZAVIMO KOKYBĖS ĮVERČIAI

Kodavimas	Metrika	Rand	Purity	Dunn	DB
Standartinis	Manheteno	0,845	0,683	0,543	3,047
	Euklido	0,823	0,676	0,469	2,956
	Čebyševo	0,823	0,676	0,469	2,956
Alternatyvus	Manheteno	0,845	0,662	0,627	3,364
	Euklido	0,788	0,583	0,614	3,27
	Čebyševo	0,788	0,583	0,614	3,27

Turint omenyje pačio objektų suskirstymo į klasterius kokybę reikia atsižvelgti į vidinius kokybės matus (4 lentelė, *Dunn*, *Davies-Bouldin*). Čia pastebimai geresnius rezultatus parodė alternatyvus kodavimas, tačiau geriausios metrikos vienareikšmiškai išskirti negalima – *Dunn* indeksas nežymią pirmenybę teikia Manheteno metrikai, o *DB* indeksas tokiu pat nežymiu santykiu sako, jog Euklido ar Čebyševo metrikos yra pranašesnės. Toliau darbe išsamiau nagrinėjama Manheteno metrika su alternatyviu kodavimu.

5 ir 6 lentelėse nurodyta, kiek kiekvienos frakcijos atstovų patenka į kiekvieną klasterį (paryškintas skaičius kiekviename klasteryje žymi, kuri frakcija tam klasteriui buvo priskirta).

5 LENTELĖ. MANHETENO METRIKA, STANDARTINIS KODAVIMAS

			Klasteriai							
			1	2	3	4	5	6	7	8
Frakcijos	Opozicija	DPF	–	–	6	–	2	1	1	–
		FTT	–	–	1	–	**11**	1	1	**3**
		KPF	–	–	–	1	–	6	–	–
		LSDPF	–	–	**15**	1	–	–	**6**	–
	Pozicija	LSF	9	–	1	1	–	–	–	–
		LCSF	2	1	–	**8**	–	–	–	–
		TS-LKDF	**16**	**26**	–	1	–	–	–	–
		Kiti	–	1	3	2	–	**10**	2	–

6 LENTELĖ. MANHETENO METRIKA, ALTERNATYVUS KODAVIMAS

			Klasteriai							
			1	2	3	4	5	6	7	8
Frakcijos	Opozicija	DPF	–	–	6	–	2	1	1	–
		FTT	1	2	–	–	**9**	–	2	**3**
		KPF	–	1	–	–	–	–	6	–
		LSDPF	–	**15**	–	1	–	**5**	1	–
	Pozicija	LSF	–	1	8	1	–	–	1	–
		LCSF	1	–	1	**9**	–	–	–	–
		TS-LKDF	**30**	–	**12**	1	–	–	–	–
		Kiti	2	1	–	2	1	2	**9**	1

42 iš 43 pozicijos branduolį sudarančios TS-LKDF atstovų buvo paskirti į 2 klasterius (26 viename ir 16 kitame). Analogiškai, opozicijos kertinė frakcija – LSDPF – kituose 2 klasteriuose turi 21 iš 22 savo narių (15 ir 6). Tai liudija, jog šių frakcijų pozicijos gana aiškios, tačiau egzistuoja vidinis susiskaldymas frakcijų viduje – priešingu atveju didžioji dauguma frakcijos narių būtų sutalpinti į vieną klasterį.

Pasinaudojant *Purity* kokybės įverčiu, galima teigti, jog pozicijos frakcijos tarpusavyje yra vieningesnės (7 lentelė). Tai sąlygoja opozicijos frakcijų aiškių pozicijų nebuvimas – visos jos tarsi sukrenta į "bendrą katilą", tuo tarpu pozicijoje TS-LKDF vidinį pasiskirstymą kompensuoja aiškesnis LSF ir LCSF vieningumas.

7 LENTELĖ. MANHETENO METRIKA

	Narių kiekis	Purity matas	
		Standartinis	Alternatyvus
Pozicija	65	0,769	0,785
Opozicija	56	0,625	0,571
Kiti	18	0,556	0,5

C. Rezultatų vizualizavimas ir analizė

Prieš išgaunant grafinius rezultatus, reikia pasirinkti matavimų, kuriuose tuos rezultatus vaizduosime, kiekį. Dviejų ir trijų matavimų sprendiniais paaiškinamos duomenų sklaidos santykiniai kiekiai pateikti 8 lentelėje.

8 LENTELĖ. PAAIŠKINTOS DUOMENŲ SKLAIDOS KIEKIAI

Kodavimas	Matavimų kiekis					
	1	**2**	**3**	4	5	...
Standartinis	17,8%	**23,4%**	**26,4%**	29,1%	31,2%	...
Alternatyvus	16,3%	**22,2%**	**24,9%**	27,5%	29,4%	...

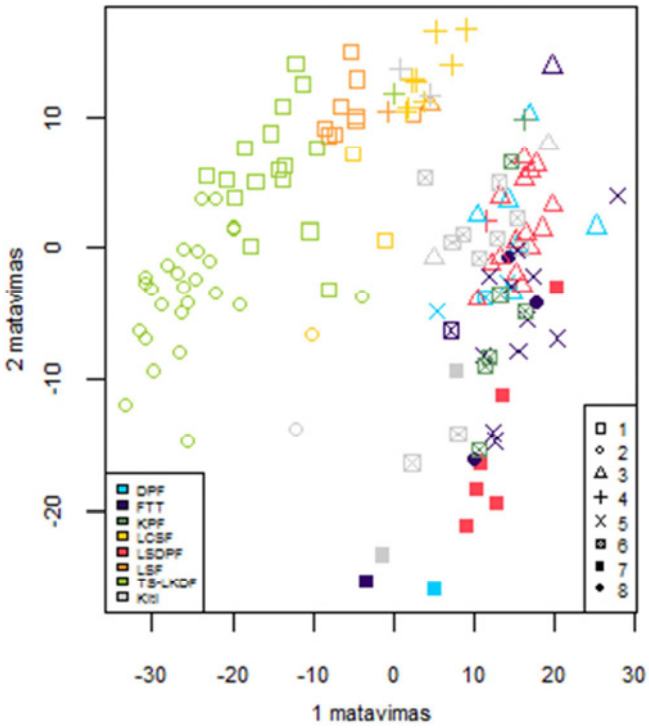
Standartinis kodavimas pasirodė šiek tiek tinkamesnis už alternatyvų išsaugomos pradinės informacijos atžvilgiu.

Trimačio sprendinio paaiškinamos duomenų sklaidos kiekis yra nedaug didesnis už dvimačio, todėl pasirinktas vaizduojamų dimensijų skaičius – 2.

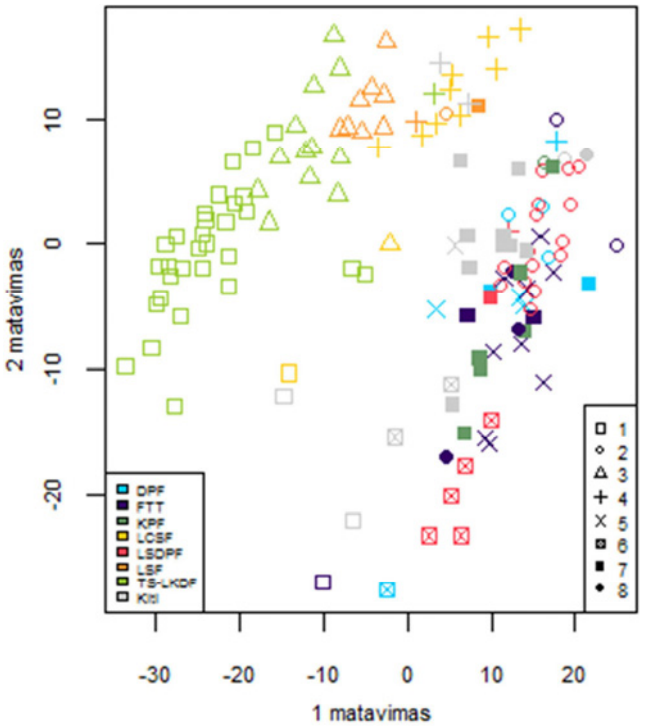
Dviejų geriausių klasterizavimų rezultatai atsispindi 1 ir 2 paveikslėliuose. Čia skirtingos spalvos reiškia skirtingas frakcijas, o skirtinga objektus žyminčių taškų forma – skirtingus klasterius.

Grafikuose matyti, jog pozicijos ir opozicijos Seimo narių balsavimas aiškiai skiriasi. Taip pat įdomu pastebėti tai, kad pozicijos ir opozicijos atstovų vieningumas, nors ir neginčytinas, bet pasireiškia skirtingai. Poziciją sudarančios frakcijos vieningesnės viduje ir išlaiko šiokį tokį atstumą viena nuo kitos, tik didžioji valdančioji frakcija – TS-LKDF – yra lyg perskirta į dvi dalis. Tuo tarpu dauguma opozicijos atstovų telkiasi apie grupės branduolio – LSDPF – daugumą (taigi skirtingų opozicijos frakcijų atstovai balsuoja gana panašiai), o vidinė frakcijų vienybė aiškiai mažesnė nei pozicijos frakcijų.

Alternatyvus balsavimų kodavimas, kaip ir buvo galima prognozuoti, šiek tiek labiau "išstumdo" parlamentarus link skirtingesnių pozicijų (tai iš dalies nulemta mažesnio galimų reikšmių kiekio šiame kodavime), tačiau sprendinio interpretacija iš esmės tokia pati kaip ir analizuojant standartinio kodavimo duomenis.



1 pav. Manheteno metrika, standartinis kodavimas



2 pav. Manheteno metrika, alternatyvus kodavimas

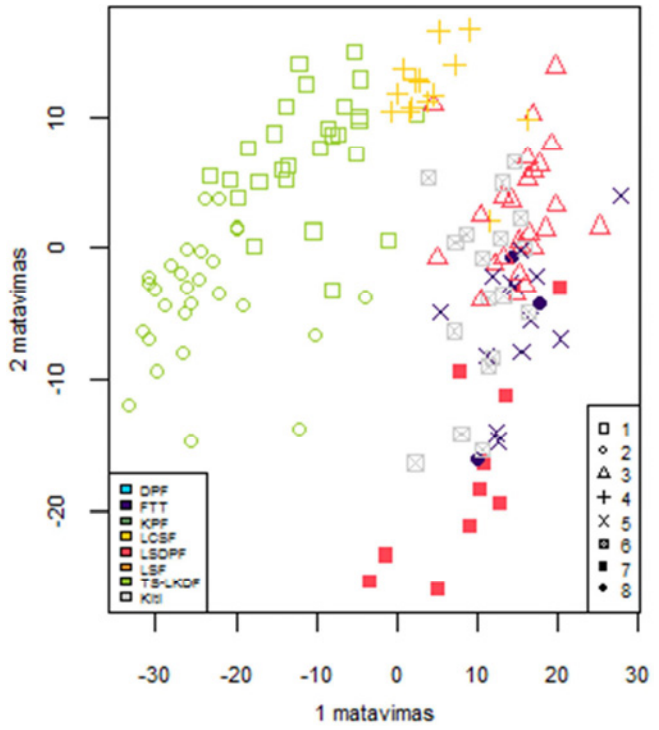
Pasinaudojant 5 lentele galima vizualizuoti prognozuojamą parlamentarų pasiskirstymą į frakcijas, atsižvelgiant į klasterizavimo procedūrų rezultatus. Tokiu atveju kiekvienam klasteriui būtų priskirta kažkuri frakcija, taigi, vertindami kokybę turime atsižvelgti į išorinius kriterijus. Pastarieji

geriausi buvo gauti naudojant Manheteno metriką ir standartinį kodavimą (9 lentelė).

LENTELĖ 9. PASISKIRSTYMO Į FRAKCIJAS PROGNOZAVIMAS (MANHETENO METRIKA, STANDARTINIS KODAVIMAS)

Klasterio nr.	Klasterio elementų skaičius	Klasteriui priskirta frakcija	Teisingai priskirta į klasterį	Realus narių kiekis frakcijoje
1	27	TS-LKDF	16	43
2	28	TS-LKDF	26	43
3	26	LSDPF	15	22
4	14	LCSF	8	11
5	13	FTT	11	17
6	18	Other	10	18
7	10	LSDPF	6	22
8	3	FTT	3	17

Grafiškai prognozavimas pateiktas 3 paveikslėlyje.



3 pav. Pasiskirstymo į frakcijas prognozavimas

Šiame grafike teisingai priskirtų frakcijoms Seimo narių yra 68,35% (*Purity* indeksas, 4 lentelė). Prognozuojant parlamentarų pasiskirstymą į frakcijas mažesnės frakcijos kartais yra nustelbiamos gausesnių ir jų nariai "išskirstomi" į didesnes frakcijas. Geriausias to pavyzdys yra klasteris nr. 1, kuriame yra susitelkę 9 iš 11 LSF narių, tačiau klasteris priskirtas TS-LKDF, kuri jame turi 16 narių.

VI. REZULTATAI IR IŠVADOS

1) **Klasterizavimas yra tinkama priemonė parlamento balsavimų analizei.** Parlamentarų suskirstymo į frakcijas pagal joms priskiriamus klasterius prognozė buvo atlikta 68% tikslumu. Tai, ypač atsižvelgiant į labai panašią opozicijos frakcijų padėtį viena kitos atžvilgiu, yra geras rezultatas.
2) **Daugiamatės skalės yra naudinga priemonė parlamento balsavimų rezultatų vizualizavimui.** Beveik nepriekaištingai atvaizduojama pozicijos ir opozicijos skirtis, taip pat gana aiškiai matosi frakcijų (ne)vieningumas.
3) Klasterizuojant *k*-means metodu tinkamiausia pasirodė esanti Manheteno metrika. Kokybiškesnis klasterizavimas gautas naudojant alternatyvų kodavimą, tačiau iškėlus prielaidą, jog kiekvienas išskirtas klasteris turėtų atitikti tam tikrą frakciją, labiau tinkamas pasirodė standartinis kodavimas.

Remiantis tyrimo rezultatais, buvo galima daryti ir politikos lauko analizei svarbias išvadas:

1) Frakcijų nariai balsuoja gana vieningai (ypač aiški pozicijos-opozicijos skirtis).
2) Pozicijos ir opozicijos frakcijų vieningumas pasireiškia skirtingai. Pozicijoje kiekviena iš 3 ją sudarančių frakcijų turi truputį skirtingą gana aiškiai pastebimą padėtį. Opozicijoje didžiausią įtaką turi LSDPF dauguma, apie kurią "susibūrę" likusių opozicijos frakcijų nariai (taigi skirtingų opozicijos frakcijų balsavimas panašus).
3) Didelių frakcijų vienybė mažesnė. Tai atsispindi TS-LKDF ir LSDPF frakcijų narių balsavimuose.
4) Nebuvo aptikta aiškių "grupių", kurių nariai sistemingai balsuotų kitaip nei "nuosava" frakcija.

Planuojami tyrimai:

1) **Skirtingi kodavimai ir panašumo matai.** Buvo ištirti 2 skirtingi kodavimai, bet tinkamiausio kodavimo paieška yra ilgas procesas, todėl reiktų patikrinti ir kitų logiškai galimų variantų atvejus. Taip pat ir su panašumo matais – šiame straipsnyje buvo aptarta tik nedidelė jų dalis.
2) **Skirtingas balsavimų svoris.** Tam tikri balsavimai Seime (pvz. kai kurios įstatymų pataisos) yra gerokai svarbesni nei likusieji. Todėl reiktų ištirti, ar suteikiant skirtingą svorį balsavimams tyrimų rezultatai pasikeistų (būtų tikslesni ir objektyvesni).
3) **Programinės įrangos tobulinimas.** Planuojama artimiausiu metu sukurti R paketą, kuris leistų tyrime naudojamus skaičiavimus ir rezultatų vizualizavimą atlikti automatiškai, įvedus kitokius (atnaujintus) duomenis. Tai gerokai paspartintų tolesnę parlamento veiklos analizę.

PADĖKA

Šis tyrimas yra iš dalies finansuojamas ESFA (VP1-3.1-ŠMM-10-V-02-025).

LITERATŪROS SĄRAŠAS

[1] Global Politician: http://www.globalpolitician.com/ default.asp?26729-democracy-voting-elections

[2] S. Hix, A. Noury, and G. Roland, "Dimensions of Politics in the European Parliament", American Journal of Political Science, vol. 50, no. 2, pp. 494-520, 2006.

[3] S. Hix, and A. Noury, "Government-Opposition or Left-Right? The Institutional Determinants of Voting in Fourteen Parliaments," Working paper, version 1, May 2008.

[4] K. T. Poole, Spatial Models of Parliamentary Voting. New York: Cambridge University Press, 2005.

[5] V. Morkevičius, Neideologinis Seimas? Statistinė svarbių 2004-2008 m. kadencijos Lietuvos Seimo balsavimų analizė. Partinės demokratijos pabaiga? Politinis atstovavimas ir ideologijos. Vilnius: Versus Aureus, pp. 53-87, 2009.

[6] T. Krilavičius, V. Morkevičius, "Mining Social Science Data: a Study of Voting of Members of the Seimas of Lithuania Using Multidimensional Scaling and Homogeneity Analysis," Intelectual Economics, vol. 5, no. 2, pp. 224-243, 2011.

[7] T. Krilavičius, V. Morkevičius, Lietuvos parlamentarų ideologinės pozicijos ir jų raida 2008-2012 m. kadencijos Seime: statistinė Seimo narių balsavimo analizė [Ideological positions of Lithuanian MPs and their dynamics in the term 2008-2012: Statistical Analysis of MPs Voting]. Annual conference of the Lithuanian Political Science Association, November 2010, Vilnius.

[8] R. Užupytė, V. Morkevičius, T. Krilavičius, Lietuvos Respublikos Seimo narių balsavimų tyrimas pasitelkiant socialinių tinklų analizę: tinklo konstravimo parametrų įtaka. Studentų mokslinė praktika, Konferencijos pranešimų santraukos, I dalis. Vilnius, 2012.

[9] T. Krilavičius, A. Žilinskas, "On Structural Analysis of Parlamentarian Voting Data," Informatica, vol. 19, no. 3, pp. 377-390, August 2008. URL: http://www.mii.lt/informatica/ pdf/INFO727.pdf.

[10] T. Krilavičius, P. Cimmperman, T. Žalandauskas, Duomenų užteks visiems [Plenty data for everyone]. Vilnius, 2010.

[11] R Core Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing," Vienna, Austria, 2012. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[12] R. Muenchen, " The Popularity of Data Analysis Software," URL: http://r4stats.com/articles/popularity/.

[13] R. I. Kabacoff, R in Action. Data analysis and graphics with R. Manning: Shelter Island, 2011.

[14] V. Čekanavičius, G. Murauskas, Statistika ir jos taikymai II. Vilnius: TEV leidykla, 2002.

[15] S. Basu, I. Davidson, K. L. Wagstaff. Constrained Clustering. Advances in Algorithms, Theory and Applications. USA: CRC Press, 2008.

[16] Stanford University [interactive]. http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html

[17] E. Rendon, I. Abundez, A. Arizmendi, E. Quiroz, " Internal versus External cluster validation indexes," International Journal of Computers and Communications, vol. 5, issue 1, 2011.

[18] Trinity University [interactive]. http://www.trinity.edu/rjensen/352wpvisual/000datavisualization.htm

[19] H. Abdi, "Metric Multidimensional Scaling (MDS): Analyzing Distance Matrices", [interactive]. URL: http://www.utd.edu/~herve/Abdi-MDS2007-pretty.pdf

[20] A. Buja, D. F. Swayne, M. L. Littman, N. Dean, H. Hofmann, L. Chen, "Data visualization with multidimensional scaling," Journal of Computational and Graphical Statistics, vol. 17, no. 2, pp. 444-472, 2008.