TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MATEMAATIKA JA STATISTIKA INSTITUUT

Sven Erik Ojavee

Päritolu hindamine geeniandmete põhjal: TÜ Eesti Geenivaramu andmete analüüs

Magistritöö matemaatilise statistika erialal (30 EAP)

Juhendaja: PhD Krista Fischer

Tartu 2018

Päritolu hindamine geeniandmete põhjal: TÜ Eesti Geenivaramu andmete analüüs

Lühikokkuvõte

Käesoleva magistritöö eesmärk on leida võimalusi andmaks geenidoonoritele tagasisidet nende päritolu kohta, lähtudes SNPde andmetest. Nendele tuginedes on leitud peakomponendid, millele rajaneb edasine analüüs. Esmalt kirjeldatakse päritolu rahvuste tasandil, mille käigus antakse doonorile tõenäosuslik hinnang kuulumise kohta 22 rahvusgrupi hulka. Sellele järgnevalt kirjeldatakse päritolu Eesti-siseselt, kus leitakse K-keskmiste klasterdamise algoritmi abil Eesti sees tekkivad klastrid, mis moodustavad geograafiliselt loogilisi tervikuid. Klasterdamise tulemusi rakendatakse selleks, et klassifitseerida tekkinud klastrite alusel ning pakkuda ka hinnang klastritesse kuulumise tõenäosustele. Ühtlasi kontrollitakse, kui hästi töötab Eesti-sisene klassifitseerimine, valides klassideks maakonnad. Klassifitseerimismeetoditest võrreldakse lineaarset diskriminantanalüüsi, tugivektormasinaid ning juhuslikke metsi.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika
Märksõnad: klasteranalüüs, klassifitseerimine, mitmemõõtmeline skaleerimine, tehisõpe, kõrgdimensionaalsed andmed, simulatsioon

Estimating Ancestry Using Genome Data: The Analysis of Estonian Genome Center Data

Abstract

The aim of this thesis is to find ways for giving feedback to gene donors about their ancestry by using SNP data. Based on the SNP data principal components are calculated which are used in further analyses exclusively. Firstly, a description about ancestry concerning different nationalities is given by yielding a probabilistic estimate about belonging to 22 nationalities. Secondly, ancestry is described within Estonia. By implementing the K-means clustering algorithm, geographically consistent clusters are constructed within Estonia. The results of the clustering are applied in order to build a classification of gene donors and to predict the probability of donor belonging to clusters. Furthermore, it is tested how well classification works within Estonia when using counties as class labels. Methods used for classification are linear discriminant analysis, support vector machines and random forests.

CERCS research specialisation: P160 Statistics, operations research, programming, actuarial mathematics
Keywords: cluster analysis, classification, multidimensional scaling, automatic learning, high-dimensional data, simulation

Sisukord

1 Sissejuhatus

Erinevate rahvuste geneetilisi erinevusi on uuritud ja uuritakse palju. Sellekohased teadmised aitavad aru saada inimkonna ajaloost ja rahvaste rändamisest aastatuhandete vältel, samuti võib nii saada selgitusi sellele, kuidas on inimorganism kohastunud erinevate elutingimustega maailma eri piirkondades. Ka Eestis tuntakse palju huvi selle vastu, kas ja mille poolest erinevad eestlased teistest lähematest ja kaugematest rahvastest ning isiklikul tasandil huvitab paljusid oma esivanemate täpsem päritolu.

TÜ Eesti Geenivaramuga on jaanuariks 2018 liitunud enam kui 53 000 geenidoonorit kõigist Eesti maakondadest. Neist ligi 50 000 DNA on genotüpiseeritud ülegenoomsete kiipidega - st nende kohta on olemas andmed genoomi varieeruva osa kohta. See võimaldab uurida, kui suurel määral saab geeniandmete põhjal hinnata inimese esivanemate päritolu. Et tegu on huvitava küsimusega ka paljude geenidoonorite enda jaoks, on plaanis lisada päritolu info ka geenidoonoritele antava tagasiside hulka. Käesoleva magistritöö eesmärk ongi uurida, kas klassikalise mitmemõõtmelise statistika või uuemate masinõppe meetodite abil saadud hinnangud oleks selliseks tagasisideks sobivad ja milliseid valikuid tehes meetodite ja nende parameetrite osas saame täpsemaid hinnanguid.

Päritolu kirjeldamiseks on paljudes töödes kasutatud üksiknukleotiidsete polümorfismide ehk SNPde andmeid ning SNPde andmetest lähtutakse ka selles töös. Pisut erinevalt suhtutakse SNPde andmete kasutamisesse. Kui mitmetes töödes on keskendutud just rahvust määravate SNPde tuvastamisele [1], siis käesolevas töös minnakse teist teed ning SNPdest lähtuvalt leitakse peakomponendid, millele järgneb edasine analüüs.

Eelnevalt on käesoleva töö autor uurinud bakalaureusetöös [2] võimalusi, kuidas kirjeldada tõenäosusi, et valitud indiviid kuulub teatavasse rahvusgruppi. Bakalaureusetöös võrreldi MixFit algoritmiga [3] saadud hinnanguid ja peakomponentanalüüsi ja lineaarse diskriminantanalüüsiga saadud hinnanguid. Sealsed tulemused andsid lootust, et peakomponente kasutav lähenemine võib olla asjakohane ning anda häid tulemusi. Siiski olid bakalaureusetöös saadud tulemused esialgsed ning mitmeid probleeme ja valikuid seal täpsemalt ei analüüsitud. Antud töö käigus vaadeldakse küsimust põhjalikumalt, kaasates enam referentsandmestikke ning vaadeldakse täpsemalt võimalusi Eesti-sisese päritolu ennustamiseks.

Käesoleva magistritöö eesmärk on päritolu hindamine mitmel tasandil. Esmalt hinnatakse päritolu rahvuse tasandil, kasutades referentsandmestikke 22 Euroopa rahvuse kohta. Seejärel uuritakse, kas tõenäoliselt eesti päritolu inimeste puhul saab ka hinnata, millisest Eesti piirkonnast pärinevad nende esivanemad.

Autor tänab väga juhendaja Krista Fischerit hindamatu abi ning asjalike nõuannete eest, mis on aidanud töö valmimisele tohutult kaasa. Autor tänab abi eest ka Toomas Hallerit, Kristi Lälli, Reedik Mägit ja Mare Vähit.

2 Töös kasutatavad andmed

2.1 Uuringuvalimid

Käesolevas töös kasutatud valimid saab liigitada kolmeks: referentsvalimid, testvalimid ja põhivalim. Mitte kõiki valimitüüpe ei rakendatud võrdsel määral. Rakendamine sõltus eelkõige ülesande vajalikkusest. Üldiselt üritati vältida eraldi testvalimi kasutamist ning vajalikud hinnangud parameetritele saadi enamasti ristvalideerimise teel.

Referents- ehk treeningvalimid on moodustatud teadaoleva (või eeldatavalt teadaoleva) päritoluga inimestest. Selle valimi pealt töötatakse välja mudelid, mille põhjal inimese päritolu prognoosida. Töö kahe ülesande jaoks on need valimid moodustatud erinevalt. Täpsemalt kirjeldatakse referentsvalimite moodustamist alapeatükkides Referentsandmestik rahvuse määramiseks ja Referentspopulatsiooni valik Eesti-sisese päritolu uurimiseks.

Olukorras, kus vaatlusega kaasnes ka kindel klass, kasutati täpsuse kontrolliks ja parameetrite hindamiseks ristvalideerimist ning eraldi testvalimit ei kasutatud. Et näiteks klasterdamise puhul pole ette teada, mis on igale vaatlusele vastav klaster ja vaatlused mõjutavad klastrite välja kujunemist, siis pärast klastrite arvu leidmist kontrolliti tulemusi ka omaette testvalimi põhjal. Eesti-sisese päritolu hindamiseks kasutati sünniaastatelt referentsist veidi nooremaid, kuid siiski võimalikult vanu isikuid, kelle puhul võib eeldada, et suurel osal neist on esivanemad pärit piirkonnast, kus need isikud sündisid. Need inimesed vastavad samadele kriteeriumitele, mis Eesti-sisese päritolu uurimise referentspopulatsioongi, ent nad on sündinud aastail 1961-1970.

Omaette testvalim moodustati simuleeritud inimeste genotüüpidest, kelle üks vanem oli suure tõenäosusega eestlane ning teine vanem teiselt Eestiga piirnevalt alalt. Seega antud testvalim koosneb genotüüpidest, mille päritolu on ligikaudu vastavalt 0,5-0,5 eesti ja siis mingist muust lähirahvusest. Analoogiliselt tekitati ka 0,75-0,25 eesti ja mingist muust lähirahvusest olevate inimeste genotüüpe. Saadud testandmete põhjal kontrolliti, kui täpselt on võimalik prognoosida teadaoleva päritoluga inimeste päritolu.

Põhivalimi moodustavad TÜ Eesti Geenivaramu andmed ligi 50 000 genotüpiseeritud inimese andmetest. Selle andmestiku jaoks leitakse prognoosid rahvusgruppi kuulumise tõenäosusele.

Statistilisi meetodeid rakendati vaid peakomponentanalüüsi abil teisendatud andmetele. Rahvuse määramise osas arvutati peakomponendid, kasutades rahvuse määramise referentspopulatsiooni ning tehes eelvalik SNP-dele (täpsemalt alapeatükis Genotüübiandmete esmane töötlus rahvuse klassifitseerimisel). Eesti-sisese päritolu kirjeldamise osas leiti peakomponendid sugulusmaatriksi pealt.

2.2 Referentsandmestik rahvuse määramiseks

Lähteandmestikuks on Toomas Halleri MixFit algoritmi [3] tarvis valitud referentspopulatsioon. Selle algoritmi töötamiseks on välja valitud 22 rahvust ning neile vastavalt umbes 45-100 esindajat igast rahvusest. Taani, Ühendkuningriigi ning Hollandi andmed on saadud GenomEUtwin uuringust [3] ning nende valik on kirjeldatud täpsemalt artiklis [4]. Ülejäänud 19 rahvuse esindajate genotüübi valik pärineb artiklist [5]. Edaspidises eeldame, et need valitud on teatavas mõttes sobivad antud rahvusrühma esindajad ning et nende alusel on

võimalik prognoosida ka teiste inimeste rahvusgruppidesse kuulumist.

Et kavandatava päritolu tagasiside eesmärk on anda hinnang Eesti Geenivaramu doonoritele, kes on valdavalt eestlased või venelased, siis ennustuse korrigeerimiseks otsustati suurendada referentspopulatsiooni nendes rahvusrühmades. Suurendamise vajadus oli eriti ilmne venelaste puhul, sest ainult senise populatsiooni kasutamine näis andvat nihke ning liialt väikse varieeruvuse võrdluses Geenivaramu venelastest doonoritega. Eelpool kirjeldatud vastav eestlaste referentspopulatsioon kattus doonorite tulemustega valdavalt hästi, kuid täpsuse kindlustamiseks otsustati siiski ka referentseestlaste hulka suurendada.

Eestlaste jaoks valiti Geenivaramu doonorite seast välja igast maakonnast seitse inimest, kes olid sündinud enne aastat 1930, ise ennast raporteerinud eestlaseks ning kes olid surnud. Valik tehti nõnda, et tagada kõikide maakondade kaetus ning vanemate inimeste esivanemad on tõenäolisemalt ka antud piirkonnas elanud kauem. Surnud inimesed valiti, sest nende rahvuse kohta tagasiside andmine pole primaarne. Venelastest Geenivaramu doonorite seast valiti välja inimesed, kellede emakeel on vene keel, kes on end raporteerinud venelaseks ning kes on sündinud nii raporteeritult kui ka rahvastikuregistri alusel Venemaal.

Eestlaste lisandunud referentspopulatsioon osutus sarnaseks varasemalt valituga peakomponentide mõttes. Venelaste lisandunud referentspopulatsioon osutus mõnevõrra nihkes olevaks peakomponentide mõttes, võrreldes esialgse referentspopulatsiooniga. Seega, võib kahtlustada, et ainult 90 inimesest koosnev valim ei suuda tõepoolest suure vene kogupopulatsiooni varieeruvust korralikult kirjeldada. Seega, et täpsustada eestlaste klassifitseerimisvõimekust ning oluliselt parandada venelaste oma kaasati mõlemad uued populatsioonid olemasolevatesse referentspopulatsioonidesse.

2.3 Referentspopulatsiooni valik Eesti-sisese päritolu uurimiseks

Kirjeldamaks Eesti-sisest päritolu, oleks tarvilik leida inimesed, kes on suurema tõenäosusega sünnikohaga juba kauem seotud ning loodetavasti seda juba mitmeid põlvkondi. Sellised inimesed peaksid eeldatavasti paremini kirjeldama piirkondlikke geneetilisi eripärasid, kui neid peaks leiduma. Selliste inimeste leidmiseks, on referentspopulatsiooni valitud Geenivaramu doonorid, kes täidavad järgmisi nõudeid:

1. On raporteerinud ennast eestlasena;
2. Sünnikohaks ei ole märgitud ükski Eesti suurem linn, st Tallinn, Tartu, Narva, Pärnu;
3. Inimene on sündinud enne aastat 1960;
4. Peatükis Tulemused I: päritolu hindamine rahvuse tasandil saadud mudelit kasutades saadud tõenäosuse prognoos näitab eesti grupi tõenäosuseks vähemalt 0,9;
5. Inimene ei ole lähisugulane ühegi teise inimesega referentspopulatsioonist.

Ilmselt on sobivaim kasutada referentsis vanemaid inimesi ning seega eelistada enne aastat 1930 sündinuid, ent on oht, et väike valimimaht ei võimalda kirjeldada päritolu täiesti adekvaatselt. Seega, selle probleemi vältimiseks on valitud

referentsi inimesed, kes on sündinud enne aastat 1960 ning sellega ühtlasi tagades piisav valimimaht. Kokkuvõttes jäi sellise valiku tulemusena referentspopulatsiooni 5785 inimest.

2.4 Genotüübiandmete esmane töötlus rahvuse klassifitseerimisel

Määramaks rahvust, on esmalt vaja leida SNP-dest lähtudes peakomponendid. SNP-de koguarv on ligikaudu 260 000. Kõik SNP-d ei pruugi antud klassifitseerimisülesande seisukohalt olla sisukad, sest ainult mõnede SNP-de väärtused erinevad rahvuseti. Vähendamaks arvutusmahtu ning võimaliku ülesobitamise ohtu, valitakse välja edasiseks analüüsiks vaid sellised SNP-d, mis põhjustavad suuremat varieeruvust rahvuste vahel. Eesmärk on leida SNP-d, kus rahvustevahelised erinevused moodustavad võimalikult suure osa SNP-i koguhajuvusest.

Eeldame, et ühe konkreetse SNP kodeeritud alleelide arv on binoomjaotusega juhuslik suurus $X \sim B(2, p)$. See on sobiv, sest X võimalikud väärtused on $0, 1, 2$ ja p tähendab vastavat alleelisagedust. Sel juhul on võimalik hinnata SNPi koguhajuvust, leides dispersiooni, mis avaldub $DX = 2p(1-p)$.

Olgu p_i alleelisagedus grupis i. Järgnevalt hindame iga SNP jaoks oodatava alleelisageduse kui erinevate riikide oodatavate alleelisageduste keskmine: $\hat{p} = \frac{1}{22}\sum_{i=1}^{22} \hat{p}_i$. Seega, ühe konkreetse SNP jaoks hinnang kogudispersioonile avaldub kui $\hat{DX} = 2\hat{p}(1-\hat{p})$.

Teisalt hindame, kui suur on varieeruvus (jällegi ühe SNP piires) erinevate rahvusgruppide keskmiste vahel. Seda hindame kui $\frac{1}{21}\sum_{i=1}^{22}(\hat{p}_i - \hat{p})^2$. Kokkuvõttes leiame nende kahe hajuvuse hinnangu suhte

$$\frac{\frac{1}{21}\sum_{i=1}^{22}(\hat{p}_i - \hat{p})^2}{2\hat{p}(1-\hat{p})},$$

mis näitab kui suure osa vastava SNPi koguhajuvusest kirjeldab ära varieeruvus rahvusgruppide vahel. Edasiseks analüüsiks valime välja 20 000 sellist SNPi, mis kirjeldavad suurima osa varieeruvusest. Paraku nendest 20 000 SNPst langesid omakorda ligi 2000 välja halva imputatsiooni kvaliteedi tõttu.

3 Statistilised meetodid

3.1 Genotüübiandmete peakomponentanalüüs

Järgnev kirjeldus peakomponentanalüüsi kohta põhineb õpikul [6]. Tihti leidub olukordi, kus tarvilik oleks vähendada andmete dimensionaalsust, samaaegselt mitte kaotades liialt vajalikku informatsiooni. Olgu vaatluse all p tunnust ning nendele vastavad juhuslikud suurused on $X_1, ..., X_p$. Peakomponentanalüüsi eesmärk on leida sellised lineaarkombinatsioonid (peakomponendid)

$$P_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p, i = 1, ..., p,$$

mille korral oleksid dispersioonid $DP_i, i = 1, ..., p$ maksimaalsed ning järjestatud kui

$$DP_1 \geq DP_2 \geq ... \geq DP_p,$$

kusjuures peavad olema täidetud tingimused

$$a_i^T a_i = 1, i = 1, ..., p$$

ning

$$Cov(P_i, P_k) = 0, \forall k < i, i = 1, ..., p,$$

kusjuures $a_i^T = (a_{i1}, a_{i2}, ..., a_{ip})$. Osutub, et eelkirjeldatud ülesanne taandub teatava omaväärtusülesande lahendamisele, ent detailsemalt on teemat selgitatud autori bakalaureusetöös [2].

Tähistame andmestikku, kus leidub N vaatlust ja p tunnust kui X. Klassikalisel juhul on $N > p$. Sellisel juhul saab näidata, et peakomponentide maatriks P avaldub kui $P = XW$, kus $p \times p$ maatriksi W veergudeks on maatriksi $\frac{1}{N-1}X^T X$ omavektorid. Kui X veerud on tsentreeritud, siis on $\frac{1}{N-1}X^T X$ näol tegemist maatriksi X kovariatsioonimaatriksiga. Geeniandmete puhul kehtib aga enamasti $p > N$. Seetõttu on võimalik leida ülimalt N sõltumatut peakomponenti. Saab näidata, et sellisel juhul on võimalik lähtuda hoopis transponeeritud maatriksist X^T, mille veerud on tsentreeritud, ning seega on vaja leida omavektorid maatriksile $\frac{1}{p-1}XX^T$, st vaatluste omavahelisele kovariatsioonimaatriksile, mida geeniandmete korral nimetatakse ka sugulusmaatriksiks. Algse maatriksi X jaoks saab peakomponendid tuletada, kasutades järgmisi seoseid [7]:

$$\lambda_k^T = \lambda_k \frac{N-1}{p-1},$$

$$a_k^T = \frac{Xa_k}{\sqrt{2\lambda_k \frac{N-1}{p-1}}},$$

kus λ_k^T on transponeeritud maatriksi korral leitud k-s omaväärtus, λ_k on esialgse maatriksi korral leitud k-s omaväärtus, a_k^T on transponeeritud maatriksi korral leitud k-s omavektor, a_k on algse maatriksi korral leitud k-s omavektor.

Selline situatsioon, kus tunnuseid on märgatavalt palju rohkem kui vaatlusi ongi väga levinud geeniandmestikes. Enne eelvalikut oleks SNPe ehk kasutatavaid tunnuseid ligi 260 000, pärast eelvalikut umbes 18 000, ent vaatlusi vaid kõigest ligi 2000. Seega näib geeniandmete jaoks peakomponentanalüüs hea viisina, kuidas vähendada märgatavalt tunnuste arvu.

Rahvuse määramise osa juures oli valimimaht üsna väike ning peakomponendid leiti lähtudes esialgsest kovariatsioonimaatriksist, kasutades tarkvarana R-i funktsiooni "prcomp". Eesti-sisese päritolu kirjeldamiseks leiti peakomponendid sugulusmaatriksilt, kasutades tarkvarana PLINK 2.0 [8], mille jaoks viis arvutused läbi TÜ Eesti Geenivaramu vanemteadur Reedik Mägi.

3.2 Lineaarne diskriminantanalüüs

Järgnev meetodi kirjeldus põhineb raamatul [9]. Klassitõenäosuste leidmisel ja lõpuks seeläbi ka klassifitseerimisel lähtutakse Bayesi valemist, millega leitakse vaatluse klassi k kuulumise tõenäosus juhul, kui vaatluse puhul on vaadeldud tunnusvektor x:

$$P(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l},$$

kus Y on objekti klassi kirjeldav juhuslik suurus juhuslik suurus, X on juhuslik vektor, mis kirjeldab prognoosiks kasutatavaid tunnuseid, $f_l(x)$ on juhusliku vektori X tihedus tingimusel, et vaatlus kuulub klassi l, π_l on klassi l kuulumise eeltõenäosus ning $\sum_{l=1}^{K} \pi_l = 1$, K on klasside arv. Lineaarse (või ka ruut) diskriminantanalüüsi korral kasutatakse tihedusfunktsioonidena f_l mitmemõõtmelise normaaljaotuse tihedusfunktsiooni

$$f_l(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_l|^{1/2}} e^{-0.5(x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l)}.$$

Lineaarne diskriminantanalüüs tekib erandjuhul, kui valida dispersioonimaatriks samaks igas klassis ehk siis $\Sigma_l = \Sigma, \forall l = 1, ...K$. Parameetreid on võimalik küllalt lihtsasti hinnata ning seejärel saabki leida vastavad klassi kuulumise tõenäosused. $\hat{\pi}_k = \frac{N_k}{N}$, kus N_k on klassi k kuuluvate vaatluste arv; $\hat{\mu}_k = \frac{1}{N_k} \sum_{y_i=k} x_i$ ning $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{y_i=k} (x_i - \mu_k)(x_i - \mu_k)^T$.

Kuna töö käigus on vaja klassifitseerida küllalt paljude erinevate klasside vahel, siis eelistatakse kasutada ainult lineaarset diskriminantanalüüsi, et vältida vajadust hinnata väga suurt arvu parameetreid.

Vaadates näiteks etteruttavalt joonist 6, kus on välja toodud mõned klassid kahe peakomponendi järgi, on näha, et väga suuri vastuolusid meetodi tuletamisel kasutatud eeldustega ei ole, punktiparvede kujud pole väga erinevad. See pole aga veenev tõend lineaarse diskriminantanalüüsi headuseks, sest mõõtmete suurenemisel muutub mitmemõõtmelise normaaljaotuse struktuur pisut keerulisemaks ning tegelikult on vaatluse all veelgi enam klasse.

Kirjanduses [9] väidetakse, et head tulemused LDA kasutamisel ei ole tihti põhjustatud sellest, et andmed oleksid ligikaudu normaaljaotusega või dispersioonimaatriksid oleksid võrdsed. Mitmes situatsioonis ei saa andmetest lugeda välja enamat kui lineaarne otsustuspiir ning sel juhul annabki LDA stabiilsema hinnangu võrreldes keerukamate alternatiividega.

R-is on realiseeritud meetod paketis "MASS"funktsioonina "lda".

3.3 Tugivektormasinad

3.3.1 Tugivektorklassifitseerija

Järgnev alapeatükk põhineb raamatul [9].

Vaatleme esialgu N vaatlusega p-mõõtmelist andmestikku, kusjuures iga vaatlus võib kuuluda kas klassi $y_i = -1$ või $y_i = 1$. Seega andmestik koosneb vaatlustest $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$, kus $x_i \in \mathbb{R}^p$. Idee on leida selline hüpertasand $f(x) = x^T\beta + \beta_0 = 0$, mille korral oleksid kaks klassi võimalikult hästi teineteisest eristatud. Lähtudes sellest kummale poole leitud hüpertasandit prognoositav vaatlus satub, prognoositaksegi vaatlus vastavalt klassi -1 või 1. Seega on vaja lahendada optimiseerimisülesanne kujul:

$$\max_{\beta,\beta_0,||\beta||=1} M$$

$$y_i(x_i^T\beta + \beta_0) \geq M, i = 1, ..., N,$$

kusjuures suurust M nimetatakse marginaaliks.

Et tihti pole aga selline täielik eralduvus võimalik või viib ebastabiilsete tulemusteni, modifitseeritakse eelkirjeldatud ülesannet suutmaks toime tulla ka vääralt klassifitseeritud vaatlustega. Endiselt maksimiseeritakse marginaali, kuid nüüd lubatakse mõndadel punktidel asetseda valel pool marginaali või ka hüpertasandit. Selleks defineeritakse abimuutujad $\xi = (\xi_1, \xi_2, ..., \xi_N)$. Nende abil muudetakse eeltoodud kitsendust järgnevalt

$$y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i), i = 1, ..., N,$$

$\xi_i \geq 0$, $\sum_{i=1}^N \xi_i \leq$ const. Idee seisneb selles, et tõkestades ξ_i -de summa, on seega tõkestatud marginaalvigade summa, sest $\xi_i \geq 0$. Defineerides $M = \frac{1}{||\beta||}$, saame optimiseerimisülesande esitada kujul

$$\min_{\beta,\beta_0} ||\beta||$$

$$y_i(x_i^T\beta + \beta_0) \geq (1 - \xi_i), i = 1, ..., N,$$

$$\xi_i \geq 0, \sum_{i=1}^N \xi_i \leq \text{const.}$$

Arvutuslikult on mugavam eeltoodud ülesanne esitada järgmisel kujul, kus eelneva konstandi rolli võtab nüüd C.

$$\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^N \xi_i$$

$$\xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq (1 - \xi_i), i = 1, ..., N,$$

Selle ülesande asemel on märksa hõlpsam lahendada vastav duaalne ülesanne.

$$\max \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N \sum_{i'=1}^N \alpha_i\alpha_{i'}y_iy_{i'}x_i^Tx_{i'}$$

$$0 \leq \alpha_i \leq C, \sum_{i=1}^N \alpha_i y_i = 0.$$

Eelnev meetod võimaldab edukalt lahendada mitmeid probleeme, kus klassid eristuvad üsna lihtsasti, kuid meetod jääb hätta juhtudel, kus ainult tavalise hüpertasandi abil pole võimalik eristust teha. Sellised on näiteks olukorrad, kus klassi 1 vaatlused asuvad kahes punktiparves, ja klassi -1 vaatlused asuvad punktiparves, mis paikneb eelnevate parvede vahel.

Osutub, et probleemist on võimalik mööda saada kasutades tuumasid. Nimelt asendatakse eelnevas maksimiseeritavas funktsioonis skalaarkorrutis $x_i^T x_{i'}$ tuumafunktsiooniga $K(x_i, x_{i'})$. Ühtlasi on võimalik näidata, et piisav ja tarvilik tingimus K valikuks on sümmeetrilisus ning positiivselt poolmääratus. Sellise muudatusega on võimalik paremini lahendada märkimisväärselt enam klassifitseerimisprobleeme. Populaarsemad valikud K jaoks on näiteks polünoomtuum ja Gaussi tuum, kuid tuumi on palju ning vastavalt situatsioonile võib leida parasjagu hästi sobiva tuuma. Käesolevas töös on valitud tuuma rolli Gaussi tuum, mis on esitatav kujul $K(x, x') = \exp(-\gamma||x - x'||^2)$. Kokkuvõttes saadakse klassifitseerija kujul $\hat{G}(x) = \text{sign}(\hat{f}(x))$, kus

$$\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0.$$

Tugivektormasinad on algselt välja töötatud kahe klassi eristamiseks, kuid meetodit on võimalik üldistada kahest enama klasside arvu eristamiseks. Kaks levinumat võimalust üldistamiseks on 1 vs 1 ja 1 vs kõik. Käesolevas töös on rakendatud 1 vs 1 meetodit. Selle kohaselt tekitatakse $\binom{K}{2}$ tugivektormasinat (K on siin klasside arv), kus võrreldakse paarikaupa omavahel kõiki klasse. Prognoositav vaatlus määratakse klassi, kuhu ta sattus kõige enim kordi kõikide võrdluste korral. [10]

3.3.2 Tõenäosuste hindamine tugivektormasinatega

Osutub, et tugivektormasinatega lahendatava optimiseerimisülesannet saab esitada kujul $\min_{\beta,\beta_0} L(x, y, \beta) + \lambda P(\beta)$, kus L on nn kaofunktsioon ja P on karistusfunktsioon. Seejuures on tugivektormasinatele vastav kaofunktsioon väga sarnane logistilisele regressioonile vastava kaofunktsioonile, mistõttu annavad nad tihti ka sarnaseid tulemusi. [10]

Osalt ka eelneva tulemuse tõttu on välja pakutud meetodeid, mille abil oleks võimalik leida tõenäosuseid analoogiliselt logistilisele regressioonile. Üks tuntumaid ja kasutatumaid meetodeid on Platti skaleerimine (Platt scaling) [11]. Selle kohaselt tekitatakse lähtuvalt saadud $\hat{f}$-st tõenäosuslikud hinnangud kui

$$P(Y = 1|\hat{f}) = \frac{1}{1 + \exp(A\hat{f} + B)}.$$

Protsessi käigus hinnatakse parameetrid A ja B suurima tõepära meetodil ning sisuliselt on tegemist logistilise regressioonimudeli hindamisega. Seega on igale vaatlusele võimalik anda paarikaupa võrdlustes tõenäosus, et ta kuulub näiteks klassi 1. Klassifitseerigem kahe klassi i ja j vahel. Klassifitseerides nende kahe klassi vahel, tähistame klassi i (ja mitte klassi j) kuulumise tõenäosuse hinnangut $r_{ij} := \hat{P}(Y = i|(\{Y = i\} \cup \{Y = j\})), x)$.

Paraku on antud juhul tegemist suurema arvu klassidega kui ainult kaks ning vaja on leida võimalus üldistamaks, et leida hinnangud kõikidele tõenäosustele

$p_i = P(Y = i|x), i = 1, ..., K$. Ühe võimaliku lahenduse sellele on välja pakkunud Wu, Lin ja Weng [12]. Lihtne on näidata, et

$$\frac{P(Y = i|(\{Y = i\} \cup \{Y = j\}), x)}{P(Y = j|(\{Y = i\} \cup \{Y = j\}), x)} = \frac{P(Y = i|x)}{P(Y = j|x)}.$$

Sellest loogikast lähtuvalt peaksid siis ka

$$\frac{r_{ij}}{r_{ji}} \approx \frac{p_i}{p_j}.$$

Seega on mõistlik lahendada optimiseerimisülesannet kujul

$$\min_p \sum_{i=1}^{K} \sum_{j:j\neq i} (r_{ji}p_i - r_{ij}p_j)^2$$

tingimustel

$$\sum_{i=1}^{K} p_i = 1, p_i \geq 0, i = 1, ..., K.$$

Kindlasti pole siin välja toodud võimalus ainukene, kuidas leida tõenäosust. Samas artiklis [12] toovad autorid välja ka teise meetodi ning ka mitmed teised autorid on välja pakkunud oma meetodeid. Sarnaselt leidub ka modifikatsioone eeltutvustatud tugivektorklassifitseerijale. Nimetatud meetodeid on esitletud siin täpsemalt, kuna need on implementeeritud ka laialt kasutatavas R paketis "e1071"vastavas funktsioonis "svm". Pakett "e1071"toetub omakorda tarkvarale LIBSVM [13]. Käesoleva töö puhul rakendati tugivektormasinate kasutamisel eelnimetatud paketti ja tarkvara.

3.4 Juhuslikud metsad

Järgnev alapeatükk põhineb raamatutel [14] ja [9].

Juhuslike metsade idee rajaneb otsustuspuudele kasutamisele, mis üksikuna võttes ei ole väga head klassifitseerijad. Eriti suur probleem on otsustuspuude puhul hinnangute võimalik suur hajuvus, ent keskmistades üle paljude mingis mõttes erinevate puude, on võimalik saada stabiilsed ja head hinnangud klassikuuluvustele. Osutub, et mida vähem korreleeritud on kasutatavad puud, seda täpsemaks võib osutuda ka lõppotsus.

Et tekitada mitte- või vähekorreleeritud puid, kasutatakse koos kahte meetodit, mille abil modifitseeritakse tavalist otsustuspuude hindamist. Iga puu sobitamise jaoks võetakse esmalt algsest valimist *bootstrap*-valim, millele hakatakse puud treenima. Kokku võetakse niimoodi B valimit. Teiseks, iga kord, kui puud jaotatakse kaheks, valitakse juhuslikult m tunnust kõigi p tunnuse hulgast, ($m \leq p$), mille põhjal tehakse otsus puu jagunemise kohta.

Ülejäänud aspektid iga puu treenimise kohta jäävad samaks. Rekursiivselt jaotatakse andmestik sammhaaval kaheks. Igal sammul otsitakse juhuslikult valitud m tunnuse seast sellist tunnust ning tunnusele vastavat kohta, mille alusel teostada järgmine andmete jagunemine kaheks. Rekursiivset protsessi jätkatakse seni kuni väikseim tekkinud alamandmestik on suurem kui mingi valitud arv n_{min}. Olgu $\hat{C}_b(x)$ b-nda puu klassiennustus. Juhusliku metsa ennustus vaatlusele x on selline klass i, mille korral enamus puid prognoosib just sedasama klassi i: $\hat{C}^B_{RF}(x) = enamusvalik\{\hat{C}_b(x)\}$.

Hinnangu saamiseks tõenäosusele kuuluda klassi i leitakse nende puude arvu osakaal, mis ennustasid klassi i [15]:

$$\hat{P}(Y = i|x) = \frac{1}{B}\sum_{b=1}^{B} I_{\hat{C}_b(x)=i}.$$

Samas rõhutab Breiman [15], et ehkki sellised hinnangud võivad anda kasulikku infot olukorra kohta, ei tohiks neid tõlgendada hinnangutena õigetele klassitõenäosustele. Et sellest probleemist vabaneda on pakutud välja mitmeid lahendusi, näiteks eelmainitud Platti skaleerimise abil korrektuuride tegemine [11]. Käesoleva töö raames nii detailselt probleemi ei süveneta.

Käesolevas töös kasutati juhuslike metsade hindamiseks R-i paketti "randomForest"ning seal vastavat funktsiooni "randomForest", mis põhineb Breimani artiklil. [16]

3.5 K-keskmiste klasterdamine

3.5.1 K-keskmiste algoritm

Järgnev alapeatükk tugineb raamatule [10].

Olgu ette antud sobitatavate klastrite arv K ning N on vaatluste arv. Olgu $C_1, ..., C_K$ indeksite hulgad, mis tähistavad vastavatesse klastritesse kuulumist. Kuulugu iga vaatlus ühte klastrisse ning olgu klastrid paarikaupa lõikumatud. K-keskmiste klasterdamise idee seisneb selles, et tekitada kogumid, millede sees oleks varieeruvus väike. Seega, kui tähistada klastrisisest varieeruvust kui $W_k := W(C_k)$, siis eesmärk on lahendada järgmine optimiseerimisülesanne

$$\min_{C_1,...,C_k} \sum_{k=1}^{K} W_k.$$

Andmed $\{x_{ij}\}, i = 1, 2, ..., N, j = 1, 2, ..., p$ koosnevad p tunnusest ja N vaatlusest. Tähistagu $d_{ii'}$ vaatluste i ja i' omavahelist kaugust, mis võib olla näiteks tavaline eukleidiline kaugus, aga võib ka olla midagi muud, näiteks absoluutne kaugus $\sum_j |x_{ij} - x_{i'j}|$. Olgu vaatlused klasterdatud K-sse klastrisse $C_1, ..., C_K$ ning klastris r olgu $n_r = |C_r|$ vaatlust. Olgu

$$D_r = \sum_{i,i' \in C_r} d_{ii'}$$

klastri r punktide paarikaupa kauguste summa ning olgu

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r.$$

Ülesanne klastrite leidmiseks lahendatakse järgmisel põhimõttel. Esiteks pannakse iga vaatlus juhuslikult mingisse klastrisse. Seejärel korratakse järgnevat sammu nii kaua kuni klastrite koosseisus enam muudatusi ei teki. Igale klastrile leitakse temale vastav tsenter (p-mõõtmeline vektor iga tunnuse keskmisest) ning iga punkt määratakse klastrisse, mille tsentrile on ta kõige lähemal.

Eelkirjeldatud algoritm väljastab tihti kõigest lokaalse miinimumi ning parema lahendi saamiseks on seega soovitatav rakendada algoritmi mitmel korral

ja seejärel valida parim lahend. Käesolevas töös valiti algoritmi kordusrakendamiste arvuks 100.

Käesolevas töös on K-keskmiste algoritmi rakendatud tarkvaras R programmis "kmeans".

3.5.2 Klastrite arvu määramine

Üks olulisemaid probleeme K-keskmiste klasterdamise algoritmi juures on olnud vajaliku klastrite arvu määramine. Kriteeriumid klastrite arvu valikuks pole kindlasti defineeritud üheselt ning võimalusi on mitmeid. Üks lihtsamatest kriteeriumitest on nn küünarnuki meetod. Meetod leiab klastrisisesed hajuvuste summad iga soovitud k korral ning soovitab valida klastrite arvuks selle, kus hajuvuste summa väheneb märgatavalt. Siiski ei pruugi see meetod töötada paljudel juhtudel, näiteks juhul, kui sellist hüpet ei esine, vaid vähenemine toimub ühtlaselt. [17]

Et eelkirjeldatud meetod ei pruugi mitmel puhul hästi töötada, on proovitud leida ka teisi võimalusi. Tibshirani, Walther ja Hastie on välja pakkunud *gap*-statistiku idee [18]. Käesoleva töö kontekstis rakendatakse klastrite arvu soovituse saamiseks seda meetodit. Järgnev ülevaade põhineb eelnimetatud artiklil.

Idee seisneb $\log(W_k)$ standardiseerimises, võrreldes $\log(W_k)$ väärtust tema keskväärtusega, mis on leitud kasutades selleks sobivat nullhüpoteesi olukorda. Sobivaks hinnanguks klastrite arvule K oleks selline $\hat{K}$, mille puhul $\log(W_k)$ oleks võrdluskeskväärtusest erinevaim. Seega defineeritakse järgnev statistik

$$\mathrm{Gap}_n(k) = E_n^*\{\log(W_k)\} - \log(W_k),$$

kus E_n^* tähistab keskväärtust, mis on saadud nullhüpoteesile vastavast jaotusest n valimipunkti genereerimisel, kasutades saadud n vaatlust keskväärtuse hindamiseks. Hinnang K-le leitakse nii, et leitakse k, mis maksimiseeriks statistiku $\mathrm{Gap}_n(k)$.

Võrdlusjaotuse valikuks on pakutud kaks varianti, mis lähtuvad teatavatest teoreetilistest kaalutlustest. Esiteks võib genereerida iga võrdlustunnuse ühtlasest jaotusest vastavatest piirkondadest, kus on tunnusel väärtuseid. Teisalt võib leida algul andmetest peakomponendid ning siis rakendada esimest meetodit juba peakomponentidele. Lõpuks on vaja ka ühtlasest jaotusest genereeritud andmed tagasi teisendada esialgsesse ruumi. Et käesoleva töö käigus tegeletaksegi klasterdamisel peakomponentidega ning et inimeste välja noppimine klastri loomiseks mõeldud referentsvalimisse ilmselt väga palju ei mõjuta andmete paiknemist, siis käesolevas töös on arvutuste kiirendamiseks eelistatud esimest meetodit.

Järgnevalt hinnatakse $E_n^*\{\log(W_k)\}$ kui B $\log(W_k^*)$ koopia keskmine, kusjuures iga $\log(W_k^*)$ on leitud (parameetrilise) *bootstrap* valimist $X_1^*, X_2^*, ..., X_n^*$, mis on valitud eelkirjeldatud eeskirja alusel referentsjaotusest.

Tähistagu $\mathrm{sd}(k)$ B $\log(W_k^*)$ koopia standardhälvet. Võtmaks arvesse simuleerimisviga $E_n^*\{\log(W_k)\}$ leidmisel, saame suuruse

$$s_k = \mathrm{sd}(k)\sqrt{1 + \frac{1}{B}}.$$

Seda kasutades on soovitatud valida hinnanguks vähim k, mille korral $\mathrm{Gap}(k) \geq \mathrm{Gap}(k) - s_{k+1}$.

Kokkuvõttes leitakse *gap*-statistik järgnevalt:

1. Klasterdada vaatlused, kasutades klastrite arvudena $k = 1, 2, ..., K$ ning leida vastavad klastritesiseste hajuvuse mõõdud W_k.

2. Luua B *bootstrap* andmestikku, genereerides vaatlused ühtlase jaotuse alusel, kusjuures ei tehta standardiseerivat teisendust peakomponentidega. Leitakse vastavad suurused $W^*_{kb}, k = 1, ...K, b = 1, ...B$. *gap*-statistik leitakse kui
$$\text{Gap}(k) = \frac{1}{B}\sum_{b=1}^{B} \log(W^*_{kb}) - \log(W_k).$$

3. Tähistagu $\bar{l} = \frac{1}{B}\sum_b \log(W^*_{kb})$ ja leida hinnang
$$\text{sd}(k) = \Big[\frac{1}{B}\sum_{b=1}^{B}\{\log(W^*_{kb} - \bar{l}\}^2\Big]^{0.5}$$
ning olgu $s_k = \text{sd}(k)\sqrt{1 + \frac{1}{B}}$.

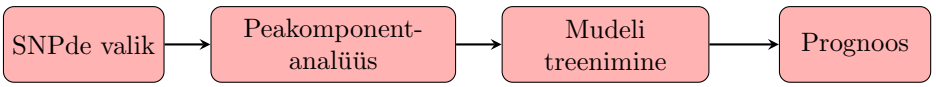
4. Klastrite arv valitakse kui
$$\hat{k} = \text{vähim } k, \text{mille korral Gap}(k) \geq \text{Gap}(k+1) - s_{k+1}.$$

Punktis 4. välja toodud kriteerium pole üheselt määratud ning mõeldavad on teistsugused valikud näiteks üks lihtne võimalus on valida k, mis maksimiseerib *gap*-statistiku väärtuse. Hilisema töö käigus on näha, et artiklis [18] pakutud meetod ei tööta antud käesolevas situatsioonis väga hästi. Alternatiivina on Dudoit ja Fridlyand pakkunud välja valida väikseim k, mis vastab *gap*-statistikule, mis ei erine maksimaalsest statistiku väärtusest enam kui $\alpha \cdot s_{k_0}$ võrra [19], kus s_{k_0} on maksimaalsele statistikule vastav standardhälbe hinnang ning α on mingi positiivne valitud kordaja. Antud töös on valitud $\alpha = 3$.

R-is on *gap*-statistiku leidmine realiseeritud paketis "cluster", kasutades funktsiooni "clusGap".

4 Tulemused I: päritolu hindamine rahvuse tasandil

Esimese sammuna on eesmärk anda tõenäosuslik hinnang inimese rahvusele. Selleks teostatakse valik toorandmetest, siis toorandmetele peakomponentanalüüs ning rakendatakse teatavat algoritmi klassifitseerimiseks.



Joonis 1: Rahvuse klassifitseerimise protsess

4.1 Peakomponentanalüüs referentsandmestikule

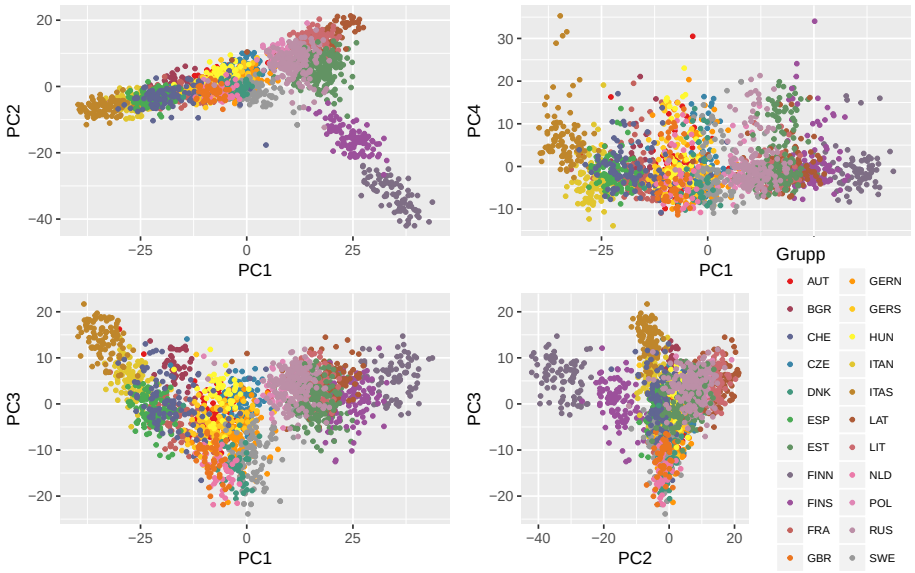
Pärast SNPde valikut tehti referentsandmetele peakomponentanalüüs, et seeläbi vähendada ülesobitamise võimalust ning et muuta ülesanne arvutuslikult teostatavamaks. Et referentsandmestikus on 2019 vaatlust, saadi ka 2019 peakomponenti. Hea ülevaate annab juba esimese kahe peakomponendi kujutamine, mida on näidatud joonisel 2. Joonisel on selguse mõttes välja toodud erinevatel joonistel erinevad rahvusgrupid ning esimesel joonisel on kõik grupid koos. On võimalik näha, et mitmed grupid eristuvad juba kahe peakomponendi alusel väga hästi, mis annab lootust neid ka klassifitseerimisel hästi eristada. Samas on ka selliseid gruppe, mis kahe peakomponendi alusel nõnda hästi ei eristu. Näiteks võib tuua hollandi, briti ja taani grupid, millede punktiparved on peaaegu samal alal. Samas tuleb arvestada, et klassifitseerimisel saab kasutada enam peakomponente, mille alusel loodetavasti saab tuvastada erinevuse ka nende gruppide puhul.

Joonisel 3 on esitatud ka iga peakomponendi jaoks protsent selle kohta, kui suure osa hajuvusest vastav peakomponent kirjeldab. Esimesed kolm peakomponenti annavad märgatavalt suurema kirjeldatuse ning alates neljandast peakomponendist väheneb kirjeldatus aeglaselt. Päris viimased peakomponendid annavad üsna halva kirjeldatuse.

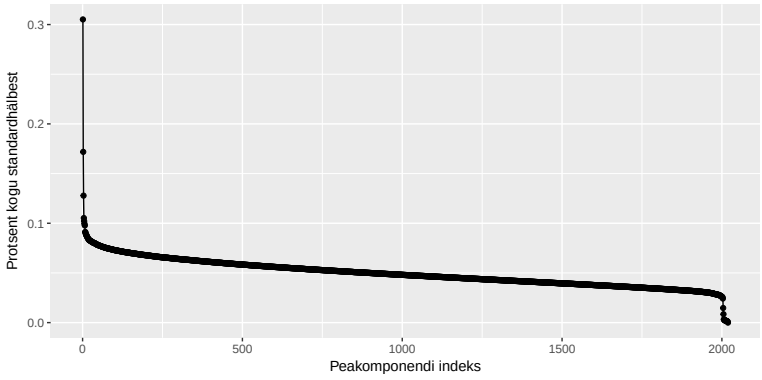
4.2 Klassifitseerimismeetodite võrdlus

Saavutamaks parim võimalik täpsus klassifitseerimisel tuleb otsustada, millist meetodit kasutada ning mitut peakomponenti oleks vaja kasutada. Liiga suur peakomponentide arv võib viia ülesobitumiseni, liiga vähe tekitaks alasobitumust.

Antud juhul võrreldi kolme meetodit: lineaarne diskriminantanalüüs, tugivektormasinad (SVM) ja juhuslikud metsad (RF). Meetodeid võrreldi 10-jaotusega ristvalideerimise abil (10-fold cross-validation). See tähendab, et andmestik jagati juhuslikult kümneks osaks. Seejärel võeti esimesed üheksa osa ning neid kasutati mudeli hindamiseks ja viimase osa andmetele leiti mudeli alusel prognoosid. Prognoose võrreldi päris tulemustega ning raporteeriti prognoositäpsus. Samamoodi kasutati testandmestikuna ükshaaval ka kõiki teisi üheksat alamandmestikku ning lõpptäpsusena raporteeriti kõigi kümne täpsuse keskmine. Saamaks hinnang täpsuse hajuvusele, võeti kümnest saadud hinnangust 300



Joonis 2: Peakomponentanalüüsi tulemused nelja esimese peakomponendiga rahvusgruppide kaupa



Joonis 3: Protsentuaalne hajuvuse kirjeldatus peakomponenditi

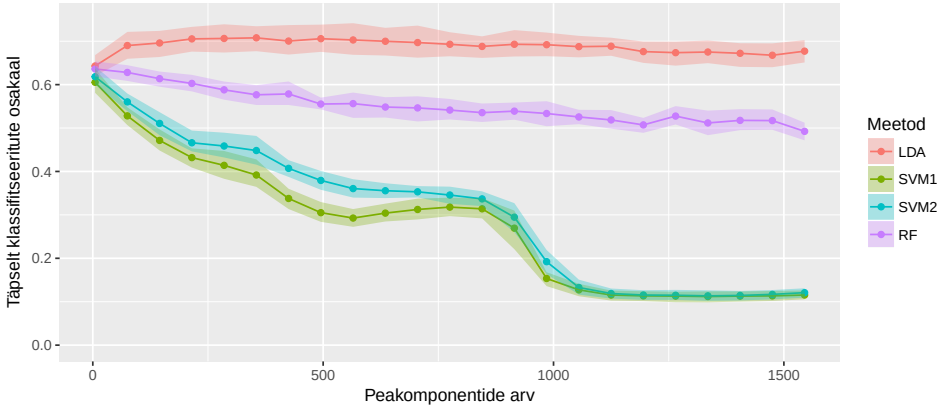
bootstrap-valimit, mille jaoks arvutati samuti välja keskmine täpsus. Saadud 300 täpsuse hinnangust leiti kvantiilid $q_{0.025}$ ning $q_{0.975}$ ehk siis eeldatavasti ligi 95% vaatlustest peaks jääma nende arvude vahele.

Oluline küsimus, millele on vaja tähelepanu pöörata, on parameetrite valik mudelites. Tugivektormasinate puhul on vaja leida karistusliige C ning Gaussi ehk RBF tuuma korral ka γ väärtus. Juhusliku metsa puhul on vaja leida puu hargnemise otsustamisel kasutatavate peakomponentide arv m, minimaalne lehe suurus n_{min} ning puude arv. Lisaks mõjutab suuresti tulemust see, kui palju peakomponente kasutada mudeli tegemiseks. Et kõikide nende otsuste tegemine ristvalideerimise abil osutuks ajaliselt praktiliselt võimatuks, siis järgnevas on valikut pisut lihtsustatud.

Tugivektormasinate puhul kasutati C parameetrina $C = 1$ või $C = 2$. Tu-

givektormasinate puhul kontrolliti ka teisi parameetrite väärtuseid, kuid need andsid kas märksa halvema tulemuse ($C < 1$) või jõudsid parimal juhul sisuliselt sama tulemuseni, mis $C = 2$ juhul ($C > 2$). Parameetrina γ otsustati kasutada väärtust $\frac{1}{p}$. Katsetades mitmel juhul läbi erinevate γ väärtustega, andis just selline valik parimaid tulemusi või parimale lähedasi tulemusi. Ühtlasi on $\frac{1}{p}$ ka vaikimisi soovitus tarkvara poolt. Nüüd ja edaspidi näitab joonistel märksõna SVM1, kui kasutatakse tugivektormasinaid parameetriga $C = 1$ ning SVM2, kui parameeter on $C = 2$.

Juhusliku metsa korral valiti puude arvuks 600. Prooviti ka muid puude arve, kuid puude arvu suurendamine ei andnud mitte mingit lisatäpsust. Peakomponentide arvuks, mille alusel tehakse puu treenimisel otsus, valiti kirjanduse [9] soovitusel $m = \lfloor \sqrt{p} \rfloor$, kus p on peakomponentide arv. Minimaalseks lehe suuruseks valiti $n_{min} = 1$.



Joonis 4: Klassifitseerimismeetodite täpsused 10-jaotuse ristvalideerimise korral sõltuvalt valitud peakomponentide arvust koos 95% *bootstrap* usaldusintervallidega

Jooniselt 4 on näha, et väikese arvu peakomponentide korral annavad kõik algoritmid keskmiselt ligilähedase tulemuse, mis seejuures on SVM1, SVM2 ja juhuslike metsade puhul parim, lineaarsel diskriminantanalüüsil aga halvim. Siinkohal tuleb märkida, et väikseim arv peakomponente, mida näidatakse joonisel, on 5. Veel väiksema peakomponentide arvu korral olid tulemused sarnased kui 5 peakomponendi juhul, kusjuures LDAst paremaid tulemusi ei saavutatud. Peakomponentide arvu suurenemisel paraneb lineaarse diskriminantanalüüsi täpsus, saavutades 200 peakomponendi juures platoo. Teised meetodid käituvad ebaühtlasemalt, kusjuures kõik ülejäänud meetodid saavutavad pigem halvemaid tulemusi peakomponentide arvu suurenemisel.

Joonisele 4 on lisatud ka *bootstrap*-meetodil leitud 95% usaldusintervallid. Huvitav on näha, et LDA hajuvus on pisut suurem kui teistel meetoditel, kuid suurte peakomponentide arvu korral väiksema hajuvuse tagavad SVM1 ja SVM2 annavad teisalt väga halbu täpsuse näitajaid.

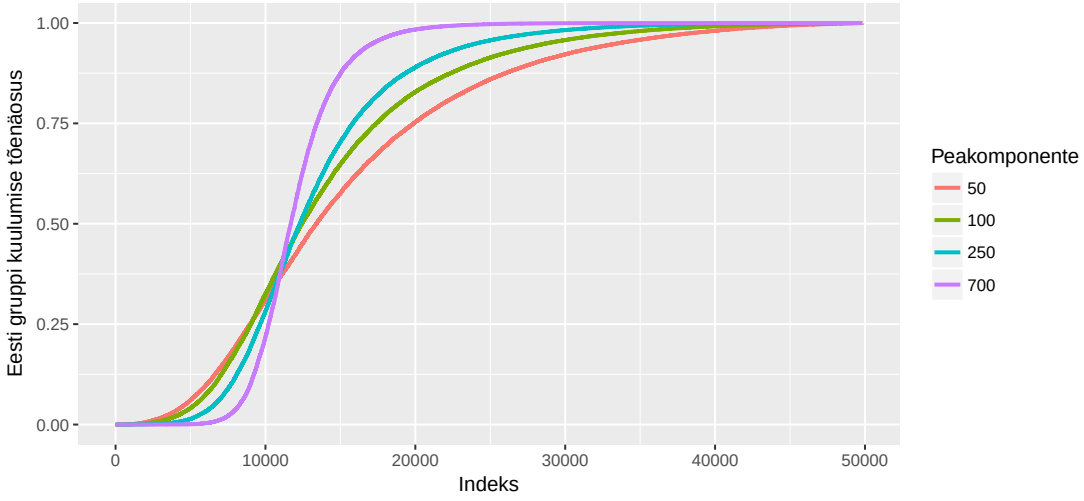
Võib öelda, et lineaarne diskriminantanalüüs saavutab antud võrdluses parimaid tulemusi. Diskriminantanalüüs vajab heaks tulemuseks juba vähe tunnuseid ning teised meetodid jäävad LDAle alla keskmise täpsuse mõttes mis tahes peakomponentide arvu korral. Tabelis 1 on välja toodud viis täpsemat meetodit.

Et LDA on robustsem võimaliku peakomponentide arvu valiku suhtes ning

eelneva ristvalideerimise korral maksimaalne keskmine täpsus tekkis just diskriminantanalüüsi rakendamisel, siis ka edasises keskendutakse rahvuse prognoosimisel lineaarse diskriminantanalüüsi rakendamisele. Pigem tuleks eelistada väiksemat peakomponentide arvu, kui on tagatud juba maksimaalne saavutatav täpsus.

Tabel 1: Viis parema täpsusega meetodit

	Meetod	PC arv	Täpsus	$q_{0.025}$	$q_{0.975}$
1	LDA	355.0000	0.7078	0.6770	0.7345
2	LDA	285.0000	0.7063	0.6739	0.7390
3	LDA	495.0000	0.7058	0.6754	0.7382
4	LDA	215.0000	0.7053	0.6761	0.7331
5	LDA	565.0000	0.7029	0.6687	0.7415



Joonis 5: LDA-ga prognoositud eesti gruppi kuulumise tõenäosused sõltuvalt ennustamiseks valitud peakomponentide arvust

Tabelis 2 on esitatud põhivalimisse kuuluvate inimeste ($n = 49\,199$) sagedustabel inimeste enda raporteeritud rahvuste ning nendele prognoositud rahvuste vahel. On näha, et üldjoontes on prognoos ootuspärane, mida on eelkõige näha raporteeritud eestlaste ja venelaste prognoosimisest. Näiteks 87,8% raporteeritud eestlastest prognoositakse samuti eestlaseks, 73,6% raporteeritud venelastest prognoositakse venelasteks. Tendents leidub ka teiste raporteeritud gruppide puhul, kuid on nõrgem. Raporteeritud soomlastest 52,7% prognoositakse soomlaseks, raporteeritud lätlastest vaid 17,4% prognoositi lätlasteks. Halvasti on prognoositud ka näiteks raporteeritud leedukaid (15,4%) ning ühtegi poolakat näiteks ei õnnestunudki prognoosida. Paraku prognoositakse viimastel juhtudel inimene tihti venelaste sekka. Kui välja arvata raporteeritud eestlased, soomlased ja sakslased, siis ülejäänud raporteeritud rahvuste korral prognoositakse inimene enamasti kuuluma vene gruppi.

Muidugi täielikult ei ole võimalik veenduda raporteeritud rahvuse õigsuses

Tabel 2: Prognoositud rahvus ja raporteeritud rahvus, kasutades 200 peakomponenti ja LDAd

	Raporteeritud											
Prognoositud	1	2	3	4	5	6	7	8	9	10	11	12
Austria	54	51	9	0	0	4	2	5	0	0	0	18
Bulgaaria	0	1	0	0	0	0	0	0	0	0	0	7
Tšehhi	82	190	118	5	0	3	1	6	1	2	0	16
Taani	0	0	0	0	0	0	0	0	0	0	0	0
Eesti	35 051	1693	15	7	100	7	9	4	20	5	5	23
Põhja-Soome	0	0	0	0	0	0	0	0	0	0	0	0
Lõuna-Soome	176	25	0	0	119	0	2	0	0	0	0	15
Prantsuse	0	0	0	0	0	0	0	1	0	0	0	0
Põhja-Saksa	23	8	1	0	0	0	0	5	0	0	0	4
Lõuna-Saksa	14	11	5	0	0	1	0	4	0	0	0	4
Hollandi	0	0	0	0	0	0	0	0	0	0	0	0
Ungari	3	1	0	0	0	0	0	0	0	0	0	1
Põhja-Itaalia	0	0	0	0	0	0	0	0	0	0	0	0
Lõuna-Itaalia	0	0	0	0	0	0	0	0	0	0	0	0
Läti	166	9	0	0	0	0	0	0	12	0	0	0
Leedu	87	12	0	0	0	0	0	0	5	1	6	0
Poola	0	0	0	0	0	0	0	0	0	0	0	0
Vene	4273	5645	473	286	7	7	43	4	31	37	28	37
Hispaania	0	0	0	0	0	0	0	0	0	0	0	0
Rootsi	1	1	0	0	0	0	0	0	0	0	0	1
Šveitsi	2	14	3	3	0	23	0	0	0	0	0	45
Ühendkuningriigi	0	0	0	0	0	0	0	0	0	0	0	0
Kokku	39 932	7661	624	301	226	45	57	29	69	45	39	171

1-eesti, 2-vene, 3-ukrainlane, 4-valgevenelane, 5-soomlane, 6-juut, 7-tatarlane, 8-sakslane, 9-lätlane, 10-poolakas, 11-leedukas, 12-muu

ajaloolises mõttes, kuid kindlasti näitab see teatavat tendentsi. Ilmselt oleks sobivam kui ka muude raporteeritud rahvuste seas prognoositaks enam vaatlusi nimigrupi hulka. Täiendavat võimalust selle kindlustamiseks vaadeldakse edasises.

Joonisel 5 on välja toodud erinevate peakomponentide arvudega saadud hinnangud eesti gruppi kuulumise tõenäosusele. Siinkohal on prognoos tehtud kõikidele Geenivaramu andmetele. Jooniselt 5 on selge see, et erinev arv peakomponente ei käitu klassifitseerimise mõttes väga erinevalt, kuid erinevused on suured vastavates tõenäosustes. On näha, et pea 50 000 doonorist ligikaudu 12 500 ei määrata ilmselgelt eestlaseks (neil on tõenäosus alla 0.5 kuuluda eesti gruppi). LDA puhul oli joonisel 4 täheldada, et väga suurt erinevust rahvuse klassifitseerimise täpsuse mõttes peakomponentide arvudel ei ole. Seega jääb küsimus selle kohta, mis võiks olla sobiv peakomponentide arv, et lähendada tõenäosuseid võimalikult hästi.

4.3 Simulatsioonikatse tõenäosuse prognoositäpsuse hindamiseks

Eelnevalt jäi selgusetuks, milline peakomponentide arv võib sobida lähendama tõenäosust kõige paremini. Näiteks pakub huvi see, et kui inimesel on üks vanematest eestlane ja teine mõnest muust rahvusest, kas siis hinnatakse tema eesti rahvusgruppi kuulumise tõenäosuseks ligikaudu 0,5 (mis oleks õige), või on tõenäolisem mingi muu tulemus. See teadmine oleks vajalik selleks, et mõista, kui hästi saab nende analüüsitulemuste põhjal hinnata inimese vanemate ja vanavanemate päritolu. Et lahendada seda probleemi, otsustati tekitada teadaoleva tõenäosusliku päritoluga inimesi, kasutades simulatsiooni. Kuna valdavalt klassifitseeritakse Eestiga seotud inimesi, siis simulatsioonis kasutati referentsi-

dena eesti ja eestile lähedasi gruppe. Teostati järgmine protseduur:

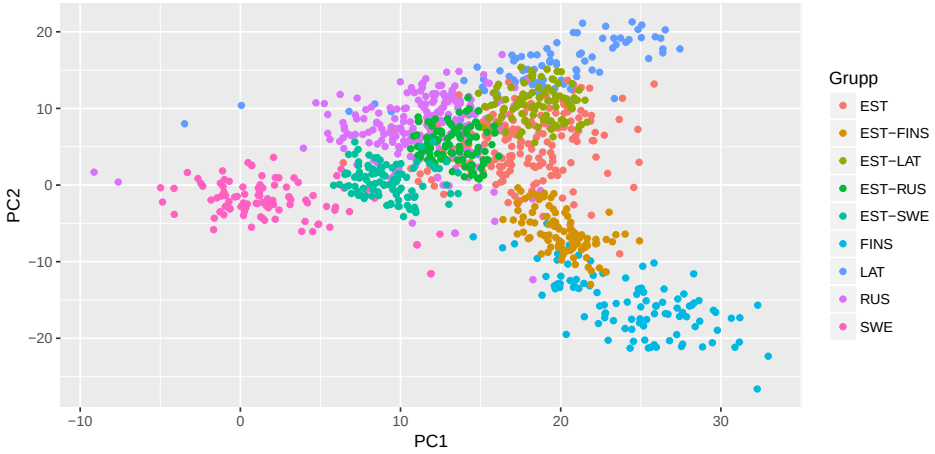
1. Valiti välja olemasolevaist eesti, läti, soome, rootsi ja vene referentsandmetest välja sellised, mille tõenäosus kuuluda vastavasse rahvusgruppi oli suurem kui 0,95. Ennustuse tegemiseks kasutati 200 peakomponenti. Igast grupist valiti juhuslikult viis inimest.

2. Valiti juhuslikult üks eestlane ja üks lätlane ning lähtudes valitud SNP andmetest tekitati sellele eestlasele ja lätlasele üks genotüüp, mis on tekitatud sama tõenäosusmudeli põhjal nagu tekiks nende indiviidide lapse genotüüp (vanema alleel antakse edasi lapsele tõenäosusega 0.5). Sammu korrati 100 korda. Täpsem kirjeldus pseudokoodis 1.

3. Sammu 2. korrati ka vanemate paari eesti-rootsi, eesti-soome, eesti-vene korral.

4. Saadud inimestele arvutati välja olemasolevate andmete põhjal peakomponendid.

Seega saadi neli korda 100 genotüüpi, kes peaksid olema ligikaudu tõenäosusega 0,5 eestlased ja tõenäosusega 0,5 mingist muust vastavast rahvusest. Jooniselt 6 on näha, et üldjoontes on tulemus tõepoolest ootuspärane: uued populatsioonid näivad vähemalt kahe peakomponendi mõttes kuuluvat lähtereferentside vahele.



Joonis 6: Referentsvalimid ning neist simuleeritud järglaspopulatsioonid

Andmed: Eesti ja muu referentspopulatsiooni SNP vektorid (iga element on 0, 1 või 2), mõlemast viis. m on SNP vektori pikkus
Tulemus: Järglaspopulatsiooni 100 SNP vektorit X_3
for $i = 1$ **to** 100 **do**
  Vali juhusliku eestlase SNP vektor X_1;
  Vali juhusliku muu rahvuse SNP vektor X_2;
  Tekita abitulemuste vektor Y_1;
  Tekita abitulemuste vektor Y_2;
  for $j = 1$ **to** m **do**
    if $X_{1,j} = 1$ **then**
      $Y_{1,j} \leftarrow Z \sim Be(0.5)$;
      ▷ Juhuslik suurus Z on Bernoulli jaotusega
    else
      $Y_{1,j} \leftarrow \frac{X_{1,j}}{2}$;
    end
    if $X_{2,j} = 1$ **then**
      $Y_{2,j} \leftarrow Z \sim Be(0.5)$;
    else
      $Y_{2,j} \leftarrow \frac{X_{2,j}}{2}$;
    end
  end
  $X_3 = Y_1 + Y_2$;
end

Pseudokood 1: Järglaste genereerimine

4.3.1 Rahvusgrupi prognoosimine 0,5-0,5 järglaspopulatsioonis

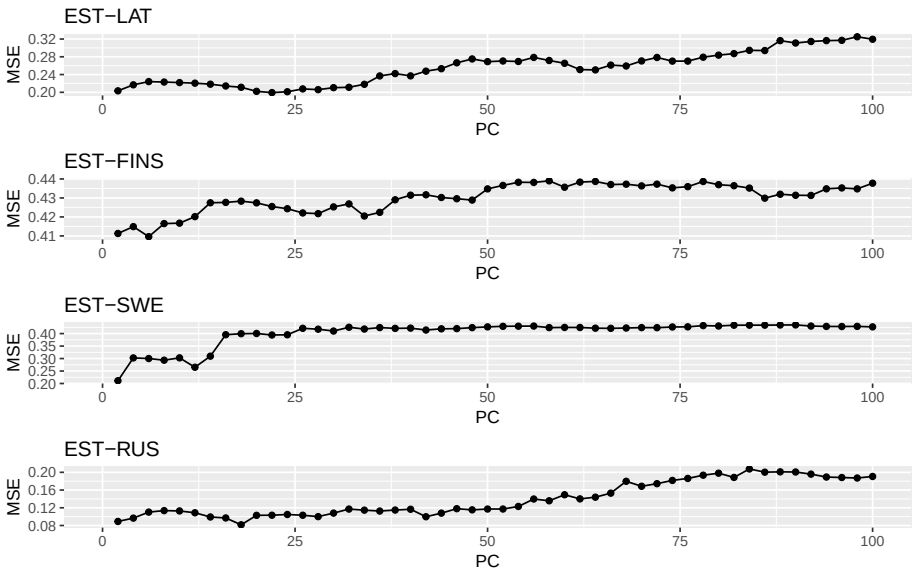
Järgnevalt kasutame LDA mudelit ning uurime, milline peakomponentide arv annaks prognoosina tõele lähedasemaid tõenäosuseid. Meenutagem, et andmed on genereeritud olukorras, kus n-ö õige hinnang inimese tõenäosuslikule päritolule vastavalt eesti ja mingist muust rahvusgrupist inimesele on 0.5-0.5.

Tulemused leitud tõenäosuste jaotumisest on esitatud jooniseil 25,26,27 ja 28. Tegemist on teatavas mõttes muretava tulemusega, sest paljudel juhtudel ei ole valdav osa tõenäosusmassist koondunud 0,5 ümbrusesse vaid kuhugi mujale. Kõikide rahvusgruppide puhul on täheldada, et peakomponentide arvu suurenedes saavutavad mudelid teatavas mõttes "enesekindlust": prognoositakse pigem kas väikest või suurt tõenäosust kuuluda vastavalt eesti või mingisse muusse gruppi. Eriti hästi on tendents näha rootsi ja lõunasoome puhul, kus 700 peakomponendi juhul on tõenäosusmassid koondunud äärtesse.

Teatavas mõttes parem on pilt läti ja vene gruppide puhul, kus on samuti täheldada äärtesse koondumist, ent suundumus pole nõnda tugev. Väiksema peakomponentide arvu puhul võib mõnel juhul (näiteks vene 5, 10, 50) pidada tulemust pigem heaks. Tulemus annab märku sellest, et suur kasutatavate peakomponentide arv ei pruugi olla hea mõte, sest võib viia ülesobitamiseni.

Kõrvutades eelmiste jooniste tulemusi joonisega 6, võib märgata huvitavat tõika. On täheldada, et lõunasoome ja rootsi grupid, mille vastavad 0.5-0.5 järglaste tõenäosusmassid koondusid äärtesse, on ka visuaalselt eesti grupist kaugemal. Eesti referentsvalim näib pigem hästi eristuvat rootsi või lõunasoome grupist, ühtlasi ka vastavad eesti-lõunasoome ja eesti-rootsi järglased jäävad kaugele klastrite keskmeist. Samas on läti ja vene referentsid kahe peakompo-

nendi alusel eesti grupile märksa lähemal ning vastavad 0.5-0.5 järglased on surutud tihedalt olemasolevate eesti ning läti või vene referentside vahele. Näib, et järglaste selline paiknemine tingib ka selle, et ennustatakse enam väärtuseid mitte äärtest vaid keskelt, mis on omakorda tõele lähedasem tulemus.



Joonis 7: Ruutkeskmised vead sõltuvalt peakompoentide arvust ja järglasgruppide vanemate rahvusest; 0,5 eesti-0,5 muu

Seega võib järeldada, et korrektsemaks prognoosimiseks peaksid referentsgrupid olema lähedasemad. Juba eesti-lõunasoome ja eesti-rootsi järglaste puhul on näha, et mudeli otsustus ei ole väga sobiv, üks põhjus selle taga võib olla see, et piirkonnas, kuhu eesti-lõunasoome ning eesti-rootsi järglased sattusid, oli äärmiselt vähe või üldse mitte vaatlusi. Teine märksõna on pigem väike peakomponentide arv, sest näib, et peakomponentide arvu suurendamine tekitab ohtu ülesobitamiseks.

Joonisel 7 on välja toodud prognooside ruutkeskmised vead sõltuvalt valitud peakomponentide arvust. Joonise lõunasoomlastega osa ei ole väga informatiivne, sest mis tahes arv peakomponente on tekitanud suure vea. Vene ja läti joonistel tekivad miinimumid sarnastel kohtadel: üks kahe peakomponendi juhul ning teine 20 peakomponendi ümbruses. Rootsi puhul tekib parim tulemus, kasutades kahte peakomponenti.

On selge, et ükski tulemus ei täitnud ootusi jõuda valdavalt 0,5-le lähedaste tõenäosusteni. Üks põhjus selle taga on kindlasti mõndade gruppide liiga suur eraldatus ning seda probleemi üritatakse järgnevas ka adresseerida.

4.3.2 Rahvusgrupi prognoosimine 0,75-0,25 järglaspopulatsioonis

Järgneva osa idee seisneb selles, et vähendada referentsgruppide kaugusi teineteisest. Üks lihtne lahendus selleks on kaasata referentspopulatsioonide hulka ka eelmises alapeatükis loodud 0,5-0,5 järglaspopulatsioonid.

Testandmestikuks tekitatakse analoogiliselt koodiga 1 igast rahvusest 100 järglast, kelle üks vanem on eesti ning teine vanem pooleldi eesti ja pooleldi

mingit muud päritolu. Kokkuvõttes peaks tõene lähedane tulemus olema, et eesti päritolu tõenäosus on 0,75 ja mingi muu päritolu tõenäosus 0,25. Ka uutele testandmetele leitakse vastavad peakomponendid ning seejärel prognoositakse LDA mudeli abil, mille referentspopulatsioonide hulka on nüüd arvatud ka 0,5-0,5 järglaspopulatsioonid, ka 0,75-0,25 järglaspopulatsioonide tõenäosused. Märgime, et lõpliku eesti gruppi kuulumise tõenäosuse arvutame kui

$$P_{ESTfinal} = P_{ESTinit} + \frac{1}{2} P_{EST-MUU_{init}},$$

kus $P_{ESTinit}$ on algsesse referentsgruppi kuulumise tõenäosus ning $P_{EST-MUU_{init}}$ on lisatud eesti-muu 0,5-0,5 järglaspopulatsioon. Analoogiliselt saab lõpliku tõenäosuse arvutada ka läti, lõunasoome, vene ja rootsi gruppidele.

Eelkirjeldatud mudeli kohaselt saadud tõenäosuste prognooside jaotused on esitatud jooniseil 29,30,31ja 32. Jooniseilt nähtub, et tulemused on pisut paremad kui 0,5-0,5 järglaste prognoosimisel saadi, st lähedasemad väärtustele 0,75 ja 0,25. Huvitava tendentsina on täheldada, et peakomponentide arvu suurenedes koonduvad prognoosid valdavalt 0,5 ümbrusesse. See näitab seda, et suure peakomponentide arvu korral määratakse inimesed enamasti lisatud 0,5-0,5 referentsgruppidesse.

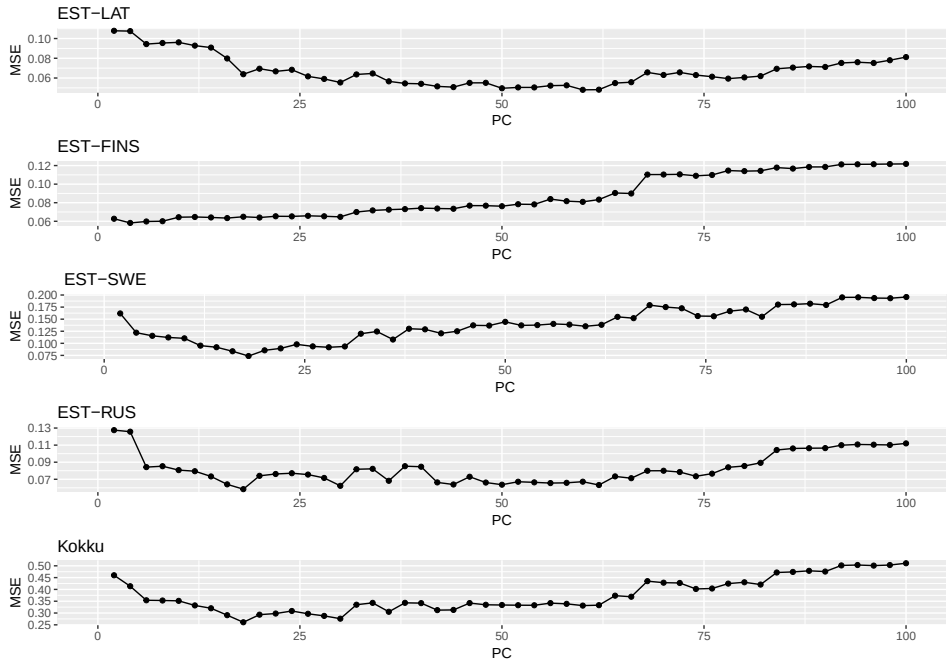
Jooniselt 8 on näha, et ruutkeskmiste vigade mõttes on käesolevad prognoosid täpsemad kui oli ainult 0,5-0,5 järglaspopulatsiooni prognoosi tulemused. Antud juhul võib öelda, et prognoosid on muutunud sisulisemaks ka lõunasoome ja rootsi gruppide lõikes, kuna väga suuri MSE väärtusi enam ei esine.

Paraku ei anna joonis 8 ühest vastust küsimusele, milline võiks olla sobiv peakomponentide arv, kõikide rahvuste lõikes. Erinevatele rahvustele näib sobivat erinev arv peakomponente: näiteks läti grupil on parim tulemus 50 ümbruses, rootsi ja vene gruppidel on parim tulemus 18 ümbruses ning lõunasoomlaste puhul on parim tulemus väga väikse arvu, näiteks 4 puhul. Samas, arvutades kõikide gruppide pealt ühise ruutkeskmise vea, on näha, et sobivaimaks osutub 17 peakomponendi kasutamine. Täpsemad tulemused ühise ruutkeskmise vea osas on esitatud tabelis 3.

Tabel 3: Kümme parima täpsusega peakomponentide arvu 0,75-0,25 järglaste ennustamisel

	MSE	PC arv
1	0.2598	17
2	0.2609	18
3	0.2761	30
4	0.2792	29
5	0.2835	19
6	0.2878	28
7	0.2909	16
8	0.2920	27
9	0.2931	20
10	0.2941	21

Võib väita, et mudelile vastavate referentsvalimite 0,5-0,5 järglastega laiendamine oli õigustatud. Ruutkeskmised vead vähenesid ja leitud tõenäosused jaotusid tõepärasemalt. Seega näib kehtivat tõdemus, et korrektse prognoosi tagamiseks peaks prognoositava vaatluse piirkonnas siiski leiduma piisavalt refe-



Joonis 8: Ruutkeskmised vead sõltuvalt peakomponentide arvust ja järglasgruppide vanemate rahvusest; 0,75 eesti-0,25 muu

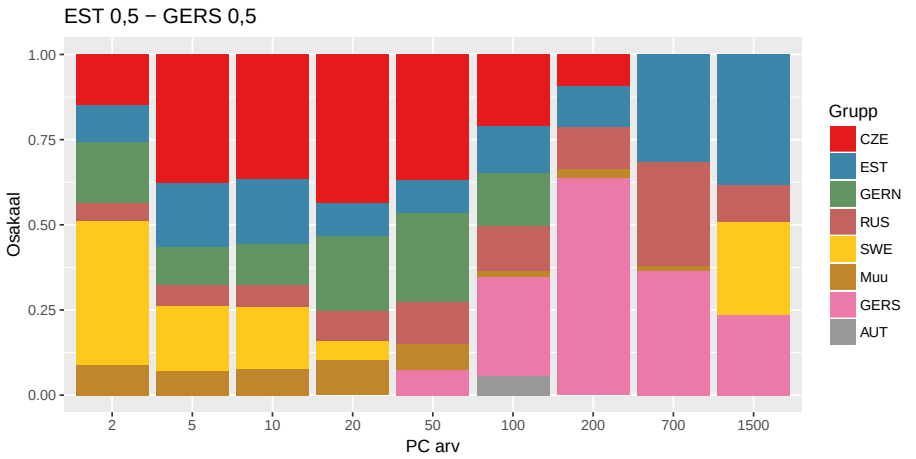
rentspopulatsiooni elemente. Seejuures vähemalt 0,75-0,25 järglaspopulatsiooni puhul osutus sobivaimaks valida peakomponentide arvuks 17.

4.3.3 Rahvusgrupi prognoosimine kaugemate rahvuste puhul

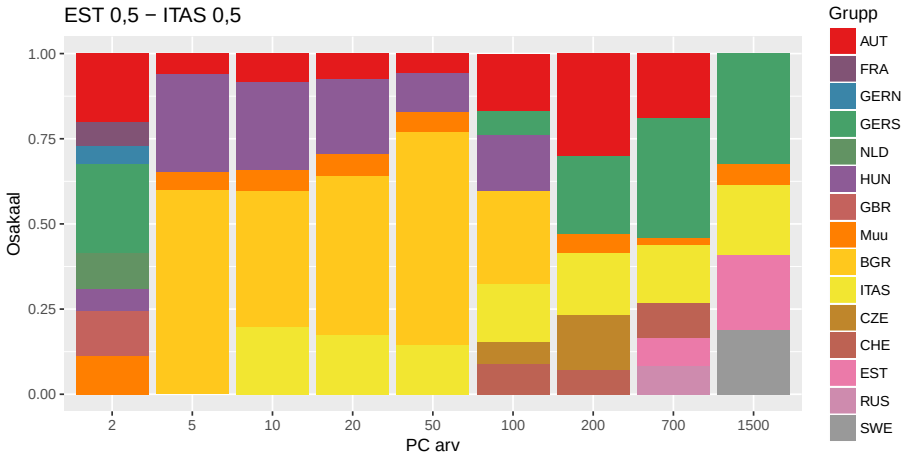
Eelmises alapeatükis järeldati, et kaasates mudeli referentspopulatsioonide hulka ka 0,5-0,5 järglaspopulatsioonid lähedasematest gruppidest, on võimalik tõsta prognoositäpsust. Ideaalvariandina võiks pakkuda, et nõnda võiks tekitada ka kõikide teiste rahvusgruppide omavahelised 0,5-0,5 järglased. Paraku viib see lähenemine absurdini, sest kokku oleks vaja nõnda tekitada juurde $\binom{22}{2} = 231$ referentsvalimit. Kui isegi piirduda vaid eestlasi sisaldavate uute 0,5-0,5 referentsvalimitega, mida on 22, jääb oht, et uued tekkivad referentsvalimid on väga sarnased omavahel või olemasolevate referentsvalimitega. Näiteks võib juhtuda, et eesti-lõunasaksa järglaspopulatsioon on sarnane põhjasaksa algse referentspopulatsiooniga. Parim variant on see, kui mudel töötaks juba algusest peale nõnda, et poleks vaja seda täiendada lisareferentsidega.

Kontrollimaks seda, kui hästi mudel prognoosib 0,5-0,5 järglaspopulatsioone, kelle vanemad on gruppidest, mille distants on suurem, tekitame järglaseid eestlaste ja lõunaitaallaste ning eestlaste ja lõunasakslaste vahel. Skeem on analoogiline sellele, mida on kirjeldatud koodis 1. Jällegi valitakse vanempopulatsiooni rahvuse esindajad, kes prognoositakse antud gruppi vähemalt tõenäosusega 0,95.

Joonistel 9 ja 10 on esitatud keskmised prognoositud tõenäosused sõltuvalt peakomponentide arvust. Keskmised tõenäosused, mis jäid alla 0,05 summeeriti kokku gruppi Muu. On näha, et väike arv peakomponente ei anna oodatule lähedaseid tulemusi. Vaadates täpsemalt joonist 9 lõunasaksa päritolu inimes-



Joonis 9: Keskmised prognoositud osakaalud 0,5-0,5 eesti-lõunasaksa järglastel sõltuvalt peakomponentide arvust



Joonis 10: Keskmised prognoositud osakaalud 0,5-0,5 eesti-lõunaitaalia järglastel sõltuvalt peakomponentide arvust

tega, on näha, et kahte kuni 20 peakomponenti kasutades ei tuvastata üldse lõunasaksa päritolu ning eesti grupi tõenäosust hinnatakse küllalt väikseks. Sarnane tulemus on näha ka lõunaitaalia päritoluga inimeste prognoosimisel joonisel 10, kus kahe või viie peakomponendi puhul ei tuvastata praktiliselt üldse eesti või lõunaitaalia päritolu. Selle asemel on huvitav näha, et mudel prognoosib väikese peakomponentide arvu korral inimesi enamasti geograafiliselt kuhugi vahepealsetesse või naabrusesse kuuluvatesse gruppidesse. Näiteks, kasutades 2-50 peakomponenti, prognoositakse lõunaitaalia-eesti järglaste puhul austria, ungari ja bulgaaria gruppi kuulumise tõenäosuseid, mis ilmselgelt ei ole geograafiliselt lähedal ei lõunaitaallastele ega ka eestlastele.

Huvitaval kombel on saadud joonistelt näha ka positiivseid tulemusi sellest, kui kasutada ennustamiseks suuremat arvu peakomponente. Lõunaitaalia päritolu järglaste puhul võib märgata, et alles peakomponentide arvu 700 ja 1500

puhul hakatakse ennustama piisavalt suuri eesti gruppi kuuluvaid tõenäosusi. Osutub, et eelneva joonise 10 põhjal otsustades annab tõepäraseima tulemuse lõunaitaalia päritolu inimestele 1500 peakomponenti; lõunasaksa päritolu inimestele 700 peakomponenti. Kahjuks mõlemal juhul tulemused ei ole siiski väga lähedased tõesele tulemusele 0,5-0,5 ning prognoositud tõesnäosused jäävad ka parimal juhul neile oluliselt alla.

Paraku näib, et ei leidu üheselt määratud peakomponentide arvu, mis sobiks ühtlaselt mis tahes juhtumil küllalt hea prognoosi saamiseks. Olukordades, kus vaatlused on osaliselt eesti päritoluga, on võimalik saavutada küllalt hea täpsus, kasutades prognoosiks ligi 20 peakomponenti. Samas, eristamaks kaugemaid gruppe on ikkagi vajalik märksa suurem peakomponentide arv. Võimalik selgitus tulemusele on see, et lähedaste gruppide korral viib väga suur peakomponentide arv ülesobitamiseni. Kaugemate gruppide puhul on ilmselt vaja mudelisse ka märksa rohkem informatsiooni, et jõuda sisukate hinnanguteni, ning seega ei jõuta ka suuremate peakomponentide arvude puhul nõnda ruttu ülesobitamiseni.

4.4 Tulemused modifitseeritud mudeli korral

4.4.1 Prognoositud tõenäosused raporteeritud rahvuseti

Eelnevast simulatsioonikatsest tehtud järelduste põhjal tehti mudelisse parandused ning saadud soovituste ja alusel tehtud prognoosid Geenivaramu doonorite kohta on esitatud järgnevalt. Tulemuste saamiseks kasutame tabelis 3 soovitatud 17 peakomponenti, mis järglastele, kes olid ligikaudu tõenäosusega 0,75 eesti päritolu ja 0,25 muud päritolu, andis väikseima ruutkeskmise vea.

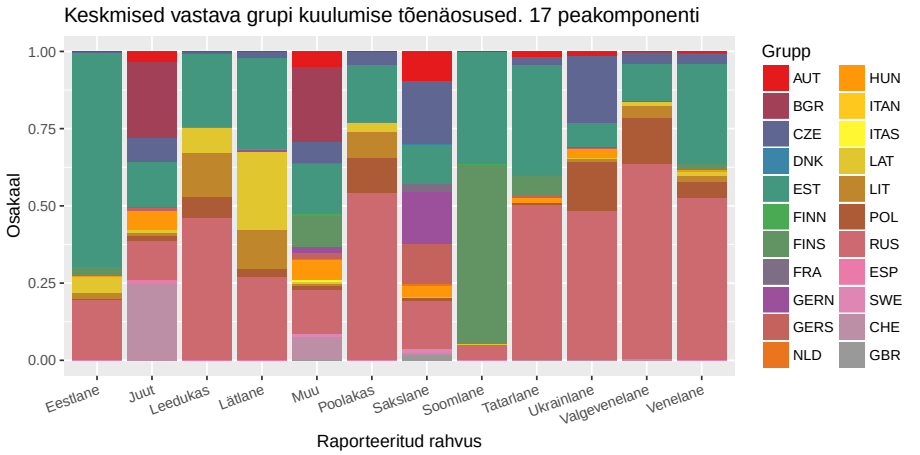
Joonisel 11 on välja toodud keskmised prognoositud tõenäosused rahvuse järgi. 17 peakomponendi korral on näha, et tulemused on pigem ootuspärased: valdavalt on keskmised tõenäosused suuremad kuuluda enda raporteeritud rahvusgruppi, ent tendents pole iga rahvuse puhul väga tugev. Selgesti on näha, et näiteks raporteeritud eestlaste või venelaste puhul on suurimad prognoositud tõenäosuste keskmised just vastavalt eesti või vene grupi puhul. Sarnane tulemus on näha ka raporteeritud soomlaste puhul, kus keskmine lõunasoome tõenäosus on suurim.

Huvitav on tulemus näiteks ise raporteeritud läti, leedu ja poola rahvuste puhul. Kõikidel juhtudel on täheldada suur vene ja eesti komponent, kuid märgatavalt on esindatud ka vastav raporteeritud rahvus. Teatavas mõttes on geograafilise läheduse mõttes ka ootuspärane, et lätlaste seas võib mudel anda ka mõningast leedu mõju, sarnaselt ka leedukate seas võib täheldada mõningast läti ja poola gruppi kuulumise tõenäosuse olemasolu.

Sakslaste puhul võib märgata, et suuremad keskmised tõenäosused on kuuluda austria, tšehhi, põhjasaksa ja lõunasaksa gruppi, mis on kõik geograafiliselt ootuspärased tulemused. Jällegi on täheldada arvestatav eesti ja vene grupi tõenäosuste olemasolu.

Teatavas mõttes on väga põnev jälgida, kuidas ennustatakse rahvused, mille esindajaid pole referentsidena otseselt välja toodud. Rõõmustav on tõdeda, et ka siinkohal on tulemused loogilised. Ukrainlaste puhul on suuremad keskmised tõenäosused prognoositud tšehhi, poola ja vene gruppi kuulumiseks; valgevenelaste puhul on suuremad keskmised tõenäosused prognoositud vene, poola ja ka leedu gruppi kuulumiseks. Tatarlased prognoositakse valdavalt eestlasteks või

venelasteks ning juutide puhul ei joonistu paraku selget suundumust, vaid üsna kaalukalt on esindatud bulgaaria, tšehhi, eesti, ungari, vene ja šveitsi grupp.



Joonis 11: Keskmised prognoositud tõenäosused kuuluda vastavasse gruppi rahvuseti, kasutades prognoosiks 17 peakomponenti

Väga huvitav on ka see, et pea kõikide raporteeritud rahvuste puhul on prognoositud tõenäosuste hulgas märkmisväärselt suured vene ja eesti tõenäosused. Sellele saab välja pakkuda kaks võimalikku seletust. Esiteks ongi võimalik, et geenidoonorid, kes on Eestiga seotud omavadki teatavat eesti või vene päritolu, sest seda päritolu peaks esinema Eestis enim.

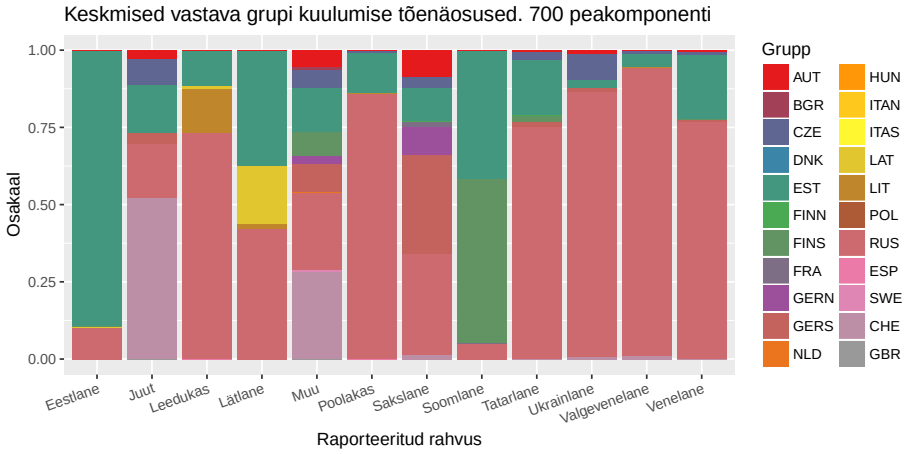
Samas, et suundumus näib nii massiline, on ka võimalik, et probleem tekib, kuna referentsvalimite maht ning esindatavus ei ole tasakaalus. Meenutagem, et eesti ja vene referentse on suurendatud tagamaks piisav võimaliku hajuvuse kirjeldatus nendes gruppides, kusjuures eriti vajalik oli see vene grupi jaoks, kus 100 inimesest koosnev referentsvalim osutus selgelt liiga väikseks. Näiteks esialgsetel tulemustel prognoositi valdav osa raporteeritud venelasi leedukaiks. Võis juhtuda, et venelaste referentspopulatsiooni suurendamisega prognoositakse nüüd ka mõned valdavalt leedu päritolu inimesed sagedasemini vene gruppi kuuluvaiks, millele võib viidata ka suur vene gruppi kuulumise tõenäosus raporteeritud leedukate puhul. Autor usub siiski, et vene ja eesti referentsvalimite suurendamine oli õigustatud, sest väga arvuka (7738 inimest) vene rahvusest doonorite korrektsem prognoos avaldab paremat mõju kui võimalik väike viga leedu rahvusest doonorite puhul, keda on kokku 116.

Selge on see, et täpsemate tõenäosuste saamiseks nii leedukate kui ka teiste rahvuste jaoks oleks kõige sobivam samm siiski suurendada tunduvalt referentsvalimite mahtu. Kui juba vene referentsi puhul on täheldada teatavaid probleeme võimaliku varieeruvuse kirjeldamisega vaid 100 inimesega, siis võib ilmselt arvata, et sama probleemiga tuleks kindlasti tegeleda ka näiteks prantsuse, ühendkuningriigi, hispaania ja poola referentsvalimite korral. Paraku käesoleva magistritöö raames seda küsimust adresseerida pole võimalik.

Võrdluse mõttes on hea vaadata ka joonist, kus on saadud tulemused kasutades 700 peakomponenti. Tulemused selle kohta on esitatud joonisel 12. On näha, et suurem varieeruvus erinevate ennustuste vahel, mida oli veel näha joonisel 11, on 700 peakomponendi prognoosi joonisel kadunud. Näiteks on muutunud

kaduvväikseiks tšehhi, poola ja leedu gruppi kuulumiste tõenäosused valgevenelastel ja ukrainlastel. Selle asemel ennustatakse neid pea täielikult vene gruppi kuuluvaiks. See näitab seda, et prognoosid muutuvad liialt enesekindlateks ning enam ei esitata reaalsuses leiduvat ebakindlust korrektselt. Ilmselt ei ole õige öelda ukrainlastele, et tegemist on peaaegu kindlasti venelastega eriti veel juhul, kui mõningane tšehhi ja poola gruppi kuulumise tõenäosus tegelikult viitaks kas või natuke sellele, et päritolu selle inimese puhul ei ole tegelikult nii lihtsalt kirjeldatav.

Seega, ehkki alapeatükis Rahvusgrupi prognoosimine kaugemate rahvuste puhul on viidatud, et kaugema päritoluga vanemate järglaste puhul võib olla suurest peakomponentide arvust kasu, siis esmapilgul tundub, et väiksem arv peakomponente sobib antud ülesande jaoks paremini. Väiksem arv ei kaota ära varieeruvust ning arvatavasti vähendab ka ohtu ülesobitamiseks. Ka järgnevas eelistame rahvusgrupi prognoosimisel kasutada 17 peakomponenti.



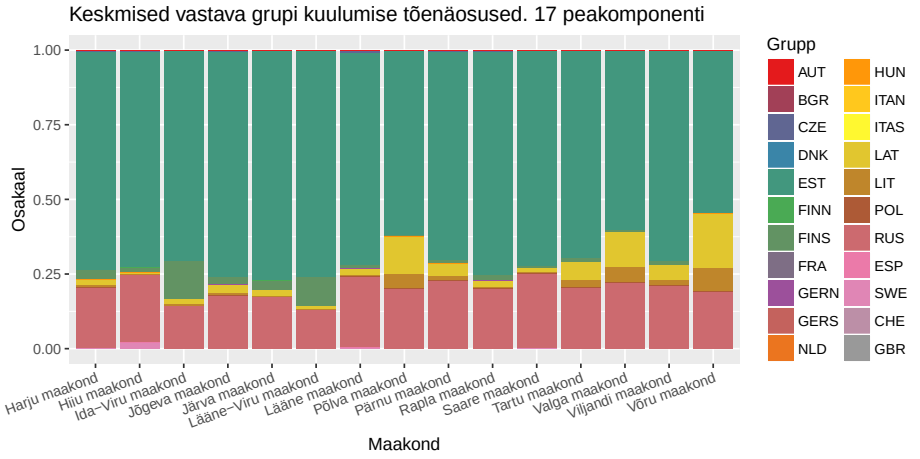
Joonis 12: Keskmised prognoositud tõenäosused kuuluda vastavasse gruppi rahvuseti, kasutades prognoosiks 700 peakomponenti

Eelnevatel joonistel anti prognoositud tõenäosustest ülevaade keskmistega, kuid lisaks on tähtis vaadata ka kuidas täpsemalt tõenäosused jaotuvad. Meenutagem, et alapeatükis Rahvusgrupi prognoosimine 0,5-0,5 järglaspopulatsioonis jõuti näiteks läti-eesti järglaste ja suure peakomponentide arvu korral olukorrani, kus näiteks nii eestlastele kui ka lätlastele prognoositi enamasti 0 või 1 lähedasi tõenäosuseid, mis keskmiselt üle populatsiooni annab küll 0,5-le lähedasi tulemusi, kuid indiviidi tasandil läheb selgelt mööda tõepärasest tulemusest. Tulemused tõenäosuste jaotumiste kohta on esituatud joonistel 33,34 ja 35, kusjuures iga raporteeritud rahvuse kohta on näidatud vaid need prognoositud rahvused, mille puhul keskmine tõenäosus jäi üle 0,01.

Hea on tõdeda, et kasutades rahvuse prognoosimiseks 17 peakomponenti, ei ole üldiselt tekkinud selliseid teineteisest eralduvaid alasid nagu eesti-läti 0,5-0,5 järglaspopulatsiooni prognoosimisel. Mõnede eranditega on valdavalt tegemist unimodaalsete jaotustega. Joonistelt nähtuvad tulemused on laias laastus sarnased tulemustega, mida on näidatud joonisel 11.

4.4.2 Rahvusgruppidesse kuulumise tõenäosused maakonniti

Järgnevalt uuritakse, kas ja kuidas on prognoositud tõenäosus kuuluda ühte või teise rahvusgruppi seotud inimese enda poolt raporteeritud sünnimaakonnaga. Siin uuritakse 33 710 inimese andmeid, kes on Geenivaramuga liitumisel andnud teada, et nad on sündinud Eestis ja nimetanud ka oma sünnimaakonna. Märgime, et nüüd ja edaspidi, rääkides maakondadest ja valdadest, mõeldakse maakondi ja valdu nende piirides enne 2017. aasta haldusreformi. Järgnevalt on joonisel 13 esitatud keskmised tõenäosused maakondade kaupa. Joonisel on kajastatud vaid doonorid, kes on end raporteerinud eestlastena, sest arvatavasti vaid selliste inimeste kuvamine joonisel võib midagi loogilist anda ka interpretatsiooni mõttes.



Joonis 13: Keskmised prognoositud tõenäosused kuuluda vastavasse gruppi maakonniti, kasutades prognoosiks 17 peakomponenti

Joonisel on näha, et pea kõikides maakondades on suurim tõenäosus kuuluda eesti gruppi, mis on eestlaste puhul ootuspärane. Ka ülejäänud tulemused on geograafiliselt loogilised. Lätile lähedasemates maakondades nagu Võru, Valga ja Põlva on täheldatav suurem Läti komponent; Võrumaal ka teatav Leedu komponent. Viimane tulemus võib tuleneda ka sellest, et Läti referentsvalim on olnud liialt väike ning seega on Võru maakonna inimesi ka määratud mõneti suurema tõenäosusega leedu gruppi kuuluvaks. Teistega maakondadega võrreldes on lõunasoome gruppi kuulumise tõenäosus on suurem Ida- ja Lääne-Viru maakondades. Samuti on pea ainukesed kohad, kus on näha väikest rootsi gruppi kuulumise tõenäosust, Hiiu- ja Läänemaa, mis on samuti ajalooliselt mõistetav tulemus.

5 Tulemused II: Eesti-sisese päritolu hindamine

Üks magistritöö eesmärke on rahvusliku päritolu kirjeldamine. Ent saadavate andmete rohkuse tõttu on ilmselt olemas suutlikkus kirjeldada ka detailsemalt inimeste päritolu Eesti-siseselt.

Antud peatükis kasutatakse prognoosimise lähteandmeiks jälle peakomponente, mis on arvutatud SNPde andmetest. Samas pole siin kasutatavad peakomponendid enam täpselt samad, mis peatükis Tulemused I: päritolu hindamine rahvuse tasandil, kus peakomponendid olid leitud üleeuroopaliste referentsandmestike pealt, vaid peakomponendid on arvutatud ainult Geenivaramu doonorite andmete pealt. Peakomponentide hindamiseks kasutati kõiki SNPe, midagi ei eiratud nagu on kirjeldatud alapeatükis Genotüübiandmete esmane töötlus rahvuse klassifitseerimisel, kusjuures peakomponendid on arvutatud sugulusmaatriksi pealt. Ülesande mahukuse tõttu leiti peakomponendid kasutades PLINK 2.0 tarkvara ning sealt väljastati 100 peakomponenti, millele järgnev analüüs tugineb.

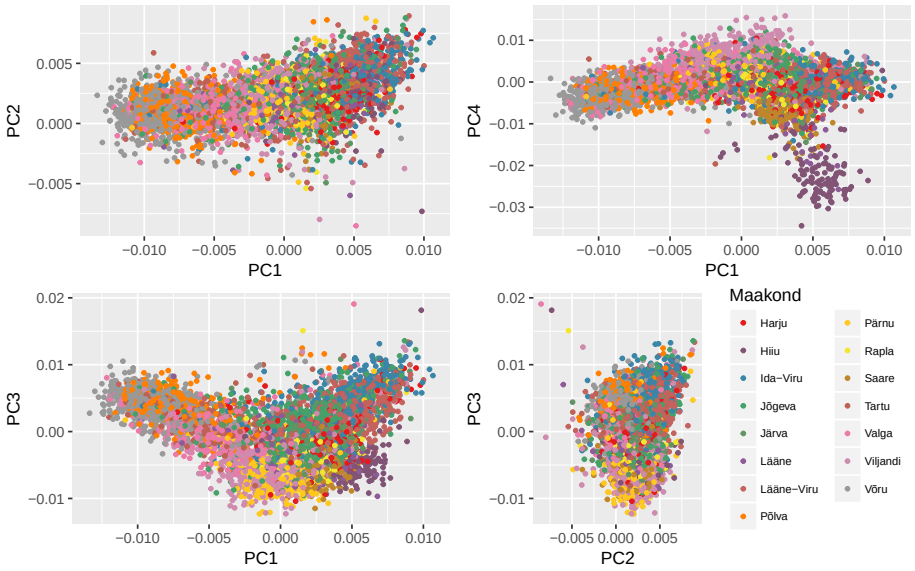
Etteruttavalt peab mainima, et peakomponentide leidmine just nõnda, kasutades vaid Geenivaramu doonoreid, osutub Eesti-sisese klassifitseerimise seisukohalt väga oluliseks. Töö käigus võrreldi, kuidas suudavad eristada rahvusvaheliste referentspopulatsioonide alusel leitud peakomponendid ja Geenivaramu doonorite pealt leitud peakomponendid inimesi Eesti-siseselt, ning osutus, et eelkirjeldatud uus meetod annab märksa loogilisemaid ja interpreteeritavamaid tulemusi.

5.1 Prognoos maakonna alusel

Üks lihtne võimalus kirjeldamaks inimeste päritolu on klassifitseerida inimeste sünnimaakonda analoogiliselt sellele, nagu oli tehtud peatükis Tulemused I: päritolu hindamine rahvuse tasandil. Erinevuseks on siinkohal see, et gruppidena vaadeldakse 15 Eesti maakonda. Autori bakalaureusetöös [2] on samuti kirjeldatud, et maakondade vahel leidub peakomponentide mõttes märkimisväärseid erisusi. Joonis 13 näitab, et maakondade vahel leidub keskmiste rahvuslike prognooside mõttes erinevusi. Ühtlasi on joonisel 14 näidatud, kuidas peakomponendid seletavad maakondlikku päritolu. Esimene peakomponent määrab eelkõige ära põhja-lõuna telje: väikestele esimese peakomponendi väärtustele vastavad enamasti Võru- või Põlvamaa vaatlused ning suurtele esimese peakomponendi väärtustele vastavad Virumaa vaatlused. Lisaks on näha, et ka järgnevad peakomponendid võivad osutuda tähtsaks klasside eristamisel. Näiteks neljas peakomponent eristab ülejäänud klassidest selgelt Hiiu maakonna. Seega võib maakondadesse klassifitseerimine anda häid tulemusi.

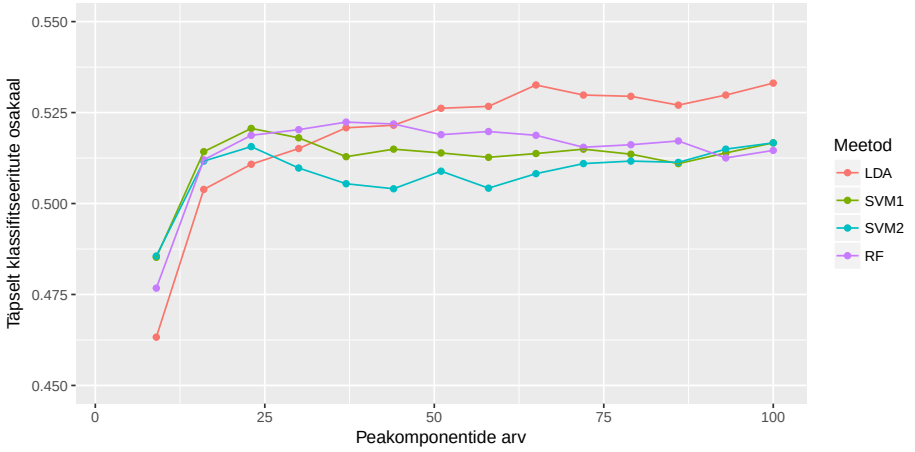
5.1.1 Meetodite võrdlus maakonna ennustamisel

Sarnaselt alapeatükiga Klassifitseerimismeetodite võrdlus teostatakse ka siin referentsandmestikule 10-jaotuse ristvalideerimine ning ühtlasi on valitud parameetrid juhuslikule metsale ning tugivektormasinatele samad. Joonisel 15 on võrreldud erinevaid klassifitseerimismeetodeid ning nende täpsusi. Võrreldes rahvusgruppide vahel klassifitseerimisega joonisel 4, on esimesena täheldada, et klassifitseerimise täpsus on märkimisväärselt langenud. Tulemus on tunduvalt halvem, arvestades, et klasse, mille vahel antud ülesandes oli vaja teha otsus,



Joonis 14: Referentsandmestik Eesti-sisese päritolu kirjeldamiseks erinevate peakomponentide lõikes maakondade kaupa

oli siinkohal seitsme võrra vähem. Üks võimalik seletus sellele on see, et klasside vaheline eralduvus ongi maakondade klassifitseerimisel halvem. Parimad tulemused koos *bootstrap* usaldusintervallidega on esitatud ka tabelis 4.



Joonis 15: Klassifitseerimismeetodite täpsused 10-jaotuse ristvalideerimise korral sõltuvalt valitud peakomponentide arvust

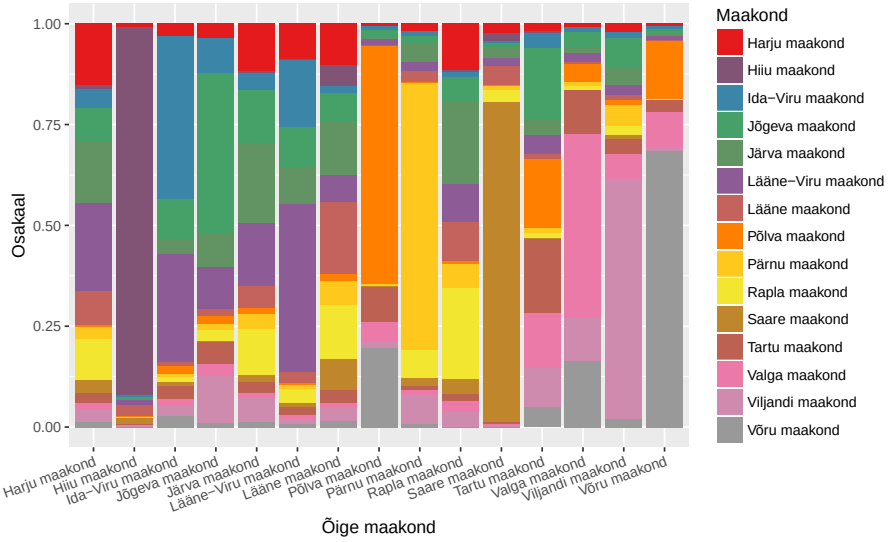
Välja joonistub jällegi tendents, et väikse peakomponentide arvu korral on prognoosid ebatäpsed, kuid kiiresti saavutatakse n-ö platoo, misjärel täpsus väga palju ei parane peakomponentide arvu kasvades. Meetodite võrdluses saavutab parima tulemuse ka siin lineaarne diskriminantanalüüs, kusjuures parimad täpsused on saavutatud 100 ja 65 peakomponenti kasutades.

Tabel 4: Viis parema täpsusega meetodit maakondade prognoosimisel

	Meetod	PC arv	Täpsus	$q_{0.025}$	$q_{0.975}$
1	LDA	100	0.5331	0.5248	0.5423
2	LDA	65	0.5326	0.5233	0.5434
3	LDA	72	0.5298	0.5194	0.5414
4	LDA	93	0.5298	0.5190	0.5397
5	LDA	79	0.5295	0.5195	0.5383

5.1.2 Prognoosid maakonniti

Et uurida täpsemalt, miks on maakonniti klassifitseerides täpsus keskpärane, vaadeldakse ennustatud tõenäosusi ka õigete maakondade lõikes. Prognooside saamiseks kasutati eelmise alapunkti tulemust, mille kohaselt parima tulemuse saab kasutades LDAd ja 100 peakomponenti. Teostati 10-jaotuse ristvalideerimine, et saada prognoosid kogu referentsandmestiku kohta. Tulemused keskmiste tõenäosustega õige maakonna kaupa on esitatud joonisel 16.



Joonis 16: Keskmised ennustatud tõenäosused õige maakonna kaupa, kasutades 100 peakomponenti ja LDAd

Paljudes maakondades pole tulemused halvad eriti arvestades seda, et mitmel juhul on valesti määratud maakondade näol tegemist õige maakonna naabermaakonnaga. Näiteks on Ida-Virumaal prognoositud suure tõenäosusega vaatlusi kuuluma Lääne-Virumaale. Sarnaseid ilminguid leidub veel palju, näiteks Võru maakonnast pärit inimestele prognoositakse ka arvestatava tõenäosusega Põlva või Valga maakonda kuulumist.

Joonise mõttes jagunevad maakonnad sisuliselt kaheks. Ühte tüüpi maakondades saavutatakse prognoosimine vastavasse maakonda kõrge tõenäosusega. Nii on näiteks Hiiu, Põlva, Pärnu, Saare, Valga, Viljandi ja Võru maakonnas ja pigem ka Viru maakondades. Teise rühma jäävad Harju, Jõgeva, Järva, Lääne,

Rapla ja Tartu maakonnad. Parema tõenäosusega esimene rühm koosneb valdavalt piiriäärsetest maakondadest ning halvema tõenäosusega teine rühm koosneb Sise-Eesti maakondadest aga ka suurematest tõmbekeskustest Harju ja Tartu maakonnast.

Täiesti rahule ei saa jääda tulemustega Sise-Eesti maakondades ja Harju ja Tartu maakonnas. Eriti halb on kirjeldatus Harjumaal, kus peaaegu võrdvääselt on esindatud Harju, Järva, Lääne-Viru, väiksemal määral ka Lääne maakonda kuulumise tõenäosused. Tulemus viitab sellele, et sellise geneetiliselt eristuva entiteedi nagu Harju maakond olemasolu on äärmiselt kaheldav, kui harjumaalasi ei suudeta hästi tuvastada. Sarnane küsimus jääb ka teiste Sise-Eesti maakondadega, kus prognoosid ei ole osutunud nii täpseiks.

Sellise probleemi tekkmine võib mitmel põhjusel olla ootuspärane. Esiteks ei pruugi maakond olla sisuliselt tähenduslik mõiste, kuna praegune haldusjaotus on kujunenud välja ümberkorralduste tulemusena Eesti NSVs, saavutades mõnede eranditega tänase 15 maakonnaga kuju aastal 1964 [20]. Teiseks, tulemus võib ka tähendada, et maakonnad pole sisuliselt parimad eritlejad, sest inimeste liikumine võis mitmel puhul olla maakondadevaheline ning tihti on oldud seotud näiteks murrete kaudu paremini naabermaakondadega. Näitena võib tuua kirderannikumurde, mida kõneldi ja vähesemal määral kõneldakse Põhja-Eesti rannikualadel Jõelähtme kihelkonnast Narva -Jõesuuni [21]. See tähendab, et selle ala inimesed võivad olla sarnased, kuid et nad kuuluvad ka kolme erinevasse maakonda, siis ilmselt põhjustavad nad ka vigu maakonna alusel klassifitseerides. Siinkohal tuleb rõhutada, et kirderannikumurde piirkonna inimeste geneetiline sarnasus on esialgu kõigest hüpotees, kuid hüpoteesi korrektsus viitaks sellele, et maakondade alusel klassifitseerimine ei pruugi olla parim variant. On võimalik, et selliseid eristuvaid gruppe, mis ületavad maakondade piire, leidub veelgi enam. See omakorda näitaks, et kirjeldamiseks ja klassifitseerimiseks on vaja välja pakkuda midagi paremat kui maakondade valimine klassifitseeritvateks objektideks.

5.2 K-keskmiste klasterdamine

Eesmärk on peakomponentide alusel tuvastada Eesti-siseselt klastrid, mis seletaksid, millised piirkonnad Eestis on paremini eristuvad. Ideaaljuhul tekivad klastrid, kuhu kuulub ligilähedaselt võrdsel arvul inimesi ning mis moodustavad teatavas mõttes geograafiliselt loogilised tervikud. Vältida tuleks selliste klastrite teket, kuhu on jäänud vaid üks või kaks vaatlust, sest sellised klastrid ei ole piisavalt üldistusjõulised.

5.2.1 Andmete puhastamine K-keskmiste klasterdamise abil

Esialgsed eksperimendid K-keskmiste klasterdamise meetodit kasutades näitasid, et suurendades K väärtust, tekkisid õige pea klastrid, kuhu kuuluski vaid üks vaatlus. On selge, et sellise klastri tekkimine pole kooskõlas eesmärgiga saada ühtlase suurusega klastreid. Tõigast, et selline inimene eristus, võib samuti järeldada vastava indiviidi erandlikkust ning seega on ilmselt parem selline inimene referentsvalimist eemaldada. Protseduuri, mille alusel andmeid puhastati, on detailsemalt seletatud koodis 2.

Protseduuri tulemusena tekitati ülimalt 25 klastrit samaaegselt eemaldades inimesi, kes moodustasid klastri üksinda. Nõnda eemaldati algsest andmestikust

üheksa vaatlust. Eelnimetatud koodi alusel eemaldati kuus vaatlust ning kuna ühte klastrisse jäi lõpuks vaid 3 vaatlust, siis otsustati ka need eemaldada. 25 klastri puhul on klastrite suurused pigem tasakaalus, mida on näha ka tabelist 5.

Tabel 5: Kirjeldavaid statistikuid klastrite suuruse kohta 25 klastri korral

Min	$q_{0.25}$	Mediaan	Keskmine	$q_{0.75}$	Max
54	117	228	231	313	481

5.2.2 Klastrite arvu valik sõltuvalt peakomponentide arvust

Järgneva arutelu motiveerimiseks on vaadeldud kahte joonist. Siin ja edaspidi uurime tulemusi Eesti kaarti kujutavate jooniste kaudu, kus igale vallale vastav ala on värvitud vastavalt sellele, kui suur protsent selles vallas sündinud geenidoonoreid kuulub vastavasse klastrisse. Joonisel 17 on kujutatud nelja klastrit, võttes aluseks kaks peakomponenti, ja joonisel 18 on kujutatud nelja klastrit võttes aluseks sada peakomponenti. Joonistelt on selgesti näha, et peakomponentide arv mõjutab tõsiselt klastrite kujunemist. Kasutades vaid kaht peakomponenti, tekkisid vähemasti territoriaalselt võrdväärsed klastrid, ent kasutades sadat peakomponenti eristab meetod eraldiseisvana väikse Hiiumaa klastri (klaster 4 joonisel 18). Kasutades kaht peakomponenti, on klastrite suurused vastavalt 1144, 1106, 2178 ja 1228; kasutades sadat peakomponenti, on klastrite suurused vastavalt 2679, 1376, 1452 ja 149.

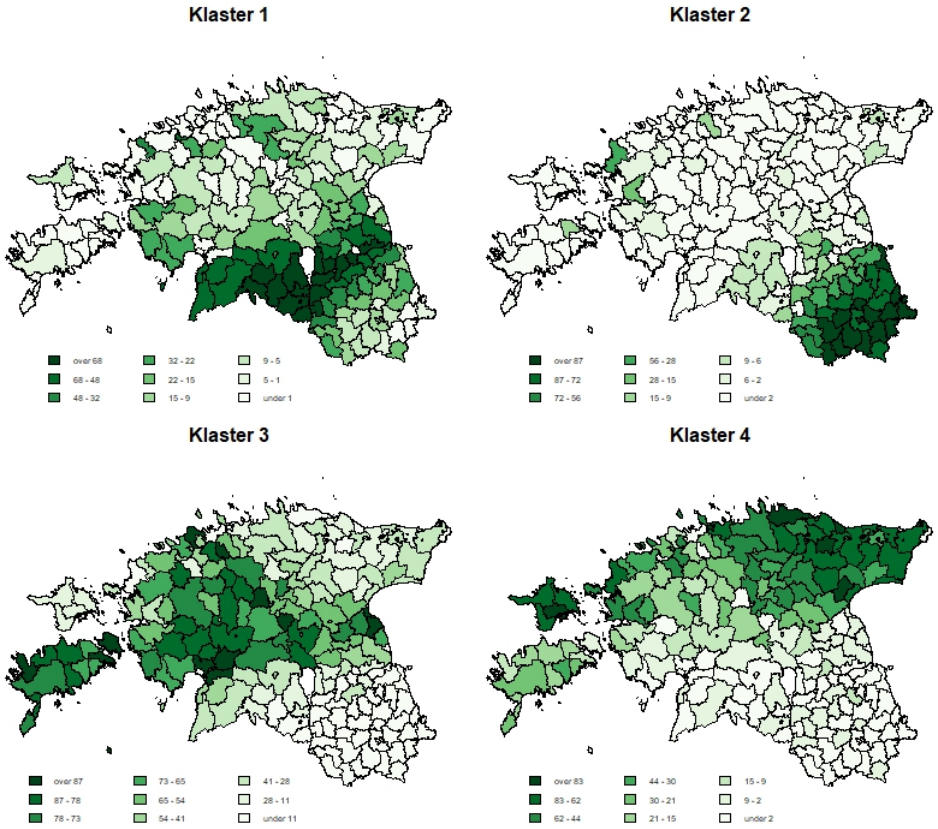
Saja peakomponendi kasutamine tekitab küll mõneti sarnaseid klastreid, võrreldes kahe peakomponendi kasutamisega, ent samas on liigendatuse tase väga erinev. Kaks peakomponenti tekitavad üldistatud pildi, kuid sada peakomponenti kirjeldavad olukorda detailsemalt.

Tõlgendatavuse seisukohalt on mõlema peakomponentide arvu kasutamine mõnes mõttes õigustatud, kuna ilmselt pakub huvi nii üldisem kui ka detailsem pilt. Suur arv peakomponente sisaldab rohkem informatsiooni ning peaks võimaldama tuvastada väiksemaid klastreid, samas kui väike arv peakomponente näitab üldisemat suundumust, kuid ei pruugi suuta eristada väikseid klastreid. Selgusetuks jääb veel vaid see, milline on siiski sobilik peakomponentide arv, et saavutada üldisuse ja detailsuse eesmärgid. Võib juhtuda, et nelja klastri kasutamine kahe peakomponendi korral on liiast, ning võib juhtuda, et kõigest nelja klastri kasutamine ei suuda kirjeldada andmetes leiduvat tegelikku detailsust.

Seega on vajalik leida sobiv suhe peakomponentide arvu ja klastrite arvu vahel. Üks võimalus selle jaoks on fikseerida klastrite arv ning leida sellele sobivaim peakomponentide arv. Seda võib teha näiteks valides sellise peakomponentide arvu, mis tagab kõige sarnasemad klastrite suurused. Märksa parema teoreetilise põhjendusega on idee fikseerida peakomponentide arv ning leida sellele vastav klastrite arv.

5.2.3 *Gap*-statistiku leidmine

Antud töös kasutatakse klastrite arvu leidmiseks lähtuvalt fikseeritud peakomponentide arvust nn *gap*-statistikut, mille olemusest on põhjalikumalt kirjutatud alapeatükis Klastrite arvu määramine.



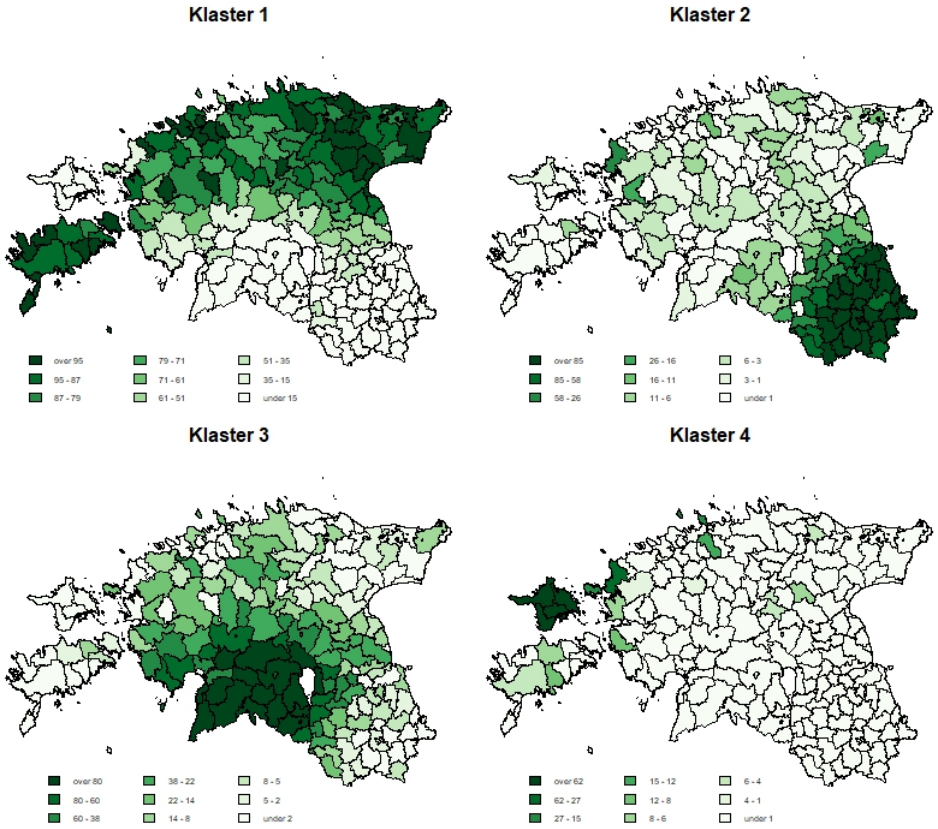
Joonis 17: Antud klastrisse kuuluvate inimeste protsent valla elanikest. Klastrid 2 peakomponendi ja 4 klastri korral

Võimalikud klastrite arvud on valitud kui $k = 1, 2, ..., 20$, sest andmeid puhastades jõuti ülimalt 25 klastrini ning suuremat klastrite arvu võib pidada keerukaks ka interpretatsiooni ja klastrite selguse mõttes. Samuti, katsetes tekitada veelgi enam klastreid hakkasid üha sagedasemini tekkima ainult väga väiksed ühe-kahe vaatlusega klastreid ning tendents jätkus ka pärast selliste väga väikeste klastrite eemaldamist. *Bootstrap* valimite arvuks valiti $B = 7$, mis on küll vähe, ent arvutusliku keerukuse tõttu oleks suurem arv nõudnud väga palju rohkem aega.

Joonisel 19 on välja toodud *gap*-statistikute väärtused peakomponentide arvu ja võimalike klastrite arvu korral. Kasutades K hindamise kriteeriumina Tibshirani et al. artiklis [18] soovitatud reeglit

$$\hat{K} = \text{vähim } k, \text{mille korral } \mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1},$$

tekiksid mõne peakomponendi arvu korral küllalt segased tulemused. Näiteks peaks siis saja peakomponendi korral olema sobiv ainult üks klaster, mis kindlasti pole sisuline valik. Teatavas mõttes on saja peakomponendi juhtum ebatavaline, sest reeglina on statistiku väärtus hakanud pärast esimest sammu siiski kasvama, aga siin ta hakkab hoopis kahanema ning suurem hüpe tuleb kaheksa klastri korral. Samas, kui lähtuda ideest, et kasulik tulemus oleks saada suur

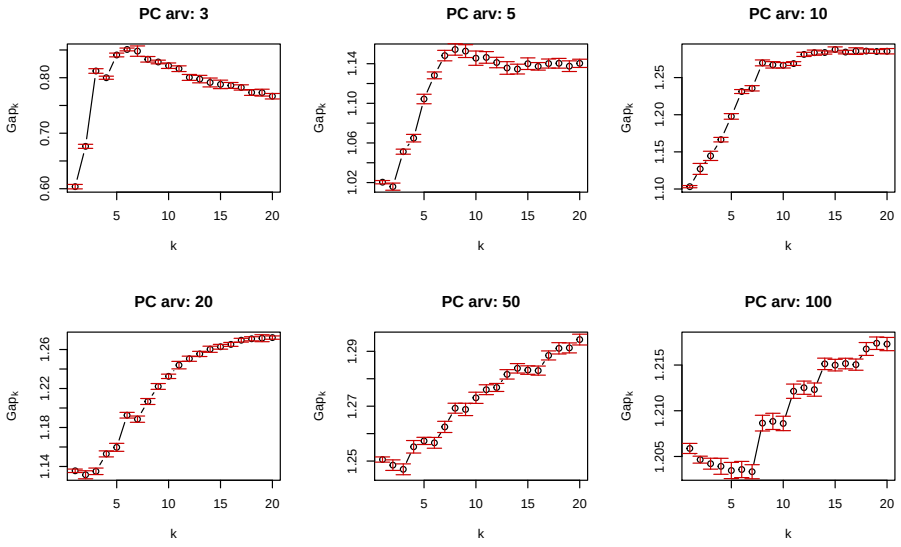


Joonis 18: Antud klastrisse kuuluvate inimeste protsent valla elanikest. Klastrid 100 peakomponendi ja 4 klastri korral

gap-statistiku väärtus, siis näiteks saja peakomponendi jaoks võiks kas või kaheksa klastrit olla selgelt parem valik kui üks klaster.

Seega võib antud situatsiooni sobida paremini teine kriteerium $\hat{K}$ hindamiseks, mis põhineb Dudoit' ja Fridlylandi artiklil [19], mille tulemusi tähistame tabelis veerus Globaalne SE Max. Kokkuvõte erinevate kriteeriumite abil saadud hinnangutest K-le on tabelis 6. Tabelist on näha, et ootuspäraseimad tulemused saadakse esimeses veerus olevate tulemuste puhul, mis on saadud globaalse *gap*-statistiku maksimumi ja sellest mitte oluliselt erinevate väärtuste otsimise abil. Tibshirani et al. välja pakututd kriteerium ei näi siinkohal andmetele sobivat, sest ehkki kolm klastrit on mõeldav kolmele peakomponendile, siis vaid ühe klastri kasutamine suurte peakompoentide arvu juures on siiski mõeldamatu. Viimases veerus on välja toodud need klastrite arvu väärtused, mille korral on tekkinud esimene hüppeline suurenemine. Esimene hüpe võib esindada seda, milline võib olla minimaalne sobiv klastrite arv antud peakomponentide arvule. Edaspidises kasutatakse enamasti Globaalse SE Max tulemusi.

Tabelist 6 on samuti näha, et tõepoolest ei näi olevat vastuolu hüpoteesiga, mille kohaselt peakomponentide arvu kasvades suureneb ka sobiv vastavate klastrite arv, ehkki klastrite arv kasvab selgelt aeglasemalt võrreldes peakomponentide arvu kasvuga.



Joonis 19: *Gap*-statistikud ja vastavad standardhälbed erinevate peakomponentide arvu korral

Tabel 6: Hinnangud k jaoks kasutades *Gap*-statistikut ja erinevaid kriteeriumeid

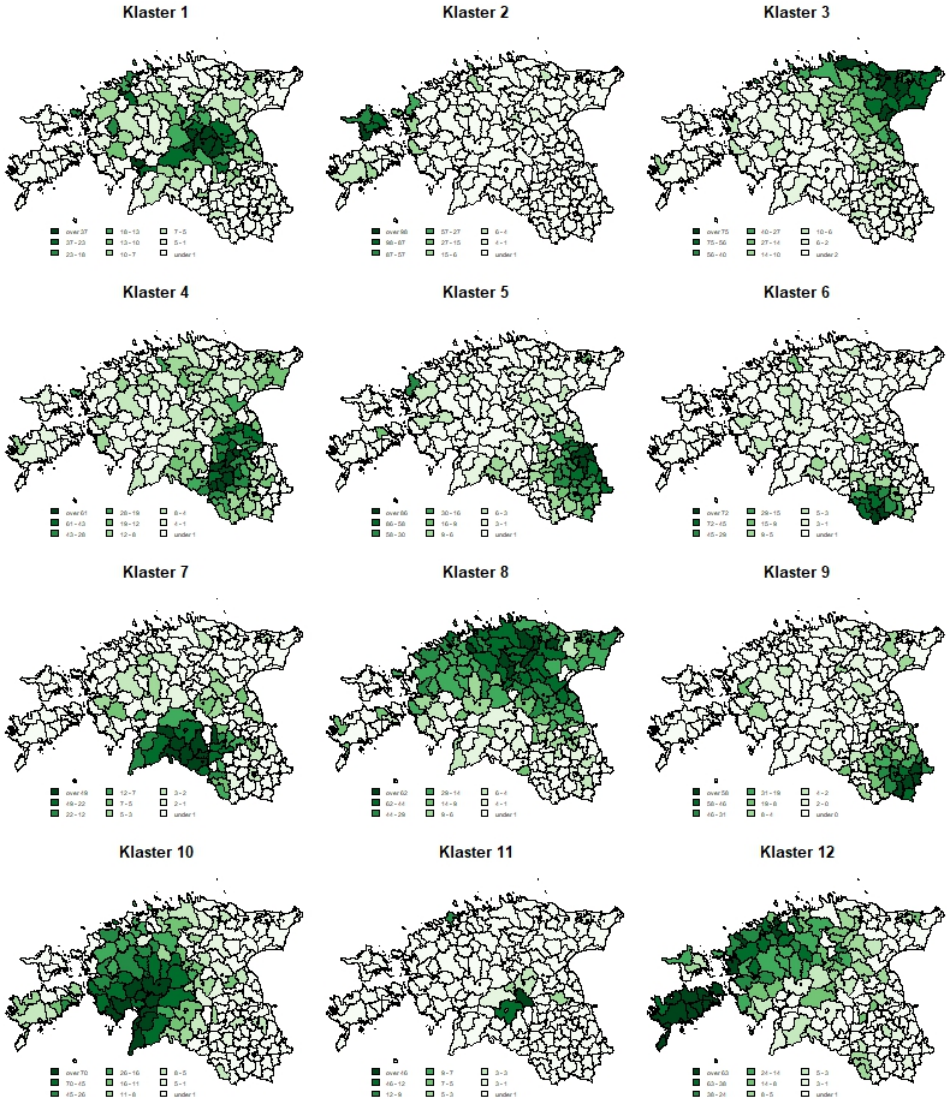
PC arv	Globaalne SE Max	Tibshirani SE max	Esimene hüpe
3	6	3	2
5	7	1	3
10	12	2	2
20	17	1	4
50	17	1	4
100	18	1	8

Saadud tulemustest lähtuvalt kasutame edasises järgnevaid kombinatsioone: 100 peakomponenti ja 18 klastrit, 10 peakomponenti ja 12 klastrit, 5 peakomponenti ja 7 klastrit ning 3 peakomponenti ja 3 klastrit. Välja on jäetud 20 ja 50 peakomponendi kasutamine, kuna mõlemale peaks vastama küllalt suur klastrite arv, mis on juba 100 peakomponendiga kaetud. Kolme peakomponendi jaoks ei kasutata kuut klastrit, sest viie peakomponendiga saadav 7 klastrit on sellele juba küllalt lähedal. Selle asemel proovitakse 3 peakomponendi jaoks 3 klastrit, sest see annab suurima hüppe ning ühtlasi Tibshirani kriteerium soovitab just seda.

5.2.4 Klasterdamise tulemused

Joonistel 36, 37, 20, 38 ja 39 esitatakse ülal valitud peakomponentide ja klastrite arvude alusel tekkinud klastrid. Joonistel on näidatud valdade kaupa antud klastrisse kuuluvate inimeste protsendid kõikidest valimis olevatest valla elanikest. Tabelis 7 on välja toodud ka vastavate klastrite suurused.

On näha, et klasterdamine on andnud loogilisi tulemusi ning valdavalt on tekkinud selgesti eristuvad klastrid. Kolme klastri puhul on eristatud vastavalt



Joonis 20: Vaatluste protsentuaalne jaotumine valdade kaupa kasutades 10 peakomponenti ja 12 klastrit

Lõuna-Eesti, Ida-Eesti ja Lääne- ning Kesk-Eesti.

Seitsme klastri puhul on eristatud vastavalt Kagu-Eesti; Lõuna-Eesti Tartu- ja Valgamaa osad; Kesk-Eesti Jõgeva-, Järva- ja Harjumaa; Lääne-Eesti Saare-, Lääne- ja Lääne-Harjumaa; Pärnu- ja Viljandimaa; Hiiumaa; Kirde-Eesti.

12 klastri korral läheb pilt veidi segasemaks, ent siiski on klastrid päris hästi eristatavad. Tekkinud klastrid on vastavalt Põltsamaa ümbrus; Hiiumaa; Kirde-Eesti; Tartu ja Otepää ümbrus; Lämmijärve ääres asuvad alad; Lõuna-Võrumaa; Mulgimaa; Põhja- ja Kesk-Eesti; Setumaa; Pärnu- ja Läänemaa; Kolga-Jaani ja Viljandi ümbrus; Saaremaa ja Lääne-Eesti.

18 klastri korral on klastrid vastavalt Lõuna-Pärnumaa; Loode- ja Kesk-

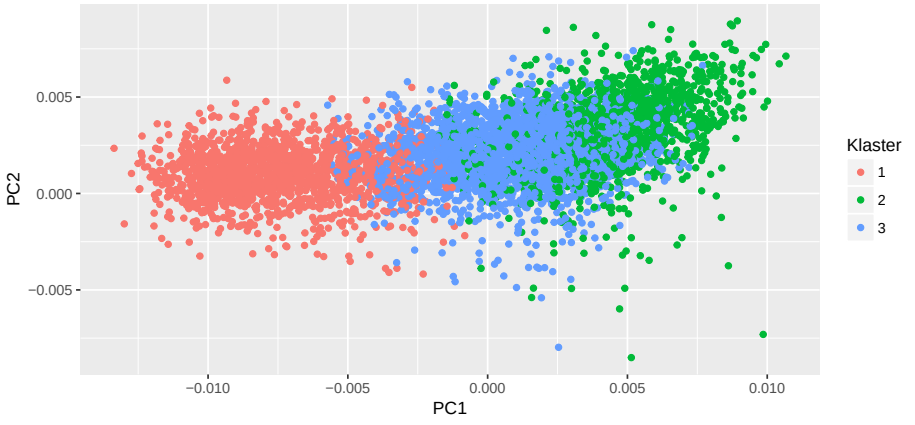
Tabel 7: Vaatluste arv klastris

	Klastri number																	
K	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
3	1483	1427	2746															
7	923	740	1185	759	1038	129	882											
12	443	134	626	624	449	305	405	958	352	705	101	554						
18	140	853	118	116	338	289	698	353	67	122	257	406	91	694	357	497	116	144

Eesti; Virumaa lõunaosa; Kolga-Jaani ja Viljandi ümbrus; Põhja-Viljandimaa; Kirderanniku alad; Setumaa; Muhu saar; Lõuna-Võrumaa, Saaremaa; Mulgimaa; Vändra ümbrus; Põltsamaa ümbrus; Antsla ümbrus; Lämmijärve kõrval olevad alad; Tõstamaa ümbrus; Hiiumaa.

Ehkki klastrid on selgesti eristuvad, siis mõneti probleemne on klastrite väga erinevad suurused. Näiteks on 18 klastriga juhul klaster 9 (Muhu klaster) vaid 67 vaatlusega, kuid klastris 2 (Loode- ja Kesk-Eesti) on vaatlusi 853. Ka 12 ja 7 klastri puhul jääb silma, et klastrite suurused pole parimas tasakaalus. Tõlgendatavuse seisukohalt on takistuseks suurte klastrite olemasolu suure klastrite arvu korral, sest kui 18 klastri korral raporteerida, et inimene kuulub klastrisse 2, saab ta märksa vähem teada oma päritolu kohta, kui inimene, kes kuulub Muhu klastrisse. Üks võimalus seda probleemi lahendada on suurendada veelgi klastrite arvu, et jagada ka suurimad klastrid alamklastreiks. Paraku näitasid katsed sellega seda, et suured klastrid jäid endiselt alles ning tegelikult tekkisid uued klastrid hoopis praeguste väikeste klastrite jagamisest. Seega teatavas mõttes on käesolev tulemus parim, mida sellise meetodiga välja pakkuda on võimalik.

Teine muretav tõik on geograafiliste erindite olemasolu. Hea tulemus oleks see, kus oleks selgelt saada aru, millistel aladel domineerib mingi klaster. Kindlasti võib leiduda ka erandeid tänu sellele, et inimesed on rännanud, ning seega võib leida klastri esindajaid ka mujal kui ainult klastri peamisel levialal. Näiteks joonisel 36 klaster 1 esindab valdavalt lõunaeestlasi, kuid päris arvestatav hulk on klastri 1 esindajaid ka teistes Eesti piirkondades.



Joonis 21: Vaatlused esitatud kahe peakomponendiga 3 klastri korral

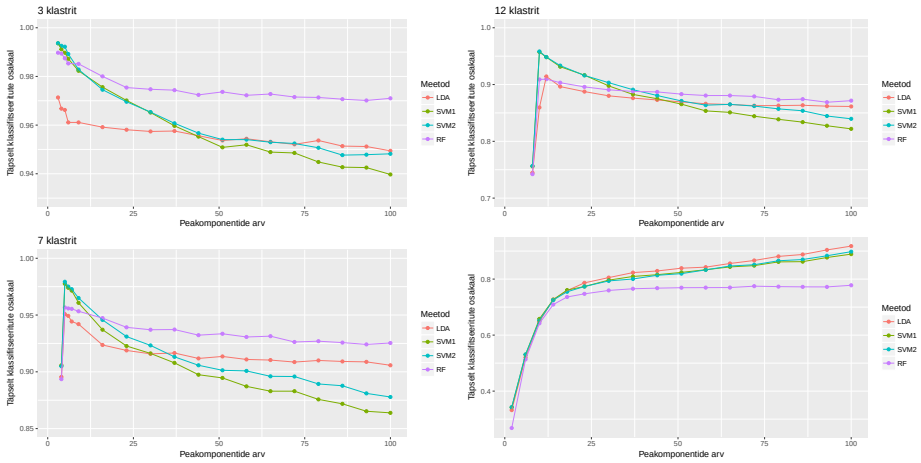
Et kirjeldada paremini erandeid, võib tagasiside raames olla hea näidata ka tulemusi peakomponentide lõikes nagu on tehtud joonisel 21. Nõnda on

võimalik inimesel tuvastada jooniselt enda võimalik erandlikkus ning sel juhul on mõistetavam ka klassifitseerimise tulemus. Klastrite keskmeist kaugemale jäävad vaatlused võivadki olla erandlikumad ning mõistetavalt võib seal tekkida enam prognoosivigu.

5.3 Klassifitseerimine klastrite alusel

5.3.1 Meetodite võrdlus klastri ennustamisel

Eelnevaga sarnaselt kontrollitakse 10-jaotuse ristvalideerimise abil, millised meetodid ja milline peakomponentide arv annavad klassifitseerimisel parimaid tulemusi. Parameetrite valik tugivektormasinate ning juhusliku metsa hindamisel on jäänud samaks. Tulemused on esitatud joonisel 22 ning tabelis 8.



Joonis 22: Klassifitseerimistäpsused sõltuvalt klastrite arvust

Joonisel ja tabelist nähtub väga loogiline tendents. Kõige paremad tulemused saavutatakse üldiselt selliste peakomponentide arvu korral, mida oli kasutatud ka klastrite loomiseks. Niisiis klassifitseerimisel kolme klastri korral peaks kasutama kolme peakomponenti, seitsme klastri korral viit peakomponenti, 12 klastri korral kümmet peakomponenti ning 18 klastri korral sadat peakomponenti.

Pisut teistsugused tulemused joonistusid välja meetodite lõikes, kus LDA ei osutunud enam alati parimaks. Nii kolme, seitsme kui ka 12 klastri korral saavutasid parimad tulemused tugivektormasinad; pisut paremad tulemused saavutati, kui valiti tugivektormasinatel C väärtuseks 2.

Täheldada on ka seda, et klasside arvu suurenedes väheneb klassifitseerimistäpsus, mis on täiesti ootuspärane tulemus. Heaks võib pidada ka saadud klassifitseerimistäpsusi, mis igal juhul ületavad 0.9 ja näiteks kolme klastri puhul on lähedal 0.99le. Samas ei tohiks eeldada, et sellise täpsusega suudetakse klassifitseerida ka uusi testisikuid, sest antud juhul olid klastrid töötatud välja otseselt kasutades referentsandmeid ning osalt selle pärast annab ristvalideerimine väga häid tulemusi. Selgemaks täpsuse hindamiseks kasutatakse eraldiseisvat testvalimit.

Seega, edasisel testimisel kasutatakse 3, 7 ja 12 klastri korral meetodina SVM2-te ning 18 klastri korral kasutakse LDAd.

Tabel 8: Viis paremat klassifitseerijat erinevate klastrite arvu korral

	Meetod	PC arv	Täpsus	$q_{0.025}$	$q_{0.975}$
3 klastrit					
1	SVM1	3.0000	0.9936	0.9922	0.9949
2	SVM2	3.0000	0.9936	0.9922	0.9950
3	SVM2	4.0000	0.9926	0.9909	0.9943
4	SVM2	5.0000	0.9922	0.9906	0.9940
5	SVM1	4.0000	0.9912	0.9890	0.9933
7 klastrit					
1	SVM2	5.0000	0.9793	0.9742	0.9840
2	SVM1	5.0000	0.9775	0.9729	0.9821
3	SVM2	6.0000	0.9751	0.9680	0.9808
4	SVM1	6.0000	0.9740	0.9680	0.9797
5	SVM2	7.0000	0.9728	0.9674	0.9778
12 klastrit					
1	SVM2	10.0000	0.9583	0.9538	0.9627
2	SVM1	10.0000	0.9574	0.9512	0.9626
3	SVM2	12.0000	0.9484	0.9404	0.9548
4	SVM1	12.0000	0.9482	0.9404	0.9561
5	SVM2	16.0000	0.9333	0.9278	0.9388
18 klastrit					
1	LDA	100.0000	0.9178	0.9126	0.9233
2	LDA	93.0000	0.9042	0.8990	0.9093
3	SVM2	100.0000	0.8980	0.8927	0.9032
4	SVM1	100.0000	0.8891	0.8811	0.8965
5	LDA	86.0000	0.8879	0.8822	0.8942

5.3.2 Klastrite prognoosimine testvalimil

Huvi pakub küsimus, kui hästi töötab klasterdamise alusel valitud klassidega klassifitseerimine mingi uue populatsiooni puhul. Et seda kontrollida, teostati kaks protseduuri määramaks igale vallale n-ö õige klaster. Ühel juhul seati vallale vastavusse rangelt üks populaarseim klaster, kusjuures viikide puhul eelistati vähima indeksiga valda; teisel juhul seati vallale i vastavusse õigete klastritena klastrid $C_k^{vald=i}$, millesse referentsvalimist kuulus vähemalt $\frac{1}{3}\max|C_k^{vald=i}|$ liiget. Viimane toiming on vajalik, sest küllalt palju leidub valdasid, mille referentspopulatsiooni liikmed klasterduvad peaaegu pooleks. Näiteks kui vaadata seitset klastrit, siis Helme valla puhul referentsvalimist 12 inimest liigitatakse klastrisse 2 (Tartu- ja Valgamaa) ning 11 inimest klastrisse 5 (Pärnu- ja Viljandimaa), mõned üksikud veel teistesse klastritesse. Esimese vastavuse alusel koostatakse joonised, teise vastavuse alusel väljastatakse täpsushinnangud.

Testisikud valiti sarnasel põhimõttel, nagu on välja toodud alapeatükis Referentspopulatsiooni valik Eesti-sisese päritolu uurimiseks ja ainus erinevus on siinkohal, et nüüd valiti sünniaastateks 1961-1970. Seega tegemist on referentsvalimist pisut nooremate inimestega, kellede puhul on endiselt oodatav, et nende sünnikoht näitab suuresti ka vanemate päritolu.

Parameetrite valikul kasutati eelnevalt saadud tulemusi. Kolme, seitsme ja 12 klastri juhul kasutatakse tugivektormasinaid parameetriga $C = 2$. 18 klastri puhul kasutatakse klassifitseerimisel lineaarset diskriminantanalüüsi. Tulemused klassifitseerimistäpsuste kohta on esitatud tabelis 9. On näha, et tulemused on mõnevõrra halvemad võrreldes sellega, mida on näha ristvalideerimise tulemustest tabelis 8. Erinevust põhjustavad mitmed asjaolud. Et testandmestikus on nooremad inimesed, siis erinevad tegurid, näiteks ränne, võivadki põhjustada testvalimi mõningast erinevust referentsvalimiga. Siiski ilmselt suurima erinevuse põhjustaja on skeem, kuidas on defineeritud vallale õige klaster: paljud vallad kuuluvad n-ö kahe või isegi kolme klastri piirile ning ka eelnev skeem, millega ühele vallale võis vastata mitu õiget klastrit, ei pruugi olla piisav.

Siiski on täheldada, et prognoositäpsused on paremad võrreldes sellega, mida oli võimalik saada maakondade alusel klassifitseerimisel ning seda isegi juhul, kui klasse on maakondadest enam. Seega näitab paranenud prognoositäpsus seda, et klastrite tekitamine on olnud õigustatud.

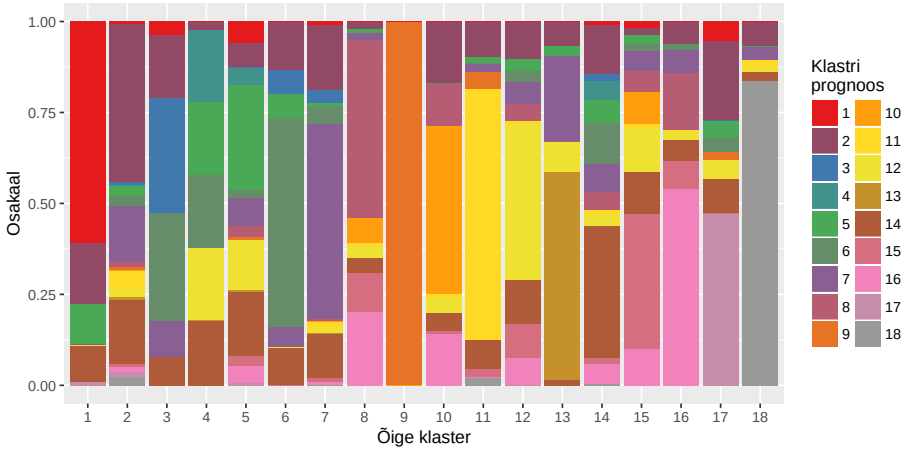
Tabel 9: Prognoositäpsused klastrite arvu kaupa testandmestikul

Klastreid	3	7	12	18
Täpsus	0.8330	0.6941	0.6875	0.6488

Järgnevalt on tulemused esitletud juhul, kus vallale on vastavusse seatud ainult üks klaster. Tabelis 10 on välja toodud keskmised prognoositud tõenäosused õige klastri kaupa, joonistel 23, 40,41 on seda sama tehtud vastavalt 18, 12 ja 7 klastri jaoks.

Tabel 10: Keskmised tõenäosused kuuluda vastavasse klastrisse õige klastri järgi 3 klastri juhul

	Prognoositud klaster		
Õige klaster	1	2	3
1	0.7431	0.1053	0.1515
2	0.0572	0.7364	0.2064
3	0.0914	0.1687	0.7398



Joonis 23: Klassifitseerimistõenäosused sõltuvalt õigest klastrist 18 klastri korral

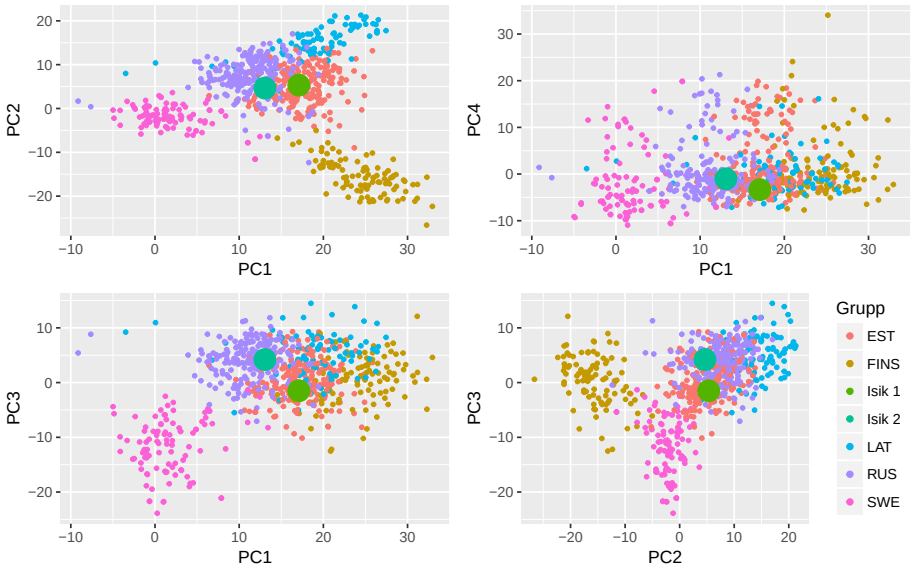
Vaadates joonist 23, on näha, et üldiselt vastavad suurimatele keskmistele tõenäosustele just õiged klastrid, st kui õige on esimene klaster, siis on keskmiselt suurim tõenäosus seal kuuluda ka esimesse klastrisse. Paraku sõltuvalt klastrist ei pruugi õigesse klastrisse kuulumise tõenäosus olla alati väga suur. Samas tuleb arvestada eeltoodud tähelepanekuga, et tihti võib vallale vastata hoopis mitu sobivat klastrit. Seda tähele pannes peaks olema ootuspärane, et suured tõenäosused võivad olla ka naaberklastritel. Kõrvutades joonise 23 tulemusi vastavate kaartidega 38 ja 39, siis pea alati on õige klastri korral tõenäosused suured

ka naaberklastritel. Seega, tegemist on väga hea ning loogilise tulemusega, mis on hästi interpreteeritav. Sarnased seosed tulevad esile ka juhul, kui vaadata detailsemalt tulemusi 7 ning 12 klastri puhul. Kokkuvõttes võib öelda, et testisiku vaatlused määratakse üldjuhul suure tõenäosusega klastritesse, mis on testisiku sünnivallale omaseimad või siiski testisiku sünnivallas väga levinud.

5.4 Näitetagasiside

Järgnevalt antakse kahe konkreetse inimese näitel ülevaade tagasiside protsessist. Nende inimeste päritolu on üldjoontes teada ning üks järgneva osa eesmärk ongi kõrvutada olemasolevat teadmist päritolu kohta töös saadud tulemustega. Inimesed on anonüümsed ning neid tähistatakse järgnevalt kui isik 1 ja isik 2. Rahvuse poolest peaksid olema nii isik 1 kui 2 eestlased, pole teada, et leiduks teistest rahvustest esivanemaid. Eesti-sisese päritolu poolest peaks ligikaudu 50% isiku 1 esivanemaist olema Viljandimaalt ning ülejäänud Jõgeva-, Järva- ja Harjumaalt, pigem Tallinn-Tartu maantee äärest paiknevatelt aladelt. Isiku 2 üks vanem on pooleldi Läänemaalt ja pooleldi Mulgimaalt ning teine vanem Läänemaalt.

Esmalt on välja toodud joonisel 24 tulemused võrreldes referentsrahvustega kahe peakomponendi teljestikus. On näha, et isik 1 satub üsna eestlaste grupi keskele, kuid isik 2 on pigem eestlaste ja venelaste piiril; esimese ja kolmanda peakomponendi teljestikus on isik 2 pigem venelastega samal alal.



Joonis 24: Näidisvaatlused võrreldes lähemate referentsrahvustega kahe peakomponendi teljestikus

Järgnevalt on ennustatud rahvusgruppidesse kuulumist. Tabelis 11 on välja toodud 6 suurema tõenäosusega gruppi. Esitatud on nii eelpool soovitatud 17 peakomponendiga kui ka 200 peakomponendiga saadud prognoos ning kasutatud on lineaarset diskriminantanalüüsi. Gruppi kuulumise tõenäosused võivad sõltuvalt peakomponentide arvust muutuda palju. Osutub, et vähemalt nende

kahe isiku jaoks osutub 200 peakomponendi kasutamine märksa loogilisemaks valikuks. See võib omakorda tähendada, et simulatsioonikatsest saadud soovitus kasutada 17 peakomponenti ei pruugi olla parim valik. Paremaks kontrolliks oleks vaja enam teadaoleva päritoluga testisikuid, sest ainult kahe inimese põhjal ei saa teha otsust. Teisalt võib probleem seisneda ka selles, et eesti ja vene grupid ongi väga lähedased ning piirialale langevad inimesed, isegi kui nad on teadaolevalt eestlased, võivad liigituda venelasiks.

Seega on ilmselt mõistlik väljastada kaks hinnangut tõenäosusele rahvusliku päritolu kohta. Vähem peakomponente kasutav ennustus väljastab laiema spektrumi erinevatest tõenäosustest, seeläbi vähendades võimalust, et mingi sarnasus teatava rahvusgrupiga jääb kirjeldamata, kuid samas ei pruugi tõenäosuste suurusjärgud olla kooskõlas reaalsusega. Rohkem peakomponente kasutav ennustus väljastab väiksema spektri erinevaid tõenäosuseid ning prognoosid on siin enesekindlamad, st suurima tõenäosusega rahvusgrupp saavutab tavaliselt suure tõenäosuse. Samas tekib oht, et hinnang on liialt enesekindel ning ei ennusta üldse päritolu teistest gruppidest.

Tabel 11: Rahvusgruppi kuulumise tõenäosused teadaolevatel isikul, kasutades prognoosiks erinevat arvu peakomponente

	Isik 1				Isik 2			
	Grupp	$\hat{P}_{PC=17}$	Grupp	$\hat{P}_{PC=200}$	Grupp	$\hat{P}_{PC=17}$	Grupp	$\hat{P}_{PC=200}$
1	Eesti	0.79833	Eesti	0.95996	Vene	0.57386	Eesti	0.84582
2	Vene	0.18959	Vene	0.04004	Eesti	0.40966	Vene	0.15393
3	Läti	0.00731	Leedu	0.00001	Läti	0.00893	Leedu	0.00020
4	Leedu	0.00406	Läti	< 0.00001	Poola	0.00384	Läti	0.00005
5	Poola	0.00066	Tšehhi	< 0.00001	Leedu	0.00369	Tšehhi	< 0.00001
6	Rootsi	0.00005	Poola	< 0.00001	Tšehhi	0.00002	Põhja-Saksa	< 0.00001

Eesti gruppi kuulumise tõenäosus on mõlemal isikul piisavalt suur, et ilmselt tasub kontrollida ka isikute Eesti-sisese päritolu kohta. Esmalt kontrollitakse, kuidas prognoositakse maakondadesse kuuluvust. Tulemused maakondade tõenäosuste prognooside kohta on esitatud tabelis 12. Siinkohal on kasutatud LDAd 100 peakomponendiga.

Tabel 12: Maakondadesse kuulumise tõenäosused teadaolevatel isikutel

	Isik 1		Isik 2	
	Maakond	$\hat{P}$	Maakond	$\hat{P}$
1	Viljandi maakond	0.58458	Järva maakond	0.27610
2	Jõgeva maakond	0.15921	Rapla maakond	0.21678
3	Järva maakond	0.10284	Harju maakond	0.17814
4	Harju maakond	0.08355	Jõgeva maakond	0.13181
5	Lääne-Viru maakond	0.02364	Lääne maakond	0.08782
6	Rapla maakond	0.02017	Valga maakond	0.03801
7	Lääne maakond	0.00953	Tartu maakond	0.03484
8	Tartu maakond	0.00877	Lääne-Viru maakond	0.02095
9	Valga maakond	0.00577	Ida-Viru maakond	0.00993
10	Ida-Viru maakond	0.00121	Viljandi maakond	0.00490

Võrreldes enda öelduga, on isiku 1 ennustused läinud väga hästi kokku prognoosiga. Tõenäosust ligi 0,5 ennustatakse Viljandimaale ning kõrged tõenäosused on ka Jõgeva-, Järva- ja Harjumaal. Isikul 2 päritolu nii hästi esile pole tulnud. Probleem võib seisneda ka selles, mida oli näha juba testimisel: Lääne maakonda on keeruline prognoosida ning selle asemel prognoositaksegi tihti naabermaakondi, näiteks Harju- või Raplamaad. Samas ei saa öelda, et tulemused oleksid täielikult ebaloogilised, sest ajalooliselt ongi pool Raplamaad kuulunud

Läänemaale ning Järvamaa lõunapoolsed osad on olnud ajalooliselt Viljandimaal. [20]

Huvitavad on tulemused välja töötatud klastrite lõikes. Kolme, seitsme ja 12 klastri puhul kasutatakse tugivektormasinaid parameetriga 2, 18 klastri puhul LDAd. Kolme klastri puhul on mõlemal inimesel suurim tõenäosus kuuluda klastrisse number 3, isikul 1 on tõenäosus 0.99999 ning isikul 2 on tõenäosus 0.95620. Lisaks on isikul 2 tõenäosus 0.04234 kuuluda klastrisse 1. Suuremate klastrite arvude jaoks on mõlema isiku jaoks tulemused koondatud tabelitesse 13 ja 14.

Tabel 13: Isik 1 klastritesse kuulumise tõenäosused erinevate klastrite arvu korral

	7 klastrit		12 klastrit		18 klastrit	
	Klastri number	$\hat{P}$	Klastri number	$\hat{P}$	Klastri number	$\hat{P}$
1	3	0.97629	4	0.35398	14	0.73439
2	5	0.01604	7	0.26062	5	0.14312
3	4	0.00561	1	0.16903	2	0.07869
4	2	0.00179	8	0.07919	12	0.04237
5	6	0.00019	12	0.04248	11	0.00139
6	7	0.00007	10	0.03691	6	0.00003
7	1	0.00001	6	0.01453	7	0.00001

7 klastri korral on isik 1 kõige tõenäolisemalt määratud klastrisse 3, milles on valdavalt Jõgeva, Järva ja Harjumaa alad. Klastris 3 on esindatud ka arvesatavalt Viljandimaa ehkki Viljandimaa valdade seas pole klaster 3 populaarseim. 12 klastri korral on isik 1 määratud suure tõenäosusega klastritesse 4, 7 ja 1. Klastris 4 on Valga- ja Tartumaa, klastris 7 Viljandimaa ning klastris 1 osa Järva- ning Jõgevamaast. 18 klastri korral on isik 1 määratud suurema tõenäosusega klastritesse 14 ja 5. Klaster 14 Jõgeva-, Järva- ja Harjumaad, valdavalt Tallinn-Tartu maantee ääres olevaid alasid ning klaster 5 on seotud enamasti Viljandimaaga. Seega, teatavate mööndustega, kuid kõikidel juhtudel andis klastrite alusel klassifitseerimine isikule 1 tema enda raporteerituga sarnase tulemuse päritolu kohta.

Tabel 14: Isik 2 klastritesse kuulumise tõenäosused erinevate klastrite arvu korral

	7 klastrit		12 klastrit		18 klastrit	
	Klastri number	$\hat{P}$	Klastri number	$\hat{P}$	Klastri number	$\hat{P}$
1	4	0.68824	12	0.39499	2	0.96880
2	3	0.26453	8	0.33781	14	0.03108
3	2	0.04216	4	0.22894	12	0.00004
4	1	0.00204	1	0.01806	15	0.00004
5	5	0.00202	5	0.00808	5	0.00002
6	7	0.00073	10	0.00594	11	0.00001
7	6	0.00028	6	0.00255	16	0.00001

Isik 2 on 7 klastri korral kõige tõenäolisemalt määratud klastritesse 4 ja 3, väiksem tõenäosus on ka klastril 2. Eriti just klaster 4 on levinud Lääne-Eestis, klaster 3 on levinud nii Kesk- kui ka Lääne-Eestis. Klaster 2 on levinud Mulgi- ja Tartumaal. 12 klastri korral on isik 2 saanud suuremad tõenäosused kuuluda klastritesse 12, 8 ja 4. Klaster 12 sisaldab jällegi eelkõige Lääne-Eesti vaatluseid, 8 sisaldab enamasti Kesk-Eesti vaatluseid, milles on ka palju Lääne-Eesti vaatluseid ning klaster 4 sisaldab enamasti Mulgi- ja Tartumaa vaatluseid. 18 klastri korral määratakse inimene väga suure tõenäosusega klastrisse 2, mis on jällegi enamasti lääne-eestlastega seotud. Seega ka isiku 2 puhul võib öelda, et suuremaid lahknevusi ei tekkinud enda raporteeritud päritolu ning klastrite abil klassifitseerimisel prognoositud päritolu vahel ja tulemused on head.

6 Kokkuvõte

Käesoleva magistritöö eesmärk oli leida võimalusi, kuidas anda TÜ Geenivaramu doonoritele tagasisidet nende päritolu kohta nii rahvuse tasandil kui ka Eesti-siseselt. Saadud tulemused näitasid, et välja pakutud võimalused päritolu ennustamiseks annavad enamasti hästi interpreteeritavaid ja loogilisi tulemusi, kuid kindlasti saaks prognoose veelgi enam täpsustada. Töö käigus lähtuti analüüside tegemisel peakomponentidest ning üks suuremaid küsimusi töö käigus oli sobiva peakomponentide arvu valik korrektseima ennustuse saamiseks.

Rahvuse klassifitseerimisel ning vastavate klassitõenäosuste leidmisel saavutas täpseima tulemuse lineaarne diskriminantanalüüs. Leidmaks korrektseks tõenäosuse arvutamiseks sobivat peakomponentide arvu, teostati simulatsioonikatse, milles võrreldi arvatavasti teadaoleva päritoluga genotüübi tõenäosuseid vastavate tõenäosuste prognoosidega sellele genotüübile. Paraku simulatsioonikatse vastus ei olnud ühene ning lõplikku vastust peakomponentide arvu valikuks ei selgunud: põhjendatuks saab lugeda nii väikse kui ka suure peakomponentide arvu kasutamist ning teema vajab täpsemat uurimist. Samas annavad mõlemad lähenemised loogilisi vastuseid ning saadud tulemust halvaks pidada ei saa.

Eesti-sisesel prognoosimisel kontrolliti esialgu võimalust klassifitseerida maakondade alusel. Sealjuures saavutas täpseima tulemuse lineaarne diskriminantanalüüs, kuid mitmete Sise-Eesti maakondade puhul oli klassifitseerimistäpsus madal. Seetõttu otsustati alternatiivina leida K-keskmiste klasterdamise abil uued klassid, mis võiksid moodustada loogilisemaid ja täpsemaid kogumeid, milledesse inimesi liigitada. Leiti seos klasterdamisel kasutatavate peakomponentide arvu ja klastrite arvu vahel ning klastrite arvudena pakuti *gap*-statistiku alusel välja 3, 7, 12 ja 18, mis kirjeldavad erineva detailsusega tekkivaid kogumeid. Leitud klastrite alusel testisikute klassifitseerimine andis häid tulemusi ning enamasti klassifitseeriti testinimesed samasse klastrisse või naaberklastrisse. Klassifitseerimismeetoditena kasutati 3, 7, 12 puhul tugivektormasinaid ja 18 klastri puhul lineaarset diskriminantanalüüsi.

Viimaseks anti näide tagasisidest kahe inimese põhjal, kelle päritolu on suuresti teada. Nende puhul oli näha, et tõenäosusliku hinnangu andmine rahvusele võib olla vägagi varieeruv sõltuvalt valitud peakomponentide arvust ning õigustatud on anda kaks hinnangut. Teisalt olid Eesti-sisesed prognoosid pigem täpsed. Nii maakonna alusel kui ka klastrite alusel klassifitseerides jõuti enamasti ootuspäraste tulemusteni.

Suurim küsimus edasise uurimise osas on ilmselt rahvusprognooside korrigeerimine. Üks osa sellest on kindlasti peakomponentide edasine uurimine ning võimaluste leidmine sobiva peakomponentide arvu valikuks. Teisalt on võimalik, et see pole piisav, sest ka käesolevas töös on näidatud, et sobivat kompromissi peakomponentide arvu valikuks ei pruugigi leiduda. Hoopis tulemuslikum võib olla referentsvalimite korrigeerimine. Ka töö käigus selgus, et vene referentspopulatsiooni on vaja kindlasti täiendada ning pole täielikult selge, kui hästi kajastab töös kasutatud vene referentsvalim varieeruvust venelaste seas. Lisaks tuleks kindlasti kaaluda Geenivaramu olemasolevate doonorite baasil ka valgevenelaste ning ukrainlaste referentsgruppide moodustamist ning ühtlasi võib saada Geenivaramu andmetest täiendust lätlaste ja soomlaste referentspopulatsioonidele.

Tasub mõelda ka selle peale, et rahvuse määramise referentsandmestikust eemaldada Põhja-Soome vaatlused, mis on ülejäänud referentsidega võrreldes

selgelt erandlikud. Selle grupi eemaldamisega tekib võimalus, et peakomponentanalüüs suudaks eristada paremini ülejäänud eurooplaste vahel eksisteerivat varieeruvust ning seeläbi jõuda klassifitseerimisel paremate tulemusteni. Teatav analoogia on olemas Eesti-sisese päritolu määramisega, kus peakomponendid on leitud vaid Geenivaramu andmestest ning nende peakomponentide kasutamine annab Eesti-siseselt sisukaid tulemusi.

Seega, edasise päritolu uurimise seisukohalt on oluline tagada, et referentsvalimid oleksid piisavalt suured ning samaaegselt esinduslikud. Eriti oluline on see rahvuse määramise kontekstis, kuid näiteks referentsvalimi suurendamine omab kindlasti positiivset mõju ka Eesti-sisese päritolu uurimise kontekstis, kus oleks seeläbi võimalik klastreid paremini defineerida. Aastal 2018 viiakse läbi 100 000 täiendava geenidoonori andmete kogumine Geenivaramusse. Kindlasti aitab ka see andmemahu suurenemine kaasa paremate referentside välja töötamisele.

Käesolevas magistritöös anti ülevaade meetoditest ning võimalustest, mille abil anda tagasisidet päritolu kohta. Tulemused näitavad, et töös pakutud võimaluste abil saab esialgsel kujul anda TÜ Geenivaramu doonoritele tagasisidet. Töös välja pakutud ideid kasutades saab meetodeid kindlasti veel parandada, et lõppkokkuvõttes tagada geenidoonoritele võimalikult täpne tagasiside.

Viited

[1] Sampson, J., Kidd, K. K., Kidd, J. R., Zhao, H., "Selecting SNPs to Identify Ancestry", *Annals of Human Genetics*, vol. 75, no.4, 2011, pp. 539-–553.

[2] Ojavee, S. E., "Geneetiliste päritolukomponentide määramine mitmemõõtmelise statistika meetodite abil", bakalaureusetöö, Tartu Ülikool, 2015.

[3] Haller, T., Leitsalu L., Fischer K., Nuotio M.-L., Esko T., Boomsma DI., et al., "MixFit: Methodology for Computing Ancestry-Related Genetic Scores at the Individual Level and Its Application to the Estonian and Finnish Population Studies", *PLOS ONE*, vol. 12, no. 1, 2017, pp. 1–14.

[4] Surakka, I., Whitfield, J.B., Perola, M., Visscher, P.M., Montgomery, G.W., Falchi, M., et al., "A genome-wide association study of monozygotic twin-pairs suggests a locus related to variability of serum high-density lipoprotein cholesterol", *Twin research and human genetics: the official journal of the International Society for Twin Studies*, vol. 15, no. 6, 2012, pp. 691-–699.

[5] Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., Toncheva, D., et al., "Genetic Structure of Europeans: A View from the North–East", *PLOS ONE*, vol. 4, no. 5, 2009, pp. 1–10.

[6] Johnson, R. A., Wichern, D. W., *Applied Multivariate Statistical Analysis*, 6th edn., Pearson Prentice Hall, 2007.

[7] Lohninger, H., *Fundamentals of Statistics*, 2012. Saadaval: Epina Bookshelf, (kasutatud 15.02.2018).

[8] Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., Lee, J.J. "Second-generation PLINK: rising to the challenge of larger and richer datasets", *GigaScience*, vol. 4, no. 1, 2015.

[9] Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning*, 2nd edn., Springer, 2009.

[10] James, G., Witten, D., Hastie, T., Tibshirani, R., *An introduction to Statistical Learning*, Springer, 2013.

[11] Platt, J. C., "Probabilities for SV Machines", in Smola, A. J. et al., (ed.), *Advances in Large Margin Classifiers*, MIT Press, 1999, pp. 61–74.

[12] Wu, T.-F., Lin, C.-J., Weng, R. C., "Probability Estimates for Multi-class Classification by Pairwise Coupling", *Journal of Machine Learning Research*, vol. 5, 2004, pp. 975–1005.

[13] Chang, C.-C., Lin, C.-J., "LIBSVM: a Library for Support Vector Machines", *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011, pp. 27:1–27:27.

[14] Zhou, Z.-H., *Ensemble Methods: Foundations and Algorithms*, CRS Press, 2012.

[15] Breiman, L., "Manual–Setting Up, Using, And Understanding Random Forests", [veebileht], 2001, `https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf`, (kasutatud 01.02.2018).

[16] Breiman, L., "Random Forests", *Machine Learning*, vol. 45, no. 1, 2001, pp. 5–32.

[17] Zimmermann, M., "Toitumismustrite analüüs Tartu Ülikooli Eesti geenivaramu andmebaasis k-keskmiste meetodi abil", bakalaureusetöö, Tartu Ülikool, 2015.

[18] Tibshirani, R., Walther, G., Hastie, T., "Estimating the number of clusters in a data set via the gap statistic", *Journal of the Royal Statistical Society: Series B*, vol. 63, no. 2, 2001, pp. 411–423.

[19] Dudoit, S., Fridlyand, J., "A prediction-based resampling method for estimating the number of clusters in a dataset", *Genome Biology*, vol. 3, no. 7., 2002.

[20] Mägi, A., *Eesti rahva ajaraamat*, Tallinn, 1993.

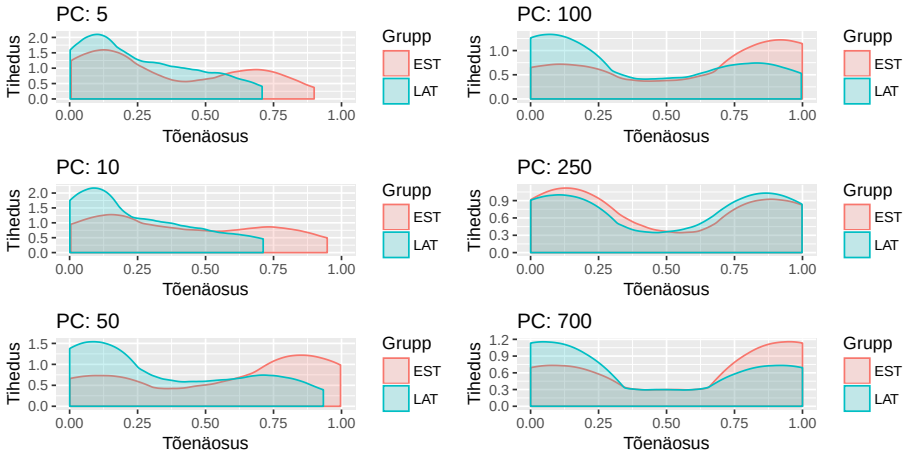
[21] Pajusalu, K., Hennoste, T., Niit, E., Päll, P., Viikberg, J., *Eesti murded ja kohanimed*, Tallinn, Eesti Keele Sihtasutus, 2002.

A Koodid

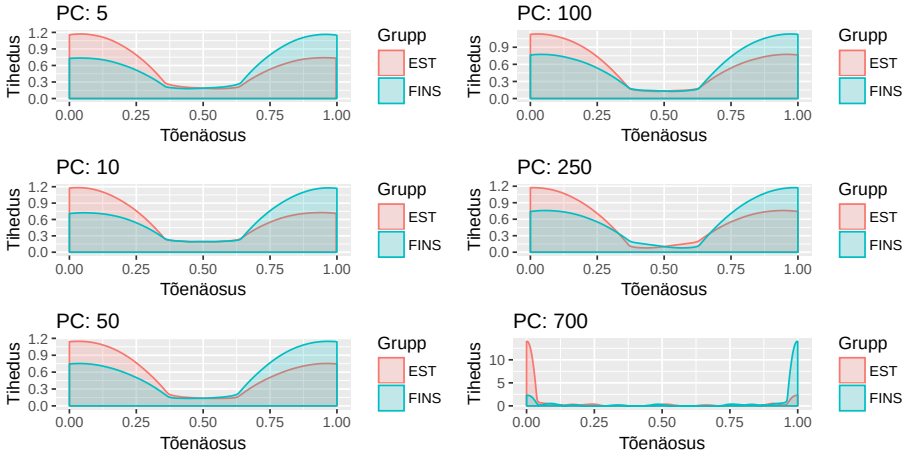
Andmed: Peakomponentide andmestik C_0, mis vastab alapeatükis Referentspopulatsiooni valik Eesti-sisese päritolu uurimiseks välja toodud kriteeriumitele, C_i tähistab peakomponentide andmestikku i-ndal sammul
Tulemus: Puhastatud peakomponentide andmestik
$k \leftarrow 2$;
while $k \leq 25$ **do**
  Tekitada k ja andmete C_i alusel k-keskmiste klasteranalüüsi mudel $\mathcal{M}(C_i, k)$;
  Mudeli alusel tekivad klastrid $C_{i1}, ...C_{ik}$;
  if $\min_j(|C_{ij}|) = 1$ **then**
    $j_0 \leftarrow \min(\arg\min_j |C_{ij}|)$;
    $C_{ij_0} \leftarrow \emptyset$;
    $C_{i+1} \leftarrow \bigcup_{j=1}^{k} C_{ij}$;
    $k \leftarrow \max(2, k-3)$;
  else
    $k \leftarrow k+1$;
    $C_{i+1} \leftarrow C_i$;
  end
end

Pseudokood 2: Andmete puhastamine

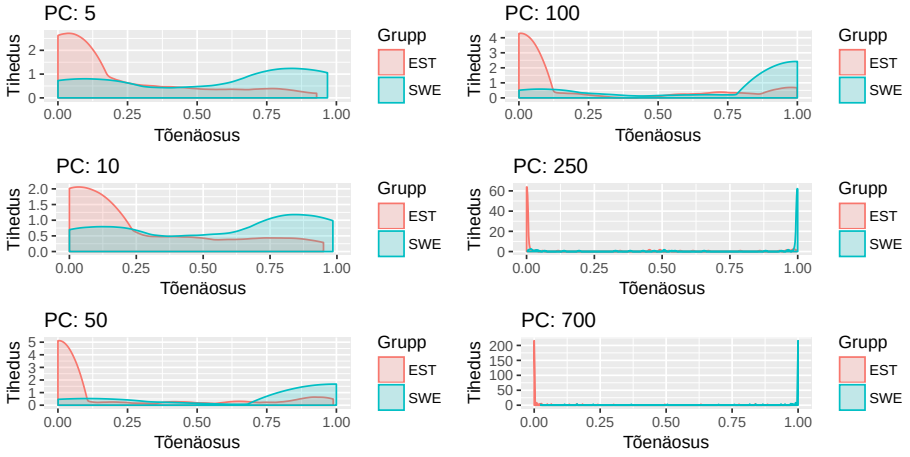
B Joonised



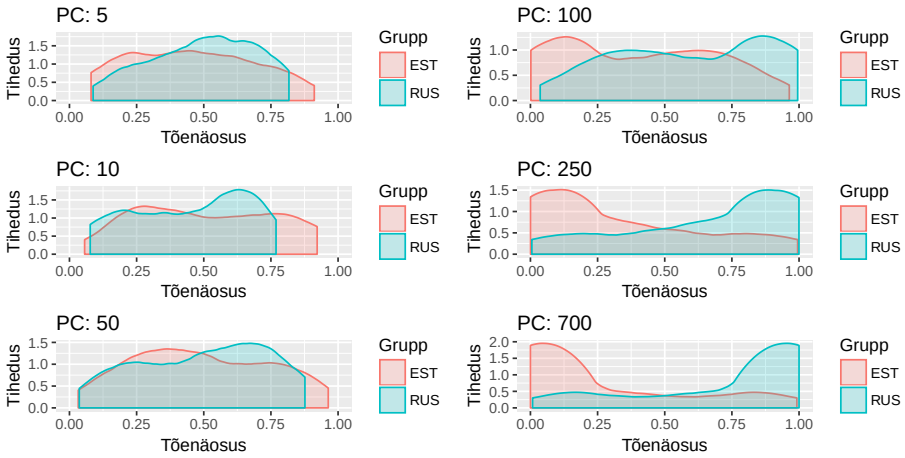
Joonis 25: Prognoositud eesti ja läti gruppide tõenäosused 0.5-0.5 eesti-läti simuleeritud andmete puhul



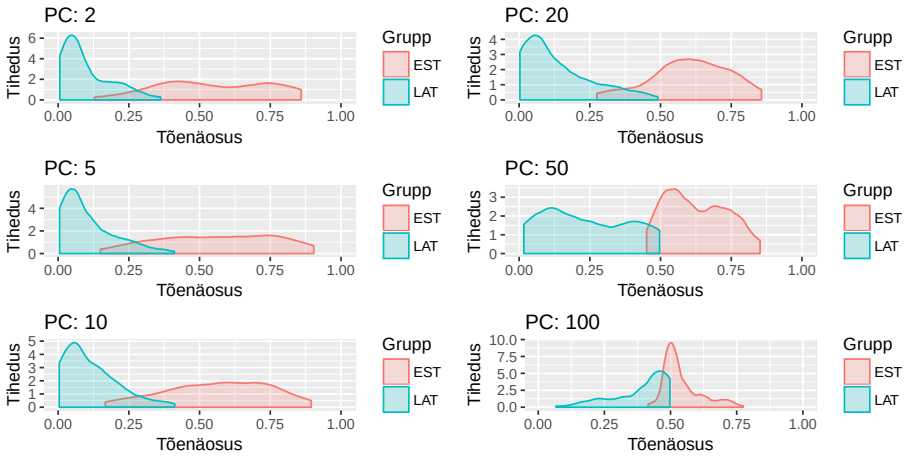
Joonis 26: Prognoositud eesti ja lõunasoome gruppide tõenäosused 0.5-0.5 eesti-lõunasoome simuleeritud andmete puhul



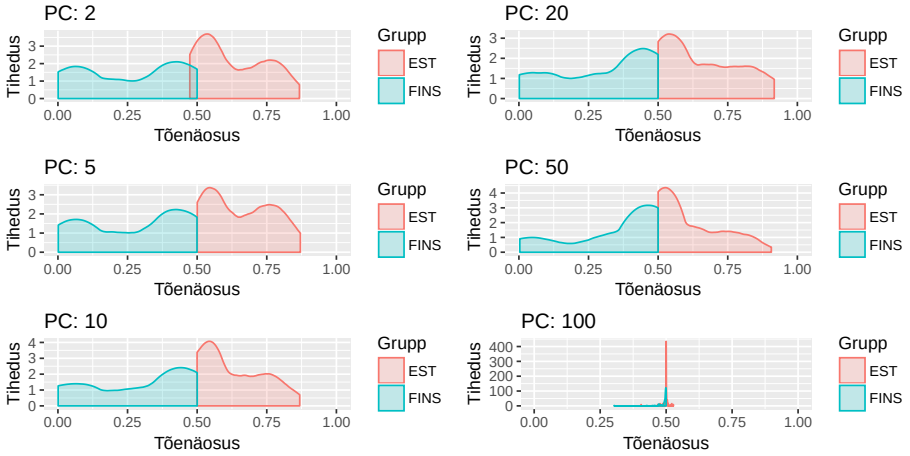
Joonis 27: Prognoositud eesti ja rootsi gruppide tõenäosused 0.5-0.5 eesti-rootsi simuleeritud andmete puhul



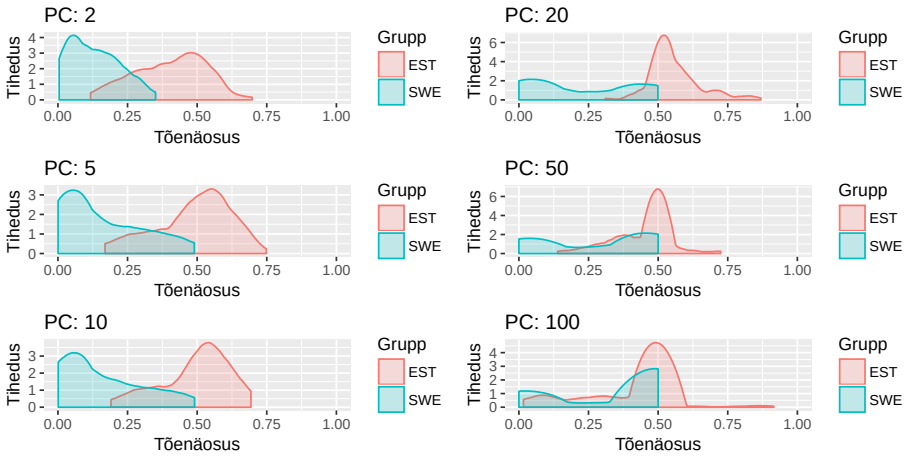
Joonis 28: Prognoositud eesti ja vene gruppide tõenäosused 0.5-0.5 eesti-vene simuleeritud andmete puhul



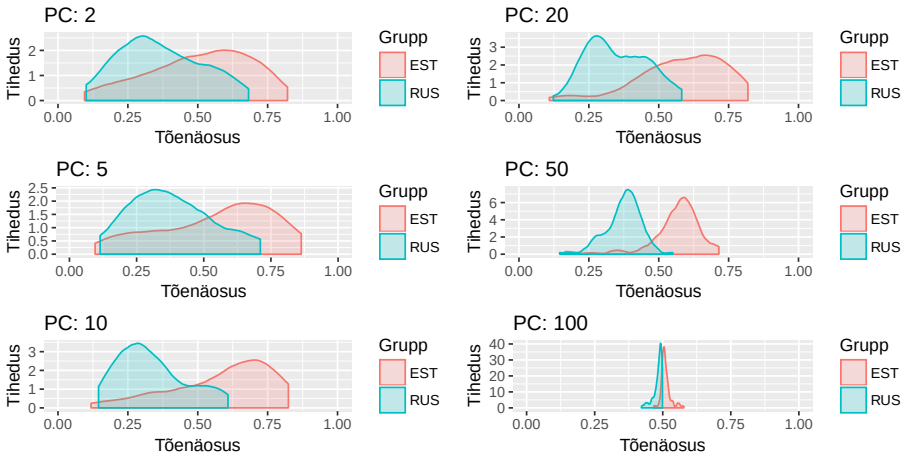
Joonis 29: Prognoositud eesti ja läti gruppide tõenäosused 0.75-0.5 eesti-läti simuleeritud andmete puhul



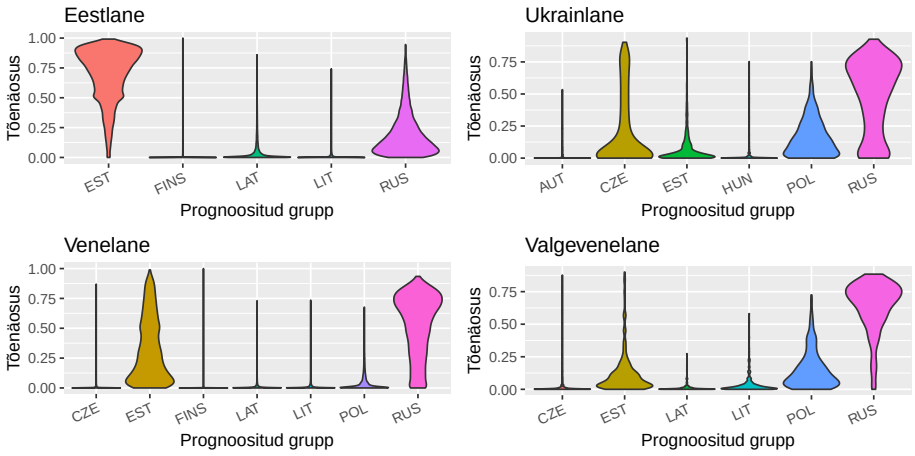
Joonis 30: Prognoositud eesti ja lõunasoome gruppide tõenäosused 0.75-0.5 eesti-lõunasoome simuleeritud andmete puhul



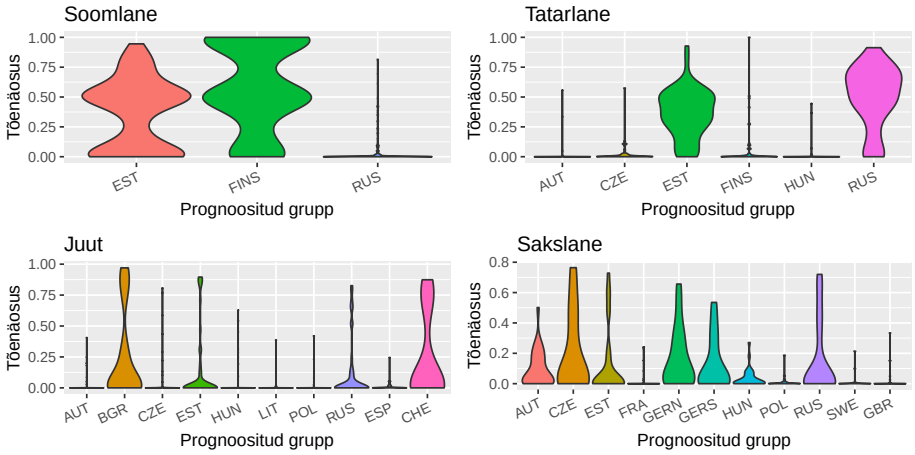
Joonis 31: Prognoositud eesti ja rootsi gruppide tõenäosused 0.75-0.5 eesti-rootsi simuleeritud andmete puhul



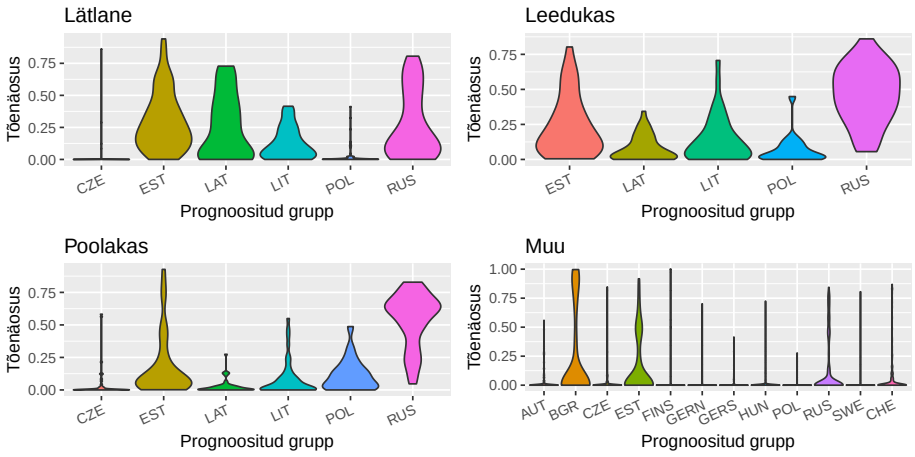
Joonis 32: Prognoositud eesti ja vene gruppide tõenäosused 0.75-0.25 eesti-vene simuleeritud andmete puhul



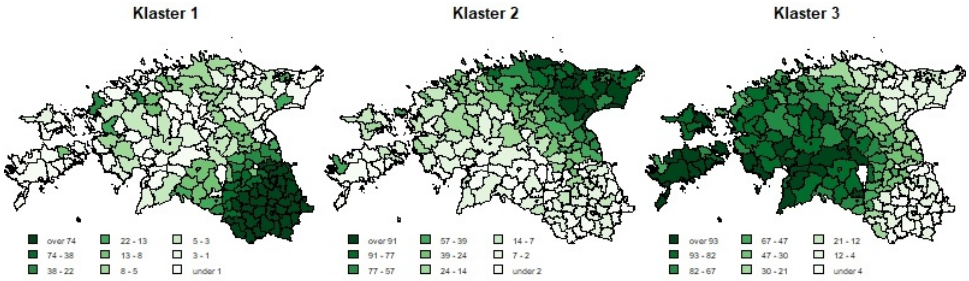
Joonis 33: Prognoositud tõenäosuste jaotumine vastavalt raporteeritud rahvusele. Eestlased, venelased, ukrainlased, valgevenelased



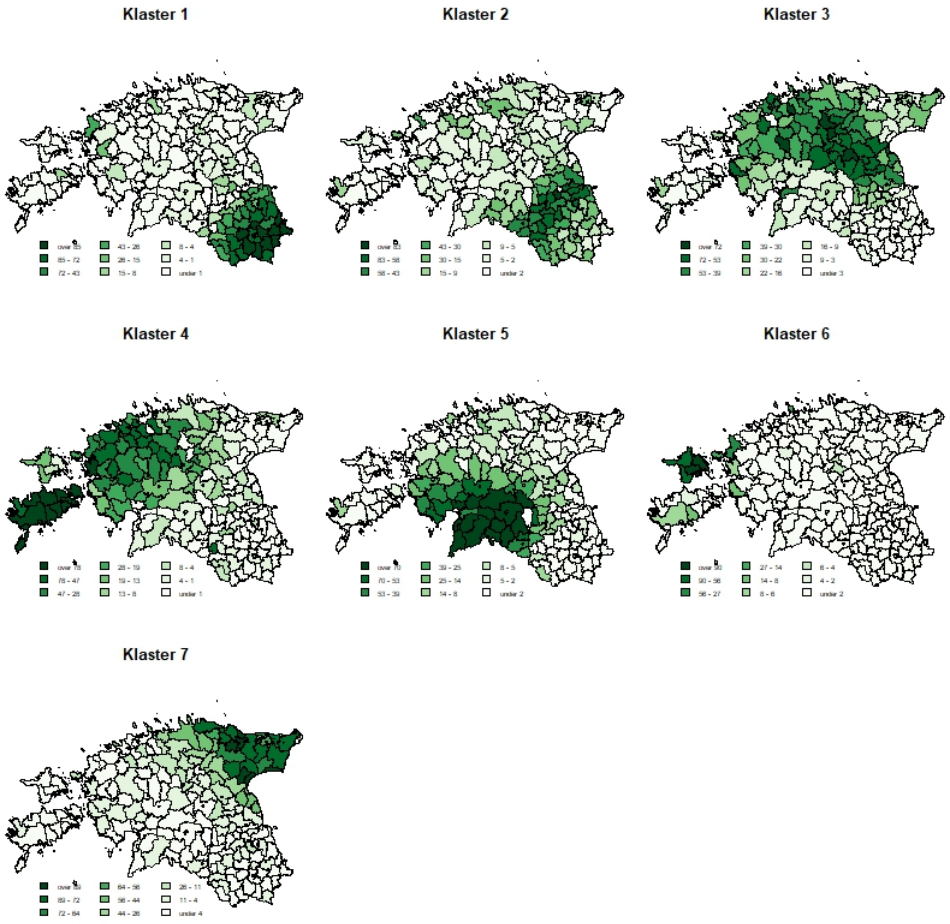
Joonis 34: Prognoositud tõenäosuste jaotumine vastavalt raporteeritud rahvusele. Soomlased, juudid, tatarlased, sakslased



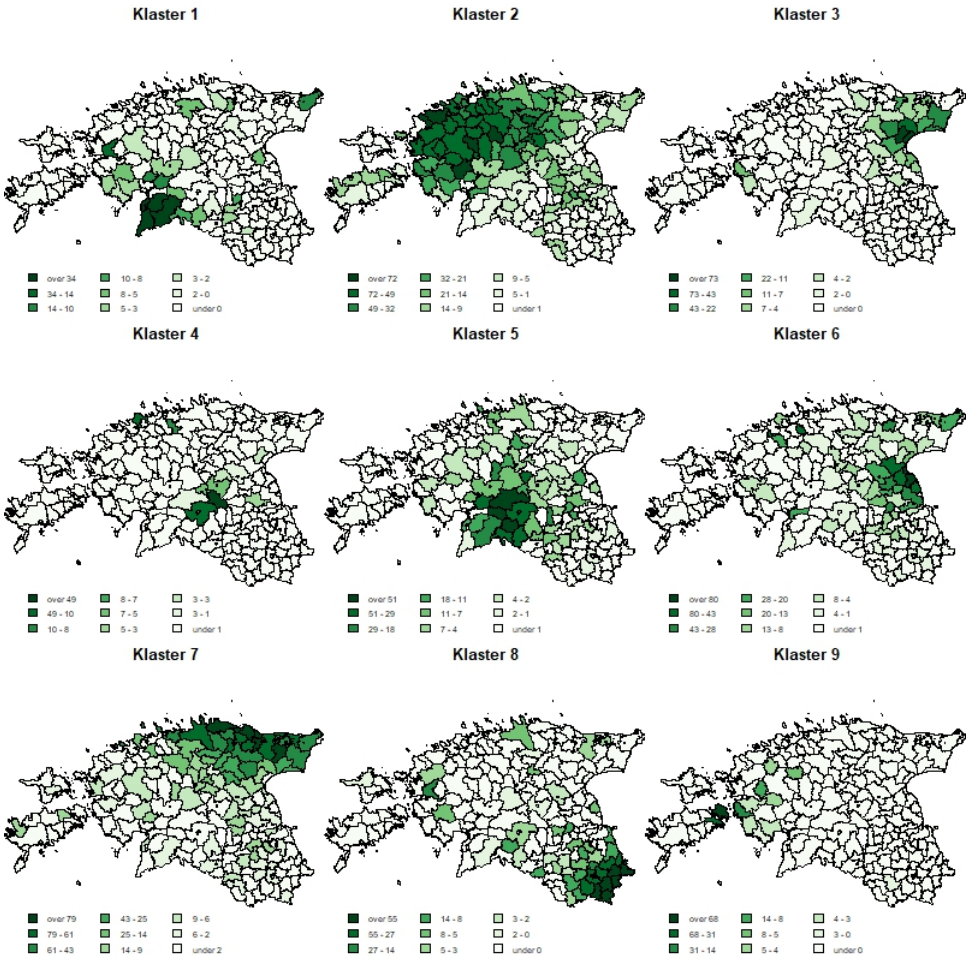
Joonis 35: Prognoositud tõenäosuste jaotumine vastavalt raporteeritud rahvusele. Lätlased, poolakad, leedukad, muud



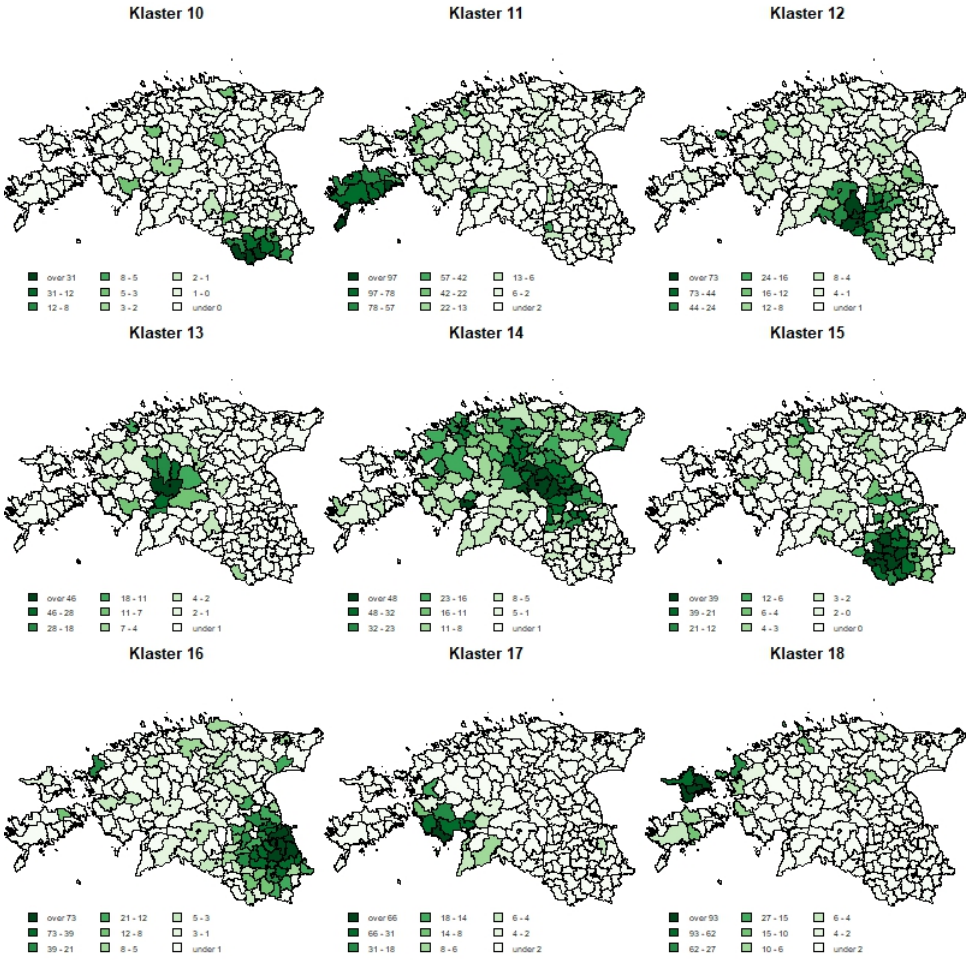
Joonis 36: Vaatluste protsentuaalne jaotumine valdade kaupa kasutades 3 peakomponenti ja 3 klastrit



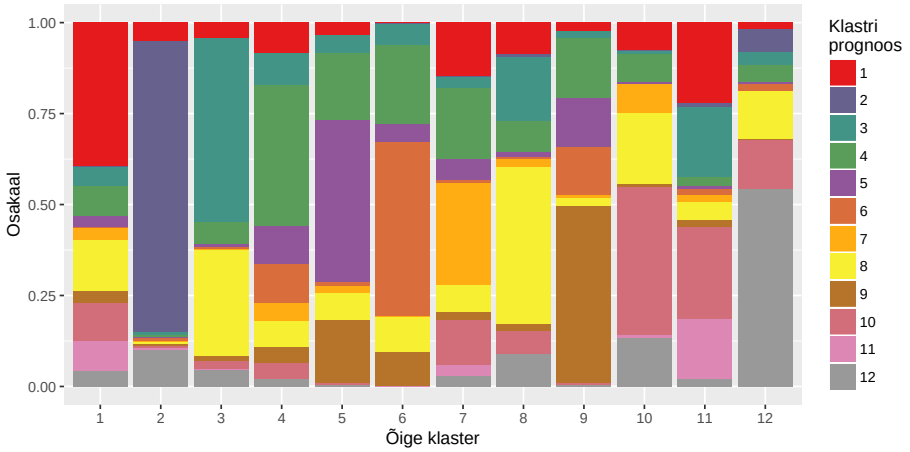
Joonis 37: Vaatluste protsentuaalne jaotumine valdade kaupa kasutades 5 peakomponenti ja 7 klastrit



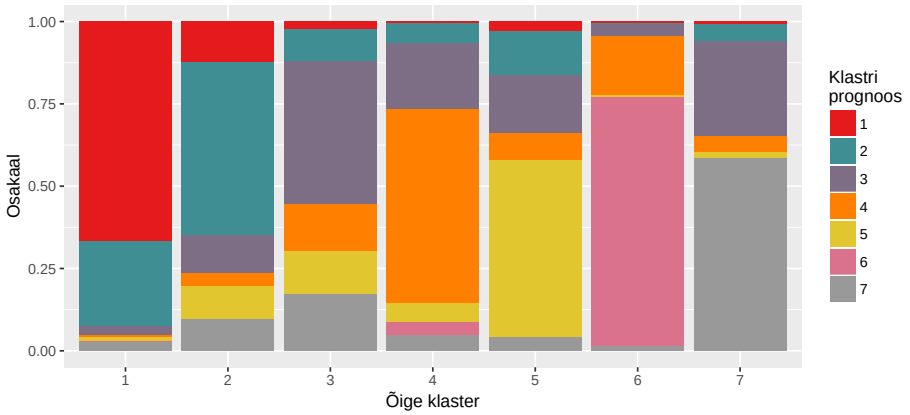
Joonis 38: Vaatluste protsentuaalne jaotumine valdade kaupa kasutades 100 peakomponenti ja 18 klastrit. Klastrid 1-9



Joonis 39: Vaatluste protsentuaalne jaotumine valdade kaupa kasutades 100 peakomponenti ja 18 klastrit. Klastrid 10-18



Joonis 40: Klassifitseerimistõenäosused sõltuvalt õigest klastrist 12 klastri korral



Joonis 41: Klassifitseerimistõenäosused sõltuvalt õigest klastrist 7 klastri korral

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Sven Erik Ojavee,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Päritolu hindamine geeniandmete põhjal: TÜ Eesti Geenivaramu andmete analüüs“, mille juhendaja on Krista Fischer,
    1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
    1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 16. veebruaril 2018