Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
62 tomas, 2021, 86–93
https://doi.org/10.15388/LMR.2021.25232

Vilnius University Press

Artimumo matų lyginamoji analizė Lietuvos populiacijos struktūros nustatymui

Alma Molytė[a], **Alina Urnikytė**[b]

[a] *Informacinių sistemų katedra, Fundamentinių mokslų fakultetas, Vilnius TECH*
Saulėtekio al. 11, LT-10223 Vilnius

[b] *Žmogaus ir medicininės genetikos katedra, Biomedicinos mokslų institutas, Medicinos fakultetas, Vilniaus universitetas*
Santariškių g. 2, LT-08661 Vilnius

El. paštas: alma.molyte@vilniustech.lt, alina.urnikyte@mf.vu.lt

Įteiktas 2021 birželio 23; publikuotas 2021 gruodžio 20

Santrauka. Darbe nagrinėjami Lietuvos populiacijos genetinės struktūros nustatymo daugiamačių skalių, pagrindinių koordinačių ir pagrindinių komponenčių metodai, kai artimumo matai yra Euklido, Gower, Bray-Curtis, Kulczynski, Jaccard ir Morisita. Analizuoti 424 lietuvių plataus masto vieno nukleotido polimorfizmo genetiniai duomenys. Atlikta artimumo matų lyginamoji analizė.

Raktiniai žodžiai: genetiniai duomenys; artimumo matas; populiacijos struktūra

AMS: 62H25, 62H30, 92D10

1 Įvadas

Lietuvos populiacijos genetinė struktūra yra vienas iš genetinių tyrimų objektų. Pirminiai populiacijos genetinės struktūros tyrimai buvo pagrįsti mitochondrinės DNR, Y chromosomos ir mikrosatelitų duomenimis. Šiuo metu metodų, skirtų populiacijos genetinių duomenų gavybai alternatyva yra naujos kartos genomo skenavimas, kuris išsprendžia daugybę ankstesnių metodikų apribojimų. Populiacijos genetinės struktūros tyrimai naudojami išaiškinti panašumus ir skirtumus tarp vienos grupės individų ar tarp skirtingų individų grupių bei veiksnius, kurie lemia tuos skirtumus. Genetiniai duomenys yra daugiamačiai, kurie gali būti analizuojami įvairiais statistikos metodais, tačiau kai duomenų kiekis yra didelis, dažnai jų nepakanka, todėl siekiant

gauti daugiau žinių iš analizuojamų duomenų, yra naudojami įvairūs duomenų tyrybos metodai: klasifikavimo, klasterizavimo, vizualizavimo ir kt. [5, 6]. Daugiamačių duomenų vizualizavimo, dar kitaip vadinamais matmenų mažinimo, metodais didelės dimensijos duomenys yra transformuojami į mažesnę matmenų erdvę taip, kad išliktų arba būtų atrastos „užslėptos“ analizuojamų duomenų savybės. Egzistuoja daugybė metodų, kuriuos galima naudoti matmenų mažinimui ir ypač n-mačių duomenų vizualizavimui: pagrindinių komponenčių analizė (PKA) [9], daugiamatės skalės (DS) [3], lokaliai tiesinis vaizdavimas (LLE) [1] ir kt. Šie metodai gali būti naudojami duomenims vizualizuoti į dvimatę ir trimatę erdvę ($d = 2$, $d = 3$). Pagrindinis šio darbo tikslas yra išanalizuoti ir ištirti artimumo matus, daugiamačių skalių, pagrindinių koordinačių, pagrindinių komponenčių metodus, kurie labiausiai atskleistų genetinių duomenų pasiskirstymą Lietuvos populiacijos struktūros nustatymui analizuojant plataus masto vieno nukleotido polimorfizmo genetinius duomenis.

Dabartinė Lietuvos populiacija yra sudaryta iš ankstesnių baltų genčių ir valstybių. Nuo neolito laikotarpio Lietuvos teritorijos gyventojų nepakeitė jokia kita etninė grupė. Tikėtina, kad dabartinės Lietuvos gyventojai išsaugojo savo senųjų protėvių genų fondą. Tiriant DNR sekos įvairovę bei pasirenkant tinkamą analizės metodą galime nustatyti didesnės skiriamosios gebos populiacijų genetinę struktūrą.

Šiame darbe taikomi daugiamačių skalių (MDS), pagrindinių koordinačių (PK) ir pagrindinių komponenčių (PKA) metodai Lietuvos populiacijos genetinės struktūros įvertinimui. Labai svarbu tinkamai parinkti artimumo matus tarp objektų. Kai atskirą objektą nusakančio parametrų skaitinių reikšmių rinkinio negalima gauti, tenka ekspertiškai, ar kokiu nors kitu būdu skaitiškai įvertinti artimumus tarp objektų porų, t. y. panašumus ar skirtingumus.

2 Medžiaga ir metodai. Mėginiai ir genotipai

Duomenų imtį sudaro 424 tarpusavyje nesusiję tiriamieji iš šešių bendros lietuvių populiacijos etnolingvistinių grupių: vakarų ($n = 79$), pietų ($n = 67$) ir rytų ($n = 78$) aukštaičių ir šiaurės ($n = 79$), vakarų ($n = 43$) bei pietų žemaičių ($n = 78$). Tiriamųjų asmenų DNR buvo išskirta iš kraujo leukocitų fenolio–chloroformo ekstrakcijos metodu. Lietuvos populiacijos genetiniai struktūrai įvertinti buvo naudojami .bed, .bim ir .fam failai sukurti PLINK v1.07 programa [1]. Bed faile saugojama genotipinė tirtų asmenų informacija, .bim faile yra visa informacija apie alelius, chromosomą, poziciją, .fam faile randasi genealoginė ir fenotipinė informacija. Vieno nukleotido polimorfizmo (VNP) duomenys buvo vizualizuoti DS ir PK metodais PAST4 programa. Populiacijos genetinei struktūrai nustatyti PKA metodu buvo naudojama EIGENSOFT 7.2.1 SmartPCA programa.

Šis tyrimas yra LITGEN projekto dalis, kurią patvirtino Vilniaus regiono tyrimų etikos komitetas Nr. 235. Iš visų tiriamųjų buvo gautas rašytinis sutikimas.

3 Daugiamačių duomenų vizualizavimo metodai

Šiame skyriuje pateikiami daugiamačių skalių (DS), pagrindinių koordinačių (PK) ir pagrindinių komponenčių (PKA) metodai naudojami Lietuvos populiacijos struktūros nustatymui.

Tarkime turime vieną konkretų analizuojamos aibės $X = \{X_1, X_2, \ldots, X_m\}$ objektą $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$, čia m yra analizuojamų objektų skaičius, n – X_i komponenčių skaičius ir i – objekto eilės numeris. Jeigu parametrų reikšmės yra skaitinės, tai $X_1, X_2, \ldots, X_m$ yra n-mačiai vektoriai. Dažnai jie interpretuojami kaip taškai n-matėje erdvėje R^n, čia n – erdvės dimensijos skaičius. Reikia rasti vektoriaus $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ transformaciją $Y_i = \{y_{i1}, y_{i2}, \ldots, y_{id}\}$ mažesnio skaičiaus matmenų projekcinėje arba vaizdo erdvėje R^d $(d < n)$.

3.1 Daugiamačių skalių metodas

Naudojant daugiamačių skalių (*angl.* multidimensional scaling) metodą n-mačiai vektoriai projektuojami į mažesnę dimensijų skaičiaus erdvę (dažniausiai į $d = 2$), siekiant išlaikyti analizuojamos aibės objektų artimumus – panašumus arba skirtingumus [3]. Tarkime, kiekvieną n-matį vektorių $X_i \in R^n$, $i \in \{1, \ldots, m\}$, atitinka mažesnio dimensijų skaičiaus vektorius $Y_i \in R^d$, $d < n$. Artumą (panašumą arba skirtingumą) tarp n-mačių vektorių X_i ir X_j pažymėkime $\delta(X_i, X_j)$, o atstumą tarp dvimačių vektorių Y_i ir Y_j – $d(Y_i, Y_j)$, $i, j = 1, \ldots, m$. Jeigu artumas yra Euklido atstumas, tai $\delta(X_i, X_j) = d(Y_i, Y_j)$. Naudojantis DS algoritmu, bandoma atstumą $d(Y_i, Y_j)$ priartinti prie atstumo $d(X_i, X_j)$. Jeigu naudojama kvadratinė paklaidos (angl. *Stress*) funkcija, tai minimizuojama tikslo funkcija $E_{MDS} = \sum_{i<j} w_{ij}(\delta(X_i, X_j) - d(Y_i, Y_j))^2$, čia w_{ij} yra svoriai. Paklaidos funkcijos reikšmė rodo, kaip tiksliai modelis atitinka pradinius duomenis.

3.2 Pagrindinių komponenčių metodas

Pagrindinių komponenčių analizė (angl. *principal component analysis*) plačiai naudojama duomenims analizuoti kaip daugiamačių duomenų dimensijos mažinimo metodas, duomenų suspaudimui, atsisakant nereikšmingų parametrų, esminių savybių suradimui ir duomenų vizualizavimui [9]. Šiuo metodu ieškoma daugiamačių duomenų mažesnės dimensijos poerdvio, kuriame būtų išlaikyta daugiau originalios erdvės duomenų savybių ir informacijos. PKA metodas plačiai naudojamas genomikoje ir genetikoje, siekiant nustatyti populiacijos struktūrą analizuojant plataus masto duomenis [11] ir/arba identifikuoti taškus atsiskyrėlius, kurie turi būti pašalinami atliekant tolimesnę duomenų analizę, pvz. plataus masto genomo asociacijų tyrimus ar nustatant gamtinės atrankos veikiamas genomo sritis [2].

3.3 Pagrindinių koordinačių metodas

Pagrindinių koordinačių ir pagrindinių komponenčių metodai yra panašūs, nes abiejų tikrinių vektorių tikrinių reikšmių apskaičiavimas vykdomas remiantis matrica sudaryta iš atstumų ar panašumų tarp visų taškų. Kai atstumų matas yra Euklido atstumas, tai gauname rezultatus panašius į rezultatus gautus pagrindinių komponenčių metodu. Pagrindinių koordinačių metodu pirmiausia domimasi objektų panašumu, o tik paskui atskirų duomenų kintamaisiais. Dėl šios priežasties pagrindinių koordinačių metodo tikslas yra matmenų skaičiaus mažinimas išlaikant kuo daugiau originalios informacijos tarp objektų.

3.4 Artimumo matai

Daugiamačių duomenų vizualizavimo metodai padeda nustatyti ar įvertinti daugiamačių duomenų struktūrą: susidariusias grupes, itin išsiskiriančius objektus. Objektai suskirstomi taip, kad skirtumai klasterių viduje būtų kuo mažesni, o tarp klasterių – kuo didesni.

Atliekant artimumo matų lyginamąją analizę buvo naudojami Euklido, Gower, Bray-Curtis, Kulczynski, Jaccard ir Morisito artimumo matai. Duomenys buvo vizualizuoti DS, PK ir PKA metodais parenkant skirtingą artimumo matą. Tarkime, kad turime objektus $X_k = \{x_{k1}, x_{k2}, \ldots, x_{kn}\}$ ir $X_l = \{x_{l1}, x_{l2}, \ldots, x_{ln}\}$, tuomet **Euklido** atstumas

$$d(X_k, X_l) = \|X_k - X_l\| = \sqrt{\sum_{i=1}^{m}(x_{ki} - x_{li})^2},$$

čia m – požymių skaičius. **Gower** panašumo matas yra apskaičiuojamas pagal formulę:

$$d_{Gower}(X_k, X_l) = \frac{1}{n}\sum_i \frac{|x_{ki} - x_{li}|}{\max_s x_{si} - \min_s x_{si}},$$

čia $\max_s x_{si}$ ir $\min_s x_{si}$ atitinkamai yra visuose palyginamuosiuose vektoriuose esančių i-ųjų elementų didžiausias ir mažiausias elementai [10]. **Bray-Curtis** panašumo matas yra modifikuotas Manheteno atstumas. Bendroji Bray-Curtis nepanašumo lygtis užrašoma taip [8]:

$$d_{BCnep.}(X_k, X_l) = \sum_{i=1}^{m}|x_{ki} - x_{li}| \Big/ \sum_{i=1}^{m}(x_{ki} + x_{li}).$$

Jeigu gauname, kad $d_{BCnep.}$ įgyja reikšmę 0, tai objektai – identiški. **Kulczynski** atstumas apskaičiuojamas pagal formulę [4]:

$$d_{Kul.}(X_k, X_l) = \sum_{i=1}^{m}|x_{ki} - x_{li}| \Big/ \sum_{i=1}^{m}\min(x_{ki}, x_{li}).$$

Jaccard panašumo indeksas yra paskaičiuojamas kaip santykis $2d_{BCnep.}/(1+d_{BCnep.})$, čia $d_{BCnep.}$ Bray-Curtis nepanašumo matas [4]. **Morisita** persidengimo indeksas yra statistinis populiacijoje esančių objektų dispersijos matas. Jis naudojamas, kai norima palyginti persidengimą tarp imčių. Apskaičiuojamas pagal formulę:

$$d_{Mor}(X_k, X_l) = 1 - \left(2\sum_{i=1}^{m}x_{ki}x_{li} \Big/ \left((S_{X_k}S_{X_l})\sum_{i=1}^{m}X_{ki}\sum_{i=1}^{m}X_{li}\right)\right),$$

čia [7]:

$$S_{X_k} = \sum_{i=1}^{m}x_{ki}(x_{ki} - 1) \Big/ \sum_{i=1}^{m}x_{ki}\left(\sum_{i=1}^{m}(x_{ki}) - 1\right),$$
$$S_{X_l} = \sum_{i=1}^{m}x_{li}(x_{li} - 1) \Big/ \sum_{i=1}^{m}x_{li}\left(\sum_{i=1}^{m}(x_{li}) - 1\right).$$

Plačiau su šiais panašumo ir atstumo matais galima susipažinti Øyvind Hammer leidinyje "Past Paleontological Statistics Version 4.06".

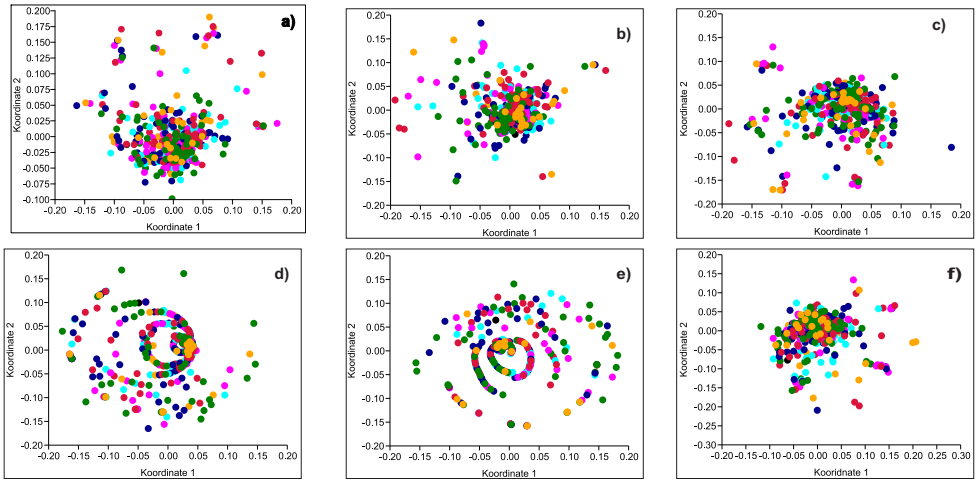
4 Rezultatai

Siekiant ištirti, kuris iš artimumo matų ir daugiamačių duomenų vizualizavimo metodų (DS, PK, PKA) yra tinkamiausias analizuojant populiacijos genetinę struktūrą buvo analizuoti 424 lietuvių plataus masto VNP genetiniai duomenys.

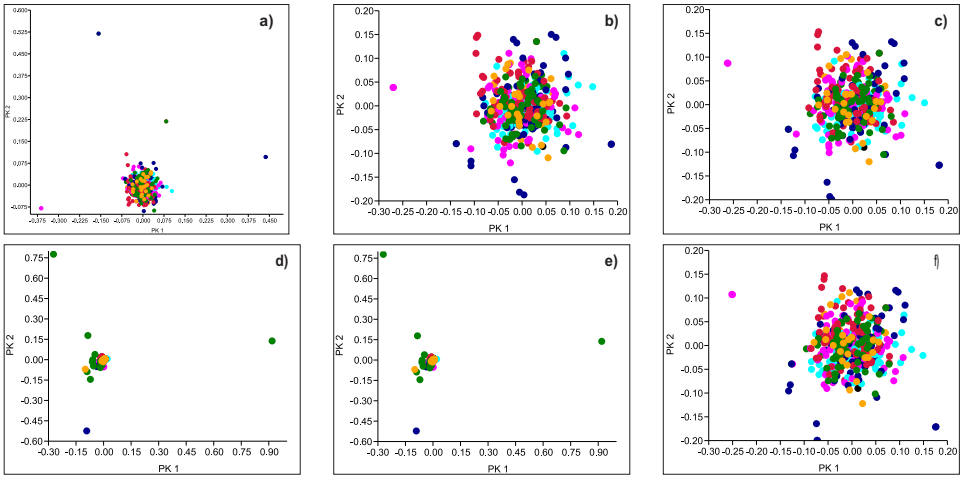
Pirmame paveiksle pateikti vizualizuoti šešių Lietuvos etnolingvistinių grupių VNP duomenys daugiamačių skalių metodu, antrame paveiksle – pagrindinių koordinačių metodu, kai panašumo matai yra Euklido, Gower, Bray-Curtis, Kulczynski, Jaccard ir Morisita. Antrame ir trečiame paveiksluose pietų žemaičiai (PA) pažymėti rožinės spalvos apskritimu, vakarų žemaičiai (VŽ) – geltonos, šiaurės žemaičiais (ŠŽ) – tamsiai raudonos, vakarų aukštaičiai (VA) – žalios, rytų aukštaičiai (RA) – mėlynos ir pietų aukštaičiai (PA) – šviesiai mėlynos spalvos apskritimais. Vizualizavimo kokybės įvertinimui DS metodu buvo skaičiuojama paklaidos funkcijos reikšmė.

Atlikus tyrimus paaiškėjo, kad taikant daugiamačių skalių metodą, paklaidos funkcijos reikšmės apytiksliai lygios, kai naudojame Euklido ($E_{MDS} = 1,374$), Gower ($E_{MDS} = 1{,}369$), Bray-Curtis ($E_{MDS} = 1{,}364$) ir Morisita ($E_{MDS} = 1{,}360$) atstumus. Naudojant Kulczynski atstumą paklaidos funkcijos reikšmė – 1,070. Paklaidos funkcija įgyja mažiausią reikšmę, kai taikome Jaccard atstumą ($E_{MDS} = 0{,}909$) (1 pav.).

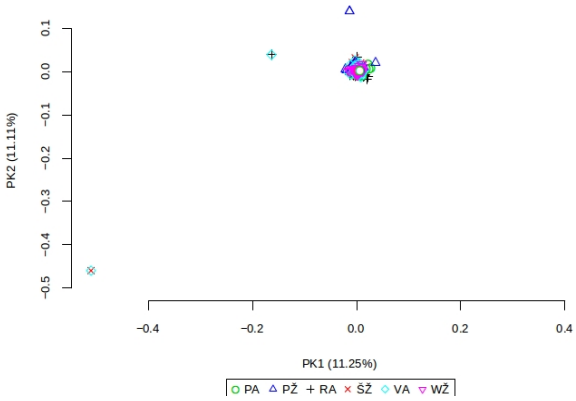
Rezultatai, gauti taikant pagrindinių komponenčių metodą, pateikti antrame paveiksle. Kai panašumo matas yra Euklido atstumas, pirma pagrindinė komponentė ($PK1$) paaiškina 0,57%, antra ($PK2$) 0,55% genetinės įvairovės duomenų. Taikant Gower ($PK1$ – 0,77%, $PK2$ – 0,76%) Bray-Curtis ($PK1$ – 0,77%, $PK2$ – 0,76%) ir Morisita ($PK1$ – 0,78%, $PK2$ – 0,77%) pirmos dvi pagrindinės komponentės paaiškina beveik vienodai genetinės įvairovės. Pirma ir antra pagrindinės komponentės, taikant Kulczynski ($PK1$ – 15,74%, $PK2$ – 6,86%) ir Jaccard ($PK1$ – 15,70%, $PK2$ – 6,87%) artimumo matus, kurios paaiškina $\approx 22,60\%$ genetinės įvairovės, parodė, kad asmenys iš šešių Lietuvos etnolingvistinių grupių suformuoja vieną bendrą klasterį, tačiau matomas ir tam tikras taškų išsibarstymas (2b pav., 2c pav., 2f pav.).



1 pav. Populiacijos genetinė struktūra, taikant DS metodą: a) Euklidas; b) Gower; c) Bray-Curtis; d) Kulczynski; e) Jaccard); f) Morisita.



2 pav. Populiacijos genetinė struktūra, taikant PK metodą: a) Euklidas; b) Gower; c) Bray-Curtis; d) Kulczynski; e) Jaccard); f) Morisita.



3 pav. Populiacijos genetinė struktūra, taikant PKA metodą (Euklido atstumas).

Iš antro paveikslo matome, kad taškai yra labiau arti vienas kito, kai taikome pagrindinių koordinačių negu daugiamačių skalių metodą, o artimumo matai yra Euklido, Kulczynski arba Jaccard.

Tą pačią tendenciją galime pastebėti, kai duomenų vizualizavimui taikome pagrindinių komponenčių metodą, o panašumo matas yra Euklido atstumas (3 pav.). Tiek DS, tiek PK metodu gauti taškai yra labiau išsibarstę, kai artimumo matas yra Gower, Bray-Curtis ir Morisita, todėl sunku įvertinti populiacijos genetinę struktūrą.

Akivaizdu, kad išskirtys yra geriau matomos, kai naudojame PK ir PKA negu DS metodą, taikant tuos pačius artimumo matus. Genetiniuose tyrimuose yra svarbu identifikuoti duomenų išskirtis bei jas pašalinti norint sumažinti klaidingai teigiamų arba neigiamų rezultatų kituose analizės etapuose, ypatingai nustatinėjant populiacijos inbrydingo ar gimininguмo koeficientą.

Iš 2e, 2d, 2f ir 3 paveikslų matome, kad visos šešios Lietuvos etnolingvistinės grupės sudaro vieną bendrą klasterį, todėl galime daryti išvadą, kad Lietuvos populiacija gali būti homogeniška.

Taip pat, tyrimų rezultatai parodė, kad pagrindinių koordinačių ir pagrindinių komponenčių metodai yra tinkamesni Lietuvos populiacijos genetinės struktūros įvertinimui, kai yra taikomi Euklido, Kulczynski ir Jaccard artimumo matai, nes duomenų struktūra labiau „atsiskleidžia“ negu taikant daugiamačių skalių metodą (2 pav. ir 3 pav.).

5 Išvados

Šiame darbe buvo analizuojami ir palyginami daugiamačių skalių, pagrindinių koordinačių ir pagrindinių komponenčių metodai, kai taikomi skirtingi artimumo matai, Lietuvos populiacijos genetinės struktūros nustatymui. Galime daryti išvadą, kad pagrindinių komponenčių ir pagrindinių koordinačų metodai gali būti naudojami VNP duomenų vizualizavimui, kai taikome Euklido, Kulczynski ir Jaccard panašumo matus, nes nebuvo pastebėta esminių skirtumų tarp gautų rezultatų palyginus su daugiamačių skalių metodu gautais rezultatais. Tyrimo rezultatai parodė, kad Lietuvos populiacija gali būti homogeniška, nes taškai yra labiau susiklasterizavę, kai taikome PKA arba PK metodus nei taikant DS metodą. Tačiau hipotezės patvirtinimui reikėtų atlikti duomenų analizę su didesniu genetinių žymenų kiekiu.

Literatūra

[1] G. Abraham, M. Inouye, Y. Zhang. Fast principal component analysis of large-scale genome-wide data. *PLoS One*, **9**(4):e93766, 2014.

[2] I. Borg, P. Groenen. *Modern Multidimensional Scaling.* Springer, New York, 2005.

[3] S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *Int. J. Math. Mod. Meth. Appl. Sci.*, **1**(4), 2007.

[4] C. Chen, W. Hardle, A. Unwin. *Handbook of Data Visualization.* Springer, Berlin, 2008.

[5] G. Dzemyda, O. Kurasova, V. Medvedev. Dimension reduction and data visualization using neural networks. In *Real World AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. Vol. 160: Frontiers in Artificial Intelligence and Applications*, pp. 25–49. IOS Press, Amsterdam, 2007.

[6] M. Greenacre, R. Primicerio. *Multivariate Analysis of Ecological Data.* 2013.

[7] Ø. Hammer. *Past Paleontological Statistics Version 3.18. Reference manual.* Natural History Museum, University of Oslo, 1999–2017.

[8] I.T. Jolliffe. *Principal Component Analysis.* 2nd edn.*, Springer Series in Statistics.* Springer, New York, 2002.

[9] M. Morisita. Measuring of the dispersion and analysis of distribution patterns. *Mem. Fac. Sci., Kyushu Univ. Ser. E: Biology*, **2**:215–235, 1959.

[10] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**:904–909, 2006.

[11] S. Purcell, B. Neale, K. Todd-Brown, *et al.* Plink: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.*, **81**(3):559–575, 2007. https://doi.org/10.1086/519795.

SUMMARY

A comparative analysis of proximity measures to determine the Lithuanian population structure

A. Molytė, A. Urnikytė

In this paper the multidimensional scaling, the principal coordinate and principal component methods for the Lithuanian population structure have investigated, taken that the proximity measures are Euclid, Gower, Bray-Curtis, Kulczynski, Jaccard and Morisita. The genome-wide single nucleotide polymorphism genetic data analyzed. A comparative analysis of proximity measures performed. The results of visualization are also presented.

Keywords: genetic data; proximity measure; population structure