Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Kaisa-Siret Hint

Seosed eksposoomi ja kõrgvererõhktõve avaldumise vahel

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendajad: PhD Jaanika Kronberg, prof. Krista Fischer

Tartu 2025

SEOSED EKSPOSOOMI JA KÕRGVERERÕHKTÕVE AVALDUMISE VAHEL

Bakalaureusetöö

Kaisa-Siret Hint

Lühikokkuvõte

Eksposoomiks nimetatakse kõiki mittegeneetilisi tegureid, mis inimese tervist mõjutavad. Käesoleva bakalaureusetöö eesmärk oli Tartu Ülikooli Eesti Geenivaramu andmete põhjal leida seoseid eksposoomi ning kõrgvererõhktõve esinemise vahel. Seoste leidmiseks kasutati Coxi võrdeliste riskide mudelit, kuhu kaasati lisaks keskkonnatunnustele ka sugu, vanus, kehamassiindeks, haridustase ja suitsetamine. Kuna keskkonnategurid on korreleeritud ja neid eraldi analüüsides on tulemuste tõlgendamine keeruline, rakendati konsensusklasterdamist, et grupeerida inimesed nende elukoha keskkonnategurite põhjal. Lähemalt uuriti kahte konsensusklasterdamise tulemusel moodustunud klastrite jaotust.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: eksposoom, elulemusanalüüs, klasteranalüüs, Coxi mudel, kõrgvererõhktõbi

CONNECTIONS BETWEEN THE EXPOSOME AND RISK OF HYPERTENSION

Bachelor thesis

Kaisa-Siret Hint

Abstract

Exposome is a term used for describing all non-genetic factors that affect human health. The objective of this bachelor's thesis was to study the associations between the external exposome and hypertension using data from the Estonian Biobank. To study the associations Cox proportional hazards model was implemented, including environmental factors alongside sex, age, body mass index, education and smoking. Since environmental factors are correlated and difficult to interpret on their own, the study used consensus clustering to group the subjects based on their exposome. Two sets of clusters were explored further.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: exposome, survival analysis, cluster analysis, Cox model, hypertension

Sisukord

Sissejuhatus

Kõrgvererõhktõbi on üks peamisi südame-veresoonkonna haiguste ja nendega seotud surmade põhjustajaid maailmas. Haiguse kõrge levimuse tagamaadeks arvatakse olevat vananev rahvastik ja ebatervislik elustiil. [1] Vererõhu seost keskkonnateguritega uurides on leitud, et pikaajaline kokkupuude madalate temperatuuride, müra ja õhureostusega soodustab hüpertensiooni teket [2].

Käesoleva bakalaureusetöö eesmärk oli kirjeldada, kuidas sõltub kõrgvererõhktõve avaldumise risk inimese eksposoomist. Selleks kasutati Tartu Ülikooli Eesti Geenivaramu andmeid, mille hulgas olid isikut, haigestumist ja eksposoomi kirjeldavad tunnused. Seoste kirjeldamiseks kasutati Cox võrdeliste riskide mudelit ja klasteranalüüsi.

Töö teoreetiline osa jaguneb kolmeks. Esmalt tutvustatakse lähemalt uuritavat haigust ja eksposoomi. Teine osa kirjeldab elulemusanalüüsi metoodikat, keskendudes rohkem Coxi mudelile. Peatüki viimases osas antakse ülevaade klasteranalüüsist ja kirjeldatakse täpsemalt klasterdamiseks kasutatavat meetodit.

Analüüsi peatükk jaotub kolmeks suuremaks osaks. Esimeses osas tutvustatakse lähemalt andmeid, valitakse eksposoomitunnused ja moodustatakse valim. Teises osas valmib esmalt üldtunnustega Coxi mudel ning seejärel proovitakse sinna lisada eksposoomitunnuseid. Peatüki lõpetab klasteranalüüs, mille tulemusel valmivad mudelid, mis käsitlevad keskkonnaandmete eraldi vaatamise asemel neid kui ühte tunnuste komplekti ehk ekspotüüpi.

Töö kirjutamiseks kasutati tekstitöötlusprogrammi LaTeX. Andmete analüüs ja joonised on tehtud rakendustarkvaraga R.

1 Taust

Siin peatükis kirjeldatakse täpsemalt, mis on eksposoom ning tutvustatakse kõrgvererõhktõve riskifaktoreid ja tagajärgi.

1.1 Eksposoom

Kõiki mittegeneetilisi tegureid, mis inimese tervist mõjutavad, nimetatakse eksposoomiks [3]. Eksposoomi moodustavateks komponentideks loetakse näiteks inimese elustiil ning teda ümbritsev väline, füüsikalis-keemiline ja sotsiaalmajanduslik keskkond [4].

Eksposoomiuuringud erinevad varasematest keskkonnatervise alal tehtud teadustöödest selle poolest, et rõhk on üksikute keskkonnatunnuste mõjude analüüsimiselt liikunud pigem tunnuste koosmõju ning interaktsioonide tuvastamisele. Uus suund on võetud ajendatult geeniuuringutest, kus on leitud, et haiguste riski on tulemuslikum prognoosida geenide koosmõjude põhjal. [3]

Selles töös analüüsiti EXPANSE projekti raames mudeldatud väliskeskkonna tegureid kui ühte komponenti eksposoomist. Kuna keskkonnategurid on omavahel võrdlemisi tugevalt seotud ning ei esine inimese elukeskkonnas iseseisvalt, siis soovitatakse eksposoomiuuringute käigus kõiki olemasolevad tunnuseid analüüsida koos [3]. Interaktsioonide analüüsi kaasamiseks on varasemates teadustöödes rakendatud erinevaid lähenemisi, näiteks peakomponentanalüüsi [5, 6] ja klasterdamist [7]. Eksposoomi arusaadavamaks kirjeldamiseks oleks otstarbekas andmeid klasterdada. Klasterdusest välja joonistuvad ekspotüübid muudaksid keerulised keskkonnaandmed hoomatavamaks ja aitaksid seeläbi tuvastada probleemseid piirkondi.

1.2 Kõrgvererõhktõbi

Kõrgvererõhktõbi ehk hüpertoonia on maailmas väga levinud haigus, mõjutades umbes 1,28 miljardit inimest. Eestis oli 2019. aastal 15-aastaste ja vanemate seas kõrge vererõhk või hüpertoonia diagnoos 25 protsendil elanikest [8]. Eelnimetatud haigus diagnoositakse inimesel, kelle süstoolne vererõhk on pidevalt üle 140 mmHg või diastoolne pidevalt üle 90 mmHg. Hüpertooniaga inimene võib kogeda tihti peapööritust, minestamistunnet, segadust, probleeme hingamise ja nägemisega, ärevust ning valu rinnus ja peas. Kõrge vererõhk kahjustab eelkõige südant ja artereid, mistõttu võib haiguse ravimata jätmise tagajärjeks olla halvimal juhul südameatakk, südamepuudulikkus või surm. Peale selle on hüpertoonia riskiteguriks ka paljudele teistele haigustele, näiteks neerupuudulikkusele. [9]

Enim uuritud hüpertensiooni riskifaktoriteks on vanus, kõrge kehakaal, istuv eluviis ja ebatervislik toitumine, eriti suures koguses soola ja alkoholi tarbimine. Lisaks nendele võib tõve esinemist soodustada ka geneetiline risk. [9] Näiteks Keaton, Kamali ja Xie leidsid oma ülegenoomse seoseuuringuga eurooplaste seas mitu geenivarianti, mis kõrge vererõhu esinemist soodustasid [10]. Viimasel ajal on hakatud rohkem uurima ka keskkonna mõju kõrgvererõhktõve esinemisele. Näiteks leidis Brook oma 2017. aasta artiklis, et pikaajaline kokkupuude madalate temperatuuride, müra ja õhureostusega võib vähendada keha võimet vererõhku kontrolli all hoida, mis omakorda võib tekitada hüpertooniat.

2 Metoodika

Siin peatükis tutvustatakse elulemus- ja klasteranalüüsi. Metoodika esimeses pooles räägitakse elukestvusanalüüsist üldisemalt ning peale seda kirjeldatakse täpsemalt Coxi võrdeliste riskide mudelit. Metoodika teises pooles tutvustatakse klasteranalüüsi, täpsemalt kauguspõhist konsensusklasterdamist ja hierarhilist klasterdamist. Peatükid 2.1 ja 2.2 põhinevad E. T. Lee ja J.W. Wangi poolt kirjutatud raamatu "Statistical Methods for Survival Data Analysis" teisel ja kolmandal verioonil ning D. Colletti raamatu "Modelling Survival Data in Medical Research" neljandal versioonil , kui ei ole märgitud teisiti. Täpsemad leheküljed on toodud kasutatud kirjanduse loetelus.

2.1 Elulemusanalüüs

Sõna "elulemus" all mõeldakse aega, mis kulub kindla sündmuse toimumiseni. Kuigi sõna ise vihjab ellujäämisele, ei pea vaatlusalune sündmus olema tingimata surm. Elulemusanalüüsi abil võib uurida ka näiteks aega haigestumiseni, haigushoo kordumiseni või hoopis tervenemiseni. Elulemusadmete hulka võivad kuuluda aeg, ravi mõju ja patsiendi näitajad, mille abil saab ennustada näiteks ravivastuse tõenäosusi ja keskmist elatud aastate arvu või võrrelda elulemusfunktsioone. Eriti oluliseks on osutunud võimalus tuvastada haiguste riskifaktoreid.

Tähistame meid huvitava sündmuse tähega A ning fikseerime uuringu algusaja t_0 ja lõppaja t_*. Olgu meil n patsienti, kes uuringus osalesid. Tähistame sündmuse A toimumise aja patsiendi i puhul kui t_i. Realiseeruda saab üks järgnevatest stsenaariumitest. Kui patsiendiga toimus sündmus A uuringu ajal, siis teame, et $t_i < t_*$. Kui patsient kadus uuringust, siis teame ainult seda, et enne kadunuks jäämist temaga sündmust A toimunud polnud ehk $t_i > t_*$. Kui uuringu lõpuks patsiendiga sündmust A toimunud polnud, teame samuti, et $t_i > t_*$. Kuna me ei oska arvata, kui palju peale vaatluse all olemist, või kas üldse viimase kahe stsenaariumi korral

patsientidega sündmus A toimus, siis kutsutakse neid vaatlusi tsenseerituteks. Sellise uuringu puhul on tegemist täpsemalt paremalt tsenseerimisega, kuna on teada uuringueelne olukord, aga on keeruline ennustada, mis juhtub pärast.

Kuna sündmuse A toimumise aeg on juhuslik, siis moodustavad ajad t_i jaotuse. Tähistame aja sündmuseni A kui juhusliku suuruse tähega T ja fikseerime aja t. Seda jaotust saab kirjeldada kolme funktsiooniga: elulemusfunktsioon $S(t)$, tihedusfunktsioon $f(t)$ ja riskifunktsioon $h(t)$. Elulemusfunktsioon kirjeldab tõenäoust, et patsiendiga juhtub sündmus A hiljem kui aeg t ehk

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t).$$

Tihedusfunktsioon $f(t)$, sarnaselt muudele pidevatele juhuslikele suurustele, on defineeritud kui

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t)}{\Delta t}.$$

Selle funktsiooni abil saame hästi arvutada tõenäosust, et patsiendiga juhtub sündmus A mingis ajavahemikus.

Riskifunktsioon $h(t)$ iseloomustab tõenäosust, et inimesega juhtub sündmus A väga väikses ajavahemikus $(t, t + \Delta t)$, kui temaga ei juhtunud sündmust kuni ajani t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t}.$$

Seega annab riskifunktsioon meile sündmuse toimumise riski ühe ajaühiku kohta. Riskifunktsiooni võib defineerida ka läbi eelmainitud tihedus- ja elulemusfunktsiooni.

$$h(t) = \frac{f(t)}{S(t)}.$$

Järelikult, kui me teame huvipakkuvas populatsioonis kehtivat T jaotust, siis saame selle alusel riskile mudeli sobitada. Näiteks, kui $T \sim Exp(\lambda)$, siis $f(t) = \lambda e^{-\lambda t}$ ning

$S(t) = 1 - F(t) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}$, mispärast

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda.$$

[14]

Praktikas ei pruugi me aga T jaotust teada, mistõttu ei saa probleemile parameetriliste meetoditega läheneda. Sellises olukorras on üks võimalik lahendus sobitada andmetele Coxi võrdeliste riskide mudel, mis töötab mitmese regressiooni põhimõttel.

2.1.1 Coxi võrdeliste riskide mudel

Olgu meil p argumenttunnust $X_1, X_2, \ldots, X_p$, siis i.-nda patsiendi jaoks on nende väärtused $X_{1i}, X_{2i}, \ldots, X_{pi}$. Võrdeliste riskide mudeli puhul peavad erinevate indiviidide kohta kehtivad riskifunktsioonid olema üksteisega võrdelised ning nende suhe ei tohi sõltuda ajast t ehk peab kehtima

$$\frac{h\left(t \mid X_{1i}, X_{2i}, \ldots, X_{pi}\right)}{h\left(t \mid X_{1j}, X_{2j}, \ldots, X_{pj}\right)} = const, \forall t \in [t_0, t_*], \forall i, j \in n.$$

Nende eelduste abil saame kirjutada riskifunktsiooni kui

$$h\left(t \mid X_1, X_2, \ldots, X_p\right) = h_0\left(t\right) g\left(X_1, X_2, \ldots, X_p\right), \tag{1}$$

kus $g\left(X_1, X_2, \ldots, X_p\right)$ on funktsioon argumenttunnustest. Funktsioon $h_0\left(t\right)$ kehtib inimese i jaoks, kelle puhul $g\left(X_{1i}, X_{2i}, \ldots, X_{pi}\right) = 1$ ning seda nimetatakse baasriskifunktsiooniks. Funktsioon $h_0\left(t\right)$ ilmestab riski muutumist ajas.

Olgu meil vaatlusalused i ja j, kelle tunnuste komplektid on $\boldsymbol{X_i} = (X_{1i}, X_{2i}, \ldots, X_{pi})$ ja $\boldsymbol{X_j} = (X_{1j}, X_{2j}, \ldots, X_{pj})$. Nende riskide suhet saame võrduse (1) abil kirjeldada kui

$$\frac{h\left(t \mid \boldsymbol{X_i}\right)}{h\left(t \mid \boldsymbol{X_j}\right)} = \frac{h_0\left(t\right) g\left(\boldsymbol{X_i}\right)}{h_0\left(t\right) g\left(\boldsymbol{X_j}\right)} = \frac{g\left(\boldsymbol{X_i}\right)}{g\left(\boldsymbol{X_j}\right)}.$$

Tähistame $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$. Enamasti kasutatakse võrdeliste riskide mudelit kujul

$$h\left(t \mid \boldsymbol{X}\right) = h_0\left(t\right) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right) = h_0\left(t\right) \exp\left(\sum_{j=1}^{p} \beta_j X_j\right), \quad (2)$$

kus β_j, $j = 1, \ldots, p$ tähistavad tunnustele $X_1, \ldots, X_p$ vastavaid koefitsiente. Nüüd saame kahe indiviidi i ja j riskide suhet kirjeldada kui

$$\frac{h\left(t \mid \boldsymbol{X_i}\right)}{h\left(t \mid \boldsymbol{X_j}\right)} = \frac{g\left(\boldsymbol{X_i}\right)}{g\left(\boldsymbol{X_j}\right)} = \frac{\exp\left(\sum_{k=1}^{p} \beta_k X_{ki}\right)}{\exp\left(\sum_{k=1}^{p} \beta_k X_{kj}\right)} = \exp\left(\sum_{k=1}^{p} \beta_k \left(X_{ki} - X_{kj}\right)\right)$$

Siit edasi tähistame $h_i\left(t\right) = h\left(t \mid X_1, X_2, \ldots, X_p\right)$.

Võrduse (2) saab teisendada kujule

$$\ln\left(\frac{h_i\left(t\right)}{h_0\left(t\right)}\right) = \sum_{j=1}^{p} \beta_j X_j = \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p$$

ning kui tähistada $y_i = \ln\left(\frac{h_i(t)}{h_0(t)}\right)$, siis saame mitmese regressiooni mudeli, kus sõltuvaks tunnuseks on logaritm riskifunktsioonide suhtest ja argumentideks võimalikud riski mõjutavad tunnused. Kuna mitmese regressiooni puhul saab olulisi tunnuseid kindlaks teha sammregressiooni meetodiga, siis saame seda kasutada ka siin. Lisaks olulisusele saab selle abil leida ka riskiindeksi, mille abil on võimalik kahe erinevate argumenttunnuste väärtustega grupi elulemust võrrelda. Selleks tuleb argumenttunnused standardiseerida ja hinnata mudel

$$\ln\left(\frac{h_i\left(t\right)}{h_0\left(t\right)}\right) = \beta_1\left(X_1 - \bar{X_1}\right) - +\beta_2\left(X_2 - \bar{X_2}\right) + \ldots \beta_p\left(X_p - \bar{X_p}\right),$$

kus $\bar{X_j}$ tähistab tunnuse X_j keskmist üle kõigi patsientide. Eelkirjeldatud võrduse vasakut poolt nimetatakse riskiindeksiks, mis kirjeldab saadud patsiendi, kellel on mingi kindel komplekt argumenttunnuste väärtuseid ja patsiendi, kelle kõik argumenttunnused on keskmise väärtusega, sündmuse A toimumise riskide suhet.

Parameetrid $\beta_j, j = 1, \ldots, p$ hinnatakse osalise tõepära meetodiga. Olgu meil r sündmuse A toimumise aega ehk t_i-d kindlalt teada. Kuna meil on n vaatlusalust, siis $n - r$ t_i-d on sellisel juhul paremalt tsenseeritud.

Järjestame sündmuse A toimumise ajad kui $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. See eeldab, et $t_i \neq t_j$, $\forall i, j$, $i \neq j$ korral. Tähistame patsientide hulga, kes on tsenseerimata ja kellega pole hetkeks $t_{(j)}$ juhtunud sündmust A kui $R\left(t_{(j)}\right)$ ja koefitsentide vektori $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$. Saame tõepärafunktsiooni

$$L\left(\boldsymbol{\beta}\right) = \prod_{j=1}^{r} \frac{\exp\left(\boldsymbol{\beta}' \boldsymbol{X}_{(j)}\right)}{\sum_{l \in R\left(t_{(j)}\right)} \exp\left(\boldsymbol{\beta}' \boldsymbol{X}_l\right)}, \tag{3}$$

kus $\boldsymbol{X}_{(j)}$ tähistab patsiendi argumentide vektorit, kellega juhtus sündmus A ajal $t_{(j)}$. Märkame, et tsenseeritud patsiente ei arvestata lugejas, aga arvestatakse nimetajas olevas summas. Kordajate vektor $\boldsymbol{\beta}$ hinnatakse võrduse (3) logaritmi maksimiseerimisel.

2.2 Klasterdamine

L. Kaufman ja P.J. Rousseeuw on öelnud: "Klasteranalüüs on andmetes gruppide leidmise kunst" [15]. Kuna sarnaselt võib käsitleda ka klassifitseerimist, siis aetakse neid tihtipeale omavahel segamini. Nende peamine erisus seisneb selles, et klassifitseerimisel jagatakse objektid juba varem defineeritud klassidesse. Klasterdamisel aga moodustuvad andmetes algoritmi kasutamise tulemusena kogumid, kus ühte kogumisse kuuluvad objektid on võimalikult sarnased ja erinevatesse kogumitesse kuuluvad võimalikult erinevad. Niiviisi saadud gruppe nimetatakse klastriteks. [16]

Peatükid 2.2.2 kuni 2.2.4 põhinevad B. Bodinieri poolt 2023. aastal kirjutatud teadusartiklil "Automated calibration of consensus weighted distance-based clustering approaches using sharp" [17], kui ei ole märgitud teisiti.

2.2.1 Klasterdamisalgoritmi valik

Kuna klasterdamisel on põhiline, et ühte klastrisse kuuluvad vaatlused oleksid võimalikult sarnased, ning, et klastrid kui tervikud oleksid üksteisest võimalikult erinevad [16], tuleb selle saavutamiseks valida sobiv algoritm. Bakalaureusetöö eesmärgiks oli klasterdada eksposoomiandmeid, seega tugineti EXPANSE projekti veebilehele, kus soovitati eksposoomiandmete jaoks kasutada R analüüsiprogrammi paketis sharp ("Stability-enHanced Approaches using Resampling Procedures") implementeeritud kauguspõhist konsensusklasterdamist [18, 19]. Sellel algoritmil on mitmeid häid omadusi, näiteks on see stabiilsem kui paljud teised klasterdamisalgoritmid. Lisaks stabiilsusele on algoritm suuresti automatiseeritud, skaleerib ise andmed ning leiab parimad parameetrid, sealhulgas klastrite arvu. [17]

2.2.2 Kauguspõhine konsensusklasterdamine

Valitud algoritmi tööks läheb vaja kahte hüperparameetrit λ ja G. Esimest neist kasutatakse kaalumiseks ja teine määrab klastrite arvu. Kuna siin analüüsis λ-t ei

kasutatud, siis kirjeldame algoritmi lihtsustatud versiooni ehk kaalumata kaugus-põhist konsensusklasterdamist.

Klasterdamisalgoritm koosneb viiest osast:

1. andmestikust K valimi võtmine;

   Olgu meil andmestik X, milles on n vaatlust. Andmestikust võetakse tagasipanekuga K valimit, millest igaüks sisaldab endas $\tau \cdot n$, $\tau \in [0,1]$ vaatlust. Parameeter τ tähistab osakaalu, täpsemalt, kui suur osa andmetest igasse valimisse võetakse. Vaikimisi kehtib $\tau = 0{,}5$. Iga vaatluste paari puhul loetakse kokku, mitmes valimis see esineb ning summadest koostatakse maatriks H. Matemaatiliselt tähendab see, et iga maatriksi H element H_{ij}, $i, j \in \{1, \ldots, n\}$ tähistab selliste valimite arvu, kus esineb nii vaatlus i kui ka vaatlus j.

2. valimite klasterdamine;

   Kõikidele valimitele rakendatakse etteantud klasterdamisalgoritmi. Käesoleva töö kontekstis on selleks hierarhiline, täpsemalt täieliku seose meetod. Selle tulemusel moodustub igas valimis G klastrit. Seejärel leitakse iga valimi $k = 1, \ldots, K$ jaoks maatriks $C^k(G)$, mis kajastab, milliste vaatluste paaride puhul kuuluvad mõlemad vaatlused samasse klastrisse. Olgu i ja j andmestikku X kuuluvad vaatlused, siis maatriksi $C^k(G)$ elemendid $C^k_{ij}(G)$ avalduvad kui

$$C^k_{ij}(G) = \begin{cases} 1, & \text{kui } i \neq j \text{ on mõlemad valimis } k \text{ ja mõlemad samas} \\ & G \text{ abil saadud klastris,} \\ 1, & \text{kui } i = j, \\ 0, & \text{mujal.} \end{cases}$$

3. maatriksi $C(G)$ moodustamine;

Summeerides üle eelmises osas leitud maatriksite $C^k(G)$, $k = 1, \ldots, K$, saab nüüd leida maatriksi $C(G)$, mis loeb kokku, mitmes valimis vaatlused i ja j ühte klastrisse sattusid. Iga $i = 1, \ldots, n$, $j = 1, \ldots, n$ jaoks avalduvad maatriksi $C(G)$ elemendid $C_{ij}(G)$ kui

$$C_{ij}(G) = \sum_{k=1}^{K} C_{ij}^k(G).$$

4. konsensusmaatriksi $\Gamma(G)$ moodustamine;

   Kuna vaatlused i ja j ei pruugi olla kõigis valimites esindatud, siis leitakse maatriksite H ja $C(G)$ abil konsensusmaatriks $\Gamma(G)$. See sisaldab endas osakaale, mis on leitud jagades iga vaatluste paari mõlemate vaatluste samas klastris esinemised samas valimis esinemistega. Ehk

   $$\forall i, j \in \{1, \ldots, n\}, \quad \Gamma_{ij}(G) = \frac{C_{ij}(G)}{H_{ij}}.$$

5. kauguspõhine klasterdamine.

   Viimaks moodustatakse kauguspõhist klasterdamist kasutades G stabiilset klastrit. Vaatluspaaride sarnasusmõõtudena kasutatakse siin eukleidilise kauguse asemel hoopis leitud konsensusmaatriksi elemente ehk osakaale. Info, kuhu iga vaatlus kuulub, talletatakse vektorisse $Z(G)$, mis sisaldab endas klastrite numbreid, kuhu vastava järjekorranumbriga vaatlus kuulub.

2.2.3 Hüperparameetrite kalibreerimine

Kasutatava algoritmi puhul on võimalik kalibreerida kaks hüperparameetrit λ ja G. Esimene neist on regulariseerimisparameeter, mida kasutatakse eelkõige suure hulga tunnuste kaalumiseseks. Teine parameeter, G, tähistab klastrite arvu. *Sharp* paketis implementeeritud meetodites kalibreeritakse need automaatselt. Seda tehakse läbi stabiilusskoori sharp (edaspidi S_c) maksimeerimise, võrkotsingu algo-

ritmi abil. See tähendab seda, et algoritmi jooksutatakse kõigi etteantud λ ja G väärtuste kooslustega, mille tulemusel saadakse neile S_c skoor. Kui parameetrit λ meetodile ette ei anta, siis tehakse võrkotsing vaid G suhtes, järelikult on jooksutuskordi vähem, meetod lõpetab töö kiiremini ning on mälusäästlikum. Kalibreeritud hüperparameetriteks saab suurima skooriga parameetrite paar. Vahemikud, kuhu parameetrid kuuluvad, sai algoritmile ette anda, kuid lõplikud väärtused leidis see ise.

2.2.4 Stabiilsusskoor S_c

Hüperparameetrite kalibreerimiseks ja meetodi töö hindamiseks on paketis kirjeldatud S_c skoor, mis näitab kui stabiilsed saadud klastrid on. Mida suurem skoor, seda stabiilsem tulemus. *Sharp* skoor arvutatakse maatriksite $C(G)$, H ja $Z(G)$ abil.

Kuna λ jäi seekord analüüsist välja, siis kirjeldame S_c leidmist sõltuvalt vaid klastrite arvust G. Olgu meil hüperparameeter G. Sellele vastava skoori leidmiseks peame defineerima muutujad:

$$
\begin{aligned}
X_w(G) &= \sum_{i<j} C_{ij}(G) \cdot I[Z_i(G) = Z_j(G)], \\
X_b(G) &= \sum_{i<j} C_{ij}(G) \cdot I[Z_i(G) \neq Z_j(G)], \\
N_w(G) &= \sum_{i<j} H_{ij} \cdot I[Z_i(G) = Z_j(G)], \\
N_b(G) &= \sum_{i<j} H_{ij} \cdot I[Z_i(G) \neq Z_j(G)].
\end{aligned}
$$

Alaindeksid w ja b tähistavad sõnu klastrisisesed (*within*) ja klastritevahelised (*between*). Muutuja X_w tähistab nende $C(G)$ elementide summat, mis on samas $Z(G)$ poolt ette antud konsensusklastris ja X_b nende elementide summat, mis on erinevates konsensusklastrites. N_w ja N_b toimivad samamoodi, aga summeeritakse

üle H elementide. Nagu valemitest paistab, siis summeeritakse vaid üle ülemiste diagonaalmaatriksite, kusjuures maatriksite diagonaale endid summadesse sisse ei arvestata. See tuleneb sellest, et $C(G)$ ja H on sümmetrilised maatriksid ning nende diagonaalid kajastavad ühe ja sama vaatluse samasse klastrisse või samasse valimisse kuulumist, mis on igal juhul garanteeritud.

Kui eeldada, et korrelatsioonid $C(G)$ elementide vahel on puhtalt $Z(G)$ struktuuri poolt põhjustatud, siis võime käsitleda maatriksi $C(G)$ elemente $C_{ij}(G)$ kui üksteisest sõltumatuid binoomjaotusega juhuslikke suurusi, tingimusel, et H ja $Z(G)$ on fikseeritud ehk

$$C_{ij}(G)|H, Z(G) \sim \mathcal{B}\left(H_{ij}, p_{ij}(G)\right).$$

Tõenäosus $p_{ij}(G)$ on defineeritud kui

$$p_{ij}(G) := \begin{cases} p_w(G), & Z_i = Z_j, \\ p_b(G), & \text{mujal.} \end{cases}$$

Binoomjaotuse aditiivsuse tõttu on ka $X_w(G)$ ja $X_b(G)$ binoomjaotusega, tingimusel, et maatriksid H ja $Z(G)$ on fikseeritud ehk

$$X_w(G)|H, Z(G) \sim \mathcal{B}\left(N_w(G), p_w(G)\right),$$
$$X_b(G)|H, Z(G) \sim \mathcal{B}\left(N_b(G), p_b(G)\right).$$

Vaatleme hüpoteeside paari

$$H_0 : p_w(G) \leq p_b(G),$$
$$H_1 : p_w(G) > p_b(G).$$

Kuna suure valimi korral läheneb binoomjaotus normaaljaotusele, siis defineeritak-

se S_c kui Z-statistik

$$S_c(G) = \frac{\hat{p}_w(G) - \hat{p}_b(G)}{\sqrt{\hat{p}_0(G)\left(1 - \hat{p}_o(G)\right)\left(\frac{1}{N_w(G)} + \frac{1}{N_b(G)}\right)}},$$

kus

$$\hat{p}_w(G) = \frac{X_w(G)}{N_w(G)},\ \hat{p}_b(G) = \frac{X_b(G)}{N_b(G)},\ \hat{p}_0(G) = \frac{X_w(G) + X_b(G)}{N_w(G) + N_b(G)}.$$

See tähendab, et mida suurem S_c on väärtus, seda kindlamalt saame öelda, et klastritesisesed tõenäosused on suuremad kui klastritevahelised.

2.2.5 Hierarhiline klasterdamine

Klasterdamisalgoritmi teises ja viiendas osas kasutati hierarhilist klasterdamist. Hierhilise klasterdamise meetodid saab jagada kaheks: aglomeratiivsed ja hargnevad. Aglomeratiivne paneb esialgu iga vaatluse eraldi klastrisse ja hakkab siis neid kokku liitma. Hargnev aga paigutab kõik vaatlused ühte suurde klastrisse ja hakkab seejärel seda tükkideks jagama. Kuna sharp kasutas hierarhilise klasterdamise implementeerimiseks stats paketist funktsiooni hclust [19], mis kasutab aglomeratiivset versiooni, täpsemalt täieliku seose klasterdamist [20], siis kirjeldame täpsemalt just seda hierarhilise klasterdamise versiooni.

Hierarhiline klasterdamine põhineb klastritevahelise kauguse mõõtmisel. Selleks, et mõõta klastritevahelist distantsi, peame kõigepealt defineerima, kuidas mõõta andmepunktide vahelist distantsi. Kasutame selleks eukleidilist kaugust, mille valem d-mõõtmelises ruumis on

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{d}(x_i - y_j)^2},$$

kus x_j ja y_j tähistavad vektorite $\boldsymbol{x}$ ja $\boldsymbol{y}$ j.-ndaid elemente.

Hierarhilise klasterdamise algoritmid erinevad üksteisest peamiselt klastritevahelise

kauguse arvutamise viisi poolest. Täieliku seose meetod kasutab selleks kaugeimate naabrite vahelise kauguse valemit. Olgu meil klastrid $C_1 = \{\boldsymbol{y}_1, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_r\}$ ja $C_2 = \{\boldsymbol{z}_1, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_s\}$, vastavalt suurustega r ja s. Tähised $\boldsymbol{y}_i$ ja $\boldsymbol{z}_i$ märgivad klastrites C_1 ja C_2 olevate vaatluste tunnuste väärtuste vektoreid. Nende klastrite kaugeimate naabrite vahelise kauguse saame leida kui

$$D(C_1, C_2) = \max_{1 \leq i \leq r,\, 1 \leq j \leq s} d(\boldsymbol{y}_i, \boldsymbol{z}_j). \tag{4}$$

Olgu meil nüüd kolm klastrit C_k, C_i ja C_j. Võrduse (4) abil oskame neid omavahel paarikaupa võrrelda. Kui aga liidame neist kaks klastrit üheks ja tahame seda võrrelda kolmandaga, siis selleks tuleb meil klastritevahelisi kaugusi uuendada. Seda saame hierarhilise klasterdamise korral teha Lance-Williamsi valemiga:

$$\begin{aligned} D(C_k, C_i \cup C_j) = {} & \alpha_i D(C_k, C_i) + \alpha_j D(C_k, C_j) \\ & + \beta D(C_i, C_j) - \gamma \left| D(C_k, C_i) - D(C_k, C_j) \right|, \end{aligned}$$

mis siinse meetodi puhul võtab kuju

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \frac{1}{2} D(C_k, C_i) + \frac{1}{2} D(C_k, C_j) - \frac{1}{2} \left| D(C_k, C_i) - D(C_k, C_j) \right| \\ &= \max\{D(C_k, C_i), D(C_k, C_j)\}. \end{aligned}$$

Meetod töötab, kuni on kõik algsed ühe vaatlusega klastrid üheks suureks klastriks liitnud. Seda, millised klastrid vahepeal moodustusid saab illustreerida näiteks dendrogrammi abil. [16]

2.3 Tunnuste skaleerimine

Nii Coxi mudel kui ka hierarhiline klasterdamine võivad olla tundlikud tunnusete väärtuste skaalade suhtes [13, 17]. Probleemide vältimiseks võiks tunnused normaliseerida. Seda saab R-is teha näiteks funktsiooniga *scale*(). Funktsioonile antakse ette tunnuse väärtuste vektor $(x_1, \ldots, x_n)$. Selle pealt arvutatakse tunnuse keskmine $\hat{\mu}$ ja standardhälve $\hat{\sigma}$. Vektori $(x_1, \ldots, x_n)$ väärtused teisendatakse valemiga

$$y_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

ning tagastatakse juba standardiseeritud kujul, y_i-dest koosneva vektori näol. [20]

3 Analüüs

Bakalaureusetöös kasutati Tartu Ülikooli Eesti Geenivaramu pseudonüümitud isikuandmeid ja EXPANSE projekti raames mudeldatud keskkonnaandmeid. Uuring toimus EBINi loa "Mõjusfääri-põhised tööriistad tervisliku linnakeskkonna heaks" (1.1-12/3435 (08.12.2020), 1.1-12/1021 (13.04.2021), 1.1-12/1021 (14.12.2021), 1.1-1/3452 (20.10. 2022), 1.1-12/1086 (13.03.2023), 1.1-12/4367 (07.12.2023), 1.1-12/388 (13.02.2024), 1.1-12/1411 (3.06.2024)) alusel, mille põhjal kogutud andmed väljastati Eesti Geenivaramu poolt andmeväljastuse 6-7/GI/1853 alusel ja analüüsiti Tartu Ülikooli serveris.

3.1 Andmed

Analüüsis kasutati TÜ Eesti Geenivaramu esimest kohorti ehk enne 2018. aastat liitunuid, keda oli kokku 52 266. Et analüüsitav aeg oleks piisavalt pikk ning selle kohta oleks saadaval ka piisavalt eksposoomiandmeid, valiti uuritavaks ajavahemikuks 01.01.2013 kuni 31.12.2023. Algusajaks valiti 2013, kuna keskkonnaandmed lingiti geenidoonoritega nende 2013. aasta aadressi alusel. Lõpuaeg oli jälgimisaja lõpp, ehk aeg, mil kõigi geenidoonorite kohta oli teada, kas nad olid jälgimisperioodil saanud hüpertoonia diagnoosi või mitte.

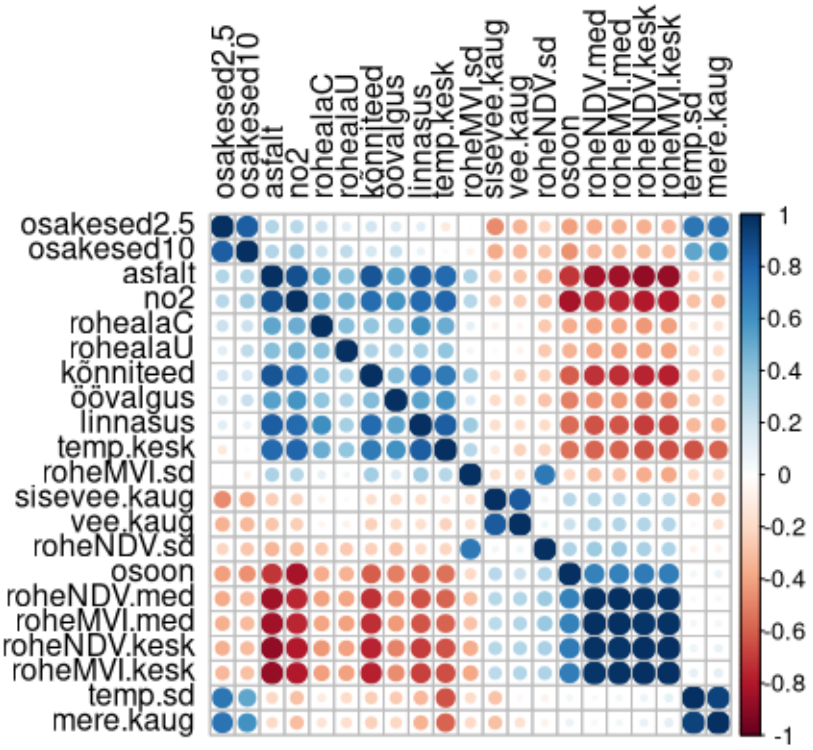
Kõrgvererõhktõbi defineeriti RHK-10 diagnooside alusel, kus juhtudel oli diagnoos 'I10' ja kontrollide hulgas ei olnud inimesi diagnoosidega 'I10', 'I11', 'I12', 'I13', 'I15', 'I27.0', 'I67.4' ei enne ega pärast vaatlusperioodi algust. Esimesest kohordist 21 502-l oli juba enne Geenivaramuga ühinemist või vaatlusperioodi algust kõrgvererõhktõbi, seega nemad jäid analüüsist välja. Ülejäänutest 6409 said diagnoosi vaatlusperioodi jooksul ning 24 355 polnud 2023. aasta detsembri seisuga haigust saanud.

3.1.1 Eksposoomitunnuste valik

Paljude geenivaramuga liitunute jaoks olid saadaval eksposoomiandmed ehk inimese elukohaga seotud väliskeskkonna tunnused EXPANSE projektist. Kuna keskkonnaandmed lingiti geenidoonoritega nende 2013. aasta aadressi alusel, siis analüüsiti samal aastal mõõdetud eksposoomitunnuste väärtusi. Argumentide puhul, mida 2013. aastal mõõdetud polnud, vaadati 2013. aastale lähimat mõõtmisaastat (2012 või 2015). Selline lähenemine andis endiselt mõistlikud tulemused, kuna suuremate vahedega mõõdetavate näitajate väärtused ei muutunud ajas nii kiiresti. Eksposoomitunnuseid oli kokku 37, lisaks olid saadaval 9 temperatuuri ilmestavat näitajat (vt lisa 1). Selleks, et temperatuur klasterdamises teisi tunnuseid üle ei kaaluks, võeti analüüsi aasta keskmine temperatuur ja standardhälve, mis sisaldasid piisavalt infot ka väljajäänud tunnuste kohta.

Mõned eksposoomitunnused sisaldasid selle uuringu kontekstis ebasobivaid või liigseid andmeid. Näiteks olid paljud tunnused mõõdetud kolmel erineval raadiusel: 300m, 500m ja 1000m. Nendest valiti tunnused raadiusega 500m. Kuna uuriti täiskasvanud inimesi, kes võivad liikuda oma elukoha ümber rohkem kui 500m, siis erandlikult valiti kergliiklusteid ja asfalti kirjeldavate tunnuste WAL ja IMP jaoks raadiuseks 1000m. Argumentidele hinnati korrelatsioonimaatriks.

Korrelatsioonimaatriksist (joonis 1) ilmnes, et paljud tunnused olid omavahel tugevalt seotud, kusjuures mõned neist väljendasid ideelt samu näitajaid. Nii $rohealaU$ (GSU_DIS) kui ka $rohealaC$ (GSC_DIS), iseloomustasid kaugust lähima rohealani ja erinesid vaid roheala definitsiooni poolest. Esimene neist oli arvutatud vaid linnades ning sisaldas seega palju puuduvaid väärtusi. Teine ei võtnud arvesse mõningaid kohti, mis võiksid olla rohealad, näiteks Tartu Botaanikaaed ja Ülejõe park (vt lisa 3). Nendest põhjustest ajendatuna jäeti mõlemad tunnused analüüsist välja. Sama omadust väljendasid ka tunnused $roheNDV$ ja $roheMVI$, mis kirjeldasid mõlemad rohelust. Nendest valiti rohelust esindama NDV. Mõlemal tunnusel olid antud nii keskväärtus, mediaan kui ka standardhälve. Kuna me-



Joonis 1: Eksposoomitunnuste korrelatsioonimaatriksi graafiline esitus

diaan ja keskväärtus olid tugevalt korreleeritud, siis jäeti mediaan välja. Edaspidi märgitakse valitud rohelisuse tunnuseid kui *rohelus.kesk* ja *rohelus.sd*. Veekogu kaugust elukohast iseloomustasid tunnused *sisevee.kaug* (BSI_DIS), *mere.kaug* (BSS_DIS) ja *vee.kaug* (BSW_DIS). Kuna veekogu kaugus on arvutatav siseveekogu ja mere kauguse abil, siis oli ka see tunnus üleliigne.

Klasterdamiseks pidid igal vaatlusel kõik argumendid olemas olema. Seega, kui mõnel tunnusel leidus teistest märgatavalt rohkem puuduvaid väärtusi, oli mõttekas argumendist loobuda. Selline lähenemine võimaldas vältida inimeste välja jätmist, kellel muud tunnused peale probleemse olemas olid. Problemaatiliste argumentide tuvastamiseks arvutati kõikidele eksposoomitunnustele puuduvate väärtuste osakaal. Tunnusel *öövalgus* (LAN) oli puuduvaid väärtusi $\sim$ 18,6%. Teiste eksposoo-

mitunnuste puuduvate väärtuste protsendid jäid $6-7\%$ vahele. Kuna öövalgusel oli ligi kolm korda rohkem puuduvaid väärtusi kui teistel, siis jäeti see tunnus edasisest analüüsist välja.

Tunnused *konniteed* (WAL) ja *rohealaU* (GSU_DIS) olid arvutatud vaid linnasiseselt. Kuna analüüsi sooviti jätta ka maapiirkonnad, siis neid tunnuseid analüüsi võtta ei saanud. Ainult linnaelanike jaoks olemas olevaid tunnuseid oleks huvitav edaspidi kasutada eraldi linnapõhistes analüüsides. Andmestikust eemaldati ka *osoon* (OZO_AAV), mida ei soovitata eksposoomiuuringutes kasutada ning siseveekogu ja mere kaugust ilmestavad tunnused.

Valituks osutusid argumendid $temp.kesk$, $temp.sd$, $rohelus.kesk$, $rohelus.sd$, $no2$, $osakesed10$, $osakesed2.5$, $asfalt$ ja $linnasus$. Edasises analüüsis võetakse arvesse, et mõned neist tunnustest on üksteisega tugevalt korreleeritud, aga rohkem ühtegi tunnust välja ei jäeta.

3.1.2 Valim

Peale eksposoomitunnuste selekteerimist koostati valim. Andmetest olid juba varasemalt välja võetud geenidoonorid, kellel oli enne analüüsiaja algust või geenivaramuga liitumist diagnoositud kõrgvererõhktõbi. Kuna mõned inimesed olid mitmel korral küsimustikele vastanud, siis oli neil ka mitu komplekti tunnuseid, mille seast valiti analüüsi 2013. aastale kuupäevaliselt kõige lähedasem.

Klasterdamisalgoritmi tööks pidid kõikidel klasterdatavatel kõik eksposoomitunnuste väärtused olemas olema, mistõttu tuli puuduvate väärtustega inimesed andmestikust eemaldada. Enamasti kehtis reegel, et ühe tunnuse puudumisel olid puudu ka teised, seega ei oleks imputeerimine teiste tunnuste põhjal võimalik olnud. Peale puuduvate väärtustega geenidoonorite eemaldamist jäi andmestikku 28 633 vaatlust. Kuna analüüsi sooviti kaasata ka sugu, suitsetamist, haridust, vanust ja kehamassiindeksit kirjeldavad taustatunnused, siis tuli nendes leiduvate puuduvate väärtuste tõttu andmestikust eemaldada veel 21% vaatlustest.

Lõpuks jäi valimisse 22 503 inimest, kellest 4844 haigus avaldus ja 17659, kellel ei avaldunud. Analüüsi hõlbustamiseks tehti haridustaset ja suitsetamist kirjeldavad tunnused binaarseteks. Faktortunnusel *haridus* olid tasemed, *korghar* ja $kuniKesk$, mis näitasid, kas inimesel on kõrgharidus või mitte. Tunnuse *suits* väärtuseks määrati 1, kui inimene suitsetas või oli kunagi suitsetanud ja 0 kui mitte. Valimi tunnuseid on kirjeldatud tabelis 1 ning pidevate tunnuste histogramme saab vaadata lisast 4.

Tabel 1: Tunnuste kirjeldus valimis ($n = 22\ 503$)

Tunnus	Näitaja väärtus	Tunnus	Näitaja väärtus
Naisi	15 421 (68,5%)	*no*2	10.26 ± 5.49
Suitsetajaid	10 070 (44,7%)	*osakesed*10	14.20 ± 2.18
Kõrgharidusega	10 310 (45,8%)	*osakesed*2.5	7.00 ± 1.18
Vanus (vahemik)	40,5 $(18 - 97)$	$asfalt$	13.65 ± 16.32
Kehamassiindeks	$24{,}49 \pm 4{,}34$	*linnasus*	19.45 ± 22.52
temp.kesk	6.72 ± 0.41	*rohelus.kesk*	5384.04 ± 927.51
temp.sd	9.16 ± 0.44	*rohelus.sd*	1393.85 ± 224.39

3.2 Klastriteta mudelid

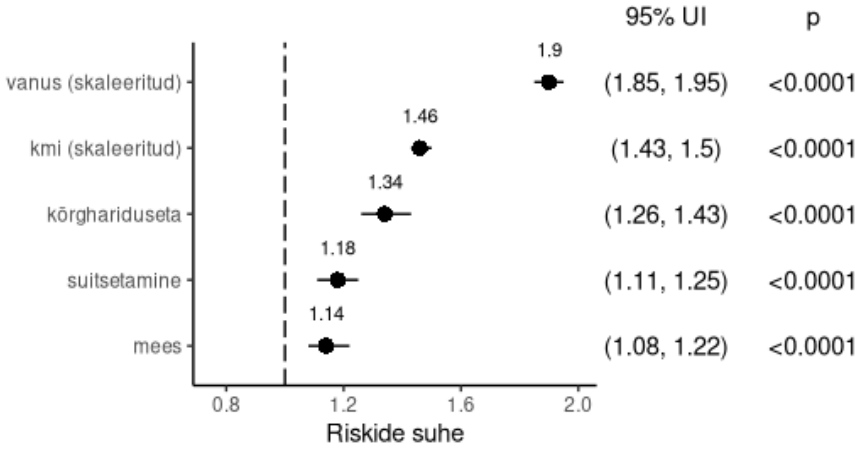
Kuna andmed olid paremalt tsenseeritud, kasutati elulemusanalüüsiks Coxi võrdeliste riskide mudelit, mida implementeeriti tarkvara R paketi *survival* abil ning koostati artiklis [21] kasutatud põhimõttel. Mudeli hindamiseks lisati andmestikku tunnused *algus*, *lopp* ja *haigestus*. Esimene neist märkis vaatlusaja algust ja kuna iga vaatlusaluse kohta olid teada tagasiminevad terviseandmed, siis isegi kui liitumine leidis aset peale aastat 2013, oli tunnuse *algus* väärtuseks 01.01.2013. Tunnus *lopp* tähistas vaatlusaja lõppu. Geenidoonoritel, kes haigestusid vaatlusaja jooksul, oli argumendi *lopp* väärtuseks diagnoosi saamise kuupäev. Nendel, kes haigust ei saanud, aga surid enne vaatlusaja lõppu, oli tunnuse väärtuseks surmakuupäev. Kui inimene oli vaatlusaja lõpuks elus ja terve, siis sai tunnus väärtuseks vaatluse lõpu kuupäeva ehk 31.12.2023. Karakteristiku *haigestus* väärtuseks märgiti 1, kui uuritav sai vaatlusperioodi jooksul kõrgvererõhktõve ja 0, kui ei saanud. Viimaks lisati andmestikku tunnus *aeg*, mis näitas geenidoonori kõrgvererõhktõveta elatud aega aastates alates vaatluse algusest. Tunnuse *aeg* väärtus arvutati valemiga

$$aeg = \frac{lopp - algus}{365}.$$

3.2.1 Taustatunnustega mudel

Kuna inimese hüpertooniasse haigestumist mõjutab palju tema eluviis [9], siis esmalt koostati elulemusmudel taustatunnustest sugu, suitsetamine, haridus, vanus ja kehamassiindeks. Selleks kasutati sammregressiooni, mille käigus vaadati esmalt kõiki argumente mudelis ühekaupa, valiti neist vähima p-väärtusega tunnus ja lisati lõplikku mudelisse. Seejärel prooviti valitud argumendiga mudelisse lisada ühekaupa ülejäänud tunnuseid. Niimoodi tekkinud mudelite võrdlusest valiti omakorda selline, mille lisatud tunnuse p-väärtus oli vähim. Protsessi korrati kuni ühegi lisatava tunnuse p-väärtus polnud enam väiksem kui 0,05. Analüüsi käigus saadud p-väärtusi on võimalik vaadata lisast 5.

Lõplikku mudelisse jõudsid kõik uuritavad argumendid. Peale mudeli hindamist kontrolliti selle võrdeliste riskide eeldust funktsiooni *cox.zph* abil, mis hindab iga argumenttunnuse kordaja puhul, kas see võiks sõltuda ajast, ja väljastab vastava p-väärtuse. Kuna kõikide tunnuste puhul oli p-väärtus suurem kui 0,05, siis võib võrdeliste riskide eelduse täidetuks lugeda. Mudeli visualiseerimiseks tehti blobogramm, millele märgiti tunnuste eksponenditud kordajad ja nende usaldusvahemikud ning p-väärtused.



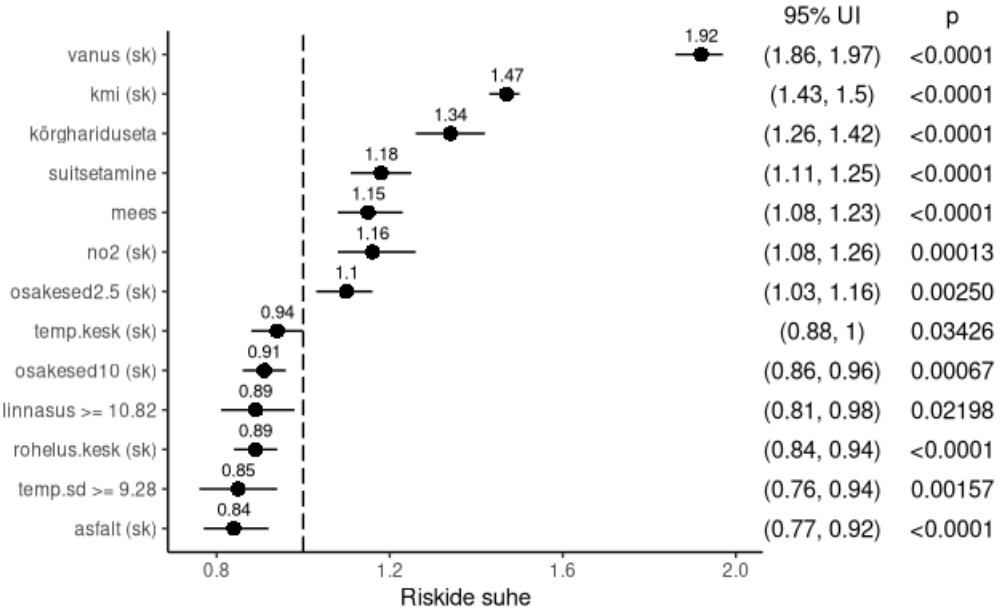
Joonis 2: Lõplik taustatunnuste mudel blobogrammina

Jooniselt 2 näeme, et kõik mudeli tunnused suurendasid kõrgvererõhktõve avaldumise riski. Teistest oluliselt suurema kordaja poolest eristus skaleeritud vanus, mis suurendas riski ligi kaks korda. Vanuse standardhälbeks oli valimis 13,2, mis tähendab, et vanuse suurenedes 13 aastat tõusis risk hüpertooniat saada 1,9 korda. Suuruselt teine kordaja oli kõrge kehamassiindeks, mille standardhälve oli valimis 4,34. See tähendab, et kehamassiindeksi suurenemine 4,34 ühiku võrra kasvatas riski umbes 1,5 korda. Peale nende tunnuste tõstsid riski suitsetamine ja meessoost olemine, mis suurendasid riski vastavalt 1,2 ja 1,1 korda. Suurem oli risk ka kõrghariduseta inimestel, kelle jaoks tõusis risk 1,3 korda.

3.2.2 Tausta- ja eksposoomitunnustega mudel

Järgmiseks lisati saadud taustatunnustega mudelisse klasterdamisel kasutatavad eksposoomitunnuseid. Seda tehti sama meetodiga, mis enne, aga nüüd olid algses mudelis taustatunnused sees. Mudeli moodustamisel saadud p-väärtusi saab vaadata lisast 6.

Lisaks taustatunnustele tulid olulised argumendid *osakesed*10, *rohelus.kesk*, *no*2, *osakesed*2.5, $asfalt$, *temp.sd*, *temp.kesk* ja *linnasus*. See tähendab, et kõik eksposoomitunnused peale roheluse standardhälbe olid hüpertoonia riskiga seotud. Mudelile tehti võrdeliste riskide eelduse kontroll, mille põhjal selgus, et tunnuste *linnasus* ja *temp.sd* puhul ei olnud riskide suhe konstantne. Kuna põhjus võis seisneda nende tunnuste jaotustes (vt lisa 4), siis tehti mõlemad binaarseks, kusjuures poolituskohtadeks valiti mediaanid. Tunnused lisati binaarsetena mudelisse tagasi ja võrdeliste riskide eelduste kontrolli korrates enam probleemi ei tuvastatud.



Joonis 3: Lõplik eksposoomi- ja taustatunnustega mudel blobogrammina

Mudelit illustreerivalt jooniselt 3 näeme, et eksposoomitunnused olid erineva mõjuga, kusjuures üldiselt mõjutasid need haigestumisriski vähem kui varem mudelisse pandud argumendid. Haigestumise riski tõstsid suurem NO_2 ja $PM_{2.5}$ kontsentratsioon õhus. Lämmastikdioksiidi kontsentratsiooni tõus 5,5$\mu g/m^3$ suurendas riski 1,16 korda ja peenosakeste kontsentratsiooni tõus 1,2$\mu g/m^3$ võrra kasvatas riski 1,10 korda. Vastupidiselt neile PM_{10} kontsentratsiooni suurenemine 2,2$\mu g/m^3$ võrra langetas riski $\frac{1}{0{,}91} = 1{,}10$ korda. Kuna *osakesed*2.5 ja *osakesed*10 olid üksteisega tugevalt korreleeritud (vt joonist 1), siis võisid nende kordajad mudelis üksteist tasakaalustada. Peale selle ei kajasta mudel nende võimalikku koosmõju. Veel vähendasid haigestumise riski suurem aastase temperatuuri keskmine ja standardhälve, keskmine rohelus, linnasus ja asfaldi protsent. Aastase keskmise temperatuuri tõus 0,4 kraadi võrra vähendas riski $\frac{1}{0{,}94} = 1{,}06$ korda. Keskmine rohelus pidi riski $\frac{1}{0{,}89} = 1{,}12$ kordseks vähendamiseks tõusma 928 võrra, sama palju vähendas riski ka 10,82 võrra suurem linnasuse indeks. Riski vähendas ka suurem temperatuuride kõikumine ehk kui temperatuuride standardhälve oli suurem kui 9,28, siis vähenes risk $\frac{1}{0{,}85} = 1{,}18$ korda. Kõige rohkem vähendas riski asfaldi protsent, mille suurenemisel 16,3 võrra vähenes kõrgvererõhktõve risk $\frac{1}{0{,}84} = 1{,}19$ korda. Ka nende hulgas oli mitu üksteisega seotud tunnust, näiteks keskmine rohelus ja asfaldi protsent. Mudelist võis välja lugeda, et nii kõrge rohelus kui ka kõrge asfaldi protsent olid riski vähendava efektiga. See võib tunduda vasturääkivana, aga suurem asfaldi protsent võis tegelikult viidata kergliiklusteede olemasolule, mis toetasid omakorda elanike liikumisharjumust. Asfaldi ja kõnniteede korreleeritust kinnitas ka korrelatsioonimaatriks (joonis 1). Järgmistes töödes oleks huvitav suurlinnu analüüsida eraldi, et kaasata ka praegu välja jäänud, vaid linnades defineeritud tunnuseid.

Mudel võib tulla parem, kui uurida ka tunnuste koosmõjusid. Neid oleks aga keeruline interpreteerida, eriti, kui olulised tuleksid rohkem kui kahe tunnuse koosmõjud. Seetõttu muudaks nende lisamine niigi keerulise mudeli veelgi raskemini hoomatavaks. Probleemi lahendamiseks kasutati klasteranalüüsi.

3.3 Klasteranalüüs

Klasterdati 28 642 geenidoonorit ehk alamvalimit $n = 30\ 764$ doonorist, kellel kõik valitud eksposoomitunnused olemas olid. Klasterdamiseks kasutati temperatuuri (*temp.kesk*, *temp.sd*), rohelust (*rohelus.kesk*, *rohelus.sd*), õhureostust (*no*2, *osakesed*10, *osakesed*2.5), kõvakattega pinda ($asfalt$) ja linnasust (*linnasus*) iseloomustavaid karakteristikuid.

3.3.1 Klasterdamisalgoritmi parameetrite valimine

Suurte andmehulkade konsensusklasterdamine on võrdlemisi mälumahukas. Seepärast arendati vajalik kood Tartu Ülikooli tundlike andmete analüüsiserveris (SA-PUs) [22], aga kogu valim klasterdamiseks kasutati suurema mäluga HPC serverit. Siin peatükis räägitakse täpsemalt, kuidas klasterdamisel kasutatud parameetrid valiti. Koodi arendamisel kasutati suunisena R. Colindresi 2024. aasta praktikumide materjale [23]. Kauguspõhise konsensusklasterdamise algoritmi kasutades oli võimalik mitu parameetrit ise valida. Peamine nendest oli valimite klasterdamiseks kasutatav meetod ehk implementatsiooni parameeter. Sobiva meetodi leidmiseks võeti klasterdatavast andmehulgast väiksemad valimid ja mõõdeti nende jooksutamiseks kulunud aega.

Tabel 2: Meetodi *Clustering()* jooksutamiseks kulunud aeg

Meetod	100 objekti	1000 objekti
K-keskmised	2.135 s	7.986 min
GMM	error	error
Hierarhiline	1.305 s	5.665 min
PAM	2.837 s	11.82 h
DBSCAN	ei tööta	ei tööta

Tabelist 2 torkab silma, et mõned meetodid töötavad isegi uuringu kontekstis väikeste andmemahtudega aeglaselt või ei tööta üldse. Edasi otsustati uurida neist kii-

remaid ehk k-keskmisi ja hierarhilist meetodit. Selleks suurendati testandmestikku 10 000 objektini, mis moodustab umbes kolmandiku klasterdada soovitava andmestiku mahust. Kumbki meetod ei hakanud SAPUs sellise andmehulgaga tööle, seega tuli mälu säästmiseks teisi vaikimisi seatud parameetrite hulki kitsamaks teha. Esmalt kohendati parameetrit n_c, mis hoidis endas klastrite arve sisaldavat vektorit. Parameetri väärtuseks seati arvud ühest kümeneni. Andmevektor oli endiselt liiga suur, seega vähendati ka parameetrit K, mis näitas mitu valimit andmestikust võeti. Parameeter vähendati esialgu kümne peale ning prooviti klasterdada mõlema meetodiga, K väärtust aina suurendades. Nõnda saadud tulemused on näha tabelites 3 ja 4.

Tabel 3: Erinevate K väärtustega klasterdamine K-keskmiste meetodiga

näitaja \ K	$K = 10$	$K = 20$	$K = 30$	$K = 40$	$K = 50$	$K = 60$
aeg (min)	5.17	6.64	10	10.93	13.12	15.36
max S_c	39	92	151	383	414	605
klastrite arv	7	8	10	3	10	10

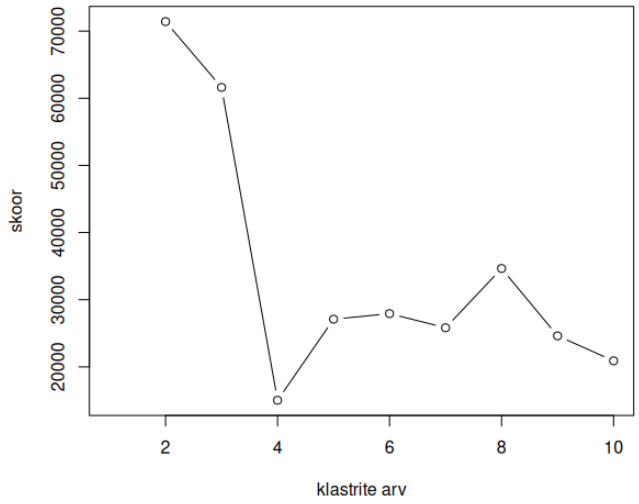
Tabel 4: Erinevate K väärtustega klasterdamine hierarhilise meetodiga

näitaja \ K	$K = 10$	$K = 20$	$K = 30$	$K = 40$	$K = 50$	$K = 60$
aeg (min)	4.66	6.77	9.21	12.02	14.42	16.19
max S_c	147	391	467	555	740	2875
klastrite arv	10	2	6	2	8	6

Tabeleid võrreldes on näha, et klasterdamiseks kulunud aja poolest meetodid palju ei erinenud. Küll aga on näha, et hierarhiline meetod suutis sama alamvalimite arvuga saavutada kõrgema S_c skooriga tulemusi. Peale selle oli ka hierarhilise meetodi kalibreerimiskõver stabiilsem, mis tähendab, et oli võimalik eristada üks teistest suurem S_c skoor. Nende põhjuste tõttu valiti klasterdamise alammeetodiks hierarhiline klasterdamine. Kuna peaaegu et kõigi, v.a hierarhilise klasterdamise $K = 60$ korral, meetod ei koondunud, siis oli oht, et parameetrit K pidi suurema andmemahu peal tõstma.

3.3.2 Klasterdamine

Parameetrid seadistati $K = 200$, $\tau = 0.5$ ja $n_c = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$ (parameetrite tähendusi on seletatud eelnevates peatükkides). Eelmises alapeatükis pöörati tähelepanu sellele, et 60 alamvalimist ei pruugi 28 642 vaatluse klasterdamiseks piisata. Kuna mõned vaatluste paarid ei sattunud $K = 60$ puhul ühtegi alamvalimisse, siis tuli K väärtust suurendada. Kuna K väärtus oli meetodis vaikimisi 200, siis prooviti järgmisena seda. Vaikimisi väärtus oli piisavalt suur ja rohkem probleeme klasterdamise meetodis ei täheldatud. Algoritmisiseste valimite klasterdamiseks kasutati hierarhilist klasterdamist, täpsemalt täieliku seose meetodit.



Joonis 4: Kalibreerimiskõvera graafik

Jooniselt 4 näeme, et algoritm valis stabiilseimaks klastrite arvuks kaks. Stabiilsuselt teisena on graafikult näha kolm klastrit ja kolmandana kaheksa klastrit. Kalibreerimiskõveralt on näha, et kaheks ja kolmeks klastriks jaotamine ei erinenud teineteisest skoori poolest eriti palju. Tegelikult võiks parim klastrite arv rohkem teistest eristuda. Edasi vaadati, kuidas inimesed klastritesse jaotuvad.

2 klastrit		3 klastrit		8 klastrit	
klaster	inimeste arv	klaster	inimeste arv	klaster	inimeste arv
1	19 011	1	19 011	1	15 337
2	9622	2	45	2	6837
		3	9577	3	2072
				4	790
				5	1562
				6	333
				7	1682
				8	20

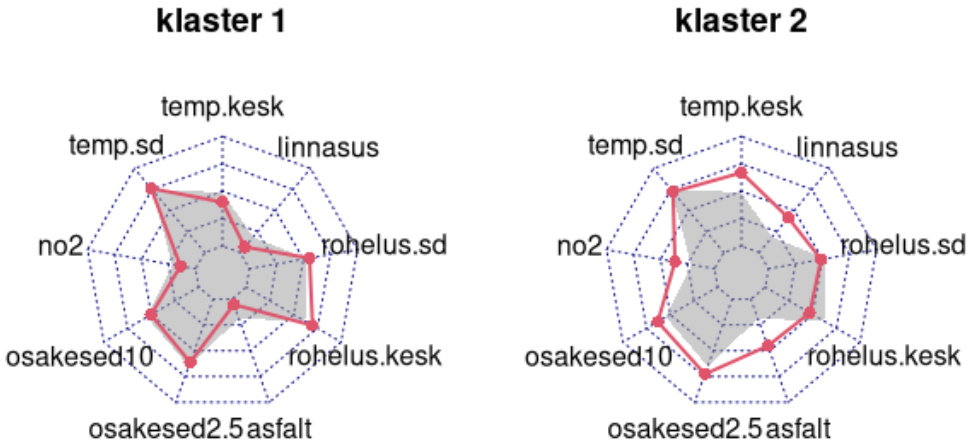
Tabel 5: Inimeste arv klastrites kahe, kolme ja kaheksa klastri puhul

Tabelist 5 on näha, et kolmeks jagunemisel jäi ühte klastrisse väga vähe inimesi. Kuna kaheks ja kolmeks jaotumine erinesid üksteisest vaid 45 inimese poolest, siis jäeti kolm klastrit edasisest analüüsist kõrvale. Oli alust arvata, et kaks klastrit jagunevad linnaks ja maaks, seega otsustati uurida ka paremuselt kolmandat, kaheksa klastriga varianti.

Varasemates teadustöödes [24, 7] on klastrite arv olnud fikseeritud. Sellise lähenemise eeliseks oli kõikides kohortides sama arvu ja seega ka üksteisega võrreldavate klastrite tuvastamine. Eeldati, et on kolm ekspotüüpi, madal, keskmine ja kõrge. Sellise lähenemisega välistati aga võimalus avastada teistsugust, potentsiaalselt haigustele mõju avaldava ekspotüübiga jaotust.

3.3.3 Kahe klastri analüüs ja mudel

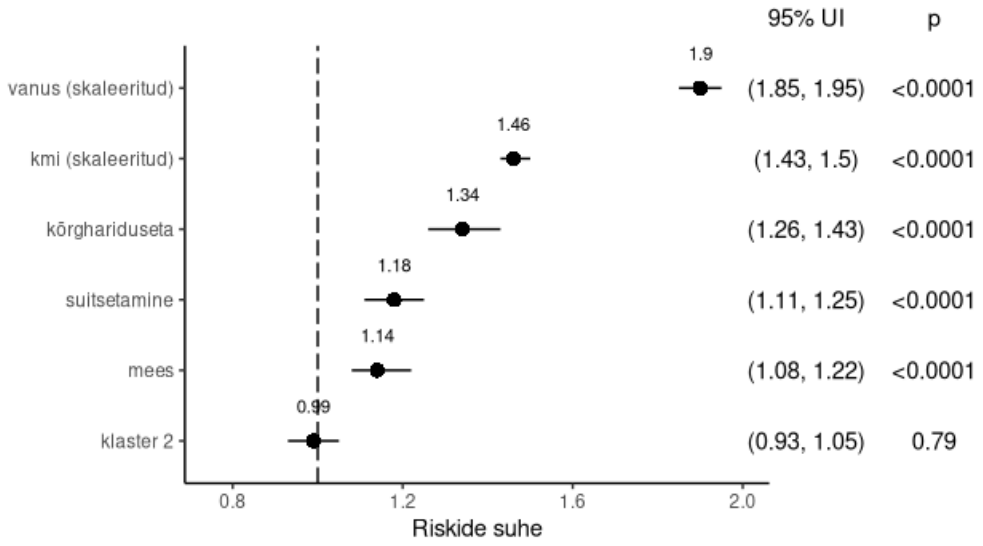
Esmalt uuriti andmestiku kaheks klastriks jaotumist. Klastrite erinevuste tuvastamiseks tehti radardiagramm, kuhu kanti tunnuste keskmised valimis (halliga) ja klastrites (punasega). Täpsemalt saab tunnuste keskmisi ja standardhälbeid uurida lisast 7.



Joonis 5: Kahe klastri radardiagramm

Jooniselt 5 näeme, et klastris 1 oli väiksem keskmine temperatuur, õhureostus ja linnalisus ning madalam asfaldi ja ehitiste protsent. Võrreldes klastriga 2 oli klastris 1 suurem keskmine aastane rohelus. Temperatuuri ja roheluse standarhälbed olid enam vähem samad, klastris 2 olid mõlemad veidi madalamad. Kuna teatavasti on linnades suurem õhureostus, rohkem asfalti ning vähem rohelust, siis selline tunnuste keskmiste võrdlus andis alust arvata, et klasterisse 2 kuulusid linnale sarnase ekspotüübiga inimesed ning klastrisse 1 linnast erineva ekspotüübiga inimesed.

Uurimaks, kas kumbki neist ekspotüüpidest mõjutab hüpertoonia riski, lisati klastri tunnus varem tehtud taustatunnuste mudelisse. Referentstunnuseks võeti klaster 1. Saadud mudeli visualiseerimiseks koostati blobogramm.



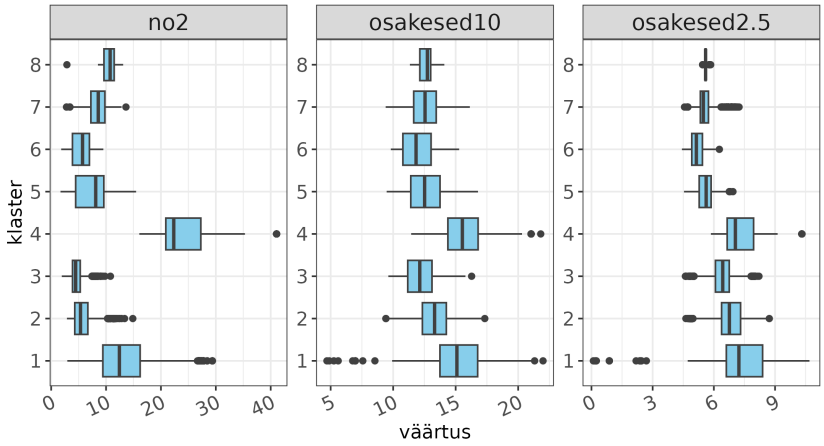
Joonis 6: Kahe klastriga mudeli blobogramm

Blobogrammilt (joonis 6) on näha, et klastrite riskide suhte usaldusvahemik hõlmas arvu üks. Kuna ka p-väärtus oli klastri tunnusel üpris suur, siis järelikult ei olnud klaster mudelis oluline. See tähendab, et ei saa öelda, et klastris 1 oleks suurem või väiksem risk kõrgvererõhktõvele kui klastris 2.

3.3.4 Kaheksa klastri analüüs ja mudel

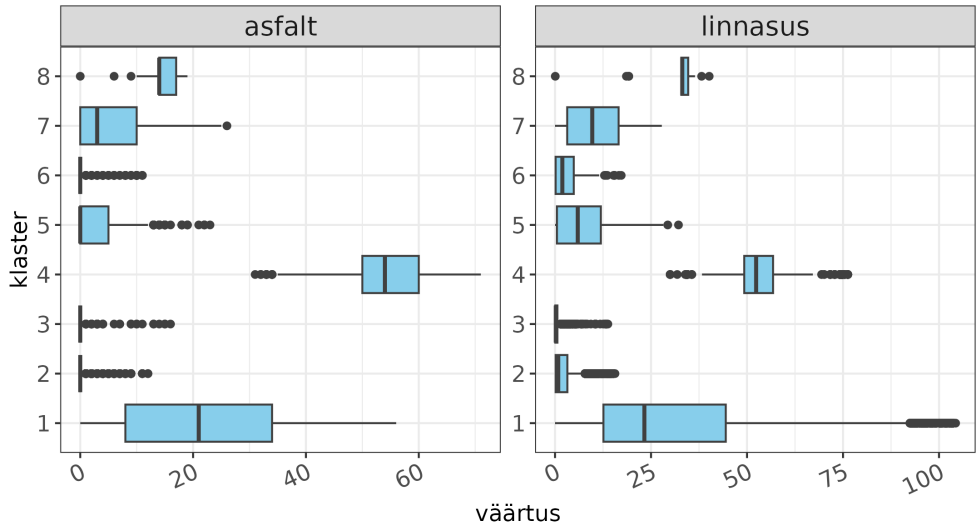
Kaheks klastriks jaotamine oli võrdlemisi üldine ja ei tuvastanud ekstreemsemaid klastreid, mis oleksid võinud osutuda hüpertoonia riskiteguriteks. Seepärast otsustati uurida ka stabiilsuselt kolmandat, kaheksaks klastriks jaotumist. Ka nende klastrite paremaks eristamiseks koostati radardiagrammid.

Radardiagrammidelt (lisa 9) eristusid teistest selgemini klastri 1 keskmised, mis olid valimi keskmistega väga sarnased ning klastrite 4 ja 8 keskmised, mis olid üldkeskmistest ekstreemsemad. Kuna klastreid 2, 3, 5, 6 ja 7 oli radardiagrammi põhjal üksteisega keeruline võrrelda, siis koostati karpdiagrammid, mille alusel võrreldi klastreid neljas kategoorias: õhureostus, rohelus, linnasus ja temperatuur.



Joonis 7: Karpdiagrammid õhureostuse tunnustele klastrites

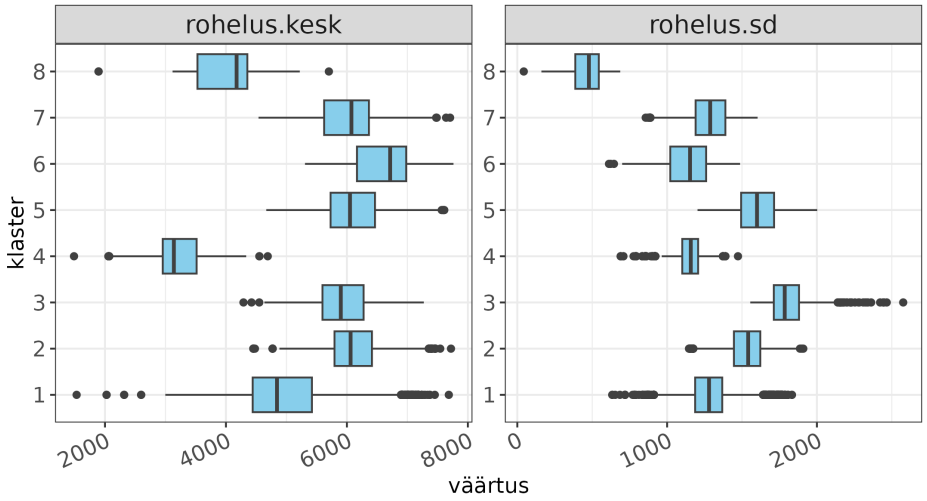
Joonisel 7 kujutatud karpdiagrammide põhjal võis öelda, et klastris 4 oli õhureostus kõige suurem, kusjuures eriti kõrge oli selle NO_2 tase. Teistest kõrgema õhureostuse poolest eristus ka klaster 1, mis oli peenosakeste kontsentratsioonilt sarnane klastriga 4, lämmastikdioksiidi kontsentratsioon jäi seal aga keskmisele tasemele. Klastrid $2, 3, 5, 7$ ja 8 olid õhureostuselt üsna sarnased. Veidi kõrgem NO_2 tase oli nende hulgas klastrites $5, 7$ ja 8, $PM_{2.5}$ oli kõrgem klastrites 2 ja 3 ning PM_{10} klastris 3. Kõige puhtama õhuga oli klaster 6, kus olid kõige madalamad peenosakeste mediaanid ja väike NO_2 tase.



Joonis 8: Karpdiagrammid linnasuse tunnustele klastrites

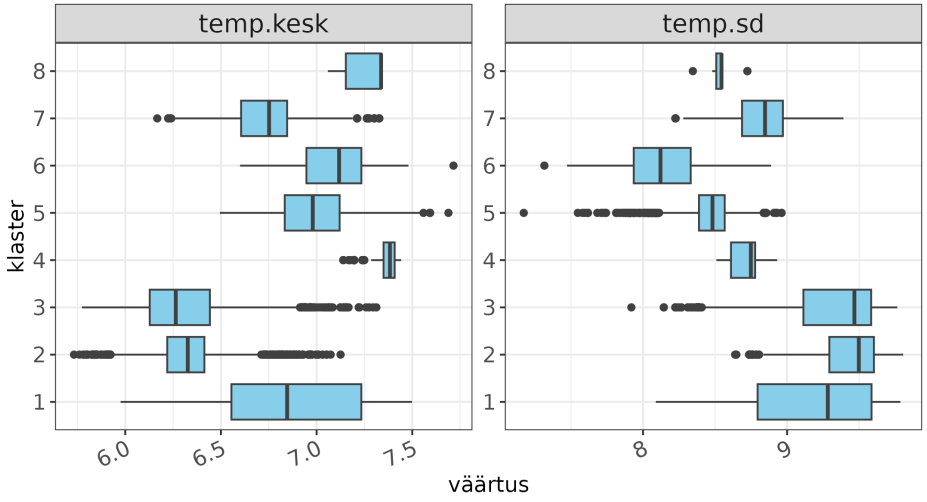
Jooniselt 8 on näha, et kõige kõrgem asfaldi protsent ja linnalisus olid klastris 4. Veidi madalama, kuid endiselt kõrge tasemega olid klastrid 1 ja 8, kusjuures klastris 1 oli ülemise ja alumise kvartiili vahe üsna suur ning väärtused olid jaotunud üle kogu võimaliku vahemiku. Teisted klastrid olid kompaktsemad ning nendesse sattunud vaatlused ei võtnud väärtusi kogu võimaliku vahemiku ulatuses. Kõige madalamad asfaldi protsenti ja linnasust kirjeldavate tunnuste mediaanid olid klastrites 2, 3 ja 6, kus nende väärtused olid nullilähedased. Nendes gruppides oli ka tunnuste hajuvus madal. Klastrid 5 ja 7 paigutusid teistega võrreldes asfaldi ja linnalisuse poolest keskele.

Joonise 9 põhjal võib öelda, et kõige madalama keskmise rohelisusega oli klaster 4 ja madal tase oli ka klastris 8. Enamus klastreid olid sarnase, kõrge rohelisuse tasemega. Kuigi klastrites 2 ja 3 olid madalad asfaldi ja linnalisuse mediaanid, siis roheluse poolest ei olnud nad klastritest 5 ja 7 kõrgemad. Teistest veidi kõrgema keskmise roheluse mediaaniga oli klaster 6. Klaster 1 paigutus rohelisuse poolest keskele. Aastase rohelisuse keskmise ja standardhälbe lõikes ei eristunud klastreid, kus kvartiilide vahe oleks võrreldes teistega palju suurem olnud.



Joonis 9: Karpdiagrammid roheluse tunnustele klastrites

Rohelisuse standardhälve oli madalaim klastris 8 ja suurim klastris 3. Madal standardhälve võib viidata igihaljastele taimedele (nt kuusepuud), suur aga näiteks põllule. Kuna klastris 2 ja 3 on väga vähe asfalti ja suur roheluse standardhälve, siis võivad need olla maakohad. Klastrites 1, 4, 6 ja 7 oli roheluse muutus aasta jooksul keskmise tasemega.



Joonis 10: Karpdiagrammid temperatuuri tunnustele klastrites

Aasta keskmise temperatuuri poolest olid madalamad klastrid 2 ja 3 ning kõrgemad 4, 6 ja 8. Kõrge temperatuur võib viidata kõrgemale asfaldi protsendile ja linnalisusele [25], aga kuna temperatuuride vahed ei ole väga suured, siis võib kõrgem temperatuur olla tingitud ka lihtsalt geograafilisest asukohast [26]. Aastaste temperatuuride standardhälbed olid kõrgemad klastrites 1, 2 ja 3. Teistest madalam standardhälve oli klastris 6. Temperatuuride vähene kõikumine võib viidata mere lähedusele.

Klastritest parema üldise ülevaate saamiseks tehti tabel, kirjeldamaks nende olulisemate tunnuste tasemeid.

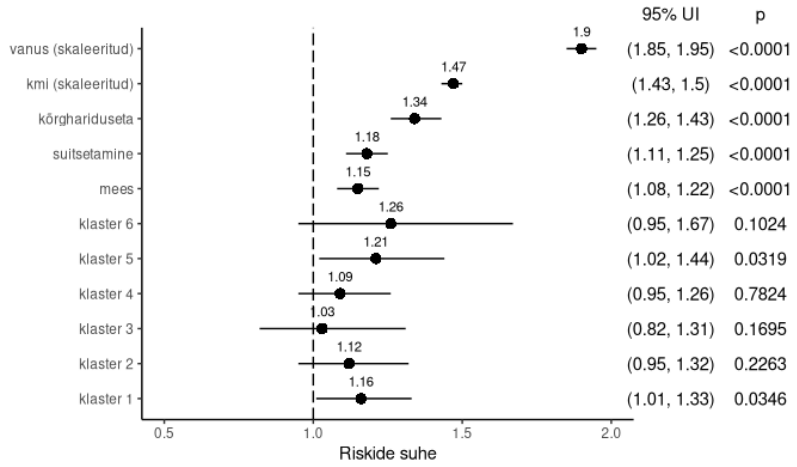
klaster	õhureostus	linnalisus ja asfalt	rohelus
1	kõrge	kõrge	keskmine
2	keskmine	madal	kõrge
3	keskmine	madal	kõrge
4	väga kõrge	väga kõrge	madal
5	keskmine	keskmine	kõrge
6	madal	madal	kõrge
7	keskmine	keskmine	kõrge
8	keskmine	kõrge	madal

Tabel 6: Olulisemate tunnuste tasemed klastrites

Klastrites 1, 4 ja 8 oli kõrge õhureostus ja linnalisus ning madal rohelus, seega võiksid nad kirjeldada suurlinnale sarnast eksposoomi. Kõige reostunum oli neist klaster 4 ning kõige puhtam klaster 8. Suure reostustaseme tõttu võiks klaster 4 olla üks hüpertoonia riskifaktoritest. Üksteisega olid sarnased klaster 5 ja 7, kus oli vähem õhureostust ja suurem rohelus kui klastrites 1, 4 ja 8. Linnalisust kirjeldavate tunnuste poolest olid neis veidi väiksemad väärtused, kusjuures klaster 7 oli natuke linnalisem kui klaster 5. Klastreid 2 ja 3 ilmestas madal õhureostuse ja linnasuse tase ning suur keskmine rohelus. Nende kahe võrdluses oli kastris 2 veidi suurem õhureostus. Kõige puhtam ja rohelisem oli klaster 6, kus oli madal õhureostus ja linnalisus ning kõrgeim rohelus.

Klastri tunnus tehti faktortunnuseks ja lisati taustatunnustega mudelisse, referents-

väärtuseks võeti näitajate järgi teiste klastritega võrreldes keskele paigutunud klaster 7. Kuna klastris 8 oli vaid 20 inimest, siis mudelisse jättes oli selle kordajal väga suur usaldusvahemik, seega jäeti see tase lõplikust mudelist välja. Mudeli võrdeliste riskide eeldus oli täidetud.



Joonis 11: Kaheksa klastriga mudeli blobogramm

Kuigi oma olemuselt olid ekstreemsemad klastrid 4 ja 6, siis nende seost kõrgvererõhktõve riskiga mudel ei tuvastanud. Nendele vastavate kordajate punkthinnangud on küll ühest suuremad, aga usaldusvahemikud on laiad ning sisaldavad ka väärtust üks. Oluliseks loeti aga klastrid 1 ja 5, mis võrreldes klastriga 7 hüpertoonia riski suurendasid. Selline tulemus oli pigem üllatav, sest mediaanide poolest klastrid 5 ja 7 palju ei erinenud. Klastrite jaoks leitud kordajate usaldusvahemikud tulid laiad (vt joonis 11), seega oli keeruline öelda, kui palju klastrisse 1 või 5 kuulumine täpsemalt riski suurendas. Ligikaudu võis öelda, et mõlemate puhul suurenes hüpertoonia risk umbes 1,2 korda. Elukohast rohkem tõstsid haigestumise riski taustatunnused vanus ja kehamassiindeks, mida oli ka märgatavalt kitsamate usaldusvahemike tõttu kergem interpreteerida. Suitsetamise, soo ja hariduse kordajate usaldusvahemikud kattuvad oluliste klastrite omadega, seega ei saanud öelda, et need oleksid haigestumisriski suurendanud rohkem kui elukoht.

4 Tulemuste arutelu

Analüüsi tulemusel selgus, et kõrgvererõhktõve avaldumise risk suureneb enim vanuse ja kehamassiindeksi kasvuga. See läheb kokku varasemate uuringutega, mis on leidnud, et vananedes kaotavad arterid elastsust, suurendades seeläbi vererõhku, kusjuures elastsuse vähenemist on täheldatud ka kõrgema kehamassiindeksiga inimeste hulgas [27]. Peale selle oli hüpertensiooni haigestumise risk suurem suitsetajate, meeste ja kõrghariduseta inimeste seas. Madalam haridustase võib olla seotud madalama sotsiaalmajandusliku staatusega, mis on varasemates uurimustes välja toodud kui üks hüpertoonia riskifaktoritest [28].

Tulemused näitasid, et eksposoomitunnustest suurendasid kõrgvererõhktõve riski kõrge NO_2 ja $PM_{2.5}$ kontsentratsioon õhus. Ka varasemates uurimustes on õhureostus osutunud keskkonnatunnuste seas üheks olulisemaks riskifaktoriks [2]. Riski vähendasid kõrgem asfaldi protsent, temperatuuride kõikumine, rohelus, linnasus ja PM_{10} kontsentratsioon. Enim vähenes risk asfaldi protsendi ja temperatuuride kõikumise tõusuga. Kuna kõva pinna protsent on suurem linnas, siis võis seos asfaldiga viidata tegelikult näiteks paremale arstiabi kättesaadavusele või elukvaliteedile, samuti sotsiaalsele eksposoomile, mida käesolavas töös ei uuritud. Linnas on ka rohkem kergliiklusteid, mis võimaldavad tervislikumaid liikumisviise.

Keskkonnatunnuste kui komplekti analüüsimiseks kasutati klasterdamist. Täpsemalt uuriti kaheks ja kaheksaks klastriks jaotunud andmestikku. Kahte klastrit omavahel võrreldes olid ühes neist valimiga sarnased või veidi madalamad tunnuste keskmised ja teises kõrgemad. Andemestiku kaheks ekspotüübiks jaotamisel eksposoomi seost hüpertoonia riskiga ei tuvastatud.

Kaheksa klastri võrdlusel eristusid kõrgema urbaniseerumise ja reostuse taseme poolest teistest kolm klastrit - 1, 4 ja 8. Suurim saastatuse tase esines klastris 4, kus oli kõrgeim õhureostuse ja linnalisuse määr ning vähe rohelust. Ülejäänud klastrid olid madalama asfaldi protsendi ja linnalisusega. Kõige vähem urbaniseerunud oli klaster 6, mida iseloomustasid kõrge rohelus ning vähene õhureostus ja linnali-

sus. Kuigi klastrid 4 ja 6 esindasid ekstreemsemaid ekspotüüpe, siis mudelis need olulised ei tulnud. Selgus, et kõrgvererõhktõve avaldumise riski tõstsid klastrid 1 ja 5. Laiade usaldusvahemike tõttu oli keeruline öelda, kui palju need ekspotüübid riski suurendasid, aga tõenäosusega 0,95 võis öelda, et klastri 1 ekspotüüp tõstis riski 1,01 kuni 1,33 korda ja klastri 5 ekspotüüp 1,02 kuni 1,44 korda. Kuna baasväärtuseks oli mudelis klaster 7, siis võrdleme seda klastritega 1 ja 5. Klastrid 5 ja 7 olid üksteisega väga sarnased ning erinesid veidi vaid roheluse standardhälbe ja temperatuuri poolest. Neis klastrites oli võrreldes teistega keskmine õhureostus, linnalisus ja asfaldi protsent. Nende sarnasuse tõttu on keeruline leida põhjendust riski tõusule. Klastris 1 oli kõrge õhureostuse ja linnalisuse tase ning suur asfaldi protsent. Keskmise roheluse ja selle standardhälbe poolest oli ekspotüüp vahepealse tasemega. Lisaks ilmestas klastrit pigem suur temperatuuride kõikumine. Tasemeid eksposoomitunnuste Coxi mudeliga võrreldes peaks klastril olema pigem hüpertoonia riski vähendav mõju. Vastuolud mudelite vahel viitavad eksposoomitunnuste kui kompleti käsitlemise väärtusele.

Edasistes uurimustes võiks eksposoomitunnuseid klasterdada teisiti. Näiteks oleks mõistlik tunnuseid klasterdada gruppide kaupa (õhureostus, temperatuur jne), mitte kõiki koos. Selline lähenemine on kasutusel näiteks artiklis [24]. Peale gruppidesse jagamise võiks kaaluda ka teise algoritmisisese klasterdamismeetodi kasutamist. Rohkem võiks uurida linnasiseseid ekspotüüpe, mille leidmiseks saaks klasterdamisel kasutada praegu analüüsist välja jäänud, vaid linnasiseselt arvutatud näitajaid. Samuti saaks ainult Tartu või Tallinna mudeli tulemusi tõlgendada mõlema linna kontekstis eraldi. Hüpertoonia riski ja eksposoomi seost võiks uurida ka 150 000 hilisema geenivaramuga liitunu seas. Kuna kasutatud paketi klasterdamise võimekus oli töö valmimise ajal limiteeritud 50 000 vaatlusele, siis poleks see siin kasutatud lähenemisega mõistlik. Huvitav oleks tulevikus mudelisse kaasta lisaks eksposoomile ka geneetiline risk ning uurida selle koosmõjusid eksposoomiga.

Kokkuvõte

Bakalaureusetöö eesmärgiks oli uurida keskkonnategurite mõju kõrgvererõhktõve esinemise riskile. Selleks kasutati Tartu Ülikooli Eesti Geenivaramu RHK-10 koodide põhiseid terviseandmeid, vastuseid geenidoonori poolt täidetud küsimustikest ja EXPANSE projekti raames mudeldatud eksposoomi keskkonna komponenti kirjeldavaid andmeid.

Esmalt koostati mudel inimest iseloomustavatest teguritest, mille põhjal selgus, et vaadatud valimi jaoks suurenes hüpertoonia risk enim vanuse ja kehamassiindeksi kasvuga. Lisaks tõstsid haigestumise riski ka kõrghariduse omandamata jätmine, suitsetamine ja meessugu. Sellised tulemused ühtivad varasemate teadmistega [1, 9, 28].

Kirjeldamaks väliskeskkonna mõju inimese haigestumisele uuriti esmalt, kuidas ekspososoomitunnused eraldi haigestumist mõjutavad. Selleks lisati need varem koostatud, inimest kirjeldavate tunnustega mudelisse. Riski vähendasid suur aastase temperatuuri kõikumine ja kõrge rohelus. Lisaks neile vähendasid riski kõrgem asfaldi protsent, linnasus, PM_{10} kontsentratsioon ja keskmine temperatuur. Riski tõstsid aga suurem $PM_{2.5}$ ja NO_2 kontsentratsioon.

Eksposoomitunnuste vahel võis leiduda palju interaktsioone, mida mudel ei kajastanud. Probleemi lahendamiseks klasterdati andmestik keskkonnatunnuste põhjal, kasutades kauguspõhist konsensusklasterdamist. Stabiilseimad klastrid moodustusid valimi kaheks ja kaheksaks jaotamisel. Kaheks jaotamisel ei tuvastatud ekspotüübi seost hüpertoonia riskiga.

Kaheksaks ekspotüübiks jaotumisel tuvastati seos neist kahe jaoks. Esimeses olulises klastris oli kõrge õhureostuse ja linnalisuse tase ning keskmine rohelus ning teises keskmine õhureostus ja linnalisus ning kõrge rohelus. Klasterdamine tuvastas ka olulistest klastritest ekstreemsemate väärtustega ekspotüüpe, aga kõrgvererõhktõve riskiga neil seost ei leitud.

Kasutatud allikad

[1] K. T. Mills, A. Stefanescu ja J. He. "The global epidemiology of hypertension". *Nature Reviews Nephrology* 16.4 (2020). DOI: 10.1038/s41581-019-0244-2.

[2] R. D. Brook. "The Environment and Blood Pressure". *Cardiology Clinics* 35.2 (2017). DOI: 10.1016/j.ccl.2016.12.003.

[3] C. P. Wild. "Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology". *Cancer Epidemiology, Biomarkers and Prevention* 14.8 (2005). DOI: 10.1158/1055-9965.EPI-05-0456.

[4] J. Vlaanderen *et al.* "Developing the building blocks to elucidate the impact of the urban exposome on cardiometabolic-pulmonary disease: The EU EXPANSE project." *Environmental Epidemiology* 5.4 (2021). DOI: 10.1097/EE9.0000000000000162.

[5] A. Saucy, F. Coloma, C. Åström S. Olmos, N. Blay, J.M.A. Boer *et al.* "Socioeconomic inequalities in the external exposome in European cohorts: The EXPANSE project". *Environmental Science and Technology* 58.37 (2024). DOI: 10.1021/acs.est.4c01509.

[6] H.-M. Kukk. "Väliskeskkonna mõju astmariskile". Bakalaureusetöö. Tartu Ülikool, 2024.

[7] A. Guillien, S. Cadiou, R. Slama ja V. Siroux. "The Exposome Approach to Decipher the Role of Multiple Environmental and Lifestyle Determinants in Asthma". *International Journal of Environmental Research and Public Health* 18.3 (2021). DOI: 10.3390/ijerph18031138.

[8] Tervise Arengu Instituut. *Tervisestatistika ja terviseuuringute andmebaas*. 2021. URL: https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__05Uuringud__01ETeU__03Haigused/ETU30.px/table/tableViewLayout2/ (vaadatud 10.03.2025).

[9] World Health Organization. *Hypertension*. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/hypertension (vaadatud 10.03.2025).

[10] J.M. Keaton, Z. Kamali, T. Xie *et al.* "Genome-wide analysis in over 1 million individuals of European ancestry yields improved polygenic risk scores for blood pressure traits". *Nature Genetics* 56 (2024), 778–791. DOI: https://doi.org/10.1038/s41588-024-01714-w.

[11] E. T. Lee ja J. W. Wang. *Statistical Methods for Survival Data Analysis*. 2. väljaanne. John Wiley ja Sons, Inc, 1992, lk. 1–4, 8–17, 250–253.

[12] E. T. Lee ja J. W. Wang. *Statistical Methods for Survival Data Analysis*. 3. väljaanne. John Wiley ja Sons, Inc, 2003, lk. 298–300.

[13] D. Collett. *Modelling Survival Data in Medical Research*. 4. väljaanne. Chapman ja Hall/CRC, 2023, lk. 55–56. DOI: 10.1201/9781003282525.

[14] M. Möls. *Elukestvusanalüüs II, slaidid*. 2024. URL: https://www-1.ms.ut.ee/mart/biostat2024/loeng7.pdf (vaadatud 14.04.2025).

[15] L. Kaufman ja P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. 2009, lk. 1–3.

[16] G. Gan, C. Ma ja J. Wu. *Data Clustering: Theory, Algorithms, and Applications*. Society for Industrial ja Applied Mathematics, 2007, lk. 3–6, 43–46, 71, 95–97, 109, 116, 120–122. DOI: 10.1137/1.9780898718348.

[17] B. Bodinier *et al.* "Automated calibration of consensus weighted distance-based clustering approaches using sharp". *Bioinformatics* 39.11 (2023). DOI: 10.1093/bioinformatics/btad635.

[18] *EXPANSE Exposome Toolbox*. 2020. URL: https://expanseproject.eu/toolbox/ (vaadatud 01.04.2025).

[19] B. Bodinier. *sharp: Stability-enHanced Approaches using Resampling Procedures*. R package version 1.4.6. 2024. URL: https://CRAN.R-project.org/package=sharp (vaadatud 16.04.2025).

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. 2023. URL: https://www.R-project.org/ (vaadatud 16.04.2025).

[21] K. Fischer *et al.* "Biomarker Profiling by Nuclear Magnetic Resonance Spectroscopy for the Prediction of All-Cause Mortality: An Observational Study of 17,345 Persons". *PLOS Medicine* 11.2 (2014), lk. 1–12. DOI: 10.1371/journal.pmed.1001606.

[22] HPC Public Documentation. *Sensitive data analysis platform*. URL: https://docs.hpc.ut.ee/public/services/SAPU/ (vaadatud 14.05.2025).

[23] R. Colindres. *Repository of Exposomics Analytics Practicals*. 2024. URL: https://github.com/Chateau-Chadeau/Exposomics_Analytics (vaadatud 02.03.2025).

[24] Z. Yu *et al.* "External exposome and incident asthma across the life course in 14 European cohorts: the EXPANSE project".

[25] United States Environmental Protection Agency. *Learn About Heat Island Effects*. 2025. URL: https://www.epa.gov/heatislands/learn-about-heat-island-effects (vaadatud 08.05.2025).

[26] Keskkonnaagentuur. *Aastakokkuvõtted*. URL: https://www.ilmateenistus.ee/kliima/aastakokkuvotted/ (vaadatud 06.05.2025).

[27] Z. Sun. "Aging, Arterial Stiffness, and Hypertension". *Hypertension* 65.2 (2015), lk. 252–256. DOI: 10.1161/HYPERTENSIONAHA.114.03617.

[28] B. Leng, Y. Jin, G. Li, L. Chen ja N. Jin. "Socioeconomic status and hypertension: a meta-analysis". *Journal of Hypertension* 33 (2 2015). DOI: 10.1097/HJH.0000000000000428.

[29] *Description of Environmental variables available through Exposome Maps*. Versioon 1.0. 2023. URL: https://surfdrive.surf.nl/files/index.php/s/uqUORDrd428H2F9 (vaadatud 01.04.2025).

[30] Copernicus Land Monitoring Service. *CORINE Land Cover*. URL: https://land.copernicus.eu/en/products/corine-land-cover?tab=overview (vaadatud 12.04.2025).

Lisad

Lisa 1: Eksposoomitunnuste kirjeldus

Tabel 7: Eksposoomitunnuste kirjeldus 1

Tunnuse lühend	Kirjeldus	Mõõtmised	Puuduvaid väärtusi $n = 30764$
OZO_AAV	osooni kontsentratsiooni ($\mu g/m^3$) aastane keskmine	iga aasta $2000 - 2019$	6,72%
$NO2\_AAV$	lämmastikdioksiidi kontsentratsiooni ($\mu g/m^3$) aastane keskmine	iga aasta $2000 - 2019$	6,72%
$P10\_AAV$	Väikeste osakeste (läbimõõduga kuni $10\mu gm$) kontsentratsiooni ($\mu g/m^3$) aastane keskmine	iga aasta $2000 - 2019$	6,72%
$P25\_AAV$	väikeste osakeste (läbimõõduga kuni $2{,}5\mu gm$) kontsentratsiooni ($\mu g/m^3$) aastane keskmine	iga aasta $2000 - 2019$	6,72%
$UR\_B15\_$	kergliiklusteede ja populatsioonitiheduse abil arvutatud linnasuse indeks	2015	6,71%
LAN_B03, LAN_B05, LAN_B10	valguse intensiivsus öösel (RAD)	iga 5 aasta tagant alates 2000	18,59%
IMP_B03, IMP_B05, IMP_B10	läbitungimatu pinna (nt asfaldi) protsent 300m, 500m ja 1000m raadiuses	iga 3 aasta tagant $2006 - 2018$	6,77%
WAL_B03, WAL_B05, WAL_B10	kergliiklusteid ilmestav tunnust	2020	6,96%
GSU_DIS	kaugus lähima rohealani (Urban Atlase järgi) (arvutatud vaid linnade jaoks)	2006, 2012, 2018	54,96%
GSC_DIS	kaugus lähima rohealani (Corine Land Coveri) järgi)	2000, 2006, 2012, 2018	6,71%

Tabel 8: Eksposoomitunnuste kirjeldus 2

Tunnuse lühend	Kirjeldus	Mõõtmised	Puuduvaid väärtusi $n = 30764$
MVI_MD1, MVI_MD3, MVI_MD5	ümbritseva roheluse indeksi mediaan 300m, 500m ja 1000m raadiuses	iga 5 aasta tagant alates 2000	6,88%
MVI_ME1, MVI_ME3, MVI_ME5	ümbritseva roheluse indeksi keskmine 300m, 500m ja 1000m raadiuses	iga 5 aasta tagant alates 2000	6,88%
MVI_ST1, MVI_ST3, MVI_ST5	ümbritseva roheluse indeksi standardhälve 300m, 500m ja 1000m raadiuses	iga 5 aasta tagant alates 2000	6,89%
NDV_MD1, NDV_MD3, NDV_MD5	ümbritseva roheluse indeksi mediaan 300m, 500m ja 1000m raadiuses	iga 5 aasta tagant alates 2000	6,88%
NDV_ME1, NDV_ME3, NDV_ME5	ümbritseva roheluse indeksi keskmine 300m, 500m ja 1000m raadiuses	iga 5 aasta tagant alates 2000	6,88%
NDV_ST1, NDV_ST3, NDV_ST5	ümbritseva roheluse indeksi standardhälve 300m, 500m ja 1000m raadiuses	iga 5 aasta tagant alates 2000	6,89%
BSW_DIS	kaugus lähima veekoguni meetrites	2013	6,71%
BSS_DIS	kaugus mereni meetrites	2013	6,71%
BSI_DIS	kaugus lähima siseveekoguni meetrites	2013	6,71%

Eksposoomitunnuste kirjeldused ja mõõtmisajad on saadud [29].

Tabel 9: Temperatuuritunnuste kirjeldus

Tunnuse lühend	Kirjeldus	Mõõtmised	Puuduvaid väärtusi $n = 30764$
TMP_MIN	aasta madalaim temperatuur	iga aasta	6,71%
TMP_MAX	aasta kõrgeim temperatuur	iga aasta	6,71%
TMP_AVG	aasta keskmine temperatuur	iga aasta	6,71%
TMP_MED	aasta temperatuuride mediaan	iga aasta	6,71%
TMP_STD	aasta temperatuuride standardhälve	iga aasta	6,71%
TMP_AVW	aasta sooja hooaja keskmine temperatuur	iga aasta	6,71%
TMP_AVC	aasta külma hooaja keskmine temperatuur	iga aasta	6,71%
TMP_STW	aasta sooja hooaja temperatuuride standardhälve	iga aasta	6,71%
TMP_STC	aasta külma hooaja temperatuuride standardhälve	iga aasta	6,71%

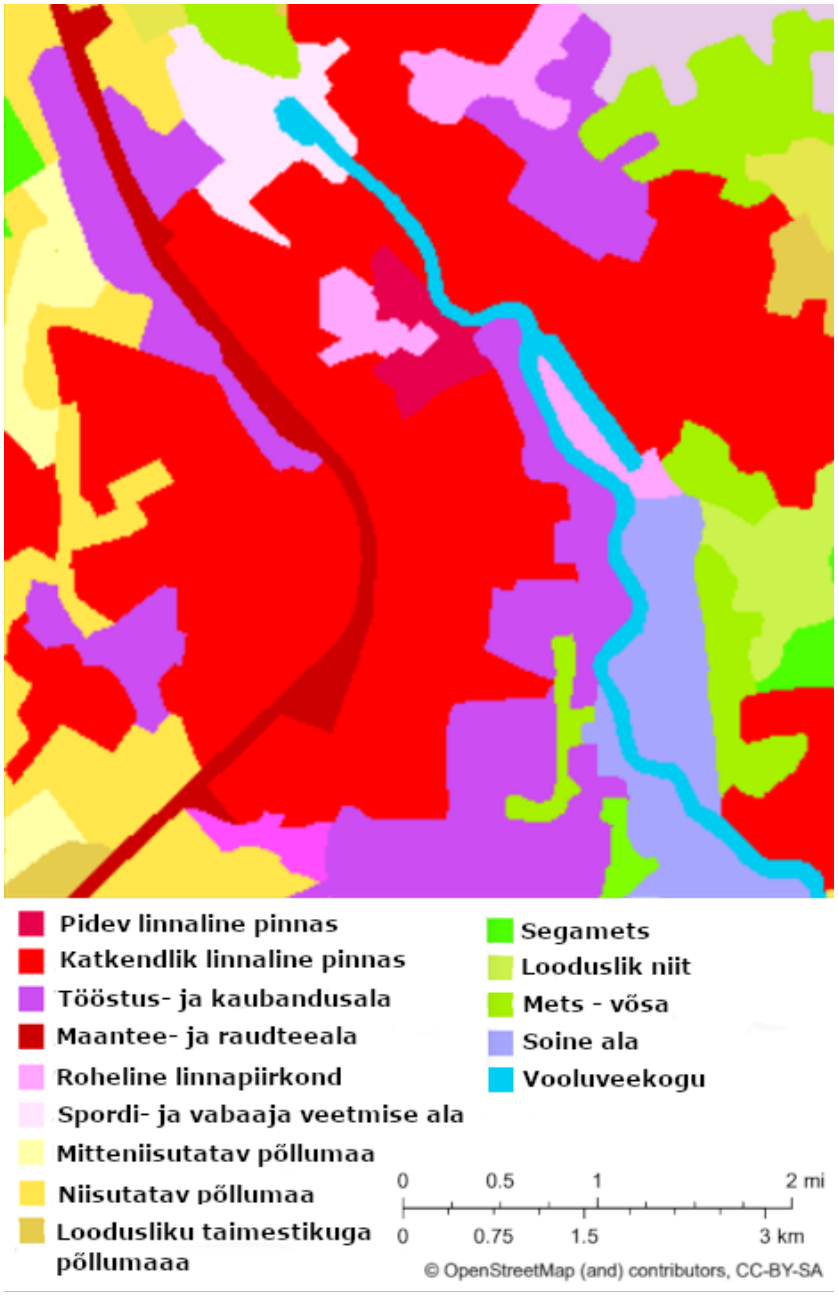
Lisa 2: Tausttunnuste kirjeldus

Tabel 10: Tausttunnuste kirjeldus

Tunnuse lühend	Kirjeldus	Puuduvaid väärtusi $n = 30764$
gender_name	sugu	0%
smoking	suitsetamise staatus (kunagi, endine, praegune, teadmata)	21,4%
bmi	kehamassiindeks	21,01%
education_class	haridustase (põhiharidus, keskharidus, kõrgharidus)	20,82%
result	kirjeldas haigestumist (1 - haigestus, 0 - ei haigestunud)	0%
earliestDate_dgn	diagnoosi saamise kuupäev (kui haigestuti)	
deathDate	surmakuupäev	
birthYear	sünniaasta	0%
liitumis_kpv	liitumiskuupäev	0%

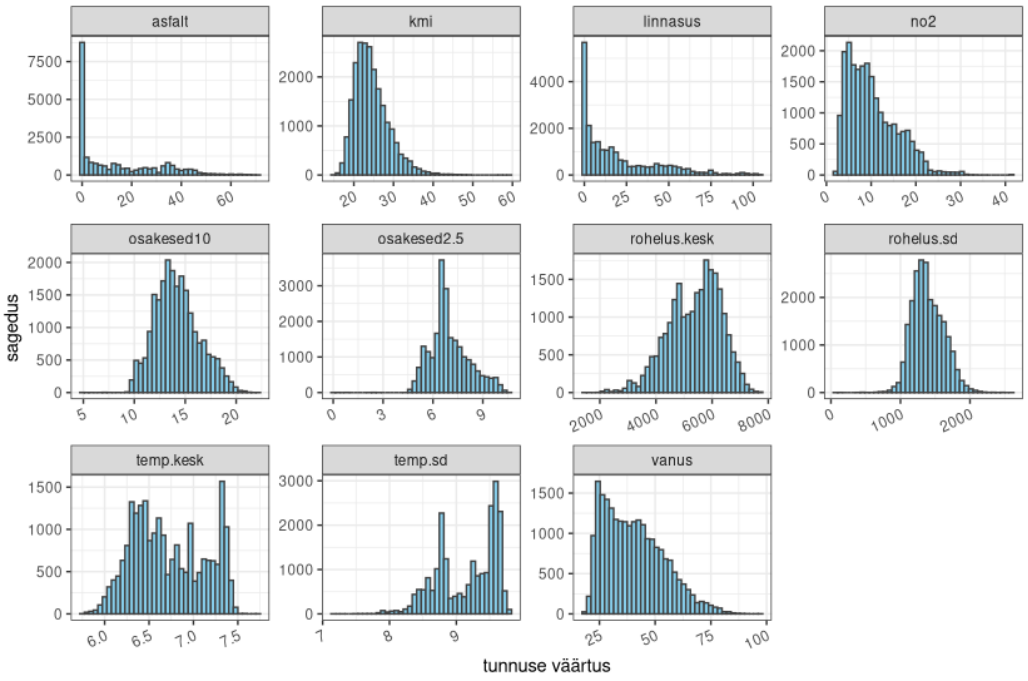
Lisa 3: Corine'i alade kaart

Illustreeriv pilt on koostatud [30] kaartide põhjal.



Joonis 12: Tartu Corine'i alasid illustreeriv kaart

Lisa 4: Pidevate tunnuste histogrammid



Joonis 13: Pidevate tunnuste histogrammid valimis MUSTAND

Lisa 5: Coxi mudel kovariaatidega p-väärtused

Tabel 11: Coxi mudelite tunnuste p-väärtused

Tunnus	1. mudel	2. mudel	3. mudel	4. mudel	5. mudel
$sugu$	0,00103	$3{,}35 \cdot 10^{-13}$	$5{,}23 \cdot 10^{-11}$	$1{,}85 \cdot 10^{-8}$	$1{,}87 \cdot 10^{-5}$
$haridus$	$< 2 \cdot 10^{-16}$	$< 2 \cdot 10^{-16}$	$< 2 \cdot 10^{-16}$		
$suits$	$2{,}09 \cdot 10^{-7}$	$< 2 \cdot 10^{-16}$	$4{,}14 \cdot 10^{-16}$	$3{,}73 \cdot 10^{-11}$	
$scale(kmi)$	$< 2 \cdot 10^{-16}$	$< 2 \cdot 10^{-16}$			
$scale(vanus)$	$< 2 \cdot 10^{-16}$				

Lisa 6: Coxi mudel eksposoomitunnustega p-väärtused

Tabel 12: Coxi mudelite tunnuste p-väärtused

Tunnus	1. mudel	2. mudel	3. mudel	4. mudel	5. mudel	6. mudel	7. mudel	8. mudel	9. mudel
$temp.kesk$	0.17	0.3583	0.0060	0.4121	0.04899	0.2933	$6.45e-5$		
$temp.sd$	0.102	0.3730	0.8784	0.0891	0.8377	0.0250			
$no2$	0.541	0.5928	0.9426	$8.45e-5$					
$osakesed10$	0.0171								
$osakesed2.5$	0.29	0.1591	0.1956	0.2518	0.0150				
$asfalt$	0.293	0.8162	0.000157						
$rohelus.kesk$	0.184	0.01532							
$rohelus.sd$	0.482	0.0638	0.2071	0.0636	0.0810	0.0758	0.0756	0.0699	0.0667
$linnasus$	0.652	0.3249	0.2928	0.0087	0.1951	0.2638	0.3396	0.0266	

Lisa 7: Tunnuste keskmised±standardhälbed kahte klastrisse jaotamisel

tunnus	klaster 1	klaster 2
$temp.kesk$	6.53 ± 0.329	7.06 ± 0.29
$temp.sd$	9.22 ± 0.435	9.11 ± 0.436
$no2$	7.16 ± 2.81	16.1 ± 4.43
$osakesed10$	13.5 ± 1.72	15.7 ± 2.26
$osakesed2.5$	6.63 ± 0.92	7.85 ± 1.24
$asfalt$	3.74 ± 6.35	32.6 ± 12.2
$linnasus$	6.79 ± 8.39	43.2 ± 21.2
$rohelus.kesk$	5873 ± 618	4432 ± 629
$rohelus.sd$	1454 ± 235	1278 ± 143

Tabel 13: Tunnuste keskmised±standardhälbed kahte klastrisse jaotamisel

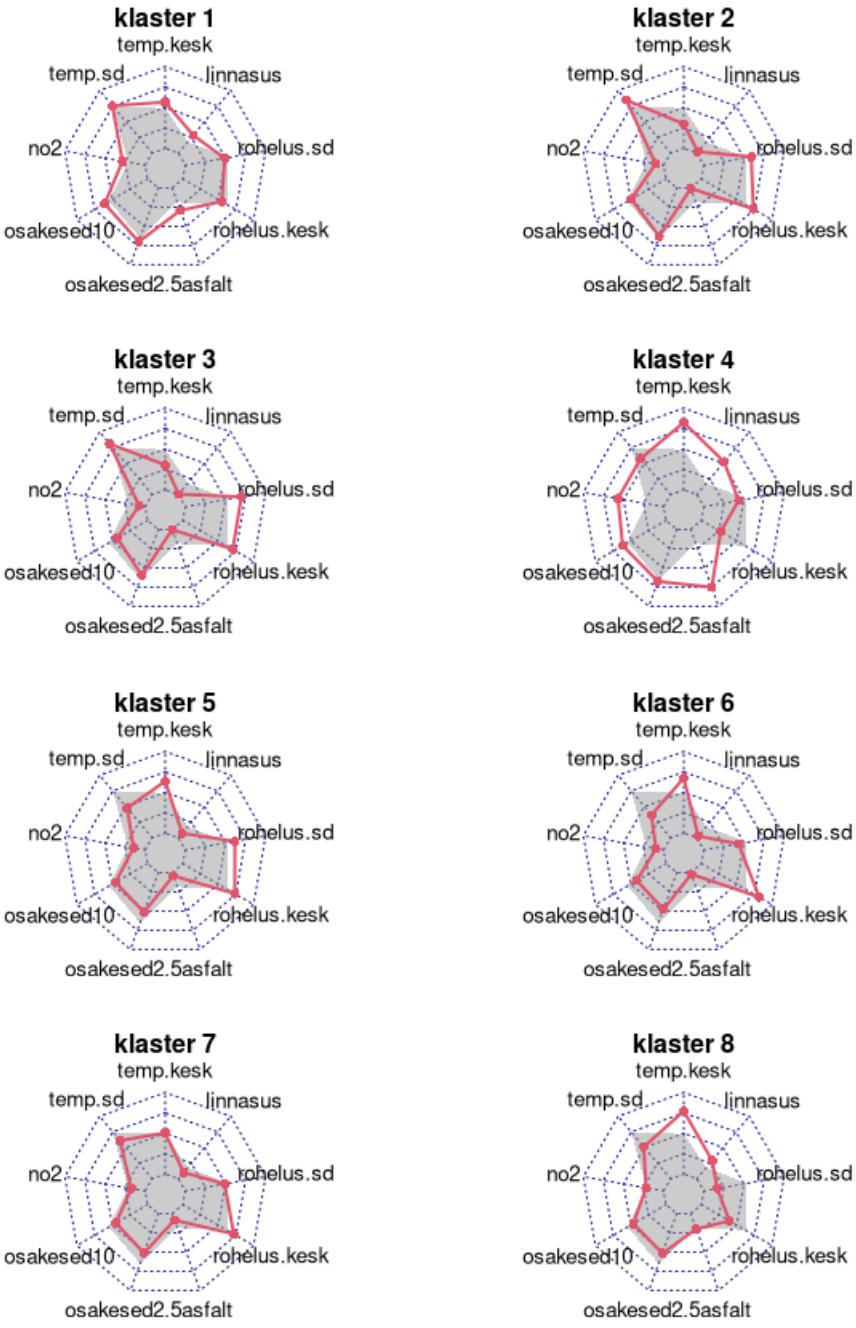
Lisa 8: Tunnuste keskmised±standardhälbed kaheksasse klastrisse jaotamisel

tunnus	klaster 1	klaster 2	klaster 3
temp.kesk	6.85 ± 0.36	6.32 ± 0.18	6.34 ± 0.30
temp.sd	9.20 ± 0.42	9.45 ± 0.21	9.34 ± 0.33
*no*2	12.81 ± 4.45	5.60 ± 1.71	4.64 ± 1.18
*osakesed*10	15.29 ± 2.07	13.28 ± 1.44	12.17 ± 1.25
*osakesed*2.5	7.55 ± 1.19	6.82 ± 0.72	6.45 ± 0.65
asfalt	21.04 ± 14.7	0.45 ± 1.39	0.13 ± 1.07
linnasus	29.63 ± 22.8	2.16 ± 2.95	0.64 ± 1.37
rohelus.kesk	4947 ± 738	6102 ± 457	5916 ± 486
rohelus.sd	1285 ± 152	1537 ± 123	1812 ± 135

tunnus	klaster 5	klaster 7
temp.kesk	7.00 ± 0.18	6.73 ± 0.19
temp.sd	8.45 ± 0.22	8.81 ± 0.22
*no*2	7.37 ± 3.05	8.45 ± 1.85
*osakesed*10	12.62 ± 1.52	12.51 ± 1.26
*osakesed*2.5	5.59 ± 0.45	5.57 ± 0.41
asfalt	2.99 ± 4.57	6.06 ± 6.76
linnasus	7.51 ± 7.72	10.54 ± 7.46
rohelus.kesk	6097 ± 534	6007 ± 538
rohelus.sd	1606 ± 151	1284 ± 136

Tabel 14: Tunnuste keskmised±standardhälbed kaheksasse klastrisse jaotamisel

Lisa 9: Kaheksa klastri radardiagrammid



Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kaisa-Siret Hint,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Seosed eksposoomi ja kõrgvererõhktõve avaldumise vahel", mille juhendajad on Krista Fischer ja Jaanika Kronberg, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaaloman-di ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kaisa-Siret Hint
15.05.2025