Tartu Ülikool

Matemaatika-informaatikateaduskond

Matemaatilise statistika instituut

Marili Zimmermann

Toitumismustrite analüüs Tartu Ülikooli Eesti geenivaramu andmebaasis k-keskmiste meetodi abil

Bakalaureusetöö (9 EAP)

Juhendaja: Krista Fischer, *Ph.D*

Tartu 2015

Toitumismustrite analüüs Tartu Ülikooli Eesti geenivaramu andmebaasis k-keskmiste meetodi abil

Käesoleva bakalaureusetöö eesmärgiks on Tartu Ülikooli Eesti geenivaramu andmebaasis olevate toitumisandmete klasterdamise kaudu toitumismustrite leidmine k-keskmiste meetodi abil. Esmalt tuuakse ülevaade klasteranalüüsist ning täpsemalt k-keskmiste meetodist. Töö teises osas antakse ülevaade kasutatavast Tartu Ülikooli Eesti geenivaramu andmestikust. Töö järgnevas osas kirjeldatakse tehtud analüüsi TÜ geenivaramu andmetel ning antakse ülevaade saadud klastritest. Ühtlasi vaadeldakse ka klastrite lõikes erinevaid tausttunnuseid nagu vanus, haridus, elukoht jms. Erinevaid taustatunnuseid vaadates tulid leitud klastrite erinevused hästi välja.

Märksõnad: klasteranalüüs, matemaatiline statistika, andmeanalüüs

Analysis of eating patterns in the database of Estonian Genome Center at the University of Tartu using the k-means method.

The goal of this thesis is to cluster and find patterns in feeding data in the database of The Estonian Genome Center with k-means cluster algorithm. Firstly, an overview of the used data set is given, which is followed by a description of k-means clustering. The second part of this thesis describes the analysis which was conducted using the data of The Estonian Genome Center. An overview of the results within the clusters will be presented. In conjunction with the clusters, the background variables, such as age, education, home country and others were analysed together with the clusters. Differences between clusters were evident.

Keywords: cluster analysis, mathematical statistics, data analysis

Sisukord

Sissejuhatus

Käesoleva bakalaureusetöö eesmärgiks on Tartu Ülikooli Eesti geenivaramu andmebaasis olevate toitumisandmete klasterdamise kaudu toitumismustrite leidmine. Samuti soovitakse uurida, milliste taustatunnustega on seotud erinevad toitumismustrid ja kas erinevalt toituvate isikute rühmad erinevad südamehaiguste riski osas. Tartu Ülikooli Eesti Geenivaramu on Tartu Ülikooli koosseisus olev teadus- ja arendusasutus, mille põhiülesanneteks on edendada geeniuuringute arengut, koguda teavet Eesti rahvastiku tervise ja pärilikkuse informatsiooni kohta ning rakendada geeniuuringute tulemused rahva tervise parandamiseks [1]. TÜ geenivaramu andmebaasiga on liitunud ligi 52000 inimest aastatel 2002-2013.

Kõik TÜ geenivaramuga liitunud geenidoonorid täitsid põhjaliku küsimustiku, kus küsiti ka 18 toiduaine tarbimise sagedust. Pidevalt räägitakse toitumise mõjust tervisele. Samas võib arvata, et toitumist ei mõjuta mitte niivõrd mõne üksiku toiduaine tarbimine vaid erinevate toiduainete osakaalud igapäevamenüüs, nn toitumismuster.

Toitumusmustrite leidmiseks saab kasutada mitmemõõtmelise statistika meetodeid, mis aitavad eristada erinevate toitumismustritega inimesi. Käesolevas töös on kasutatud k-keskmiste meetodil põhinevat klasteranalüüsi, mis aitab eristada sarnaselt toituvate isikute rühmi nii, et rühmadevaheline erinevus toitumises oleks võimalikult suur. Klasteranalüüsi meetodit rakendati 18 tunnusest ja enam kui 45000 vaatlusest koosnevale andmestikule.

Töö esimeses peatükis on toodud ülevaade antud töös kasutavatest statistilistest meetoditest. Teises peatükis on kirjeldatud kasutatavat andmestikku. Kolmandas peatükis rakendatakse klasteranalüüsi Tartu Ülikooli geenivaramu andmetele ning tuuakse ülevaade analüüsi tulemustest.

Bakalaureusetöö kirjutamiseks on kasutatud programmi Tex. Analüüsid on läbi viidud statistikapaketiga R.

Käesolevaga tänab autor bakalaureusetöö juhendajat Krista Fischerit rohkete paranduste ja nõuannete eest.

1 Teooria

1.1 Klasteranalüüs

Sõna klasterdamine viitab erinevatele tehnikatele, mida kasutatakse andmestiku väiksemateks rühmadeks või teise nimega klastriteks jaotamiseks [2]. Klasteranalüüs grupeerib andmestikus olevad vaatlused, kasutades selleks ainult andmestikus olevat informatsiooni vaatluste ja nendevaheliste erinevuste kohta. Eesmärgiks on leida grupid (klastrid) nii, et vaatlused grupisiseselt oleksid võimalikult sarnased üksteisele ja erineksid vaatlustest teistes gruppides. Mida suurem on vaatlustevaheline sarnasus grupisiseselt ja mida väiksem gruppide vaheliselt, seda parem on klasterdamise meetod.[3] Klastrisisese ja klastritevahelise sarnasuse mõõtmine erineb olenevalt andmetest või klasterdamise meetodist [2].

Andmete klasterdamine on populaarne erinevates valdkondades ning seetõttu on olemas väga mitmeid klasteranalüüsi meetodeid. Antud töös keskendutakse ühele tuntumatest klasterdamise meetoditest - k-keskmiste meetodile. Klasteranalüüs teostati k-keskmiste meetodi abil, kuna see on tarkvarapaketi R abil rakendatav suurtele andmebaasidele. Teised võimalikud klasteranalüüsi meetodid (nt hierarhiline klasterdamine) osutusid liiga arvutusmahukaks, sest need eeldavad esmalt kõikide vaatluste vahelise distantsimaatriksi arvutamist, mis nõuaks väga suurt arvutimälumahtu.

1.2 K-keskmiste meetod

Järgnev alapeatükk põhineb raamatul *An Introduction to Statistical Learning with Applications in R* [2], v.a seal, kus on märgitud teisiti.

k-keskmiste meetod eraldab andmed K-sse erinevasse, mittekattuvasse klastrisse. k-keskmise meetodi kasutamiseks on esmalt vaja teada soovitud klastrite arv K. Seejärel määrab k-keskmiste algoritm iga vaatluse täpselt ühte K-st klastrist.

Olgu meil vaatlused $\mathbf{x_1}, \ldots, \mathbf{x_n}$, kus iga vaatlus on p-mõõtmeline reaalarvude vektor ning olgu x_{il} i-nda vaatluse l-is tunnus ($i = 1, \ldots, n$; $l = 1, \ldots, p$)[2].

Märkigu $C_1, \ldots, C_K$ hulki, mis täidavad järgmisi tingimusi:

- $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, ..., n\}$ - Iga vaatlus kuulub vähemalt ühte K-st hulgast
- $C_i \cap C_j = \emptyset \; kui \; i \neq j$ - Hulgad ei ole kattuvad, iga vaatlus kuulub ainult ühte hulka.

k-keskmiste meetodi eesmärgiks on andmestiku selline klasterdamine, kus klastrite sisene hajuvus on nii väike kui võimalik.

Klastrisisene hajuvus klastri C_k jaoks on suurus $W(C_k)$, mis näitab kui palju vaatlused klastri siseselt erinevad üksteisest. Vaja on jagada vaatlused K-sse klastrisse nii, et kogu klastrite sisene hajuvus summeerituna üle kõigide klastrite oleks minimaalne. Seega tahame lahendada probleemi:

$$\underset{C_1,\ldots,C_K}{\text{minimeerida}} \{\sum_{k=1}^{K} W(C_k)\}.$$

Klastrisisese hajuvuse defineerimiseks on mitmeid variante. Kõige sagedamini kasutatakse eukleidilise kauguse ruutu. Eukleidiline kaugus p-mõõtmelises ruumis kahe vektori vahel on defineeritud kui $\|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$. Seega saame:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,j \in C_k} \sum_{l=1}^{p} (x_{il} - x_{jl})^2, \tag{1}$$

kus $|C_k|$ on vaatluste arv k-ndas klastris;
x_{il} on i-nda vaatluse l-s element $(i = 1, \ldots, n; \; l = 1, \ldots, p)$;
p-tunnuste arv (vektori $\mathbf{x_i}$ pikkus).
Teisisõnu on k-nda klastri klastrisisene hajuvus summa üle kõigi k-ndas klastris olevate vaatluste paariviisiliste eukleidiliste kauguste ruutude, mis on omakorda veel jagatud vaatluste arvuga k-ndas klastris. Kaht viimast valemit kombineerides saamegi defineerida k-keskmiste meetodi idee:

$$\underset{C_1,\ldots,C_K}{\text{minimeerida}} \{\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} \sum_{l=1}^{p} (x_{il} - x_{jl})^2\}. \tag{2}$$

Edasi kirjeldatakse algoritmi, millega saab jaotada vaatlused K-sse klastrisse nii, et viimane võrrand on minimeeritud. Olemas on küllaltki lihtne algoritm, saamaks kätte lokaalse ekstreemumi K-keskmiste leidmiseks.

Algoritm (k-keskmiste klasterdamine)

1. Kõigepealt määratakse esmased klastrite keskpunktid, olgu nendeks vektorid $\mathbf{m_1^{(1)}}, \ldots, \mathbf{m_K^{(1)}}$. Võimalusi esmaseid klastrite keskpunkte leida on mitmeid. Üks variant on määrata kõikidele vaatlustele juhuslikult number 1 ... K ning seejärel arvutada iga klastri jaoks keskpunkt (keskväärtuse p-mõõtmeline vektor). Teine variant on valida andmestikust juhuslikud n vaatlust ja määrata need esmasteks klastrite keskpunktideks.

2. Järgmisi tegevusi korratakse, kuni klastritesse jaotamine enam ei muutu:

   (a) Iga vaatlus $\mathbf{x_i}$ määratakse klastrisse $C_k^{(t)}$, mille keskpunkt on sellele vaatlusele lähim. Kaugus on määratud kasutades eukleidilist kaugust.

   $$C_k^{(t)} = \{\mathbf{x_i} : \|\mathbf{x_i} - \mathbf{m_k^{(t)}}\|^2 \leq \|\mathbf{x_i} - \mathbf{m_j^{(t)}}\|^2 \ \forall j, 1 \leq j \leq K\} \text{ [2]},$$

   kus t tähistab iteratsiooni sammu ja $\|\mathbf{x_i} - \mathbf{m_k}\|^2 = \sum_{l=1}^{p}(x_{il} - m_{kl})^2$ on eukleidilise kauguse ruut.

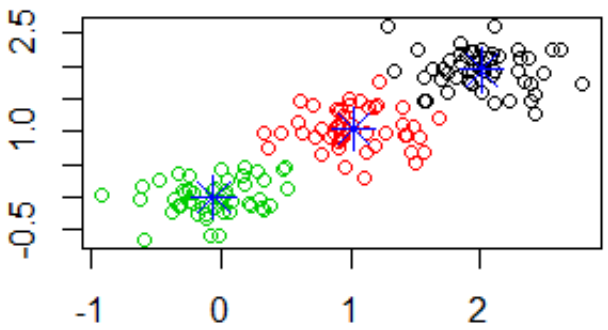
   (b) Iga klastri jaoks arvutatakse uus klastri keskpunkt (ingl. k. centroid). Klastri keskpunktiks on keskväärtuste p-mõõtmeline vektor

   $$\mathbf{m_k^{(t+1)}} = \frac{1}{|C_k^{(t)}|} \sum_{\mathbf{x_i} \in C_k^{(t)}} \mathbf{x_i} \text{ [2]}.$$

Algoritmi sammuga 2(a) määrame vaatlused klastrisse, mis asub sellele vaatlusele kõige lähemal. See vähendab klastrisisest hajuvust. Sammuga 2(b) leitud keskväärtused on konstandid, mis minimiseerivad klastrisisese hajuvuse. Seetõttu ei saa klastrite klastritesisese hajuvuse summa (väärtus (2)) algoritmi kasutades kordagi suureneda.

Kui klastritesse jaotus enam ei muutu, on leitud lokaalne ekstreemum. Kuna k-keskmiste meetod leiab lokaalse, mitte globaalse ekstreemumi, sõltub tulemus vaatluste esmasest (juhuslikust) klastritesse jaotamisest. Seetõttu on vajalik algoritmi kasutada mitmeid

kordi, erineva esmase jaotamisega, ning valida nende seast parim (näiteks kus väärtus (2) on kõige väiksem)



Joonis 1: Näide kahemõõtmelise andmestiku klasterdamise tulemustest, värv näitab klstrisse kuuluvust ning tärn klastri keskpunkti.

Joonisel 1 on näha k-keskmiste meetodiga klasterdamise tulemus genereeritud kahemõõtmelisl andmestikul. Erineva värviga on tähistatud klastritesse kuuluvus ning tärn näitab antud klastri keskpunkti [5].

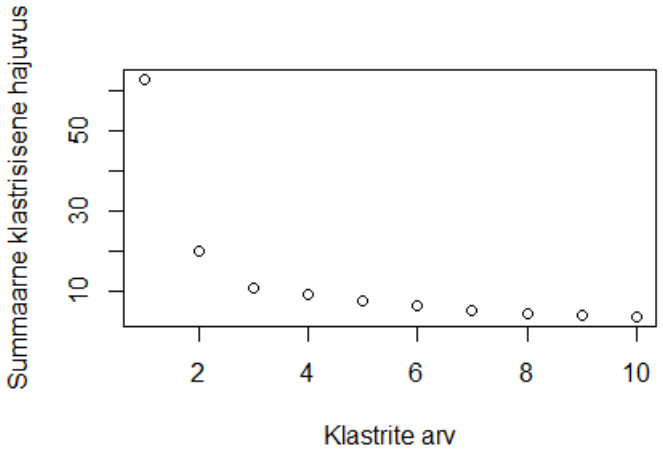
1.3 Klastrite arvu valimine

K-keskmiste meetodi kasutamiseks on vajalik eelnevalt määrata, et mitmesse klastrisse algoritm vaatlused jaotab.Teisisõnu on vaja leida meetod, mis arvutab klastrite arvu K. Klastrite optimaalse arvu hindamiseks on mitmeid meetodeid.

Intuitiivselt leitakse klastrite arvu K optimaalse valiku korral tasakaal maksimaalse täpsuse (määrates iga vaatlus eraldi klastrisse) ja maksimaalse kokkupakkimise (määrates kõik vaatlused ühte klastrisse) vahel.

1.3.1 Küünarnuki meetod

Küünarnuki meetod (ik. *elbow method*) on kõige sagedamini kasutatav meetod leidmaks klastrite arvu. See on visuaalne meetod, klastrite arv K valitakse graafikult. Klastrite arvu K valimiseks tuleb teha analüüs läbi erinevate klastrite arvuga K, alustades K=2 ja suurendades K-d igal järgneval analüüsil ühe võrra. Iga kord tuleb leida klastritesse jagamise jaotus ja klastrisisene hajuvus. Kui mingi K väärtuse korral väheneb klastrisiseste hajuvuste summa märgatavalt, siis see ongi otsitav klastrite arv. Tihti siiski see meetod ei tööta, kuna klastrisisene hajuvus väheneb ühtlaselt ning hüppelist hajuvuse vähenemist ei ole jooniselt näha. [6]



Joonis 2: Klastrite klastrisisene hajuvus

Joonisel 2 on kujutatud erinevate klastrite arvu korral summaarset klastrisisest hajuvust. Kasutatud on sama andmestikku, mis joonisel 1. Näheme, et pärast klastrite arvu 3 ei muutu oluliselt summaarna klastrisisene hajuvus. Seetõttu võibki arvata, et kolme klastriga saab kirjeldada kasutatud andmestikku.

1.3.2 Muud meetodid klastrite arvu leidmiseks

Statistikapaketis R on pakett nimega NbClust, kus on klastrite arvu K leidmiseks 30 eri algoritmi. Siiski on enamike idee seotud klastrisisese ja klastrite vahelise hajuvusega.

1.4 Üldine lineaarne mudel

Kuna edasipidises analüüsis kasutatakse väheselmääral ka lineaarset regresioonanalüüsi antakse siinkohal ülevaade üldisest lineaarsest mudelist. Järgnev alapeatükk põhineb E. Kääriku Andmeanalüüs II loengukonspektil [7].

Vaatame lineaarset mudelit maatrikskujul

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

kus $\mathbf{y}$ on $n \times 1$ uuritava tunnuse vektor,
$\mathbf{X}$ on $n \times p$ plaanimaatriks, $p = k + 1$,
β on $k \times 1$ tundmatute parameetrite vektor ja
ϵ on $n \times 1$ juhuslike vigade vektor.
Lineaarne mudel sisaldab nii pidevaid argumente, diskreetseid argumente kui ka koosmõjusid. Tundmatud parameetrid leitakse vähimruutude meetodil normaalvõrrandisüsteemi $(\mathbf{X}^T\mathbf{X})\beta = \mathbf{X}^T\mathbf{y}$ lahenditena.
Lineaarse mudeli eeldusteks on:

- lineaarne seos uuritava tunnuse ja argumendi vahel;
- vigade sõltumatus;
- vigade normaaljaotus;
- vigade konstantne hajuvus.

Kui eelpool nimetatud eeldused on täidetud, on uuritav tunnus normaaljaotusega $Y \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$.

1.5 Korrespondentsanalüüs

Korrespondentanalüüs (inglise keeles *Correspondence analysis (CA)*) on analoogne analüüsimeetod statistilistes analüüsides pidevalt kasutatavale peakomponentanalüüsile. Korrespondentanalüüsi tegemiseks peavad andmed olema positiivsed ja mõõdetud samal skaalal. Tihti kasutatakse antud analüüsimeetodit sagedustabelitel. CA analüüsib

andmestiku rea- ja/ või veeruprofiile, mis saadakse, kui andmestikus olevate ridade iga element jagada antud rea reasummaga või veeru iga element jagada veerusummadega.

Iga rida ja veerg andmestikus kaalutakse nende massiga. Mass on defineeritud, kui reasumma ja kogusumma või veerusumma ja kogusumma suhe. Kaugused reaprofiilide või veeruprofiilide vahel on defineeritud hii-ruut kaugusega. Populaarne meetod, kuidas CA tulemust kirjeldada, on sümmeetrilise kaardi abil, kus nii reaprofiilid kui veeruprofiilid on kujutatud samaaegselt.

Analüüsimeetodiga täpsemalt tutvumiseks soovitab autor G.Aru bakalaureusetööd [9] ja raamatut *„Multivariate Analysis of Ecological Data"* [8].

1.6 Logistiline regressioonanalüüs

Järgnev alapeatükk põhineb E. Kääriku Andmeanalüüs II loengukonspektil [7].

Logistilist regressioonanalüüsi kasutatakse juhul, kui funktsioontunnusel on ainult kaks võimalikku väärtust 1 ja 0, kus 1 tähistab sündmuse esinemist. Logistilise mudeliga hinnatakse šansi logaritmi:

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

kus $\pi = P(Y = 1)$ on sündmuse esinemise tõenäosus, $\frac{\pi}{1-\pi}$ on sündmuse esinemise šanss ja $\beta_1, \ldots, \beta_k$ on parameetrid.

Logistilise regressiooni parameetrite hindamiseks kasutatakse suurema tõepära meetodit, kusjuures tõepärafunktsiooni maksimeerimine toimub enamasti Newton-Raphsoni iteratiivse algoritmi abil.

2 Andmestiku kirjeldus

Tartu Ülikooli Eesti geenivaramu (edaspidi TÜ geenivaramu) andmebaasis on üle 51 000 geenidoonori andmed. Kõik geenidoonorid täitsid küsimustiku, milles on küsimusi nii haiguste, toitumise kui ka sugupuu kohta. Käesolevas töös kasutatud andmestikus on 45814 geenidoonori andmed ning vaadeldakse 28 tunnust. Neist 10 on taustatunnused ja 18 toitumistunnused. Algsest geenivaramu andmebaasist, kus oli üle 51 000 geenidoonori andmed, jäeti välja inimesed, kellel mõne toitumistunnuse väärtus oli puudu.

Taustatunnustena on töös kasutatusel geeniproovi andmise aasta, geenidoonori sugu, vanus geenidoonoriks liitumisel, kehamassiindeks (edasipidi ka KMI), sünnimaakond ja elukohamaakond, haridustase ja suitsetamisstaatus (praegune, endine või pole üldse suitsetaja). KMI on arvutatud kaalu (kg) ja pikkuse (m) ruudu suhtena. Haigustest vaadeldi antud töös vaid, kas inimesel oli geenivaramuga liitudes diagnoositud südame isheemiatõbi või kas ta on haigestunud sellesse haigusesse hilisemalt (kuni 10 aasta jooksul pärast liitumist). Toitumistunnustena on tabelis 14 toiduaine tarbimise sagedus nädala jooksul. Lisaks uuriti, mitu tassi kohvi ja teed joob ning mitu viilu saia ja leiba sööb geenidoonor päevas.

2.1 Taustatunnused

TÜ geenivaramu andmestikus, mille põhjal antud töö on tehtud, on kokku 45814 indiviidi andmed. Mehi on 15282, mis moodustab 33,36 % kõigist kasutatavas andmestikus olevatest indiviididest, ja naisi on 30532, mis moodustab 66,64% kõigist kasutatavas andmestikus olevatest indiviididest (vt. Tabel 1). Andmestikus olevate inimeste keskmine vanus on 41,8 aastat. Naiste keskmine vanus on 42,7 aastat ja meeste keskmine vanus on 39,8 aastat. Geeniproovi andmise ajal oli noorim geenidoonor 13aastane naissoost isik ja kõige vanem 101aastane meessoost isik.

TÜ geenivaramuga sai mõnede eranditega liituda aastast 2002 kuni aastani 2013. Aastal 2006 ei saanud geenidoonoriks liituda ja aastal 2005 liitus ainult 9 inimest. Liitumisaastaid vaadeldi kolmes grupis - enne 2006. aastat liitujad, aastatel 2007-2008 liitujad ja pärast 2008. aastat liitujad. Enne 2006. aastat liitus 19,0 % (8686 inimest), aastatel 2007-2008 liitus 37,9% (17380) ja pärast 2008. aastat 43,1% (19748) andmestikus

Tabel 1: Vanuse, KMI, hariduse ja suitsetamise näitajad sugude lõikes

	Mees n=15282	**Naine** n=30532	**Kokku** n=45814
Keskmine vanus (standardhälve)	39,9 (16,5)	42,7 (16,2)	41,8 (16,3)
Keskmine KMI (standardhälve)	26,2 (5,5)	26,0(6,8)	26,1 (6,4)
vähemalt keskharidusega, %	78,9	85,7	83,5
kõrgharidusega, %	20,8	27,1	25,0
Praeguseid suitsetajaid, %	41,5	23,9	29,8
Endisi suitsetajaid, %	17,9	10,0	12,6

olevatest inimestest.

Inimese kehamassiindeksit loetakse normaalseks, kui see jääb vahemikku 19-25 [4]. Tabelist 1 on näha, et antud andmestikus olevatel inimestel on keskmine kehamassiindeks 26,1, mis ületab normaalseks peetavat kehamassiindeksi piiri. Keskmine kehamassiindeks meestel ja naistel on sarnane (vastavalt 26,2 ja 26,0). Normaalseks peetava kehamassiindeksi piiridesse jääb 45,5 % andmestikus olevatest inimestest.

Geenidoonorite vastatud küsimustikus oli ka küsimus inimeste suitsetamisharjumuste kohta. Geenidoonoritest 29,8% olid küsimustiku täitmise hetkel suitsetajad, 12,6% olid endised suitsetajad ja 57,5% ei ole kunagi suitsetajad olnud. Meeste seas oli suitsetajate osakaal palju suurem kui naiste seas. Meestest olid hetkel suitsetajad 41,5% vastanutest, sama näitaja naiste seas on 23,9%. Neid, kes pole kunagi suitsetajad olnud, oli meeste seas 40,5% ning naiste seas 66,0%.

Andmestikus on veel tunnused geenidoonorite elukoha ja sünnikoha kohta maakonna tasemel. Kõik maakonnad olid andmestikus ära mainitud nii sünnikohana kui ka elukohana. Andmestikus olevatest inimestest oli kõige vähem pärit Hiiumaalt (472 inimest) ning kõige rohkem inimesi oli pärit Harjumaalt, sh Tallinn (7484 inimest). Samas oli kõige vähem inimesi märkinud elukoha maakonnaks Saaremaa (210 inimest). Harjumaad märkis elukohana peaaegu 3000 inimest rohkem kui sünnikohana, 10350 inimest. Elukoha oli vastamata jätnud 9659 inimest ja sünnikoha 9706 inimest.

Taustatunnusena vaadeldi antud töös ka inimeste haridustaset. Küsimustikus oli küsitud kõrgeimat omandatud haridustaset ning vastusevariantideks olid alghariduseta, algharidus, põhiharidus, keskharidus, keskeriharidus, rakenduslik kõrgharidus, kõrgha-

ridus, teaduskraad või ei tea. Analüüsi lihtsustamiseks kodeeriti antud töös andmed ümber. Haridustaset vaadeldi kolmes grupis - alla keskhariduse, keskharidus(k.a. keskeri) ning kõrgharidus. Kõrgharidus oli 25% vastanutest. Geenidoonoriks olemise hetkel oli kõrgeima haridustasemena omandanud keskhariduse 58,4% vastanutest ning põhihariduse 16,5% vastanutest. Kui keskharidusega inimeste osakaal sugude lõikes märgatavalt ei erinenud (vastavalt 58,0% meestel ja 58,6% naistel), siis kõrgharidusega inimesi oli naiste hulgas rohkem (vastavalt 20,8% meestest ja 27,1% naistest).

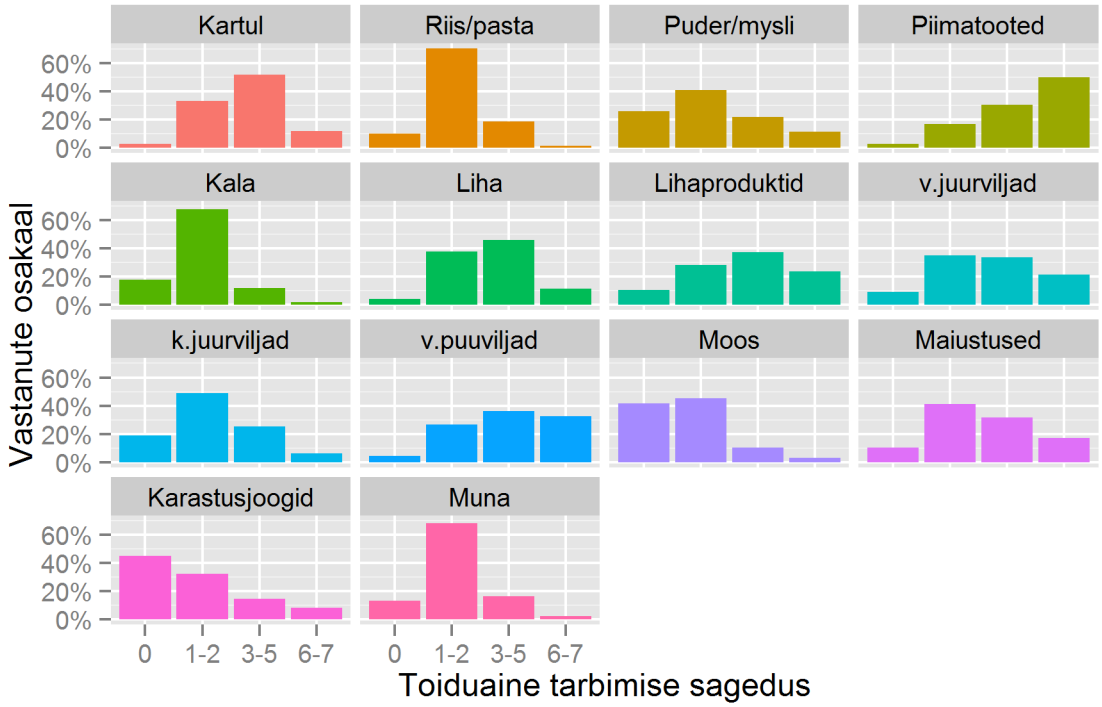
Kuigi geenivaramu andmebaasis on andmed mitmete haiguste kohta, vaadeldakse antud töös vaid seda, kuidas on südame isheemiatõbi seotud leitud toitumismustritega. Südame isheemiatõbi on südant verega varustavate pärgarterite muutustest ning südamelihase vere- ja hapnikuvaegusest tekkinud haigus, mille korral südamelihas ei saa piisavalt verd. Südame isheemiatõbi on meestel ja vanemaealistel naistel kõige sagedasem südamehaigus. Südame isheemiatõve ennetamiseks tuleb toidus vähendada loomsete rasvade ja suhkru osa, regulaarselt tegelda füüsilise tegevusega ning mitte suitsetada. [11] Inimesi, kellel oli TÜ geenivaramuga liitumise hetkel diagnoositud südame isheematõbi oli 2406. Neid, kellel on südame isheemiatõbi avastatud peale geenivaramuga liitumist, oli 1164.

2.2 Toitumistunnused

Geenidoonorite täidetud küsimustikus oli küsimus 14 toiduaine söömise kohta nädala jooksul. Uuriti kartuli, riisi/makaroni, pudru/müsli/helveste, piimatoodete, kala, liha, lihaproduktide, värskete juurviljade, keedetud juurviljade, värskete puuviljade/marjade, kompottide/keediste, maiustuste, karastusjookide ja munade tarbimist ühe nädala jooksul. Vastusevariantideks olid „mitte kordagi", „1-2 päeval", „3-5 päeval" ja „6-7 päeval". Edaspidi kasutatakse kompottide ja keediste asemel terminit moos.

Jooniselt 3 selgub näiteks, et kõige levinum nii kartuli kui ka liha söömise sagedus on 3-5 päeval nädalas, kuid nii riisi/pasta kui ka kala, muna ja keedetud juurviljade puhul on levinumaks söömise sageduseks 1-2 päeval nädalas

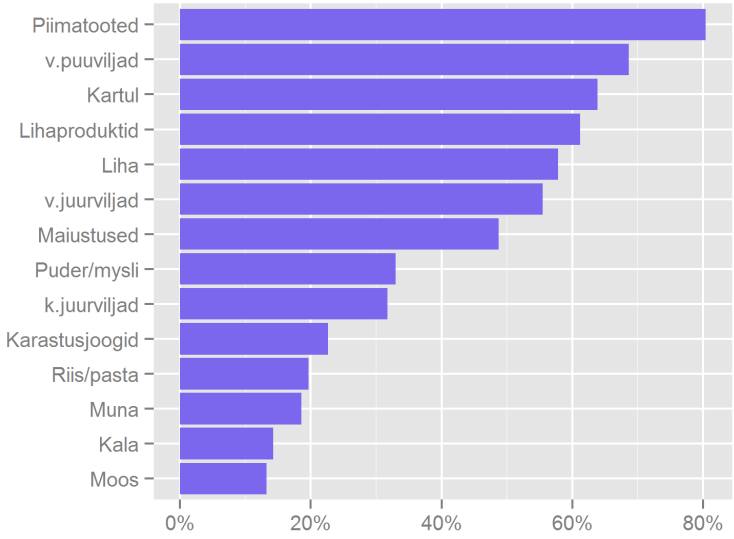
Toiduainete tarbimise vahelised korrelatsioonid osutusid kõik väga väikesteks, jäädes enamasti absoluutväärtuselt alla 0,2 (vt ka Lisa 1). Kõige kõrgem korrelatsioon oli värskete juurviljade ja värskete puuviljade vahel (Spearmani korrelatsioonikordaja $r =$



Joonis 3: Toiduainete protsendiline jaotus toiduainete kaupa

$0,37$). Sellest võib eeldada, et inimesed, kes söövad tihedamini värskeid juurvilju söövad ka tihedamini värskeid puuvilju. Teistest tugevamad korrelatsioonid esinesid veel ka lihaproduktide ja saia ($r = 0,30$), kartuli ja liha ($r = 0,25$) ning katuli ja lihaproduktide ($r = 0,22$) vahel, pudru/müsli ja keedetud juurviljade ($r = 0,23$) ning pudru/müsli ja moosi ($r = 0,22$), lihaproduktide ja karastusjookide ($r = 0,23$), värskete ja keedetud juurviljade ($r = 0,25$) ning keedetud juurviljade ja kala ($r = 0,22$) vahel. Kõik mainitud seosed on positiivsed ehk ühesuunalised. See tähendab, et eelpool mainitud toiduainete korral toob ühe toiduaine sagedasem tarbimine kaasa teise toiduaine sagedasema tarbimise.

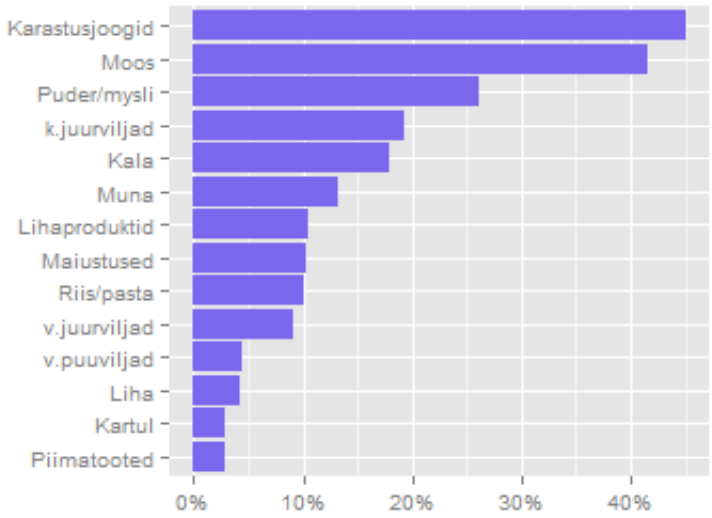
Mitmed eelpool mainitud seostest on ka väga loogilised, kuna kartulit tarbitakse sageli koos liha või lihaproduktidega, kala keedetud juurviljadega, saia vorstiga ja putru moosiga. Ei saa öelda, et pudru/müsli ja keedetud juurviljade vaheline seos tähendaks, et neid toiduaineid tarbitakse koos, aga seose põhjuseks võib olla asjaolu, et need toiduained ei ole väga sagedasti tarbitavad ja et neid tarbivad rohkem inimesed, kes toituvad tervislikumalt. Sama on ka lihaproduktide (vorsti) ja karastusjookidega, ehk neid võivad tarbida rohkem inimesed, kes toituvad ebatervislikult.



Joonis 4: Vähemalt kolm korda nädalas tarbitavate toiduainete protsentuaalne jaotus

Jooniselt 4 on näha, et vähemalt 3 korda nädalas tarbis piimatooteid 80% vastanutest. Vähemalt 3 korda nädalas tarbis värskeid puuvilju 69% inimestest ja kartulit 64% inimestest. Ainult 13% doonoritest tarbis moosi ning 14% kala kolmel või rohkemal päeval nädalas.

Toiduained, mis olid kõige sagedamini menüüst välja jäetud, on karastusjoogid, moos, puder/müsli/helbed, keedetud juurviljad, kala ja muna. Karastusjooke ei tarbi 45% inimestest, moosi ei tarbinud 42% inimestest ning putru/müslit/helbeid ei tarbinud 26% inimestest. Piimatooteid ja kartulit ei tarbinud ainult 3% inimestest. Seda on näha jooniselt 5.



Joonis 5: Toituainete mittetarbimise protsendiline jaotus

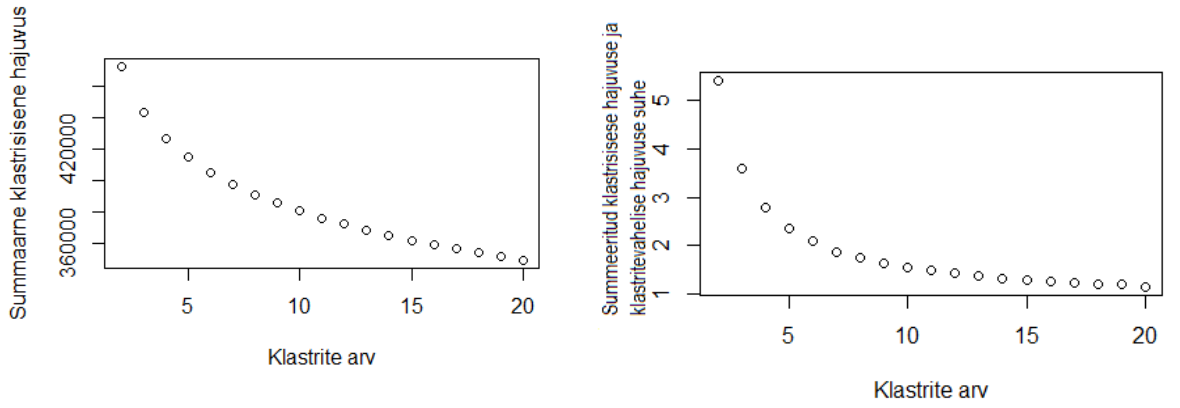
Joonis 6: Kohvi,tee, leiva ja saia päeva jooksul tarbimise protsendiline jaotus

Algses andmestikus olid kohvi, tee, leiva ja saia päevase tarbimise kogus ning skaala võis ulatuda 0st kas 30 või 40ni. Seetõttu ei olnud need võrreldavad ülejäänud toitumistunnustega. Nende tunnuste ühtlustamiseks kodeeriti kohvi, tee, leiva ja saia tarbimine ümber. Väärtus 1 näitas, et viimati nimetatud toiduained ei tarbita üldse, päevas 1-2 tassi/viilu tarbimine sai väärtuse 2, päevas 3-5 tassi/viilu tarbimine väärtuse 3 ning

rohkem kui 5 tassi kohvi, tee või leiva, saia viilu tarbimine sai väärtuse 4.

Jooniselt 6 on näha, mitu tassi kohvi ja teed ning mitu viilu leiba ja saia tarbis geenidoonor päevas. Kohvi ei tarbinud üldse 18% geenidoonoritest, seevastu 57% jõi kohvi 1-2 tassi päevas ning 2,3% vastanutest jõi rohkem kui 5 tassi kohvi päevas. Teed ei joonud üldse 25,7% geenidoonoritest ning 58% jõi teed 1-2 tassi päevas. Saia tarbis 41% vastanutest 1-2 viilu päevas. Leiba tarbis kuni 2 viilu päevas ja 3-5 viilu päevas 42% vastanutest.

3 Analüüs



(a) Summaarne klastrisisene hajuvus erineva arvu klastrite korral

(b) Summaarse klastrisisese hajuvuse ja klastritevahelise hajuvuse suhe

Joonis 7: Hajuvused erinevate klastrite arvu korral

Klasteranalüüs k-keskmiste meetodil viidi läbi programmiga R. k-keskmiste algoritm vajab eelnevalt teadmist klastrite arvust K. Selle väljaselgitamiseks prooviti andmestiku klasterdada 19ks eri arvuks klastriks (klastreid 2 kuni 20). Iga arvu klastri kohta kanti joonisele 7a summaarne klastrisisene hajuvus. Kahjuks antud andmete põhjal ei esine olukorda, kus mingi K korral klastrisisene hajuvus hüppeliselt väheneks ning K-d suurendades oleks muutused oluliselt väiksemad. Mitmed teised klastrite arvu kindlaks tegemise meetodi vajasid nii suure andmestiku korral suuremat arvutimahtu kui antud analüüsi käigus oli võimalik kasutada. Seetõttu prooviti leida sobilik klastrite arv ka ainult andmesikust võetud väiksemas juhuvalimis. Tarkvara R poolt võimaldatud

algoritmid pakkusid sobivaks klastrite arvuks kaks kuni 18. Autor teostas edaspidise analüüsi erinevate klastrite arvu korral, keskendudes rohkem juhule, kus andmestik jaotati 8 klastriks.

3.1 k-keskmiste meetodi rakendamine tarkvarapaketi R abil

k-keskmiste klasterdamine programmiga R toimib funktsiooni *kmeans(x, centers)* abil. Funktsiooni argumentideks on numbriliste väärtustega andmestik x ja *centers*, mis võib olla klastrite arv või esmaste keskmistena kastutatav maatriks. Antud funktsioon tagastab klastritesse jaotuse, klastrite keskpunktid, klastrisisese, klastrite vahelise ja koguhajuvuse ning klastrite suurused. Parameetriga *nstart* saab määrata, mitu korda jooksutab programm k-keskmiste algoritmi erinevate esmaste klastrite keskpunktidega. Funktsioon tagastab ainult kõige parema klastritesse jaotuse ehk sellise, kus klastritesisese hajuvuse summa oleks kõige väiksem. Parameetriga *iter.max* saab määrata maksimaalse iteratsioonide arvu (vaikimisi on selle parameetri väärtuseks 10). Viimasena saab veel määrata, et mis algoritmi täpselt k-keskmiste klasterdamiseks kasutatakse. Valikuteks on „Hartigan-Wong", „Lloyd", „Forgy" ja „MacQueen", kus „Lloyd" ja „Forgy" viitavad tegelikult samale algoritmile. Vaikeväärtuseks R-is on „Hartigan-Wong", mida kasutatakse ka antud töös. [5]

Klastrite visualiseerimiseks on tarkvaras R mitu võimalust. Antud töös on kasutatud kahte erinevat jooniste tegemise viisi. Funktsiooni *pheatmap* kasutati selleks, et võrrelda üksikute toiduainete keskmist tarbimist eri klastrites (algsete andmetega joonised Lisades 2-5 ja skaleeritud andmetega joonised 8 ja 16). Viimati nimetatud joonised on küll väga informatiivsed, kuid raskemini mõistetavad valdkonnaga vähem kokkupuutunud inimestele. Lisaks kasutat funktsiooni *stars*, mis on lihtsamini arusaadav, ent samas mitte nii informatiivne (joonised 15 ja 17).

3.2 Klastrite kirjeldused

Tabelist 2 on näha, et k-keskmiste klasterdamise meetodi kasutamisel tekkinud klastrid on suhteliselt sarnae suurusega.

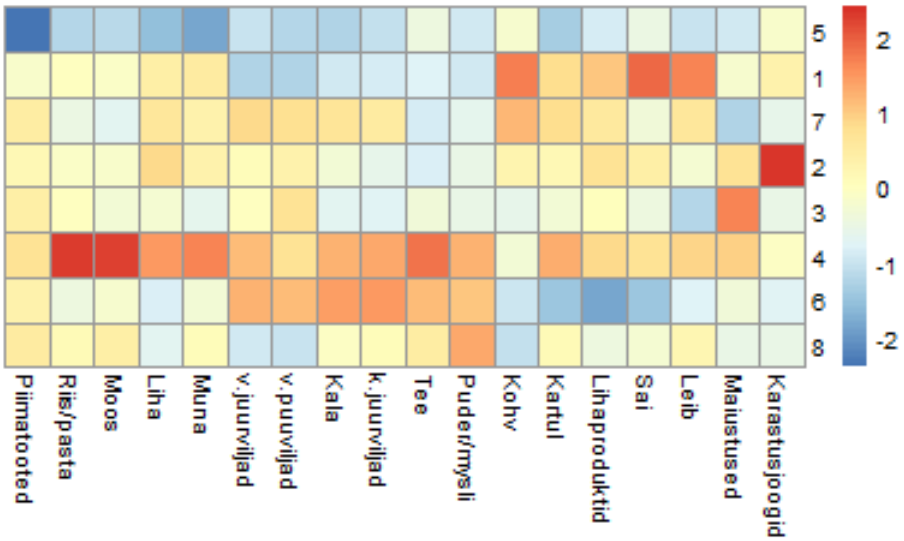
Tabel 2: Klastrite suurused

Klastrite arv	Klaster 1	2	3	4	4	6	7	8
klastreid 2	23265	22549						
klastreid 4	9531	12268	11131	12884				
klastreid 6	8656	7193	7241	7522	7125	8077		
klastreid 8	6071	5915	5395	5787	6065	5814	5498	5269

Klasterdamise k-keskmiste meetodi tulemusel jaotatakse andmestik soovitud arvuks klastriteks ning igat klastrit jääb iseloomustama 18-mõõtmeline vektor, mis näitab toiduainete keskmist tarbimist antud klastris.

Kuna erinevad klastrite arvu kindlaks tegemise algoritmid andsid erinevaid tulemusi klastrite arvu osas, kirjeldatakse antud töös klastreid küll erinevate klastrite arvu korral, aga keskendutakse peamiselt klastrite arvule 8.

3.3 Klastrite iseloomustus klastrite arvu 8 korral



Joonis 8: Toiduainete tarbimine klastrite lõikes: klastrikeskmise erinevus üldkeskmisest, juhul kui klastrite arv on 8

Joonis 8 kirjeldab toiduainete tarbimise sagedust võrreldes teiste klastritega. Mida tumedam punane on ruut joonisel, seda sagedamini tarbivad vaadeldavasse klastrisse kuuluvad inimesed antud toiduainet võrreldes teistesse klastritesse kuuluvate inimestega. Mida tumedam sinine on ruut joonisel, seda harvem tarbitakse antud toiduainet võrreldes teistesse klastritesse kuuluvate inimestega. Helebeež tähistab keskmist tarbimist võrreldes teiste klastritega. Skaleerimata andmetele on analoogne joonis toodud Lisas 5.

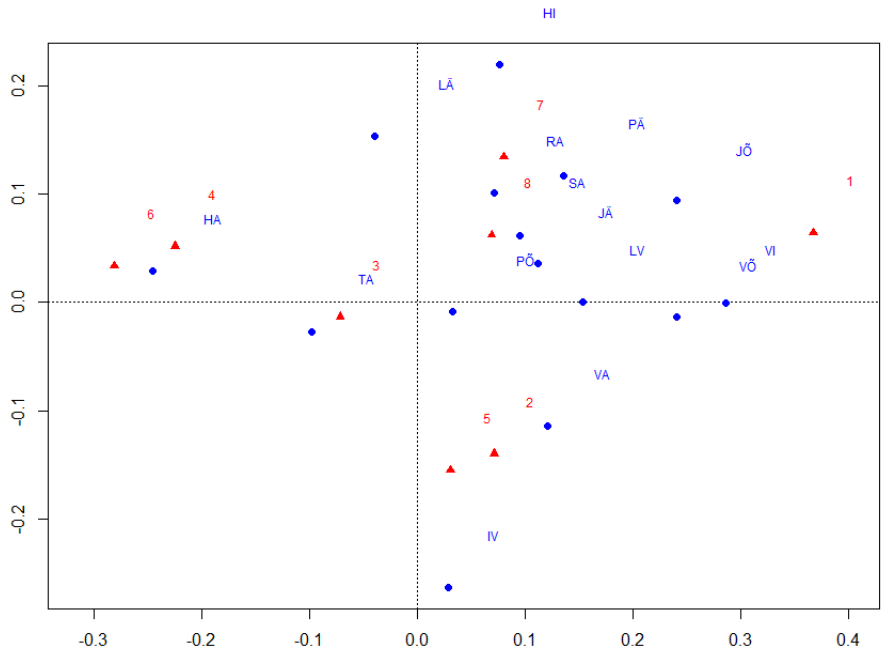
Klastrisse 5 kuuluvad inimesed tarbivad pea kõiki toiduaineid harvem kui teistesse klastritesse kuuluvad inimesed. Võrreldes teiste klastritega tarbitakse märgatavalt harvem piimatooteid, vaid kuni 2 korda nädalas, mune ja liha. Klastrisse 1 kuuluvad inimesed tarbivad teiste klastritega võrreldes rohkem kartulit, lihaprodukte, kohvi(keskmiselt rohkem kui kaks tassi päevas), saia (rohkem kui 5 viilu päevas) ja leiba (3-5 viilu päevas). Teiste klastritega võrreldes tarbivad klastrisse 4 kuuluvad inimesed sagedamini

kõiki toiduaineid peale kohvi ja karastusjookide. Kõige sagedamini tarbivad klastrisse 4 kuuluvad inimesed piimatooteid, värskeid puu- ja juurvilju ning lihaprodukte. Kõige tervislikumalt toituvad klastrisse 6 kuuluvad inimesed, kes söövad teistest sagedamini värskeid juur- ja puuvilju, kala, keedetud juurvilju, teed ning putru või müslit. Karastusjooke ei tarbita üldse, saia tarbitakse keskmiselt 1-2 viilu päevas ning lihaprodukte vaid 1-2 päeval nädalas. Teistest klastritest eristuvad klastrisse 2 kuuluvad inimesed igapäevase karastusjookide ning sagedasema liha, lihaproduktide ja maiustuste tarbimisega. Klastri 3 inimesed tarbivad erinevaid toiduaineid keskmisest pigem harvem. Erandiks on maiustuste sage tarbimine, värskeid puuvilju tarbitakse ka veidi rohkem. Klastris 7 olevate inimeste toiduainete tarbimine jääb toiduainete lõikes keskmise lähedale. Rohkem juuakse kohvi ning vähem teed. Lisaks süüakse rohkem kartulit, lihaprodukte, leiba, juur- ja puuvilju ja kala. Märgatavalt vähem süüakse maiustusi. Klaster 8 puhul jääb samuti paljude toiduainete tarbimine keskmise lähedale. Keskmisest vähem tarbitakse värskeid puu- ja juurvilju ning kohvi. Rohkem süüakse putru või müslit ja juuakse teed.

3.3.1 Seosed klastritesse jaotuse ning elu- ja sünnikohtade vahel

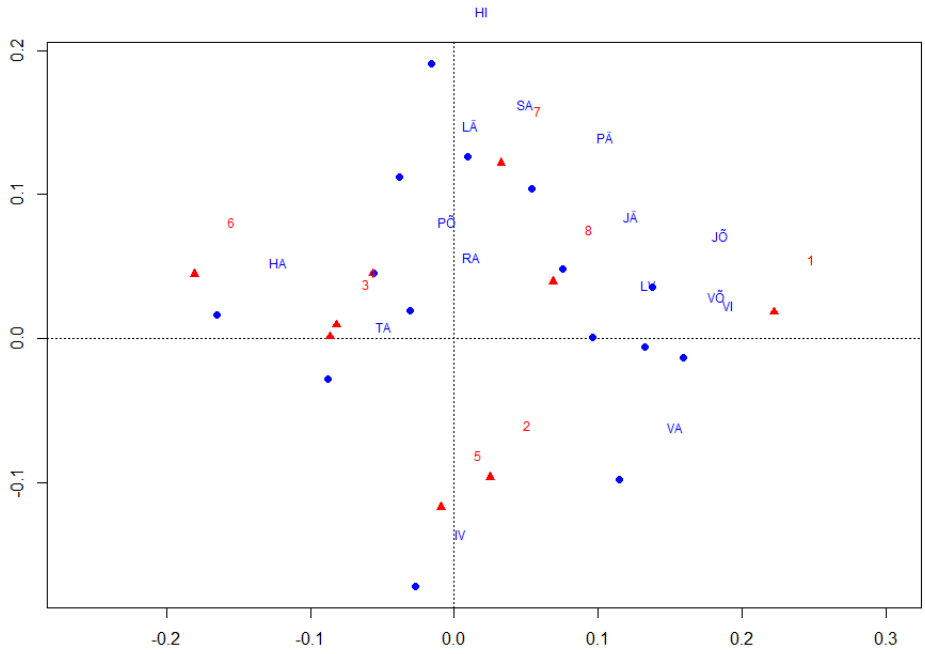
Avastamaks seoseid inimeste klastritesse jaotuse ja nende elukohtade vahel, teostati vastavate tunnuste vahel korrespondentsanalüüs. Tulemused kanti joonisele 9, kus sinisena on märgitud maakonnad ja punasena klastrid.

Korrespondentsanalüüsi tulemustest on näha, et Harjumaa ja Tartumaa inimesed eristuvad teistest ja on omavahel mingil määral sarnaste toitumisharjumustega. Siiki kirjeldavad Harjumaa inimeste toitumisharjumusi kõige paremini klastrid 4 ja 6, Tartumaa inimeste oma klaster 3. Samamoodi eristuvad teistest Ida-Virumaa ja Valgamaa, mida kirjeldavad klastrid 2 ja 5. Jooniselt 9 on näha veel, et Hiiumaa ja Läänemaa asetsevad joonisel lähestikku, ehk nende toitumisharjumused on sarnased.



Joonis 9: Korrespondentsanalüüsi iseloomustav graafik klastrite ja elukoha maakondade vahel; HA-Harju, HI-Hiiu, IV-Ida-Viru, JÕ-Jõgeva, JÄ-Järva, LV-Lääne-Viru, PÕ-Põlva, PÄ-Pärnu, RA-Rapla, SA-Saare, TA-Tartu, VA-Valga, VI-Viljandi maakond

Sarnasused Saaremaa, Hiiumaa, Läänemaa ja Pärnumaa inimeste toitumistes tulid paremini välja, kui teostada korrespondentsanalüüs inimeste sünnimaakondade ja klastritesse jaotuse vahel. Joonisel 10 on näha viimati nimetatud korrespondentsanalüüsi tulemused. Joonisel asetsevad lähestikku veel Harjumaa, Tartumaa, Põlvamaa ja Raplamaa ehk nendes maakondades sündinud inimesed toituvad sarnasemalt. Üheltpoolt iseloomustab neid maakondi ning eriti Harjumaad klaster 6 ehk mingi hulk inimesi toitub tervislikult. Teisalt iseloomustavad antud maakondi klastrid 3 ja 4 ehk teine grupp inimesi sööb maiustusi või üldse kõiki toiduaineid keskmisest rohkem. Teistest maakondadest eristuvad ka siin joonisel Ida-Virumaa ja Valgamaa. Viimati nimetatud maakondi iseloomustavad kõige paremini klastrid 2 ja 5.



Joonis 10: Korrespondentsanalüüsi iseloomustav graafik sünnimaakonna ja klastrite vahel; HA-Harju, HI-Hiiu, IV-Ida-Viru, JÕ-Jõgeva, JÄ-Järva, LV-Lääne-Viru, PÕ-Põlva, PÄ-Pärnu, RA-Rapla, SA-Saare, TA-Tartu, VA-Valga, VI-Viljandi maakond

3.3.2 Seosed toitumise klastrite ja kehamassiindeks vahel

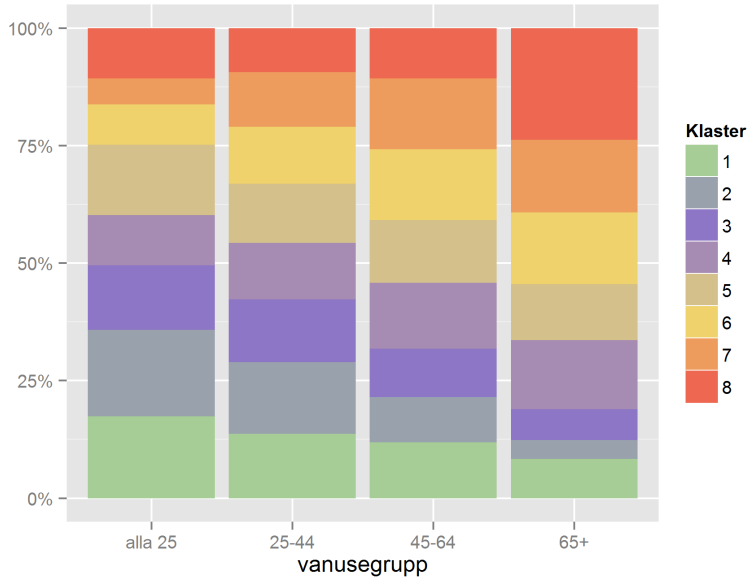
Vaatamaks, kas inimeste kehamassiindeks (KMI) on seotud sellega, millisesse klastrisse ta kuulub, teostati lineaarne regressioonanalüüs. Üldisesse lineaarsesse mudelisse võeti sugu, vanus, soo ja vanuse koosmõju ning faktorina klastrisse kuuluvus.

Tabel 3: KMI erinevused võrreldes 1.klastriga

klastri nr	β	standardviga	p-väärtus
2	0,14	0,11	0,21
3	-0,53	0,12	$4,73 * 10^{-6}$
4	-0,38	0,11	$7,54 * 10^{-4}$
5	0,28	0,11	$1,34 * 10^{-2}$
6	-0,13	0,12	$0,26$
7	0,93	0,12	$5,85 * 10^{-16}$
8	-0,43	0,12	$2,07 * 10^{-4}$

Kui saadud mudeli järgi võrrelda 8 inimest, kes on samast soost ja sama vanad, aga kuuluvad kõik erinevatesse toitumisklastritesse, siis klastrisse 2 kuuluval inimesel on kehamassiindeks mudeli järgi keskmiselt 0,14 võrra suurem, klastrisse 5 kuuluval inimesel 0,28 võrra suurem ning klastrisse 7 kuuluval inimesel 0,93 võrra suurem kui klastrisse 1 kuuluval inimesel. Samas on klastrisse 3, 4, 6 ja 8 kuuluvatel inimestel kehamassiindeks antud mudeli järgi väiksem, vastavalt 0,53, 0,38, 0,13 ja 0,43 võrra, kui klastrisse 1 kuuluval inimesel (vt Tabel 3). Ehk keskmiselt on kõige madalamad kehamassiindeksid klastrisse 3 kuuluvatel inimestel ja kõige kõrgemad kehamassiindeksid klastrisse 7 kuuluvatel inimestel. Klastritesse 2 ja 6 kuuluvad inimesed ei erine kehamassiindeksi järgi oluliselt klastrisse 1 kuuluvatest inimestest.

3.3.3 Klastrid vanusegruppide lõikes



Joonis 11: Klastritesse jaotus vanusegruppide lõikes

Kõige rohkem noori ehk alla 18aastaseid inimesi kuulub klastrisse 2 (18,4%). Peaaegu samapalju kuulub ka klastrisse 1 (17,4%). Klaster 2 paistis silma keskmisest sagedama karastusjookide tarbimisega ning klaster 1 tarbis rohkem kohvi, kartulit, vorsti, saia ja

leiba. Kõige vähem alla 18aastaseid inimesi kuulub klastritesse 7 (5,6%) ja 6 (8,6%). Kuna klastrisse 6 kuuluvad inimesed toitusid keskmisest tervislikumalt, siis vaid vähesed noored toituvad tervislikult. Klastrisse 7 kuuluvad inimesed toituvad toiduainete lõikes keskmiselt, aga pigem üsna tervislikult. Samas joovad nad ka kohvi, mida klastrisse 6 kuuluvad inimesed ei joo.

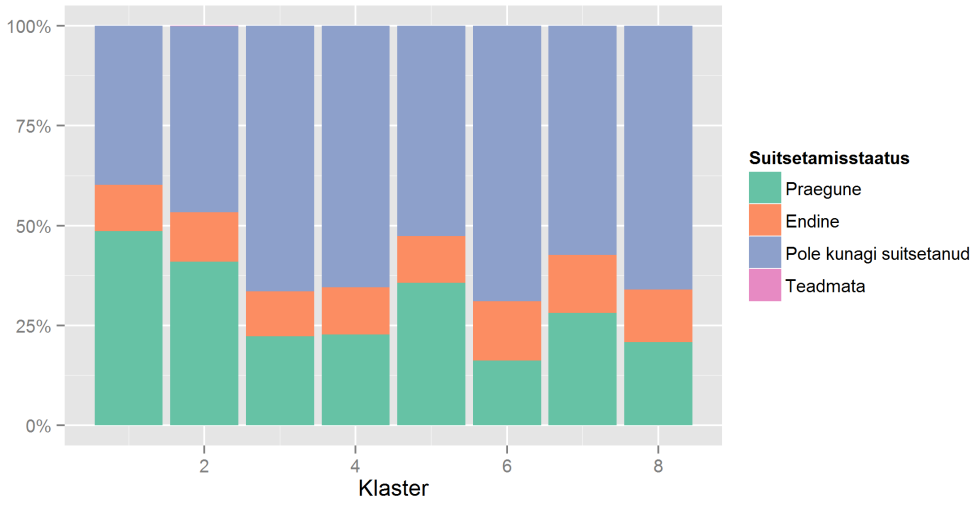
Nooremad tööealised ehk vanusegruppi 25-44 kuuluvad inimesed jagunevad kõikide klastrite vahel enamvähem võrdselt. Erandiks on klaster 2, kuhu kuulub kõige rohkem selle vanusegrupi inimesi (15,3%) ning klaster 8, kuhu kuulub kõige vähem selle vanusegrupi inimesi (9,3%). Klastrile 2 on iseloomulik igapäevane karastusjookide tarbimine.

Vanematest tööealistest ehk vanusegruppi 45-64 kuuluvatest inimestest kuulub klastrisse 6 ja 7 võrdselt 15,1%. See vanusegrupp sööb rohkem tervisliku toitu ehk tarbib keskmisest rohkem puu-ja juurvilju, kala ja putru ning joob teed. Klastrisse 7 kuuluvad inimesed joovad lisaks ka kohvi. Kõige vähem kuulub selle vanusegrupi inimestest klastrisse 2 (9,7%).

Kõige vanemast vanusegrupist ehk 65aastastest ja vanematest inimestest kuulub lausa 23,7% klastrisse 8. Klastris 8 tarbitakse keskmisest palju rohkem putru/müslit ning keskmisest palju vähem kohvi. Ülejäänud toiduaineid tarbitakse enamvähem võrdselt. Veel kuulub natuke üle 15% selle vanusegrupi inimestest nii klastrisse 6 kui 7. Samas ainult 4% inimestest kuulub klastrisse 2 ehk igapäeva menüüsse ei kuulu neil karastusjook. Veel kuulub vähe pensionealisi inimesi klastritesse 3 (6,6%) ja 1 (8,3%).

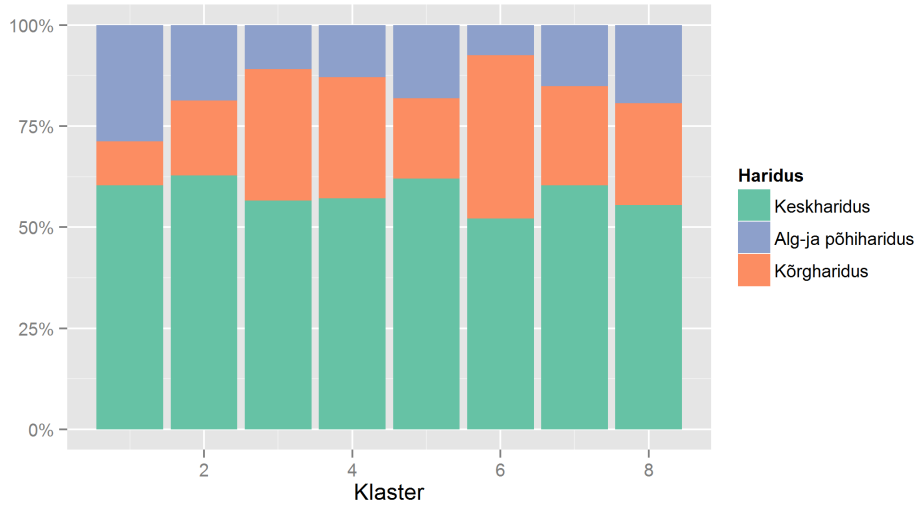
3.3.4 Klastrid ja suitsetamine

Suitsetajate osakaal klastrites on erinev. Jooniselt 12 on näha, et klastritesse 1, 2, 5 ja 7 kuuluvad inimesed suitsetavad rohkem. Klastris 1 olevatest inimestest 48,6%, klastrisse 2 kuuluvatest inimestest 41,0% , klastrisse 5 kuuluvatest inimestest 35,7% ja klastrisse 7 kuuluvatst inimestest 28,0% on praeguseid suitsetajad. Kõige vähem on praeguseid suitsetajaid klastris 6 (16,2%). Klastrisse 6 kuuluvad inimesed olid ka toitumiselt kõige tervislikumad. Ülejäänud klastrites on praegusi suitsetajaid umbes viiendik. Endiste suitsetajate osakaal on kõigis klastrites peaaegu sama suur, jäädes 11-15% vahele.



Joonis 12: Suitsetajate osakaal klastrites

3.3.5 Klastrid ja haridus

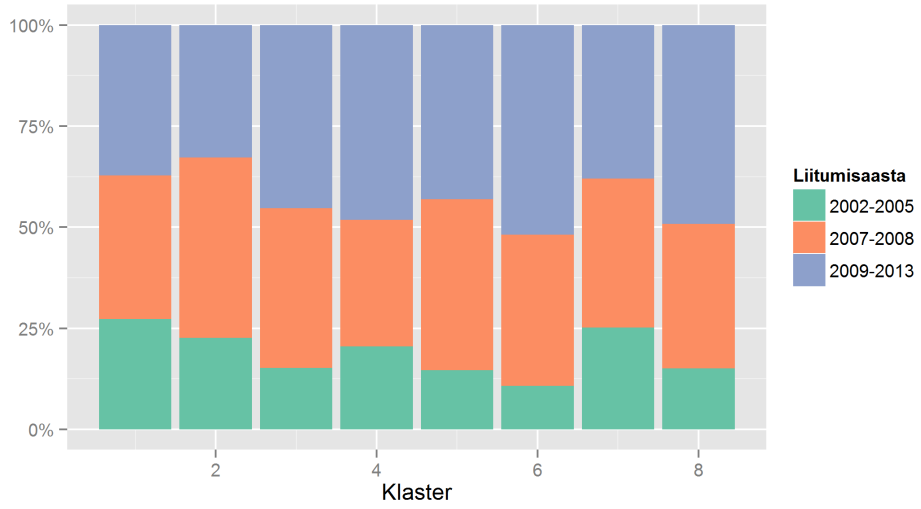


Joonis 13: Haridustasemed klastrite lõikes

Kõrgharidusega inimeste osakaal varieerub klastrite kaupa. Kõige vähem kõrgharidusega inimesi on klastris 1 (10,9%). Kõige suurem kõrgharidusega inimeste osakaal on klastris 6 (40,4%). Suur on kõrgharidusega inimeste osakaal veel ka klastris 3 (32,5% ja klastris 4 (29,8%). Eelneva põhjal võib öelda, et haritumad inimesed jälgivad rohkem, mida söövad, sest klaster 6 tarbis teistest enam tervislikke toiduaineid. Klastrisse 4 kuuluvad inimesed tarbivad üldiselt kõiki toiduaineid sagedamini kui teistesse klast-

ritesse kuuluvad inimesed ning klastrisse 3 kuuluvaid inimesi iseloomustab maiustuste sage tarbimine. Samas klastrisse 1, kuhu kuulus kõige vähem kõrgharidusega inimesi, kuuluvate inimeste toitumisharjumusi iseloomustab kohvi, saia, leiva, vorsti ja kartuli pidev tarbimine ning juur- ja puuviljade vähene tarbimine. Keskharidusega inimeste osakaal erinevates klastrites oli võrdlemisi ühtlane, jäädes 52-62% vahele. Klastrites, kus kõrgharidusega inimeste osakaal on suur, on alg - ja põhiharidusega inimeste osakaal väike.

3.3.6 Klastrid ja liitumisaasta



Joonis 14: Liitumisaasta klastrite lõikes

Ei saa öelda, et inimeste toitumismustrid oleksid kümne liitumisaasta jooksul palju muutunud. Klastrite võrdlemine liitumisaasta lõikes näitas vaid väikeseid erinevusi. Kõige rohkem aastatel 2002-2005 liitunutest kuulub klastrisse 1 ning kõige vähem klastrisse 6. Selle järgi saab öelda, et tervislikult toitujaid oli geenivaramuga liitumise algusaastatel veel vähe. Teisel perioodil liitunute osakaal on klastrites enamvähem võrdne. Viimasel perioodil liitunuid on kõige rohkem klastris 6 ehk inimesed on hakanud mõnevõrra tervislikumalt toituma.

3.3.7 Klastrid ja südame isheemiatõbi

Südame isheemiatõbi on eestlaste seas küllaltki levinud haigus. Südame isheemiatõve ennetamiseks soovitatakse ka jälgida toitumist ning mitte liialdada loomsete rasvade ja suhkruga. Vaatamaks, kas saadud klastrite ja südame isheemiatõve vahel on seost, teostati logistiline regresioonanalüüs. Huvi pakkus, et kas nendel isikutel, kellel liitumisel südame isheemiatõve ei olnud, mõjutab toitumise klaster tõenäosust, et isheemiatõbi tekib. Logistilise regresioonimudelisse lisati sugu, vanus, suitsetamisstaatus, kehamassiindeks ning faktorina klastrisse kuuluvus.

Tabel 4: Südame isheemiatõve esinemise šansside erinevus võrreldes klastriga 3

klastri nr	β	standardviga	p-väärtus
1	0,42	0,15	$5,3 * 10^{-3}$
2	0,67	0,15	$1,4 * 10^{-5}$
4	0,29	0,15	$0,05$
5	0,43	0,15	$3,9 * 10^{-3}$
6	0,23	0,16	$0,13$
7	0,42	0,15	$4,8 * 10^{-3}$
8	0,32	0,15	$3,5 * 10^{-2}$

Kõige väiksem šanss haigestuda südame isheemiatõppe on antud mudeli järgi klastrisse 3 kuuluvatel inimestel (Tabel 4). Südame isheemiatõppe mitte haigestumiseks soovitatakse muuhulgas tarbida ka vähe suhkrut, kuid antud mudeli järgi tuli kõige väiksem šanss haigestuda just neil, kes tarvitavad igapäevaselt maiustusi. Antud analüüsi põhjal ei saa öelda, et arstide soovitused on valed, vaid soovitatav on täpsemalt uurida, miks just sellesse klastrisse kuuluvatel inimestel šanss haigestuda on madal. Oluliselt ei erine haigestumise šanss klastrisse 4 ja 6 kuuluvatel inimestel klastrisse 3 kuuluvatest inimestest. Küll aga on märgatavalt suurem šanss haigestuda kõikidesse teistesse klastritesse kuuluvatel inimestel. Kõige suurem on antud näitaja klastri 2 korral, kuhu kuulusid inimesed, kes tarbisid igapäevaselt karastusjooke.

Kuna kõige suurem risk haigestuda on klastrisse 2 kuuluvatel inimestel ja kõige väiksem risk klastrisse 3 kuuluvatel inimestel, siis vaadatakse hetkel edasi ainult neid klastreid. Tabelis 5 on nende toiduainete keskmine tarbimine, mille korral erinevus oli üle 0,1. Võib tunduda, et kuna klastrisse 3 kuuluvad inimesed söövad rohkem maiustusi, tarbivad nad seetõttu ka rohkem suhkrut. Tegelikult sisaldavad paljud karastusjoogid väga suurel

määral suhkrut. Seega võib klastrisse 2 kuuluvate inimeste suhkrutarbimine olla palju suurem.

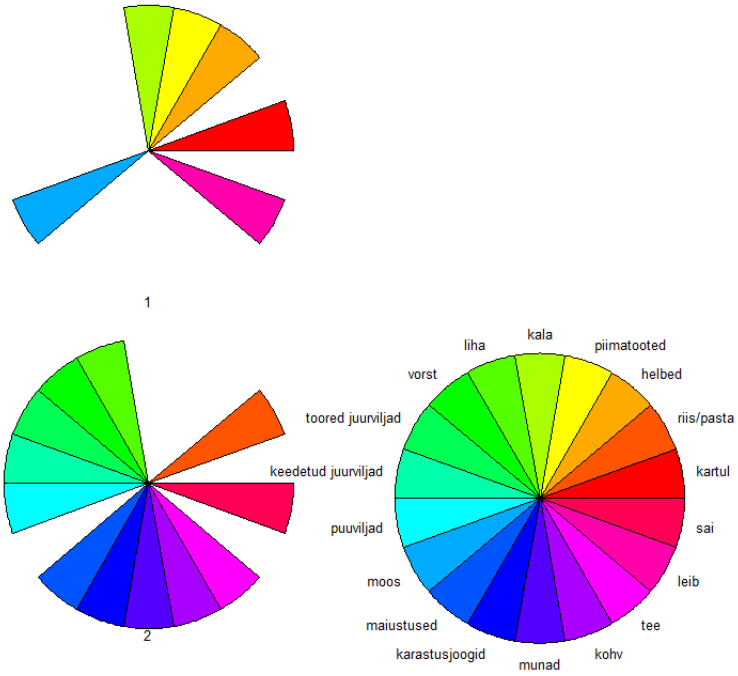
Tabel 5: Toiduainete keskmised tarbimised klastrite 2 ja 3 vahel, tabelis on ainult need toiduained, kus erinevus oli üle 0,1

	Sai	Leib	Liha	Lihaprodukt	v.puuvili	Maiustused	Karastusjoogid
klaster 2	2,4	2,5	2,9	3,2	3,1	2,9	3,5
klaster 3	2,0	2,3	2,6	2,8	3,4	3,5	1,5

Suurima haigestumise riskiga ehk klastrisse 2 kuuluvad inimesed tarbivad rohkem ka jahutooteid (sai, leib), liha ja lihatooteid ning vähem värskeid puuvilju. Kõige suurem erinevus on siiski karastusjookide tarbimises. Klastris 2 leidus ainult üks inimene, kes tarbis karastusjooke alla 3 korra nädalas. Samas tarbis 98% klastrisse 3 kuuluvatest inimestest karastusjooke alla 3 päeva nädalas. Klastrite 2 ja 3 erinevus jääb logistilises regressioonanalüüsis oluliseks ka siis, kui kõik seitse Tabelis 5 toodud toiduainet eraldi mudelisse panna. Seetõttu võib arvata, et klastrite mõju haigestumisele tuleneb pigem üldistest toiduharjumustest ja/või sellega seotud teguritest, mitte üksikute toiduainete tarbimisest või mittetarbimisest.

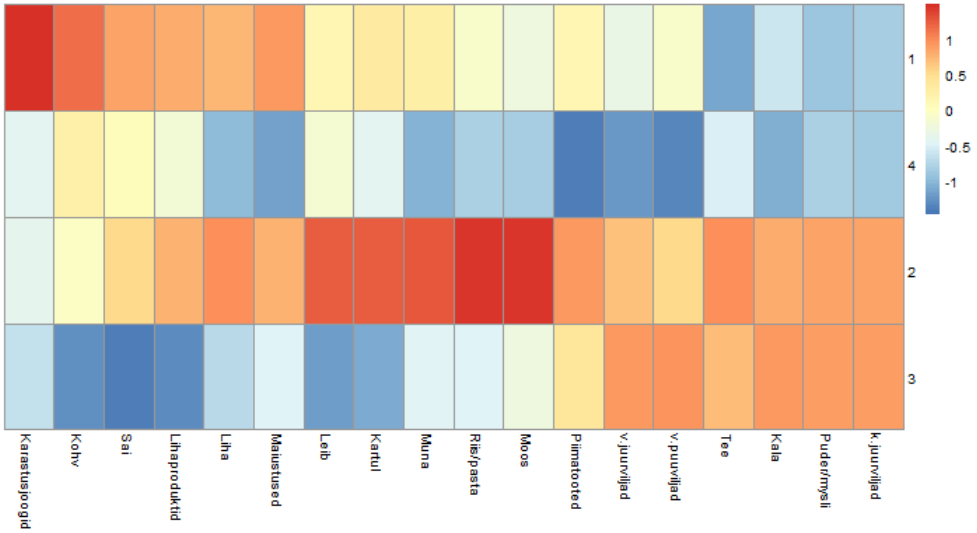
3.4 Klastrite iseloomustus klastrite arvu 2 korral

Toitumistunnuste põhjal inimeste kahte klastrisse jagamisel tulid klastrite erinevused toiduainete tarbimise sagedustes väga väikesed. Kõige suurem erinevus saadud klastrite korral on karastusjookide ja saia tarbimise sagedus. Kui klastrisse 1 kuuluvad inimesed tarbivad keskmiselt saia 3-5 viilu päevas ja karastusjooke 1-2 päeval nädalas, siis klastrisse 2 kuuluvad inimesed tarbivad saia kuni 2 viilu päevas ning karastusjooke pigem ei tarbita üldse. Esimesse klastrisse kuuluvad inimesed tarbivad rohkem ka leiba, kartulit, kohvi ja vorsti (joonis 15). Skaleerimata andmetele on analoogne joonis toodud Lisas 2.



Joonis 15: Toitumistunnuste jagunemine 2 klastri korral

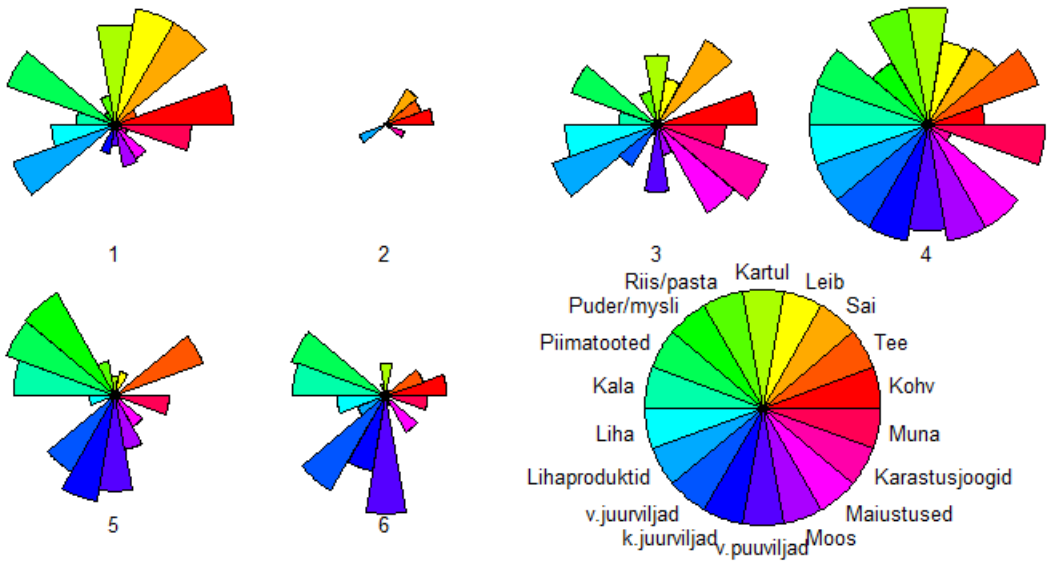
3.5 Klastrite iseloomustus klastrite arvu 4 korral



Joonis 16: Toitumistunnuste tarbimise sagedus võrreldes teiste klastritega, andmestik jagatud 4 klastriks

Jagades andmestiku nelja klastrisse, on klastrid väga eriilmelised (joonis 16). Skaleerimata andmetele on analoogne joonis toodud Lisas 2. Klastritesse 2 ja 3 kuuluvad inimesed tarbivad teistest klastritest sagedamini keedetud ja värskeid juurvilju, putru/müslit, kala, teed, värskeid puuvilju ning piimatooteid. Kui ülejäänuid toiduaineid tarbivad klastrisse 3 kuuluvad inimesed harvemini kui teistesse klastritesse kuuluvad inimesed, siis klastrisse 2 kuuluvad inimesed tarbivad neid toiduaineid ka keskmisest tunduvalt sagedamini. Klastrisse 4 kuuluvad inimesed tarbivad pea kõiki toiduaineid harvemini kui ülejäänud klastritesse kuuluvad inimesed. Antud klastris tarbitakse üle 2 päeva nädalas vaid kartulit, piimatooteid ja lihaprodukte ning üle kahe viilu leiba päevas. Moosi süüakse äärmiselt vähe. Klastrisse 1 kuuluvad inimesed tarbivad teiste klastritega võrreldes palju sagedamini karastusjooke (3-5 päeval nädalas), kohvi, saia, lihaprodukte, liha ja maiustusi ning harvemini teed.

3.6 Klastrite iseloomustus klastrite arvu 6 korral



Joonis 17: Toitumistunnuste tarbimise sagedus võrreldes teiste klastritega, andmestik jagatud 6 klastriks

Andmestiku jaotamisel 6 klastriks on taaskord näha üks klaster inimesi, kes sööb peaaegu kõiki toiduaineid sagedamini teiste klastritega võrreldes (klaster 4) ning üks klaster inimesi, kes sööb kõiki toiduaineid harvemini kui teistesse klastritesse kuuluvad inimesed (klaster 2). Klastritesse 5 ja 6 kuuluvate inimeste toitumisharjumused on küllaltki sarnased, teiste klastritega võrreldes sagedamini tarbitakse keedetud ja värskeid juurvilju, värskeid puuvilju, kala, ja piimatooteid. Klastrisse 5 kuuluvad inimesed tarbivad veel ka igapäevaselt putru või müslit ning tee joomine on ka sagedasem kui klastrisse 6 kuuluvatel inimestel. Klastreid 1 ja 3 iseloomustab pigem ebatervislik toitumine, kuna teistest sagedamini tarbivad nendesse klastritesse kuuluvad inimesed kohvi, leiba, saia ja lihaprodukte ning harvemini süüakse keedetud juurvilju ja kala (vt joonis 17 ning skaleerimata andmetele joonis Lisas 4) .

Kokkuvõte

Bakalaureusetöö eesmärgiks oli leida levinumaid mustreid inimeste toitumisharjumustes kasutades selleks klasteranalüüsi, täpsemalt k-keskmiste meetodit.

Kasutatud andmed pärinevad Tartu Ülikooli Eesti geenivaramust. Andmestikus oli üle 45 000 inimese andmed, kelle kohta oli teada 18 erineva toiduprodukti tarbimise sagedus. Taustatunnustena teati iga inimese kohta tema sugu, vanust, kehamassiindeksit, sünnimaakonda ja elukohamaakonda, haridustaset ja suitsetamisstaatust. Antud andmestikus olid ka tunnused südame isheemiatõppe haigesumise kohta.

Korrelatsioone uurides ei olnud toitumistunnused omavahel tugevalt seotud. Kuigi töös vaadatakse andmestikku erinevate arvu klastrite korral, keskendutakse klastrite arvule 8. Toitumismustrite järgi olid klastrid väga eriilmelised. Hästi eraldusid inimeste grupid, kes tarbivad vaadeldud toiduaineid sagedamini või harvemini, kes toituvad tervislikult või vähem tervislikult.

Ka erinevate taustatunnuste juures tulid klastrite erinevused välja. Näha oli sarnast toitumismustrit Lääne-Eestist pärit inimeste seas. Lisaks toitusid Tartu ja Tallinna elanikud sarnaselt. Kõige madalam kehamassiindeks oli inimestel, kelle klastrile oli iseloomulik igapäevane maiustute tarbimine, kuid samas tarbiti keskmisest rohkem ka värskeid puuvilju ja keskmisest vähem karastusjooke ja lihatooteid. Kõige tervislikumalt toitusid 45-64-aastased inimesed ja karastusjooke tarbisid kõige rohkem noored. Kõige väiksem oli suitsetajate osakaal klastris, kus tõituti ka kõige tervislikumalt. Samas klastris oli ka alg-ja põhiharidusega inimeste osakaal kõige väiksem. Liitumisaja järgi suuri erinevusi toitumises ei tulnud. Ainsa haigustunnusena vaadati antud töös südame isheemiatõppe haigestumise riski inimestel, kellel ei olnud haigust geenivaramuga liitumisel. Saadi, et kõige suurem risk haigestuda on neil, kes tarbivad igapäevaselt karastusjooke.

Tulevikus on võimalik tööd edasi arendada, vaadates eelkõige erinevate toitumisharjumuste ja haiguste vahelisi seoseid.

Viited

[1] Põhikiri. Tartu Ülikooli Eesti geenivaramu, URL (vaadatud: 27.04.2015) http://www.geenivaramu.ee/et/pohikiri

[2] Gareth, J.; Witten, D.; Hastie, T.; Tibshirani, R. (2013) *An Introduction to Statistical Learning with Applications in R*, Springer

[3] Tan, P.-N.; Steinbach, M.; Kumar, V. (2005) *Introduction to Data Mining*, Pearson

[4] k-keskmiste meetod.Wikipedia, URL (vaadatud: 14.04.2015) http://en.wikipedia.org/wiki/K-means_clustering

[5] K-Means Clustering. R Documentation, URL (vaadatud: 17.04.2015) https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html

[6] Kodinariya, T.; Dr. Makawana, R. (2013). Review on determing number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*
Saadaval: http://www.ijarcsms.com/docs/paper/volume1/issue6/V1I6-0015.pdf

[7] Käärik, E. (2014) Andmeanalüüs II. Loengukonspekt. Tartu: Tartu Ülikool, matemaatilise statistika instituut.

[8] Greenacre, M.; Primicerio, R. (2013) *Multivate Analysis of Ecological Data*
Saadaval: http://www.multivariatestatistics.org/

[9] Aru, G (2013) *Korrespondentsanalüüs ja andmete dubleerimine*, Bakalaureusetöö, juhendaja K.Pärna, Tartu

[10] Kehamassiindeks. Wikipedia, URL (vaadatud: 12.04.2015) http://et.wikipedia.org/wiki/Kehamassiindeks

[11] Südame isheemiatõbi. kliinik.ee, URL (vaadatud: 27.04.2015) https://www.kliinik.ee/haiguste_abc/sudame-isheemiatobi/id-1749

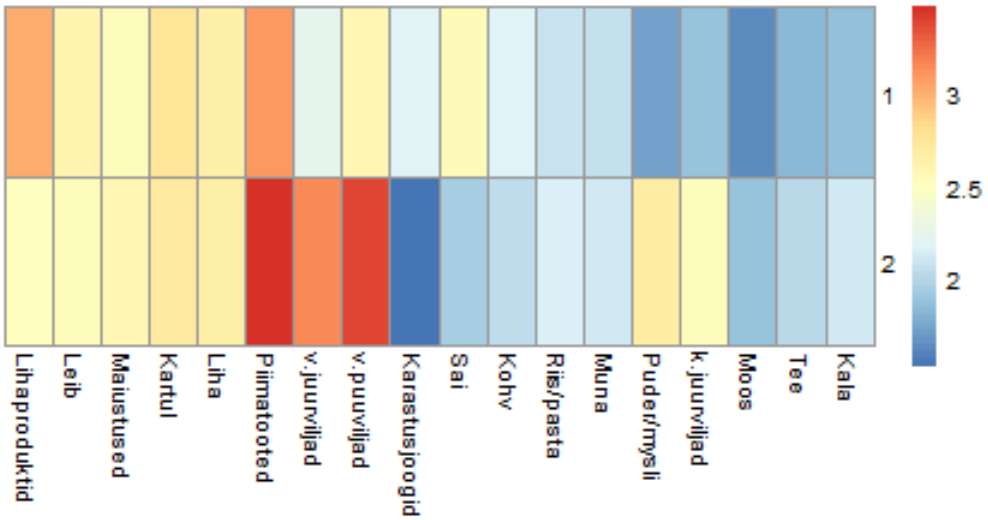
Lisad

Lisa 1 - Korrelatsioonimaatriks toitumistunnuste vahel

Tabel 6: Korrelatsioonimaatriks toitumistunnuste vahel

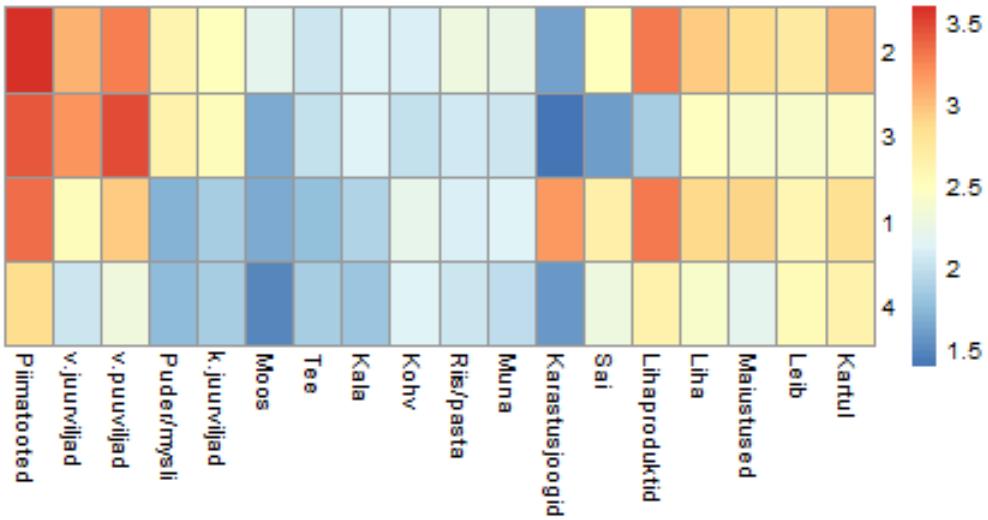
	Kohv	Tee	Sai	Leib	Kartul	Riis/ pasta	Puder/ mysli	Piima- tooted	Kala	Liha	Liha- produktid	v.juur- viljad	k.juur- viljad	v.puu- viljad	Moos	Maius- tused	Karastus- joogid	Muna
Kohv	1.00	-0.19	0.07	0.08	0.04	-0.02	-0.08	0.00	0.02	0.02	0.08	-0.01	0.00	-0.01	-0.04	-0.02	0.04	0.04
Tee	-0.19	1.00	0.03	0.04	-0.05	0.08	0.13	-0.07	0.07	-0.03	-0.05	0.10	0.11	0.03	0.09	0.02	-0.07	0.03
Sai	0.07	0.03	1.00	0.14	0.18	0.07	-0.11	-0.01	-0.06	0.08	0.30	-0.14	-0.10	-0.15	0.12	0.07	0.20	0.08
Leib	0.08	0.04	0.14	1.00	0.15	-0.01	0.02	0.06	0.06	0.05	0.13	-0.03	0.04	-0.04	0.09	-0.05	0.01	0.08
Kartul	0.04	-0.05	0.18	0.15	1.00	-0.01	0.00	0.13	0.06	**0.25**	**0.22**	0.01	0.10	0.01	0.13	0.03	0.04	0,11
Riis/pasta	-0.02	0.08	0.07	-0.01	-0.01	1.00	0.12	0.01	0.03	0.07	0.09	0.06	0.05	0.01	0.10	0.09	0.03	0,10
Puder/mysli	-0.08	0.13	-0.11	0.02	0.00	0.12	1.00	0.17	0.11	-0.02	-0.12	0.18	**0.23**	0.15	**0.22**	0.05	-0.16	0,06
Piimatooted	0.00	-0.07	-0.01	0.06	0.13	0.01	0.17	1.00	0.05	0.10	0.08	0.12	0.10	0.18	0.10	0.11	-0.02	0,09
Kala	0.02	0.07	-0.06	0.06	0.06	0.03	0.11	0.05	1.00	0.07	-0.06	0.16	**0.22**	0.13	0.12	-0.03	-0.05	0,13
Liha	0.02	-0.03	0.08	0.05	**0.25**	0.07	-0.02	0.10	0.07	1.00	0.19	0.11	0.08	0.05	0.11	0.11	0.12	0,12
Lihaproduktid	0.08	-0.05	**0.30**	0.13	**0.22**	0.09	-0.12	0.08	-0.06	0.19	1.00	-0.04	-0.08	-0.02	0.07	0.13	**0.23**	0,08
v.juurviljad	-0.01	0.10	-0.14	-0.03	0.01	0.06	0.18	0.12	0.16	0.11	-0.04	1.00	**0.25**	**0.37**	0.08	0.04	-0.07	0,06
k.juurviljad	0.00	0.11	-0.10	0.04	0.10	0.05	**0.23**	0.10	**0.22**	0.08	-0.08	**0.25**	1.00	0.17	0.16	-0.03	-0.15	0,11
v.puuviljad	-0.01	0.03	-0.15	-0.04	0.01	0.01	0.15	0.18	0.13	0.05	-0.02	**0.37**	0.17	1.00	0.06	0.12	-0.07	0,06
Moos	-0.04	0.09	0.12	0.09	0.13	0.10	**0.22**	0.10	0.12	0.11	0.07	0.08	0.16	0.06	1.00	0.17	0.01	0,15
Maiustused	-0.02	0.02	0.07	-0.05	0.03	0.09	0.05	0.11	-0.03	0.11	0.13	0.04	-0.03	0.12	0.17	1.00	0.14	0,03
Karastusjoogid	0.04	-0.07	0.20	0.01	0.04	0.03	-0.16	-0.02	-0.05	0.12	**0.23**	-0.07	-0.15	-0.07	0.01	0.14	1.00	0,04
Muna	0.04	0.03	0.08	0.08	0.11	0.10	0.06	0.09	0.13	0.12	0.08	0.06	0.11	0.03	0.15	0.03	0.04	1,00

Lisa 2 - Toiduainete keskmine tarbimine, andmestik jagatud 2ks klastriks



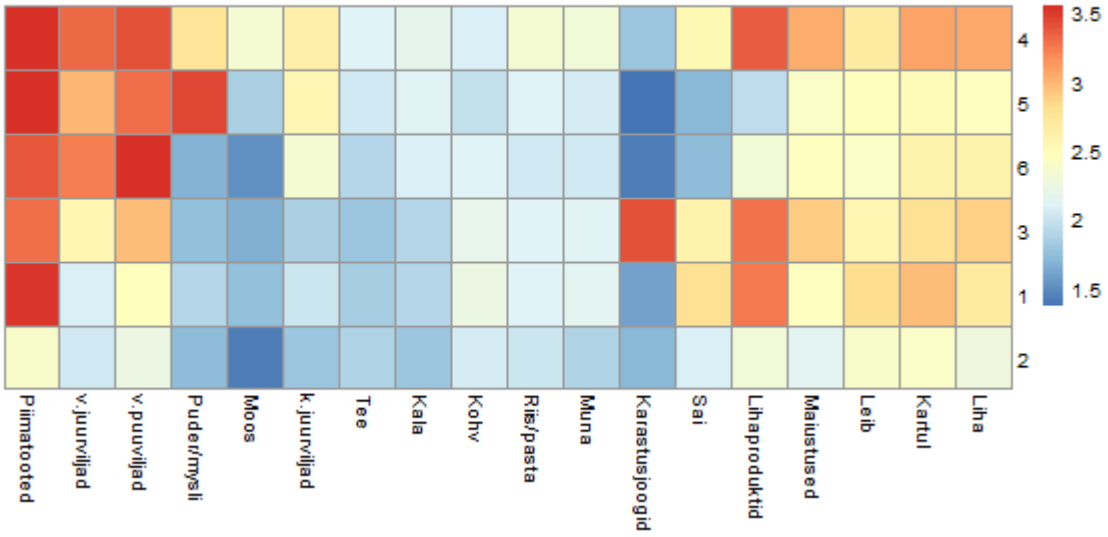
Joonis 18: Toiduainete keskmine tarbimine klastrite lõikes, andmestik jagatud 2ks klastriks

Lisa 3 - Toiduainete keskmine tarbimine, andmestik jagatud 4ks klastriks



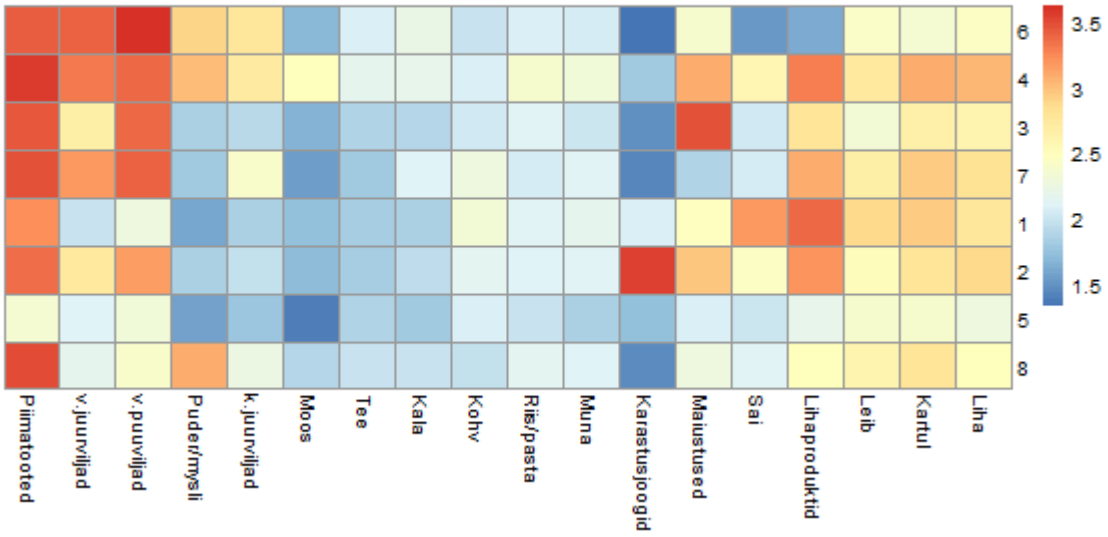
Joonis 19: Toiduainete keskmine tarbimine klastrite lõikes, andmestik jagatud 4ks klastriks

Lisa 4 - Toiduainete keskmine tarbimine, andmestik jagatud 6ks klastriks



Joonis 20: Toiduainete keskmine tarbimine klastrite lõikes, andmestik jagatud 6ks klastriks

Lisa 5 - Toiduainete keskmine tarbimine, andmestik jagatud 8ks klastriks



Joonis 21: Toiduainete keskmine tarbimine klastrite lõikes, andmestik jagatud 8ks klastriks

Lihtlintsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Marili Zimmermann** (sünnikuupäev: 08.10.1993)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Toitumismustrite analüüs Tartu Ülikooli Eesti geenivaramu andmebaasis k-keskmiste meetodi abil", mille juhendaja on Krista Fischer,

   (a) reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

   (b) üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguste 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, 29.04.2015