Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Gyvenamų patalpų oro taršos rodiklių daugiamatė analizė

Baigiamasis magistro projektas

Gintarė Zuzevičiūtė
Projekto autorė

Doc. Dr. Tomas Ruzgas
Vadovas

Kaunas, 2018

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Gyvenamų patalpų oro taršos rodiklių daugiamatė analizė

Baigiamasis magistro projektas

Taikomoji matematika (621G10003)

Gintarė Zuzevičiūtė
Projekto autorė

Doc. Dr. Tomas Ruzgas
Vadovas

Lekt. Dr. Mindaugas Kavaliauskas
Recenzentas

Kaunas, 2018

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Gintarė Zuzevičiūtė

Gyvenamų patalpų oro taršos rodiklių daugiamatė analizė

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Gintarės Zuzevičiūtės, baigiamasis projektas tema „Gyvenamų patalpų oro taršos rodiklių daugiamatė analizė“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

_____	_____
(vardą ir pavardę įrašyti ranka)	(parašas)

Turinys

Paveikslų sąrašas

Lentelių sąrašas

Zuzevičiūtė, Gintarė. Gyvenamų patalpų oro taršos rodiklių daugiamatė analizė. Magistro baigiamasis projektas / vadovas doc. dr. Tomas Ruzgas; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.
Studijų kryptis ir sritis: Fiziniai mokslai, matematika.
Reikšminiai žodžiai: duomenų tyryba, klasifikavimas, klasterizavimas, klasterinė analizė, patalpų oro tarša, statistinė analizė.
Kaunas, 2018. 110p.

Santrauka

Patalpų oro tarša yra viena svarbiausių aplinkos sveikatos problemų, nuo kurios priklauso žmogaus savijauta ir gyvenimo kokybė. Taršos šaltiniais gali būti įvarūs, pavyzdžiui statybinės medžiagos, dujinė, baldai ar net grindys. Šie kiekvieno namuose sutinkami taršos šaltiniai, atrodantys nepavojingi, skleidžia toksiškas dujas ir lakiuosius organinius junginius tokius kaip: formaldehidas, anglies dioksidas, toluenas, benzenas, toluenas, etilbenzenas, azoto dioksidas ar ksilenas. Radono dujos yra pačios gamtos keliamas pavojus, galintis sukelti vėžį. Šiame darbe apibrėžiama didelių duomenų sąvoka, aptariamos ir palygintos technologijos, apžvelgiami didelių duomenų tyrybos metodai naudojami analizei. Naudojant SAS programinę įrangą atlikta duomenų apie patalpų oro taršą Lietuvoje ir Suomijoje analizė. Hierarchiniais ir nehierarchiniais metodais nustatytas optimalus klasterių skaičius, tiriamos imties duomenys sugrupuoti į klasterius. Nustatyta, kad egiztuoja priklausomybė tarp butuose išmatuotų cheminių junginių koncentracijų ir butus aprašančių faktorių.

Zuzevičiūtė, Gintarė. In-door air polution multivariate analysis. Master's thesis in applied mathematics / supervisor assoc. doc. dr. Tomas Ruzgas; The Faculty of Mathematics and Natural Science, Kaunas University of Technology.
Study field and area (study field group): natural sciences, mathematics
Keywords: data mining, classification, cluster analysis, in-door air polution, multivariate analysis.
Kaunas, 2018. 110p.

Summary

Indoor air polution is one of the most important environmental health problem that affects people's well-being and quality of life. There are many sources of indoor pollution: building materials, furniture, gas stoves and eaven floors. These pollution sources can be found in everyone's house. At first sight they are not hazardous but they emit such toxic gases as carbon dioxide, benzene, toluene, ethylbenzene, xylene, nitrogen dioxide and formaldehyde. One of the potential dangers of nature - the radon gas that can cause cancer. In this paper defines the concept of big data, discusses and compares technologies, reviews the methods of big data research used for analysis. An analysis of indoor air pollution in Lithuania and Finland was carried out by using SAS software. Optimal number of clusters is determined by using clustering criteria. Hierarchical and partitive methods are used to group sample data into clusters. It was found that concentration of chemical compounds depends on describing factors of apartamens.

Santrumpos

BI - Pažangios duomenų analitikos

CCC - Šarlio kubinis klasterizavimo kriterijus.

CO_2 – Anglies dioksidas

CRISP-DM – Duomenų tyrybos proceso pavadinimas

DDM - Paskirstytoji duomenų tyryba

DM - Duomenų tyryba

ES - Europos Sąjunga

HDFS – Programinės įrangos Hadoop paskirstytųjų failų sistema

IBM – Tarptautinė verslo mechanizmo korporacija

KDD - Žinių radimas duomenų bazėse MS – Microsoft paketas

NO_2 – Azoto dioksidas

$PAST^2$ – Pseudo T kvadrato kriterijus

PSF – Pseudo F kriterijus

PB - Petabaitas

SAS – Statistinės analitikos sistema

SEMMA – Prognozavimo procesas

SPSS – Statistinės analizės programinis paketas

Įvadas

Oro kokybė yra viena svarbiausių aplinkos sveikatos problemų, nuo kurios priklauso žmogaus savijauta ir gyvenimo kokybė. Daugelis mūsų didžiąją dalį dienos praleidžiame uždarose patalpose taigi oro kokybė esanti patalpų viduje gali tiesiogiai įtakoti mūsų sveikatos būklę.

Darbo aktualumas – nacionalinės visuomenės sveikatos priežiūros laboratorijos specialistai sako, kad nekreipiant dėmesio į oro kokybę patalpose galime susidurti su erzinančiais sveikatos sutrikimais, tokiais kaip galvos, akių, nosies skausmai, gerklės dirginimas, sauso kosulio kamavimas, sloga, odos išsausėjimas ir niežėjimas, pykinimas, taip pat miego sutrikimas, nuovargis, negebėjimas susikoncentruoti, sutelkti dėmesio. Oro užterštumas gali būtų ir tokių ligų kaip plaučių vėžys pasėkmė.

Darbo tikslas - apžvelgus didelių duomenų tyrybos metodus, skirtus duomenims analizuoti, atlikti klasterinę duomenų analizę.

Tiriamajai daliai naudojami duomenys apie 96 butų patalpų oro taršą. Šie butai yra išsidėstę per 20 daugiabučių namų Lietuvos miestuose. Butuose matuotos dujinių oro teršalų koncentracijos ore, siekiant identifikuoti taršos šaltinius, į kiekvieno buto aprašymą įtraukti butus charakterizuojantys buitiniai faktoriai. Tokia pati analizė buvo atlikta ir su patalpų oro taršos duomenimis daugiabučiuose Suomijoje. Analizė atlikta naudojant SAS (angl. *Statistical Analysis System*) programinį paketą, kuris turi plačias statistinės analizės galimybes, yra patikimas, palankiai vertinamas naudotojų.

Baigiamojo **darbo problema** – identifikuoti taršos šaltinius, t.y. nustatyti kokie butuose esantys daiktai ar veiksniai sukelia didžiausią oro taršą. Žinant taršos šaltinius, atitinkamas gyventojų elgesys gali sumažinti patalpų oro taršą.

Nagrinėjamas **objekta**s – gyvenamų patalpų oro tarša. Šiame magistro baigiamajame darbe analizuojamos cheminės medžiagos ir faktoriai, kurie lemia oro, esančio patalpose, kokybę.

Pagrindiniai uždaviniai:

- nustatyti priklausomybių buvimą ar nebuvimą tarp butuose išmatuotų cheminių junginių koncentracijų ir butus aprašančių faktorių;
- nustatyti optimalų klasterių skaičių;
- taikant klasterinės analizės metodus sugrupuoti tiriamus duomenis į klasterius.

1. Analitinė dalis

Pirmajame skyriuje bus aprašoma patalpų oro taršos problematika ir apžvelgta tyrimui naudojama programinė įranga.

1.1. Patalpų oro tarša

Oro kokybė yra viena svarbiausių aplinkos sveikatos problemų visame pasaulyje, nuo kurios priklauso žmogaus sveikatos būklė, darbo produktyvumas, bendra savijauta, nuotaika ir gyvenimo kokybė. Didžiąją dalį dienos žmonės įprastai praleidžia uždarose patalpose, tokiose kaip namai, biuras, laisvalaikio leidimo centrai, transportas ar mokymo įstaigos. Tyrimai rodo, kad laikas praleistas patalpose sudaro nuo 65% iki 90% paros [1]. Aplinkos apsaugos agentūra nustatė, kad oras patalpose gali būti nuo dviejų iki penkių kartų labiau užterštas už orą lauke [2]. Žmonės privalo įvertinti ir stengtis sumažinti vidaus patalpų oro taršą, kadangi ji pavojinga ne tik dabar gyvenantiems žmonėms, bet ir ateities kartoms.

Ilgai būnant užterštose, nevėdinamose patalpose sveikatos sutrikimai gali pasireikšti gana greitai. Nekreipiant dėmesio į oro kokybę patalpose galime susidurti su erzinančiais sveikatos sutrikimais, tokiais kaip galvos, akių, nosies skausmai, gerklės dirginimas, sausas kosulys, sloga, odos išsausėjimas ir niežėjimas, pykinimas, taip pat miego sutrikimas, nuovargis, negebėjimas susikoncentruoti, sutelkti dėmesio. Neretai šie simptomai dingsta žmogui išėjus iš užterštos patalpos [3]. Pasaulio sveikatos organizacijos (PSO) duomenimis, namų ūkio oro taršos poveikis gali būti net tokių rimtų ligų kaip virškinamojo trakto sutrikimai, išeiminės širdies ligos tuberkuliozė, ūminių ir lėtinių kvėpavimo takų sutrikimai, insultas, katarakta, gimdos kaklelio, plaučių ir kitų vėžio formų priežastis. Pasak PSO per metus miršta 4,3 milijono žmonių būtent nuo patalpų oro taršos poveikio [4].

1.2. Gyvenamosiose patalpose randami oro teršalai, jų šaltiniai, poveikis sveikatai

Cheminė medžiaga arba kitaip teršalas – medžiaga ar jų mišinys, kurie dėl žmogaus veiklos patenka į aplinką ir kenkia aplinkai, joje esantiems žmonėms ar turtui [5]. Cheminių medžiagų koncentracijos normas, esančias gyvenamosse patalpose, nustato Lietuvos higienos norma HN 23:2007. Cheminiai organiniai junginiai yra vieni iš didžiausių patalpų oro teršalų. Tokie teršalai į patalpas paprastai patenka iš patalpose esančių medžiagų arba kartu su oru iš lauko. Toliau pateikiamos dažniausiai patalpose aptinkamas cheminės medžiagos ir galimas jų poveikis žmogaus sveikatai.

Anglies dioksidas CO_2 yra bekvapės ir bespalvės bei paprastai nenuodingos dujos, tačiau jos yra sunkesnės už orą, todėl išstumia deguonį ir taip įtakoja žmonių kvėpavimą. Anglies dioksidą išskiria

žibalo ir dujiniai šildytuvai, dujiniai prietaisai, židiniai, nesandarūs kaminai, tabako dūmai, automobilių išmetamosios dujos. Anglies dioksido koncentracija priklauso nuo vėdinimo tipo, gyventojų skaičiaus ir paros laiko [6]. Šis cheminis junginys dirgina akių, nosies ir gerklės gleivines, gali sutrikdyti plaučių funkciją, pasunkinti kvėpavimo takų ligas, sukelti bronchitą, plaučių vėžį, galvos skausmus, kosulį, gerklės uždegimą, į gripą panašius simptomus.

Gyvenamosiose patalpose labiausiai paplitę lakieji organiniai junginiai yra **benzeno, tolueno** ir **etilbenzeno** dujos. Jie gali būti lėtai skleidžiami iš tabako dūmų, baldų, dažų, tirpiklių ar grindų dangos. Šaltiniai taip pat gali būti medienos konservantai, aerozoliniai purškalai, valymo ir dezinfekavimo priemonės, kandys, repelentai, oro gaivikliai, sauso valymo drabužiai. Šie junginiai gali sukelti astmą, akių, nosies ir gerklės dirginimą, galvos skausmą, koordinacijos netekimą, pykinimą. Žaloja inkstus ir centrinę nervų sistemą. Gali sukelti vėžį gyvūnams ir žmonėms.

Formaldehidas – bespalvės ir aštraus, dirginančio kvapo dujos, kurių šaltiniais gali būti kiliminė danga, presuoti medienos gaminiai, tokie kaip kietmedis, faneruotos sienų plytelės, medžio drožlių plokštės, medienos plaušų plokštės, naudojami pastatuose ir balduose, karbamido - formaldehido putų izoliacija, klijai, tekstilė, transporto priemonių išmetamosios dujos, krosnys, židiniai, tabako dūmai, dezinfekuojančios priemonės. Vidaus oro taršos lygis priklauso nuo temperatūros, drėgmės ir vėdinimo lygio [6]. Formaldehido koncentracijos taip pat gali svyruoti priklausomai nuo dienos ir metų laiko. Šios dujos pažeidžia akių, nosies ir gerklės gleivinę, sukelia kosulį, gali būti bėrimo, alerginės reakcijos ir net vėžio pasekmė. Didesni jo kiekiai gali pažeisti žmonių DNR ir turėti įtakos mūsų palikuonių sveikatai ir reprodukcinėms savybėms. Taip pat galima mirtis.

Azoto dioksido NO_2 dirginama akių, nosies, gerklės ir viršutinių kvėpavimo takų gleivinės, gali sukelti dusulį, kvepavimo takų infekciją. Šis cheminis junginys gali sutrikdyti plaučių funkciją, sukelti galvos skausmus, svaigulį ir pykinimą, bronchitą, plaučių vėžį, į gripą panašius simptomus.Sergantieji kvėpavimo ir/ar kraujotakos sistemų lėtinėmis ligomis iškart pajunta sveikatos pablogėjimą. NO^2 yra išskiriamas dujiniais prietaisais šildant patalpas ir net gaminant maistą. Taip pat gali į patalpas patekti iš lauko.

Radonas – tai radioaktyviosios dujos, kurių įkvėpus gali būti pažeisti plaučiai. Gali būti plaučių ar skrandžio vėžio pasekmė. Būtent statybinės medžiagos yra pagrindinis radono šaltinis daugiabučiuose ir daugiaaukščiuose namuose.

Sunkieji metalai sukelia galvos skausmą, padidėjusį prakaitavimą, dirgina burnos ertmę, kenkia inkstams, gali būti kūno bėrimo priežastis. Taršos šaltiniai yra dažai, automobilių ir tabako dūmai, dirvožemis ir dulkės.

Mikromicetai (pelėsiniai grybai) turi kancerogeninių savybių, gali įjautrinti žmogaus organizmą, sukelti alergines ligas. Taip pat yra nervų ir imuninės sistemos nusilpimo priežastis. Pelėsinių grybų šaltiniai yra žmonės, naminiai gyvūnai, drėgnas paviršius, drėkintuvai, ventiliacijos sistemos, nutekėjimai, lauko oras ir augalai.

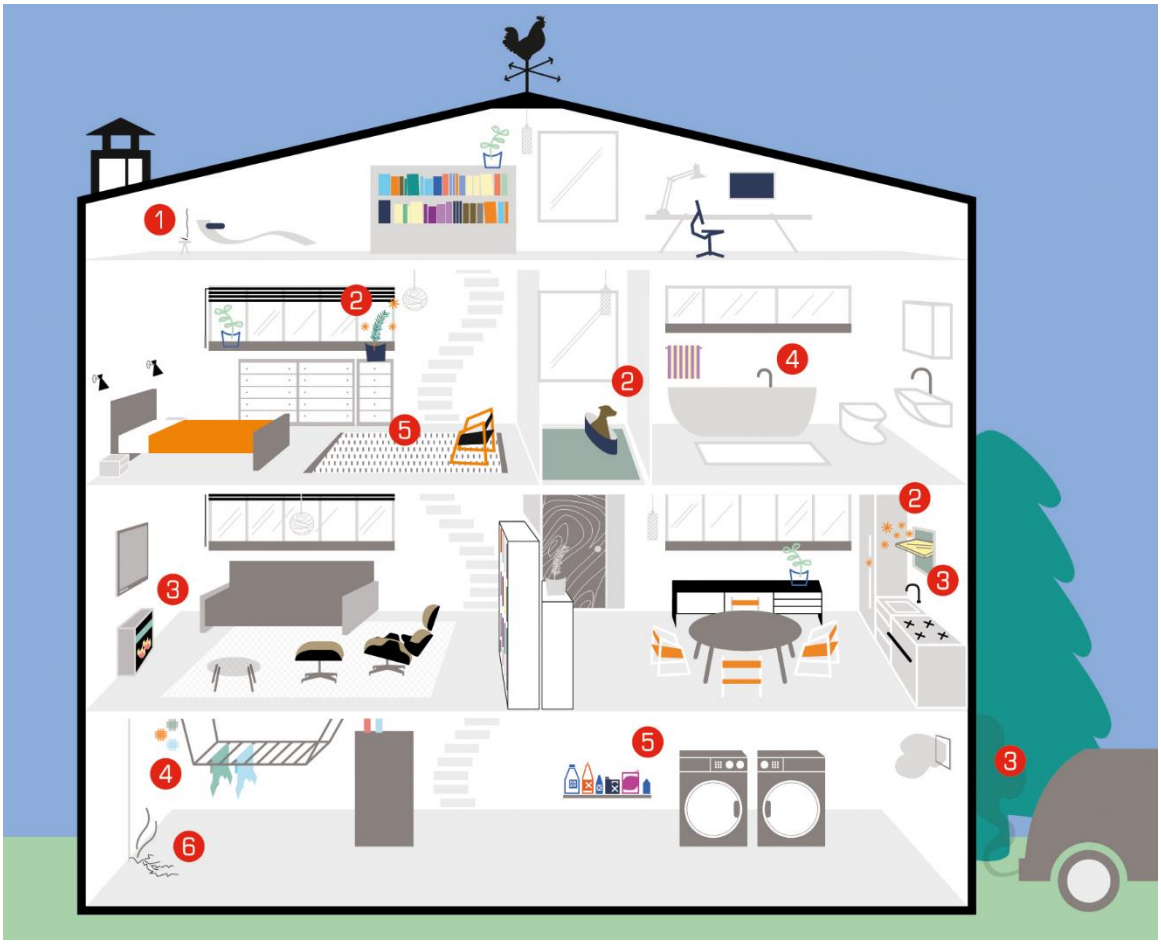
Alergenai (įskaitant žiedadulkes) gali pasunkinti kvėpavimo takų ligas ir sukelti kosulį, spaudimą krūtinės srityje, kvėpavimo sutrikimus, akių dirginimą ir odos bėrimą. Šaltiniai paprastai yra lauko oras, kambariniai augalai.

Labai svarbu pastoviai stebėti patalpų oro užterštumą, ieškoti efektyviausių metodų, kaip būtų galima pagerinti oro kokybę gyvenamojoje aplinkoje. Gyventojai turi atidžiai rinktis medžiagas, baldus, valymo ar kitas buitines priemones ir įvertinti kokį poveikį, patalpų oro taršai, gali turėti jų veikla.

Patalpų oro taršą sumažinti galima:

- Reguliariai plaunant grindis galima surinkti dulkes, kurių nesugeba surinkti dulkių siurblys;
- Palaikant pastovų 40–65 proc. drėgmės lygį namuose galima neleisti veistis erkutėms ir pelėsiams. Tam rekomenduojama naudoti drėgmės rinkiklius ir oro kondicionierius. Oro kondicionierius sumažina ir patalpose esančių žiedadulkių bei kitų alergenų kiekį;
- Reguliariai vėdinant patalpas;
- Vengiant sintetinių oro gaiviklių bei valiklių, kurie į orą išskiria šimtus cheminių medžiagų;
- Nerūkant namuose, kadangi cigarečių dūmuose yra per 4000 įvairių cheminių medžiagų, kurios didina riziką sirgti ausų ir kvėpavimo takų infekcijomis, astma, vėžiu, širdies ligomis;
- Atidžiai renkantis patalpų vidaus apdailos medžiagas.

Oro taršos poveikis sveikatai gali būti naudingas kaip rodiklis, parodantis, kad patalpose egzistuoja oro kokybės problemos (1.1 pav.). Ypač tai pasijunta po sąveikos su pesticidais, atlikus namų remontą, įsigijus naujus baldus, persikėlus į naujus namus. Remiantis Jungtinių Amerikos Valstijų aplinkos apsaugos agentūros duomenimis išskiriami tokie pagrindiniai oro teršalai bei jų šaltiniai ir su jais galimai susiję simptomai.



1.1 pav. Oro taršos šaltiniai namuose

1. Tabako dūmai.
2. Alergenai.
3. Anglies monoksidas ir azoto dioksidas.
4. Drėgmė.
5. Cheminės medžiagos.
6. Radonas.

1.3. Programinė įranga

Šiuolaikiniame pasaulyje sunku būtų atrasti žmogaus veiklos sritį, kurioje nebūtų kaupiami ir analizuojami duomenys. Besivystant naujoms technologijoms, duomenų apimtys sparčiai didėja, kartu auga ir poreikiai analizuoti turimus duomenis [7]. Didelių duomenų (angl. *Big Data*) sąvoka vartojama gana dažnai, tačiau ne visada ji apibrėžiama teisingai [8]. Dideliais duomenimis vadinami tokie duomenų rinkiniai, kuriuos dėl jų dydžio ir sudėtingos struktūros apdoroti paprastomis duomenų apdorojimo programomis ir įrankiais tampa gana sudėtinga ar net neįmanoma [9,10].

Šiandieninėje visuomenėje statistika yra neatsiejama nuo kompiuterinės duomenų analitikos, kuri padeda efektyviai ir greitai spręsti iškilusias problemas ir uždavinius [11]. Išplėstinė kompiuterinė

duomenų analizė gali būti labai informatyvi ir atskleisti tyrimui svarbius aspektus, nepastebimus įprastose užklausose ar ataskaitose. Naudojant specialią programinę įrangą bei kiekybinius matematinės statistikos metodus galima sukurti analitinius modelius bet kokiam uždaviniui spręsti. Vienas svarbiausių aspektų yra tai, kad pažangios programinės įrangos padedami galime daryti prielaidas ar net sprendimus iš esmės akimirksniu, svarbiausia tinkamai pasirinkti įrankių paketą.

Didžioji dalis duomenų tyrybos metodų yra grindžiami matematine statistika. Dažnai vienu metu sprendžiami klasifikavimo, klasterizavimo ir prognozavimo uždaviniai, siekiant gauti kuo daugiau žinių apie analizuojamus duomenis. Šiems uždaviniams spręsti naudojamos programos skirstomos į grupes:

- duomenų analizės programos (*Pascal, C++* ir kitos);
- universalios matematinių uždavinių sprendimo sistemos (*MathCad, MatLab* ir kitos);
- universalios duomenų analizės sistemos (*SAS, SPSS, Statistica* ir kitos);
- ekspertinės duomenų analizės sistemos (*Table Curve, ABP*);
- kitos sistemos (*MS Excel* ir kitos) [11].

Statistinė duomenų analizė atliekama naudojant programinius paketus, kurie leidžia atlikti skaičiavimus, pateikti duomenis ir rezultatus grafiškai. Yra begalė statistinei duomenų analizei tinkamų programų: *MS Excel, Statistika, SAS, Matlab* [12]. Skaičiuojant tik aprašomosios statistikos charakteristikas ir atliekant grafinę analizę, visiškai pakanka *MS Excel* galimybių, tačiau sudėtingesnių tyrimų duomenims apdoroti ir analizuoti dažniausiai taikomos specializuotos programinės įrangos.

Pasaulyje pirmaujanti informacinių technologijų tyrimų ir konsultacijų bendrovė „Gartner“ periodiškai atlieka pažangių analitinių platformų tiekėjų analizę ir skelbia lyderes. Vertinant statistinės duomenų analizės programines įrangas buvo remiamasi apklausos, kurioje dalyvavo 600 įmonių, duomenimis. Siekiant įvertinti funkcionalumą buvo analizuojamos tiriamųjų programų naudojimo instrukcijos. Į sąrašą pateko tokios programinės įrangos kaip *SAP, SAS, IBM, Tableau, Targit* ir kitos [13]. Lydere išrinkta *SAS*, kadangi vartotojai palankiausiai vertina dėl statistinės analizės sistemos nuolatinių naujinimų ir tobulinimų.

Siekiant nustatyti, kuris iš minėtų programinių įrankių tinkamesnis tolesniam tyrimui, atliekamas *Apache Hadoop, IBM, Tableau* ir *SAS* palyginimas.

Apache Hadoop

1.2 pav. *Apache Hadoop* programinės įrangos logotipas

Apache Hadoop – tai atvirojo kodo programinė įranga, skirta didelės ir kintamos apimties duomenų rinkinių apdorojimui kompiuterių klasteriuose, naudojant paprastus programavimo modelius. *Apache Hadoop* apima tris modulius:

- ***HDFS.*** Paskirstytųjų failų sistema, kuri suteikia aukšto našumo prieigą prie duomenų.
- ***MapsReduce.*** Didelių duomenų rinkinių lygiagretaus apdorojimo programinės įrangos programavimo modelis.
- ***YARN.*** Darbo planavimo ir klasterių išteklių valdymo sistema.

Privalumai

- **Aukštas klaidų toleravimas**. Skirtingai nei tradicinėse paskirstytųjų failų sistemose, kurios naudoja duomenų apsaugos mechanizmus, *HDFS* saugo duomenų kopijas keliuose duomenų mazguose ir gali, aptikus klaidą, atkurti duomenis iš kitų duomenų mazgų.
- **Didelės apimties duomenų naudojimas.** *Hadoop* klasteriai gali talpinti PB (Petabaitas) dydžio duomenų rinkinius.

Trūkumai

- **Saugumas.** Programinė įranga nenaudoja šifravimo metodų duomenims saugoti, o tai yra pagrindinis aspektas daugeliui įmonių, norinčių savo duomenis laikyti paslaptyje.
- **Nepritaikyta mažos apimties duomenų analizei.** Dėl savo galimybės talpinti didelio dyžio duomenų rinkinius *Hadoop* stokoja gebėjimo efektyviai palaikyti atsitiktinį mažų rinkmenų skaitymą.
- **Pažeidžiamumas.** Programinis paketas parašytas *Java*, viena plačiausiai pasaulyje naudojama, tačiau prieštaringai vertinama kalba. *Java* buvo eksploatuota kibernetinių nusikaltėlių ir dėl saugumo pažeidimų yra padidintos rizikos programinė įranga.

IBM

1.3 pav. *IBM* programinės įrangos logotipas

IBM yra įsikūrusi Jungtinėse Amerikos Valstijose, Niujorke. Įmonė siūlo statistinės analizės įrankius *SPSS Statistics* ir *SPSS Modeler*, kurie yra gerai žinomi dėl savo prieinamumo įvairių įgūdžių lygių bei tipų vartotojams. *IBM* sprendžia daugybę analizės iššukių susijusių su klientais, operacijomis, materialiuoju turtu bei rizika.

Privalumai

- **Lengvai suprantama.** *IBM* demonstruoja didelį atsidavimą šiai rinkai. Jos produktų galimybės yra gerai suprantamos potencialiam klientui.
- **Didelė rinkos dalis.** Įmonė turi didelę vartotojų bazę bei bendruomenę, kuri padeda jai pasamdyti patyrusius analitikus bei didinti rinkos supratimą apie jos produkciją.
- **Stipri vizija.** Analitikos intergacija į verslo vartotojams patogius įrankius bei *SPSS* įdiegimas leidžiantis naudotis juo debesyse.

Trūkumai

- **Pasitenkinimas**. Nors kai kurie klientai deklaravo aukštą pasitenkinimą *IBM* produktais, bendras pasitenkinimo įvertinimas yra mažesnis nei vidutinis.
- **Prastas valdymas.** Trūksta ar nepakanka dokumentacijos, silpna techninė pagalba, nepakankamas mokymas, prastas atsakas į pastebėtus trūkumus.
- **Integracija.** Sudėtingas įdiegimas ir integracijos tarp IBM siūlomų produktų trūkumas.

Tableau

1.4 pav. *Tableau* programinės įrangos logotipas

Programinė įranga sukurta amerikiečių kompanijos, įsikūrusios Sietle, Vašingtone. Tai pažangi ir sėkminga programinė įranga, lyderė duomenų vizualizavimo erdvėje. *Tableau* lengvai jungia duomenų gausybę ir greitai duomenų rezultatus paverčia grafika. *Tableau* yra puiki pagalba analitikui, padedanti lengvai suprasti duomenis [13].

Privalumai

- **Duomenų vizualizacija.** *Tableau* išsiskiria ir stipriai lenkia savo konkurentus duomenų virtualizacijos galimybėmis. Net 70% klientų naudojasi *Tableau,* nes paprastas valdymas puikiai suderintas su sudėtingomis vizualizacijomis.
- **Integracija.** Galimybė lengvai įkelti duomenis iš įvairiausių šaltinių.
- **Mobiliojo telefono palaikymas.** Vartotojo sąsaja puikiai suderinama su mobiliaisiais įrenginiais.
- **Vartotojų forumai ir klientų aptarnavimas.** *Tableau* turi labai aktyvų vartotojų ir programuotojų forumą, kuriame operatyviai gaunami atsakymai, sprendžiami nesklandumai. Tableau taip pat pasižymi ypač puikiais atsiliepimais apie klientų aptarnavimą.
- **Atnaujinimai.** 90% Tableau vartotojų naudojasi naujausia programos versija, kas parodo, jog įrangos atnaujinimai yra lengvai pasiekiami [13].

Trūkumai

- **Pirminis duomenų paruošimas.** Kompanijos turi turėti stiprius techninius įgūdžius, kad sukurti pradinę struktūrą. Importuoti duomenis iš kitų duomenų bazių sugebės tik patyręs *SQL* naudotojas.
- **Atskirti įrankiai.** Norint atlikti pažangią duomenų analizę, efektyviai valdyti duomenis bei kurti vizualizaciją, reikia pirkti kitus papildomus programinius įrankius, kas, nėra patogu, nes kaip žinia, visada yra pigiau turėti kuo mažiau programinių paketų.
- **Kaina.** Palyginus su kitomis didžiųjų duomenų programomis, Tableau išsiskiria palankia mokesčio sistema. Vartotojams leidžiama pasirinkti iš kelių skirtingų licencijavimosi parinkčių. Jie gali naudoti tik pradinį paketą už žemesnę kaina, o norint papildomų galimybių greitai ir paprastai jas prisijungti. Taip pat yra ir nemokama Tableau versija skirta asmeniniam naudojimui. Tačiau jos galimybės labai ribotos, o pilnas programinis paketas kainuoja 999$ metams. Norint pirkti programinės įrangos serverį įmonės darbuotojams, kaina prasideda nuo 10000$ dešimčiai vartotojų [13].

SAS

1.5 pav. *SAS* programinės įrangos logotipas

SAS buveinė yra Šiaurės Karolinos valstijoje Kerio mieste, JAV. Programinė įranga yra viena iš plačiausias galimybes turinčių duomenų analizės sistemų. Turėdama daugiau nei 40000 klientų ir didžiausią vartotojų bei partnerių ekosistemą ji yra dažniausias pasirinkimas tarp organizacijų, siekiančių pažangios analizės. *SAS* programinis paketas yra stiprus bankininkystės, draudimo, verslo, paslaugų bei valdžios sektoriuose. *SAS* apima naujausius duomenų sandėliavimo, informacijos išgavimo iš duomenų sprendimus, analitinius, integravimo ir ataskaitų ruošimo modulius [11].

Privalumai

- **Programų grupė.** Programinis paketas turi didelį integruotų komponentų rinkinį, kurį sudaro *SAS Visual Analytics, SAS Entreprise Guide, SAS Office Analytics* ir kitos. Šie produktai naudojami duomenų integravimui, duomenų valdymui ir gavybai, bei prognozavimo modeliavimui. Visa programų grupė sukurta skiriant labai daug dėmesio į produkto kokybę ir funkcionalumą.
- **Kalba.** *SAS* paremta nesudėtinga, paprasta ir lengva programavimo kalba, kurios sintaksė yra labai panaši į kitų progromavimo kalbų sintaksę. Sistema taip pat gali būti valdoma funkcijų, komandų, ir operatorių pagalba, kas užtikrina norimo rezultato pasiekimą.
- **Duomenys.** Apdoroja itin didelius duomenų kiekius. Galima dirbti su daugeliu populiarių formatų duomenų failais.
- **Naudojimas.** *SAS* sistema veikia plačiausiai naudojamose operacinėse sistemose, tokiose kaip *Microsoft Windows, Unix, Linux* ir kt.
- **Statistinės galimybės.** *SAS* programa yra vertinama dėl paprastumo naudotis, kuris suderinamas su analizės sudėtingumu ir plačiomis galimybėmis, tokiomis kaip regresijos, laiko eilučių ir Anova metodų naudojimas, vizualizacija, kuri užtikrina duomenų pateikimą lentelių, grafikų ir diagramų pavidalu.
- ***SAS* pagalbos sistema.** Programinė įranga turi integruotą teorijos skyrių, kuriame pateikiama daug informacijos apie *SAS* funkcijas ir galimybes su pavyzdžiais ir patarimais. Taip pat siūlo informatyvesnes mokymo programas, dėl to naudotojui reikia mažiau žinių programavimo srityje.

- **Interaktyvi sąsaja.** Galimybė susikurti interaktyvią vartotojo sąsają sąveikai su duomenimis, duomenų vadybai, analizei ir pristatymui [13].

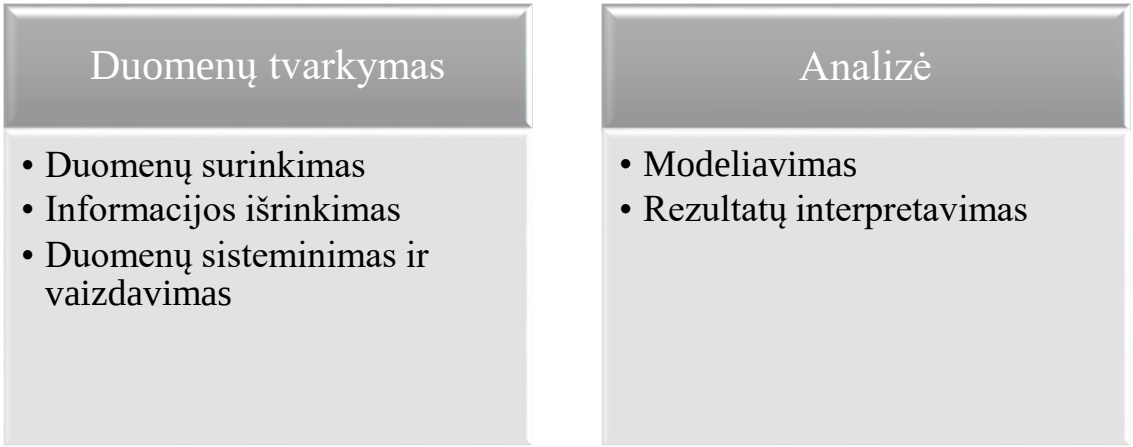
Trūkumai

- **Integracija.** *SAS* produktų pakete pasitaiko produktų su panašiomis galimybėmis (pavyzdžiui prognozavimo modeliavimas). Vartotojai nurodo, kad trūksta integracijos tarp *SAS* siūlomų paketų.
- **Licencija ir kaina.** Pagrindinė priežastis kodėl įmonės pasirenka ne *SAS* yra dėl brangios lincencijos ir sudėtingos licencijavimo struktūros.

Ir vis dėlto visos apžvelgtos įrangos stipriai nusileidžia tokiam programiniam paketui kaip *SAS*. Dėl šių visų minėtų privalumų duomenų analizė bus atliekama naudojantis būtent *SAS* programinės įrangos paketu.

1.4. Duomenų tyrybos procesas

Duomenų analizė – pagrindinis didelių duomenų uždavinys. 1.6 paveikslėlyje pavaizduotas didelių duomenų apdorojimo principas.



1.6 pav. Didelių duomenų apdorojimas

Pirmajame žingsnyje turi būti nuspręsta iš kokių šaltinių bus paimti duomenys. Šiame etape turi būti pasirūpinta duomenų nuasmeninimu ir kitais duomenų tvarkymo teisiniais dalykais. Duomenų analizės etapui turi būti atrinkta reikiama informacija, išgryninta, susieta su galbūt jau turimais duomenimis, surinktais iš kitų šaltinių, ir viskas susisteminta. Skelbiami rezultatai turi būti patikimi, patikrinti bei teisingi.

Pažangus duomenų apdorojimas ir saugojimas skiriasi nuo tradicinių duomenų analizės būdų. Kai pavienių kompiuterių išteklių neužtenka, yra pasitelkiama paskirstytoji duomenų tyryba DDM (angl. *Distributed Data Mining*). Analizuojamus duomenis suskirsčius tam tikrais būdais, duomenų tyrybos uždavinys lygiagrečiai sprendžiamas kompiuterių klasteriuose ar tinkluose.

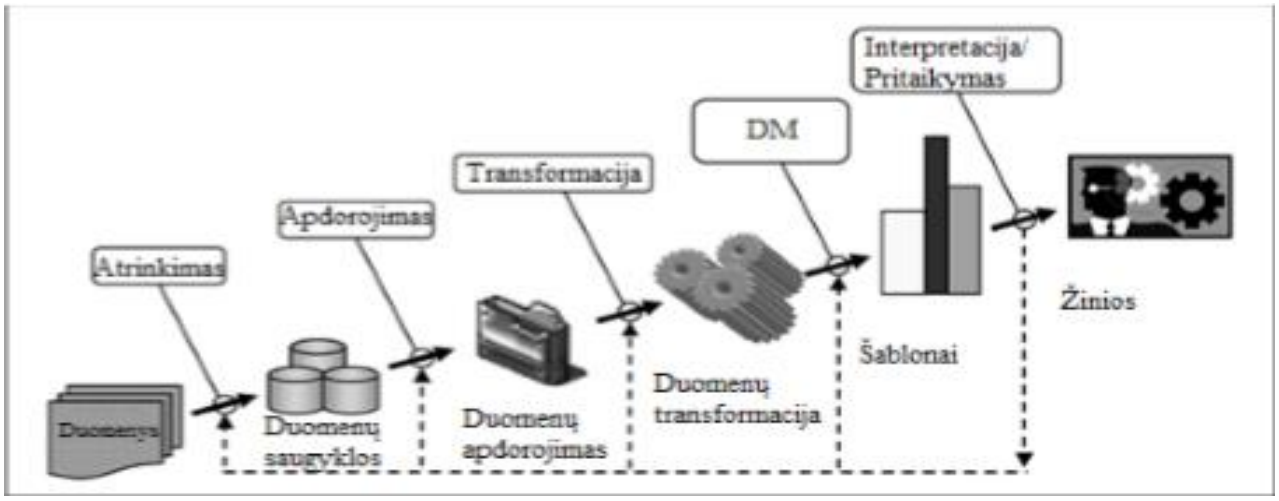
Pagrindinis šių sistemų principas yra didelę užduotį suskaidyti į smulkesnes, žymiai lengviau išsprendžiamas ir nepriklausomas dalis, kurios yra vykdomos lygiagrečiai. Tarpiniai rezultatai sujungiami, ir gaunamas galutinis rezultatas.

Pažangios duomenų analitikos BI (angl. *Business Intelligence*) padeda kaupti ir valdyti sukauptus duomenis, analizuoti didelius informacijos kiekius, iš jų ištraukti naudingą informaciją, pagrindžiant sprendimų priėmimo procesą [14].

Viena iš aktualių BI technologijų yra duomenų tyryba DM (angl. *Data Mining*), kuri yra viena iš žinių radimo duomenų bazėse KDD (angl. *Knowledge Discovery in Databases*) proceso etapų [15]. Tai yra, dideliuose duomenų kiekiuose ar didelėse duomenų saugyklose ieškantis naujos, paslėptos anksčiau nežinomos ir potencialiai naudingos informacijos procesas. Technologija sugeba faktiškus duomenis paversti naudinga informacija ir žiniomis, tinkamomis analizuojant duomenis veiklos valdymui. DM yra ilgas ir sudėtingas procesas. Siekiant pritaikyti DM technologijas efektyviai, kad gauta informacija būtų tiksli, svarbu ištirti ir įvertinti veiklą, procesus ir problemines sritis.

KDD procesas susideda iš penkių etapų (1.7 pav.):

- **Duomenų išrinkimas.** Iš įvairių duomenų šaltinių atrenkami analizuojami duomenys.
- **Pradinis duomenų apdorojimas.** Analizuojami duomenys turi būti išvalyti, išfiltruoti, transponuoti, atrinkti pagal požymius ir normuoti.
- **Duomenų transformavimas.** Duomenys, surinkti iš skirtingų informacijos šaltinių, turi būti pateikiami vienoda tinkama forma.
- **DM.** Duomenų apdorojimui taikomi pasirinkti duomenų tyrybos metodų algoritmai.
- **Interpretacija/Pritaikymas.** Gaunami duomenų analizės rezultatai [16].

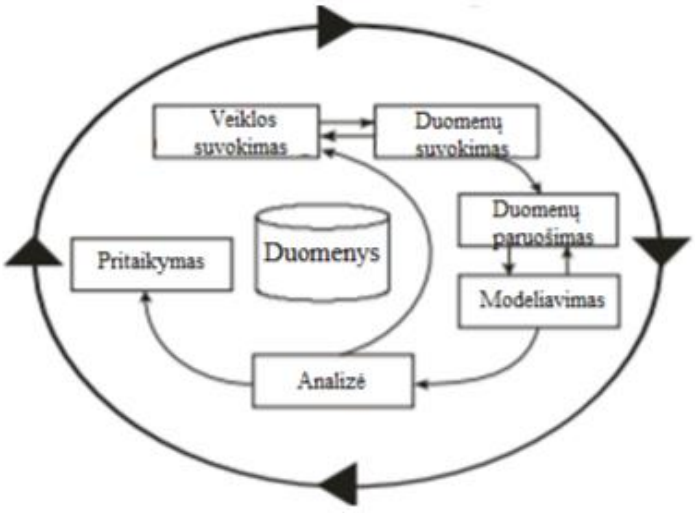


1.7 pav. KDD proceso etapai

DM įrankiai skiriasi savo savybėmis, leidžiamais naudoti metodais bei algoritmais [14]. Apžvelgsime du modelius: CRISP-DM (angl. *Cross-Industry Standart Pricess for Data Mining*) ir SEMMA (angl. *Sample, Explore, Modify, Model, Assess*), kadangi jie yra išrinkti populiariausiais ir labiausiai naudojamais praktikoje.

CRISP-DM

Standartinį duomenų tyrybos proceso modelį CRISP-DM 1996 metais sudarė DM pradininkai – *Daimler Chrysler, SPSS, Teradata* [15]. Modelis buvo kuriamas remiantis ne tik teorija, bet ir praktika. Dėl šios priežasties yra labai paplitęs ir dažnai naudojamas. Jį sudaro šeši nuosekliai išsidėstę etapai (1.8 pav.).



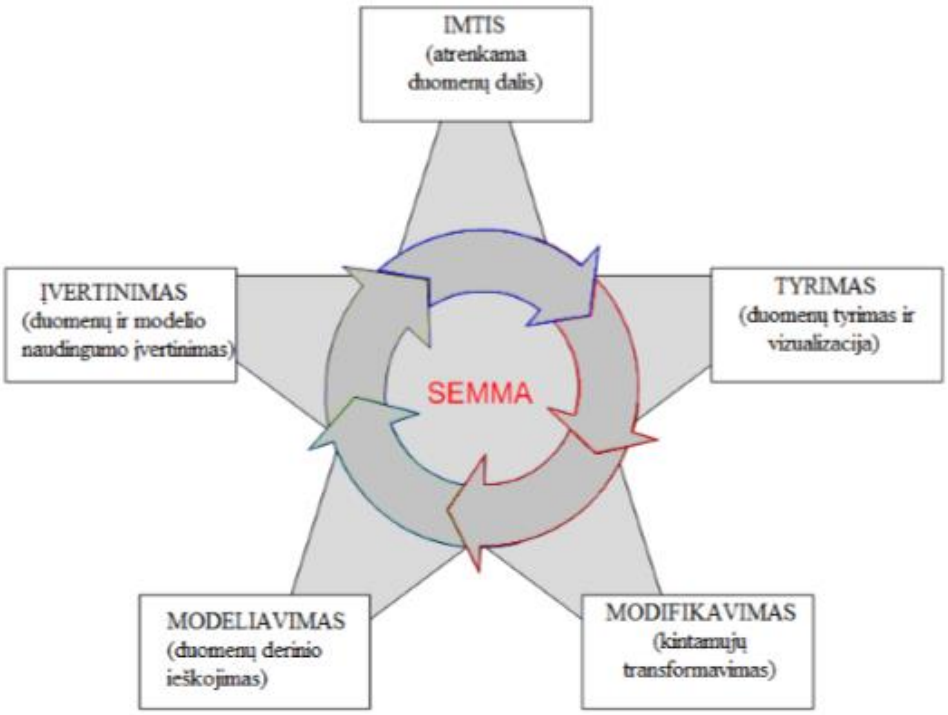
1.8 pav. CRISP-DM proceso modelis

- **Veiklos suvokimas.** Šio etapo dėmesys skiriamas suprasti projekto tikslus ir reikalavimus iš verslo perspektyvos. Tuomet sudaryti preliminarų planą tikslams įgyvendinti ir turimas žinias panaudoti duomenų tyrybos problemai spręsti.

- **Duomenų suvokimas.** Ši fazė apima pradinių duomenų surinkimą ir susipažinimą, duomenų kokybės problemų nustatymą ir hipotezių iš paslėptos informacijos formulavimą.
- **Duomenų paruošimas.** Duomenų ruošimo etapas apima visus veiksmus reikalingus paruošti pradinius duomenis iki galutinės tinkamos jų būsenos. Duomenų paruošimo užduotys gali būti atliekamos pakartotinai ir tik tam tikra tvarka.
- **Modeliavimas.** Šiame etape yra parenkami įvairūs modeliavimo metodai ir taikomi pagal optimalias parametrų reikšmes. Paprastai yra keletas skirtingų būdų gauti duomenis, tačiau kiekvienas turi konkrečius reikalavimus duomenų formatui. Todėl dažnai reikia grįžti į duomenų paruošimo etapą.
- **Analizė.** Prieš galutinio modelio diegimą, svarbu nuodugniai peržvelgti visus pereitus etapus ir patikrinti ar tikslai yra pasiekti. Pabaigai turi būti pristatytas sprendimas kaip naudotis duomenų tyrybos rezultatais.
- **Pritaikymas.** Modelio pritaikymas nėra projekto pabaiga. Net jeigu modelis suteikia daugiau žinių apie duomenis, jis turi būti pristatytas ir paruoštas naudojimui [16].

SEMMA

Procesas buvo sukurtas *SAS* instituto [16]. Pavadinimas susideda iš penkių komponenčių, kurios reiškia skirtingus proceso etapus (1.9 pav.):



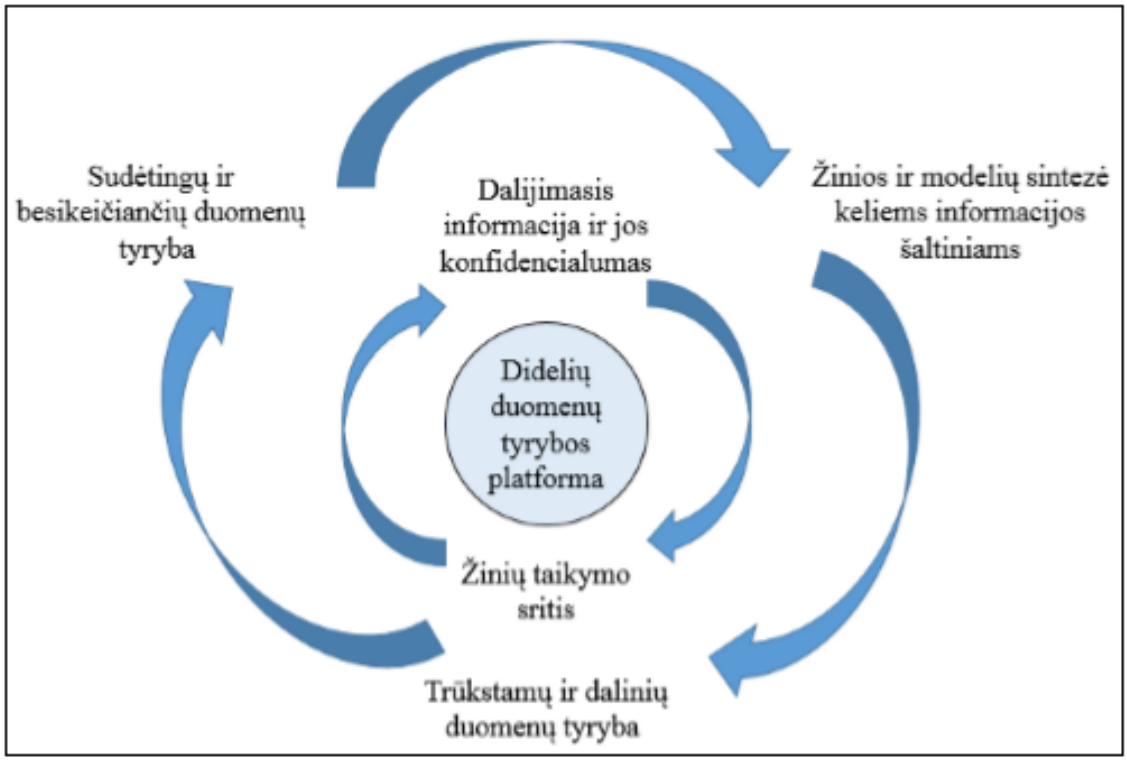
1.9 pav. SEMMA procesas

- **Imtis.** Iš didelio kiekio duomenų atrenkama dalis jų. Pakanka turėti mažą reikšmingų duomenų kiekį, kad procesas vyktų greitai. Nėra privalomas.
- **Tyrimas.** Duomenų peržiūrėjimas, ieškojimas nenumatytų tendencijų ir anomalijų, siekiant įgyti supratimą ir suformuoti idėjas.
- **Modifikavimas.** Duomenų modifikavimas, kuriant, renkant ir transformuojant kintamuosius į pasirinkto proceso pavyzdį.
- **Modeliavimas.** Automatinis duomenų modeliavimas, ieškant duomenų derinio, kuris patikimai prognozuoja norimą rezultatą.
- **Įvertinimas.** Tai duomenų įvertinimas, atsižvelgiant į naudą ir patikimumą, bei duomenų tyrybos išvados.

Siekiant kuo tikslesnių duomenų tyrybos rezultatų, etapų vykdymo metu gali tekti keletą kartų grįžti į prieš tai esančio etapo žingsnį. Nuolat papildant duomenų tyrybos etapų parametrus bei kartojant jų vykdymą, siekiama kuo tikslesnio duomenų tyrybos rezultato [15].

1.5. Didelių duomenų tyrybos platforma

Didelių duomenų apdorojimo sistemos modelis, sudarytas iš trijų pakopų: duomenų prieiga ir skaičiavimai (1 pakopa), duomenų atskyrimas ir žinių sritis (2 pakopa), didelių duomenų tyrybos algoritmai (3 pakopa) (1.10 pav.).



1.10 pav. Didelių duomenų tyrybos platforma

Duomenų prieiga ir skaičiavimai.

Įprastai duomenų tyryboje užtenka pavienių kompiuterių išteklių, tačiau didelių duomenų apdorojimo sistemos yra sudarytos iš kompiuterių klasterių, kuriuose lygiagrečiai vykdomos duomenų tyrybos užduotys. Tokios struktūros pavyzdys galėtų būti *MapReduce*.

Duomenų atskyrimas ir žinių sritis.

Dideliuose duomenyse žinių taikymas remiasi keliais aspektais, susijusiais su taisyklėmis, strategija, vartotojo ir duomenų taikymo informacija. Svarbiausi šios pakopos aspektai:

- **Dalijimasis informacija ir jos konfidencialumas.** Norint užtikrinti duomenų saugumą reikia riboti prieigą prie duomenų, užtikrinant prieigos kontrolę;
- **Žinių taikymo sritis.** Ji suteikia reikiamą informaciją, reikalingą didelių duomenų tyrybos algoritmų ir sistemų kūrimui.

Didelių duomenų tyrybos algoritmai.

- **Žinios ir modelių sintezė keliems informacijos šaltiniams.** Duomenų tyrybai didelius duomenis iš nepriklausomų šaltinių surinkti į vieną rinkinį dėl duomenų perdavimo sąnaudų ir privatumo yra sudėtinga. Tokiu atveju duomenų tyrybos procesas gali būti skaidomas į dviejų žingsnių procesą: duomenų, šablono ir žinių lygius. Duomenų lygyje kiekviename duomenų šaltinyje apskaičiuojama duomenų statistika, siekiant gauti bendrą visų duomenų pasiskirstymo vaizdą. Su kiekvienu duomenų šaltiniu atliekami duomenų tyrybos veiksmai, siekiant sudaryti duomenų modelius. Modelių koreliacinės analizės pagalba siekiama nustatyti, kaip duomenų šaltiniai yra susiję vienas su kitu ir kaip formuoti tikslius sprendimus.
- **Trūkstamų ir dalinių duomenų tyryba.** Trūkstamų ir dalinių duomenų tyryba yra vienas iš didelių duomenų analizės ypatybių. Trūkstami ir daliniai duomenys suprantami kaip per mažas objektų skaičius, reikalingas išvadoms gauti. Dauguma šiuolaikinių duomenų tyrybos sistemų tobulinama, siekiant sukonstruoti tokius algoritmus, kurie nustatytų trūkstamų duomenų reikšmes, tokiu būdu patobulinant duomenų tyrybos proceso metu gaunamus modelius.
- **Sudėtingų ir besikeičiančių duomenų tyryba.** Didelių duomenų augimas skatinamas sparčiai augančių sudėtingų duomenų ir jų pokyčių apimtis ir pobūdis. Dokumentai publikuoti www serveryje, socialiniai tinklai ir kt. yra sudėtingi duomenys. Jie yra apibūdinami daugeliu aspektu, įskaitant duomenų tipus (struktūrizuoti, nestruktūrizuoti duomenys, kurie gali būti pateikiami kaip tekstas, paveikslėliai, audio ir video duomenys ir pan.) įvairumą, priklausomybių ryšius tarp duomenų ir kt.

2. Metodinė dalis

Šiame skyriuje bus apžvelgiami duomenų paruošimo ir klasterinės analizės metodai.

2.1. Duomenų paruošimas

Duomenų paruošimas bei grafinis atvaizdavimas yra svarbios prielaidos daugumai statistinės analizės metodų. Jie ypač svarbūs kai kalbama apie klasterinę analizę. Žmogaus akis vis dar yra vienas efektyviausių įrankių vertinant klasterių stuktūrą. Nepaisant tariamo paprastumo gauti prasmingus rezultatus nėra paprasta. Klausimai ką išmatuoti, kaip kiekybiškai įvertinti panašumą, kokius grupavimo metodus naudoti ir svarbiausia kaip įvertinti rezultatus yra esminiai aspektai. Be to, manyti, kad klasterinė analizė visada imtį sugrupuos į prasmingas kategorijas, yra klaidinga. Todėl labai svarbu tinkamai pasiruošti duomenų analizei [17, 18, 19].

2.1.1. Skaitinės charakteristikos

Norint apibūdinti ir palyginti surinktų duomenų aibes naudojamos skaitinės duomenų charakteristikos, kurios sudaro dvi pagrindines grupes, tokias kaip padėties charakteristikos ir sklaidos charakteristikos. Pirmosios apibūdina duomenų reikšmių didumą, o antrosios – duomenų reikšmių išsisklaidymą. Galima išskirti svarbiausias padėties charakteristikas. Tai yra vidurkis, mediana, moda ir kvartiliai. O pagrindinės sklaidos charakteristikos yra imties plotis, dispersija, standartinis nuokrypis, kvartilių skirtumas ir variacijos koeficientas.

Padėties charakteristikos

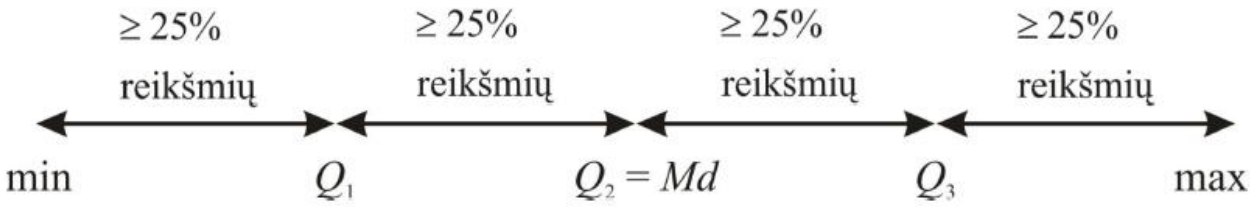
- **Vidurkis** – vidutiniškai artimiausias taškas visiems statistinės eilutės elementams. Imties vidurkis $\bar{x}$ skaičiuojamas taip:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \tag{2.1}$$

čia n – imties dydis, x_i – kintamojo reikšmė.

- **Mediana** – tai vidurinioji variacinės eilutės reikšmė. Mediana skaičiuojama:
  1. jei stebinių skaičius n nėra lyginis, tai mediana yra vidurinė variacinės eilutės reikšmė, kuri atitinka $\frac{1}{2}\cdot(n+1)$ poziciją;
  2. jei n yra lyginis, tai mediana yra dviejų vidurinių variacinės eilutės reikšmių, kurios atitinka $\frac{n}{2}\cdot\frac{(n+1)}{2}$ pozicijas, aritmetinis vidurkis.
- **Moda** – variacinėje eilutėje dažniausiai pasitaikanti reikšmė, kurioje tankio funkcija įgyja didžiausią reikšmę.

- **Kvartiliai** – jie duomenis dalija į ketvirčius. Q1, Q2, Q3 − kvartiliai; Md – mediana (2.1 pav.).



2.1 pav. Padėties charakteristikos

Sklaidos charakteristikos

- **Imties dispersija.** Ji parodo duomenų apie vidurkį sklaidą, o skaičiuojama yra taip:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad (2.2)$$

čia n – imties dydis, x_i – kintamojo reikšmė, $\bar{x}$ - vidurkis.

- **Standartinis nuokrypis.** Jis matuojamas tokiais pačiais vienetais kaip ir duomenys. Taip pat yra dažniausiai naudojama duomenų sklaidos charakteristika, kuria skaičiuojama taip:

$$s = \sqrt{s^2}; \quad (2.3)$$

čia s^2 – imties dispersija.

- **Variacijos koeficientas** arba kitaip vadinamas kitimo koeficientas yra skaičiuojamas tik kintamiesiems, kurie priklauso santykių skalei ir kurie turi tik eigiamus vidurkius. Gali būti naudojamas skirtingų duomenų aibių sklaidos ar skirtingais vienetais matuojamų duomenų aibių palyginimui. Formulė:

$$CV = \frac{s}{\bar{x}}; \quad (2.4)$$

čia s – standartinis nuokrypis, $\bar{x}$ - vidurkis.

- **Asimetrijos koeficientas** – tai skirstinio simetrijos matas, nusakantis skirstinio formą, rodantis, kur yra susitelkę dauguma stebinių – ar variacinės eilutės viduryje, ar kraštuose. Skaičiavimo formulė:

$$g_1 = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^3}{s^3}; \quad (2.5)$$

čia x – vidurkis, n – imties dydis, x_k – kintamojo reikšmė, s – standartinis nuokrypis.

- **Eksceso koeficientas** – tai taip pat skirstinio simetrijos matas, kitaip vadinamas dar skirstinio viršūnės lėkštumo matu. Jis parodo, kaip susitelkę dauguma stebinių, ar jie yra ties keliomis reikšmėmis, ar jos išsisklaidžiusios visoje variacinėje eilutėje. Apskaičiavimo formulė:

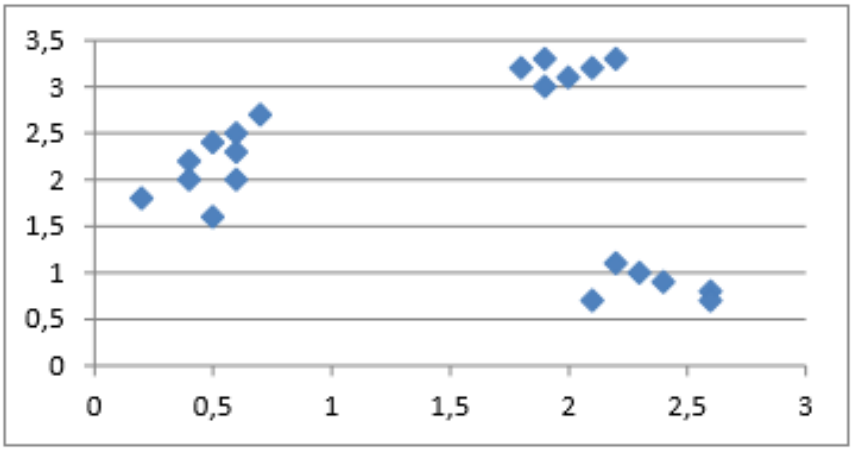
$$g_2 = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^4}{s^4} - 3; \quad (2.6)$$

čia x – vidurkis, n – imties dydis, x_k – kintamojo reikšmė, s – standartinis nuokrypis.

2.1.2. Grafinis vaizdas

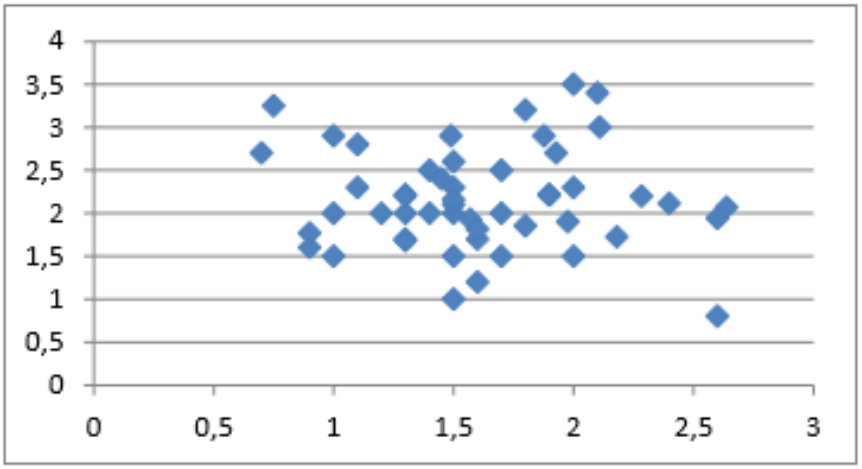
Sklaidos diagram (*Scatter Plot*) - labiausiai paplitęs būdas atvaizduoti pradinių duomenų išsidėstymą ir tarpusavio santykį. Tai galime realizuoti su SAS programine įranga atlikus *proc plot* arba *proc gplot* procedūras.

(2.2 pav.) sklaidos diagramoje aiškiai matomos trys atskiros duomenų grupės.



2.2 pav. Duomenys puikiai atsiskiriantys į klasterius

Deja, (2.3 pav.) iliustruota labiau tipiška situacija, nes aiškios grupės pasitaiko dažniausiai vienetiniais avejais.



2.3 pav. Duomenys prastai atsiskiriantys į klasterius

Klasterių dydis skiriasi, yra neaiški ir forma. Klasteriai galimai sutampa, kadangi vienas stebinys gali priklausyti visoms grupėms tuo pačiu metu tik su tam tikra tikimybe [20].

Daugelis klasterizavimo metodų randa klasterius, kurie yra tam tikros formos arba dyžio, arba dispersijos. Pavyzdžiui, mažiausių kvadratų metodas yra linkęs rasti apskritus klasterius su apytiksliai tokiu pat skaičiumi stebinių [20].

Dar viena problema, kad neretai analizuojamas duomenų rinkinys turi daugiau nei du išmatavimus, kas duomenų atvaizdavimą padaro sudėtingu procesu.

2.1.3. Duomenų standartizavimas

Duomenų standartizavimas reikalingas tam, kad duomenų matavimo skalė būtų suvienodinta [20]. Norint palyginti skirtingose matavimo skalėse išmatuotus kintamuosius tarpusavyje, standartizuojama priartinant stebinių reikšmes prie nustatyto diapozono kitų reikšmių. Kintamųjų standartizavimas supaprastina išsidėstymą, mastelį bei koreliacines struktūras [21]. Standartizuotiems kintamiesiems būdinga tai, kad jų vidurkis lygus maždaug nuliui, o dispersija - vienetui. Duomenų reikšmių standartizavimas nekeičia kintamųjų sklaidos nei pobūdžio, nei pasiskirstymo [19]. Paprastai kintamųjų standartizavimui naudojami standartinis nuokrypis arba imties plotis [20].

2.1.4. Požymių bei klasterizavimo mato parinkimas

Pirmiausia yra pasirenkami požymiai pagal kuriuos bus klasterizuojami duomenys, o tai lemia konkretus tyrimo tikslas ir uždavinys. Visais atvejais skirstymas į klasterius prasideda tada, kai jau turime objektų aibę ir kiekvieną objektą aprašančių skaitinių rodiklių aibę. Tiriami stebiniai sudaro baigtinę n objektų aibę, kurios kiekvienas narys turi m požymių. Paprastai tai yra išmatuojamų savybių reikšmės, apibūdinančios kiekvieną objektą. Tokiu būdu, kiekvieną objektą galima nusakyti daugiamačiu dydžiu $x_j = (x_1, x_2, \ldots, x_m)$, kur x_{kj} suprantamas, kaip k – tojo požymio reikšmę j – tąjam objektui. Taigi kiekvieną objektą traktuojam kaip m-matės erdvės tašką. Norint objektų aibę apibūdinti kaip galima geriau, reikia matuoti kuo daugiau požymių, apibūdinančių pasirinktą reiškinį. Tačiau didelis požymių skaičius apsunkina gautų rezultatų interpretavimą [18,19,21]. Šiame tyrime požymiai pagal kuriuos bus klasterizuojami duomenys yra oro teršalai.

Kitas žingsnis – panašumo mato pasirinkimas. Panašumo mato pasirinkimą įtakoja požymių matavimo skalė. Jeigu požymiai matuojami intervalų arba santykių skalėse - taikomi metriniai atstumo matai, jeigu požymiai yra kokybiniai naudojame koreliacijos ir asociatyvumo koeficientus [18].

Metriniai atstumo matai dažniausiai yra metrikos. Metrika – tai skaitinė dviejų neneigiamų objektų X ir Y funkcija $d(X,Y)$, tenkinanti šias sąlygas [19]:

- Simetriškumas: $d(X,Y) = d(Y,X)$;
- Trikampio nelygybė: $d(X,Y) \leq d(x,z)+d(y,z)$;
- Netapačių objektų atskiriamumas: jei $d(X,Y) \neq 0$, tai $X \neq Y$;

- Tapačių objektų neatskiriamumas: jei $X = Y$, tai $d(X,Y) = 0$.

2.1 lentelė. Dažniausiai naudojami atstumo matai

Atstumo matai	Formulės
Euklido atstumas	$\|X-Y\| = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$
Euklido atstumo kvadratas	$\|X-Y\|^2 = \sum_{i=1}^{m}(x_i - y_i)^2$
Minkovskio atstumas	$\left(\sum_{i=1}^{m}\|x_i - y_i\|^l\right)^{1/l}$, l>0
Manheteno (blokinis) atstumas	$\sum_{i=1}^{m}\|x_i - y_i\|$
Čebyševo atstumas	$\max_i\|x_i - y_i\|$
Vektorių kampo kosinuso atstumas	$\sum_{i=1}^{m}(x_i y_i)\left(\sum_{i=1}^{m}x_i^2\sum_{i=1}^{m}y_i^2\right)^{-1/2}$
Mahalanobio atstumo kvadrato atstumas	$(x-y)'V^{-1}(x-y)$

Koreliacijos koeficientas naudojamas objektų panašumui įvertinti jei duomenys yra kiekybiniai. Koreliacijos koeficiento formulė:

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\bar{x})^2\sum_{i=1}^{m}(y_i-\bar{y})^2}}; \quad (2.7)$$

čia x_i - *X i*-ojo požymio reikšmė, y_i - *Y i*-ojo požymio reikšmė, *m* - požymių skaičius.

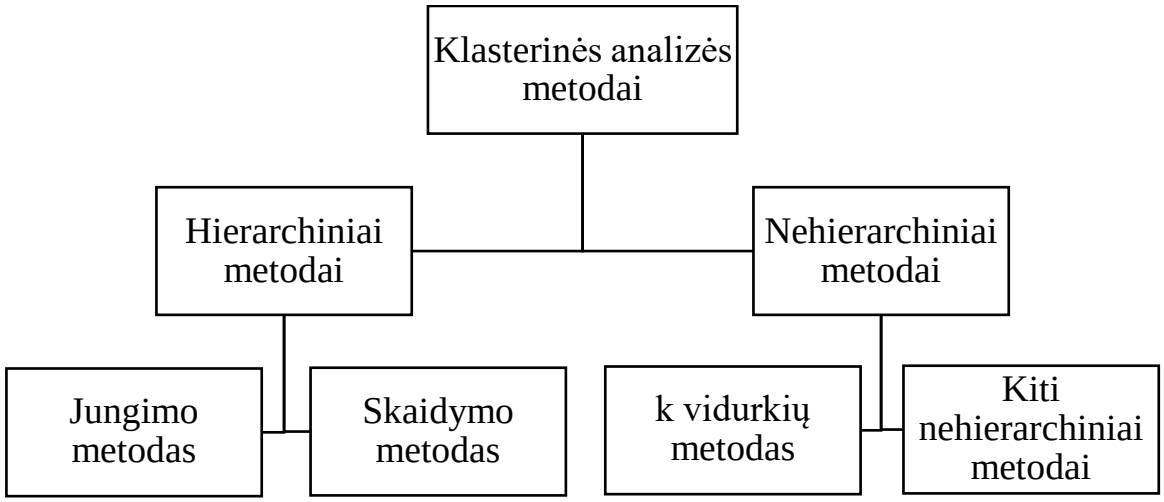
Koreliacijos keoficientas turi keletą pranašumų bei trūkumų. Visų pirma jis neturi aiškios statistinės prasmės, nes vidurkis skaičiuojamas pagal įvairius kintamuosius, o ne pagal stebinių aibę. Jis nejautrus kintamųjų reikšmių išsibarstymui bei poslinkiui. Šis matas parodo ryšį tarp dviejų kintamųjų rinkinių krypties ir stiprumo. Labai svarbu, kad kintamieji būtų matuojami toje pačioje skalėje. Kitu atveju, lyginimas tampa beprasmis. Kuo ši reikšmė arčiau vieneto tuo nagrinėjami objektai panašesni. Kai koreliacija lygi 0, vadinasi tarp kintamųjų nėra jokio panašumo [18, 22].

2.2. Klasterinė analizė

Duomenų klasifikavimas – vienas dažniausių duomenų tyryboje sprendžiamų uždavinių. Tai tiriamų objektų jungimas į klases pagal giminingumą, panašumą, atstumą ar koreliacinio ryšio stiprumą [20]. Objektų suskirstymas į klasterius atliekamas taikant vadinamąją klasterinę analizę, kuri dar vadinama

klasterizavimu ar klasterizacija. Objektus siekiama klasifikuoti taip, kad skirtumai klasteryje būtų kuo mažesni, o skirtumai tarp klasterių kuo didesni. Klasifikuojant duomenis neišvengiamas tam tikros dalies informacijos praradimas. Norint prarasti kuo mažiau informacijos, reikia sudaryti tuo daugiau klasių. Kai gaunamas minimalus klasių skaičius su sąlyga, kad informacijos praradimas neviršija iš anksto nustatyto dydžio, klasifikacija laikoma optimalia [23, 17].

Skiriamos dvi pagrindinės klasterizavimo metodų klasės: hierarchiniai ir nehierarchiniai metodai (2.4 pav.). Tai priklauso nuo panašumo matų parinkimo, atstumo tarp klasterių nustatymo kriterijų bei kokia skirstymo į klasterius strategija. Kiekvienas jų yra naudojamas priklausomai nuo situacijos ir tiriamos duomenų aibės.



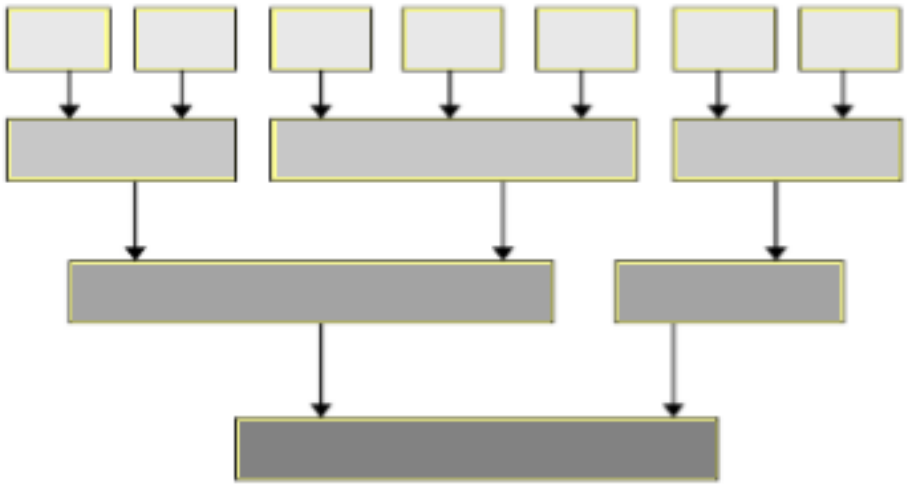
2.4 pav. Klasterinės analizės metodų klasifikavimo schema

Hierarchiniai analizės metodai yra tikslūs, tačiau didėjant stebinių skaičiui, reikia didelių skaičiavimo pajėgumų atstumų matricai apskaičiuoti ir jos elementams išsaugoti. Todėl dideliam stebinių kiekiui apdoroti geriau naudoti nehierarchinius metodus. Pastarieji nepriklauso nuo ankščiau rastų klasterių, tad šia prasme jie yra pranašesni už hierarchinius. Didelis nehierarchinių metodų trūkumas yra tai, kad jie yra stipriai įtakojami išskirčių ir netgi stebinių išsidėstymo tvarkos klasteryje. Taip pat jie reikalauja dar prieš tyrimą nustatyti duomenyse egzistuojančių klasterių skaičių. Dėl šių trūkumų vis dėlto hierarchiniai metodai yra dažniausiai naudojama grupavimo technika.

2.2.1. Hierarchiniai metodai

Hierarchinis klasterizavimas naudojamas nedideliam objektų skaičiui klasterizuoti. Jis padeda nustatyti bendrą visų klasterių tarpusavio priklausomybių struktūrą ir tik po to sprendžiama, koks klasterių skaičius būtų optimalus. Hierarchiniai metodai dar skiriami į jungimo ir skaidymo metodus.

- **Skaidymo metodas.** Hierarchija pradedama formuoti nuo stebinių visumos kaip vienos klasės, o baigiama, jog kiekvienas objektas sudaro atskirą klasę.
- **Jungimo metodas.** Tai atvirkštinis procesas skaidymo metodui. Pirmiausia kiekvienas objektas sudaro atskirą klasę, tada visi objektai jungiami tarpusavyje ir galiausiai visi objektai sudaro vieną klasę (2.5 pav.).



2.5 pav. Hierarchinio klasterizavimo jungimo metodas

Hierarchinio klasterizavimo jungimo metodų procedūros etapai [22]:

1) Turima N klasterių po 1 objektą ir NxN simetrinę atstumų matricą $(d_{i,j})_{i,j}$.
2) Pagal atstumų matricą nustatomi du klasteriai *U* ir *V*, tarp kurių atstumas yra mažiausias.
3) Sujungiami klasteriai *U* ir *V*. Naujas klasteris pavadinamas (*UV*). Tada atstumų matrica pakeičiama taip: pirmiausia išbraukiami stulpeliai ir eilutės, atitinkantys klasterius *U* ir *V*, o tada pridedamas stulpeli ir eilutė su atstumais tarp (*UV*) ir likusiųjų klasterių.
4) Kartojami 2 ir 3 žingsniai (N-1) kartų. Procesas baigiamas, kai visi objektai yra viename klasteryje.

Šios procedūros schemą atvaizduoja grafikas vadinamas dendrograma. Kuriuo etapu objektų paskirstymas į klasterius yra optimalus nusprendžia pats tyrėjas.

Nustačius atstumuus tarp objektų ir juos suskirsčius į klasterius, panašius klasterius reikia sujungti. Panašių klasterių nustatymui naudojami atstumo matai [22] (2.2 lentelė).

2.2 lentelė. Dažniausiai naudojamos klasterių artumo metrikos

Atstumo matas	Formulė
Vienetinės jungties (artimiausio kaimyno)	$d(U,V) = \min_{X_i \in U, Y_j \in V} d(X_i, Y_j)$ $X_{i,}$ - i – tasis U objektas, Y_j - j – asis objektas
Pilnosios jungties (tolimiausio kaimyno)	$d(U,V) = \max_{X_i \in U, Y_j \in V} d(X_i, Y_j)$
Vidutinio atstumo (vidutinės jungties)	$d(U,V) = \sum_{X_i \in U} \sum_{Y_j \in V} d(X_i, Y_j) / (n_U n_V)$ n_U, n_V - klasterių objektų skaičius
Centroidų (centrų)	$d(U,V) = d(\overline{U}, \overline{V})$ $\overline{U}, \overline{V}$ - objektų požymių vektorių vidurkiai
Vordo	$d(U,V) = \left\|\overline{U} - \overline{V}\right\|^2 / \left(1/n_U + 1/n_V \right)$

Artimiausio kaimyno atstumas apibrėžiamas kaip atstumas tarp klasterių artimiausių kaimynų.

Tolimiausio kaimyno atstumas apibrėžiamas kaip atstumas tarp klasterių tolimiausių kaimynų.

Vidutinio atstumo matas tai atstumas tarp dviejų klasterių, apskaičiuotas kaip vidutinis atstumas tarp visų objektų porų, esančių dvejuose skirtingose grupėse.

Centroidų atstumas apibrėžiamas kaip Euklido atstumo tarp klasterių centrų kvadratas.

Vordo atstumas tarp dviejų klasterių apibrėžiamas kaip Euklido atstumų tarp visų įmanomų klasterius sudarančių objektų porų kvadratų suma. Kiekviename hierarchinio klasterizavimo etape atliekama dispersinė analizė.

2.2.2. Nehierarchiniai metodai

Nehierarchinio klasterizavimo pagrindinis skirtumas yra tas, kad dar prieš pradedant analizę, tyrėjas jau turi žinoti norimą klasterių skaičių. Programinis įrankis suformuoja užduotą skaičių klasterių. Dažniausiai naudojamas nehierarchinio klasterizavimo metodas yra k – vidurkių metodas.

k - vidurkių metodo algoritmas klasterizuoja objektus skaidydamas juos į *k* klasterių. Daroma prielaida, kad stebiniai yra daugiamačiai. Ieškoma suskaidymo, kuris minimizuoja dispersijas klasterių viduje arba funkciją [21]:

$$\sum_{i=1}^{k} \sum_{X \in S_i} \|X - m(S_i)\|^2; \quad \textbf{(3.8)}$$

čia S_i, i=1,2,…, k yra klasteriai, o $m(S_i)$ yra S_i, sudaryto iš stebinių $X \in S_i$, „svorio centras“.

Atstumui matuoti dažniausiai naudojamas Euklido matas arba jo kvadratas. K – vidurkių klasterinė analizė atliekama kiekybiniams, t.y. intervalų ir santykių skalės, kintamiesiems. Kintamiesiems, matuojamiems kitose skalėse, naudojami hierarchiniai metodai [17, 21]. Algoritmo procedūros etapai:

1) Stebiniai suskaidomi į *k* pradinių klasterių;
2) Apskaičiuojami kiekvieno klasterio vidurkiai bei randami centrai;
3) Stebinius priskiriant artimiausiems centrams, atliekamas naujas suskaidymas;
4) Naujų klasterių centrai yra perskaičiuojami;
5) Žingsniai 3, 4 kartojami kol centrai stabilizuojasi, t.y. stebiniai nebekeičia klasterių.

Pateiksime tik vieną iš galimų **k artimiausių kaimynų metodo** variantų, nors jų atmainų yra ne viena. Pirmiausiai kiekvienas stebinys priskiriamas atskiram klasteriui, o tuomet du klasteriai yra taip pat sujungiami į klasterį [21]:

1) Sudaromos visos galimos poros iš dviejų elementų;
2) Skaičiuojamas kiekvienos poros tankis:

$$f_i(x) = \frac{n_i(x)}{n \cdot V(x)}; \quad (2.9)$$

čia $n_i(x)$ - tojo klasterio kaimynų ir stebinio x skaičius, $V(x)$ - $n_i(x)$ stebinių hipertūris.

3) Sujungiami didžiausią bendrą tankį turintys du klasteriai.

Toliau nagrinėjamas kiekvienas stebinys kartu su įvertintu tankiu. Tiriamasis stebinys gali priklausyti bet kuriam klasteriui, taip randami jo *k* artimiausi kaimynai. Taigi, šis stebinys sujungiams su tuo klasteriu, prie kuriuo jį priskyrus tankis būna didžiausias, bet ir ne mažesnis nei sujungus su bet kuriuo kaimynu. Kai klasteriai nusistovi ir jų struktūra nebekinta, klasterių tikslinimas baigiamas [21].

2.3. Klasterių skaičiaus nustatymo algoritmai

Klasterinėje analizėje viena iš problemų, su kuria dažnai susiduriama yra klasterių skaičiaus nustatymas arba modelio adekvatumo tyrimas. Yra nemažai algoritmų skirtų klasterių skaičiui nustatyti. Apžvelgsime tris plačiausiai naudojamus algoritmus.

Pseudo F kriterijus matuoja klasterių atsiskyrimą ir taip pat bando nustatyti klasterių skaičių imties pagrindu. Psiaudo F (PSF) kriterijus naudojamas ir kaip klasterių skaičiaus indikatorius. Ši statistika pasiskirsčiusi lyg Fišerio atsitiktinis dydis su *d(q-1)* ir *d(n-q)* laisvės laipsniais, jei tenkinamos dvi prielaidos: naudojant klasterizavimo metodą stebiniai klasteriams priskiriami atsitiktinai ir jie yra tarpusavyje nepriklausomi, pasiskirstę pagal daugiamatį normalinį skirstinį - kurios labai retai pasitvirtina [21].

PSF kriterijus:

$$PSF = \frac{\left(\sum_{t=1}^{n}\left\|X(t)-\overline{X}\right\|^2\right)/(q-1)}{\left(\sum_{k=1}^{q}\sum_{i\in C_k}\left\|X(i)-\overline{X_k}\right\|^2\right)/(n-q)}; \quad (2.10)$$

čia $\overline{X}$ - imties vidurkis, o $\overline{X_k}$ yra *k*-tojo klasterio vidurkis.

Išreiškiant statistikos R^2 nariais:

$$PSF = \frac{R^2/(q-1)}{(1-R^2)/(n-q)}; \quad (2.11)$$

Kriterijus interpretuojamas ieškant PSF reikšmės lokaliųjų minimumų ir sprendžiant apie klasterių skaičių [21].

Pseudo T^2 kriterijaus statistika (PST^2) lygina dviejų daugiamačių aibių vidurkius. Šis kriterijus naudojamas norint išspręsti, kurie klasteriai turi būti sujungti. Dažniausiai PST^2 kriterijumi matuojamas atskyrimas tarp dviejų vėliausiai sujungtų klasterių. Jeigu kriterijaus reikšmė yra didelė, tai klasterių vidurkiai skiriasi reikšmingai ir klasteriai nėra sujungiami, bei atvirkščiai – klasterius galima sujungti jei PST^2 reikšmė yra maža [18, 21].

Pseudo T^2 statistikos skirstinys yra Fišerio skirtinys su *d* ir $d(n_k + n_l - 2)$ laisvės laipsniais. Pseudo T^2 kriterijus apskaičiuojamas sujungus klasterius C_k ir C_l į klasterį C_m [18, 21]:

$$PST^2 = \frac{w_m - w_k - w_l}{(w_k + w_l)/(n_k - n_l - q)}; \quad (2.12)$$

Čia n_k yra *k*-tojo klasterio objektų skaičius, o $w_k = \sum_{i\in C_k}\left\|X(i) - \overline{X_k}\right\|$.

Pagrindinė PST^2 interpretavimo taisyklė yra mažėjimo kryptimi ieškoti klasterių skaičiaus pirmos didesnės reikšmės, negu buvusi, ir parinkti klasterių skaičių, kuris atitinka prieš tai esančią statistikos reikšmę [18, 21]:

Šarlio kubinis klasterizavimo kriterijus yra vienas populiariausių ir informatyviausių klasterizavimo kriterijų, paprastai žymimas CCC. Tai kvadratų sumos tarp klasterių stebinių vertinimas. Taikant šį kriterijų tikrinamos tokios hipotezės [16]:

H_0: stebinių skirstinys yra daugiamatis tolygusis;
H_1: stebinių skirstinys yra daugiamačių Gauso skirstinių mišinys.

Teigiami CCC dydžiai reiškia, kad H_0 yra atmesta. Norint apskaičiuoti CCC, pirmiausia įvertinama dispersija:

$$E(R^2) \cong 1 - \left[\sum_{j=1}^{d^*} \frac{1}{n+u_j} + \sum_{j=d^*+1}^{d} \frac{u_j^2}{n+u_j}\right]\left[\frac{(n-q)^2}{n}\right]\left[1 + \frac{4}{n}\right] / \sum_{j=1}^{d} u_j^2; \qquad (2.13)$$

čia s_j yra hiperkubo j-osios kraštinės ilgis, $u_j = \frac{s_j}{c}$, kai $c = \left(\frac{v}{q}\right)^{1/d}$ ir $v + \prod_{i=1}^{d} s_j$;

d^* - didžiausias sveikasis skaičius mažesnis už q, bet toks, kad u_{d*} būtų ne mažesnis už vienetą.

CCC, priklausantis nuo R^2, apskaičiuojamas:

$$CCC = \log\left[\frac{1-E(R^2)}{1-R^2}\right]\frac{\sqrt{nd^*/2}}{(0{,}001+E(R^2))^{1{,}2}}; \qquad (2.14)$$

čia $R^2 = 1 - \left[d^* + \sum_{j=d^*+1}^{d} u_j^2\right] / \sum_{j=1}^{d} u_j^2$.

Ieškoma CCC reikšmės lokaliųjų maksimumų. Klasterių struktūra yra galima, jei CCC reikšmės yra tarp 0 ir 2, tačiau interpretuoti ją reikia atsargiai. Klasterių skaičius parinktas tinkamai, jei reikšmės yra didesnės už 2, kad [16,19]. CCC interpretuojamas panašiai kaip ir PSF: ieškoma CCC reikšmės lokaliųjų maksimumų ir sprendžiama apie klasterių pasirinkimą.

Optimalų klasterių skaičių galima nustatyti ekspertiškai iš dendrogramos. Vienoje dendrogamos ašyje atidedami stebinių numeriai, kitoje dendogramos ašyje atidedami atstumai. Kokie objektai sujungti į klasterius ir koks atstumas tarp jų parodo objektus jungianti laužtė. Pats tyrėjas gali nuspręsti kada suskirstymas į klasterius yra optimalus.

3. Tiriamoji dalis

Šiame skyriuje bus pristatomi nagrinėjami duomenys, aprašomi tyrimai, kurie buvo atlikti, bei gauti rezultatai ir išvados.

3.1. Duomenys

Tiriamąją imtį sudaro duomenų failai .xlsx formatu apie 96 butų, kurie išsidėstę per 20 daugiabučių namų Lietuvoje, ir 172 butų, kurie išsidėstę per 32 daugiabučius namus Suomijoje, oro taršą. Pirmajame faile duomenys rinkti prieš atliekant Lietuvos daugiabučių renovaciją, antrajame - po jos. Lygiai taip pat duomenys atvaizduoti ir apie patalpų oro taršą Suomijoje. Matuoti tokie dujiniai oro teršalai kaip: anglies dioksidas, benzenas, toluenas, etilbenzenas, ksilenas, formaldehidas, azoto dioksidas, radonas. Vertinant oro kokybę taip pat buvo išmatuota oro temperatūra, drėgmė, bei įvertintas ventiliacijos efektyvumas. Siekiant identifikuoti taršos šaltinius į kiekvieno buto aprašymą įtrauktos tokios charakteristikos: viryklės rūšis, ventiliacijos tipas, buto aukštas, atstumas iki gatvės, gyventojų skaičius, baldų bei remonto būklė, grindų dangos rūšis.

Toliau pateikiamas pradinių duomenų fragmentas (3.1 pav.). Pateikiami tokie duomenys apie tiriamus butus: namas, butas, temperatūra, drėgmė, anglies dioksido, benzeno, tolueno, etilbenzeno, ksileno, formaldehido, azoto dioksido, radono kiekiai ore, viryklės tipas, ventiliacijos tipas, buto aukštas, atstumas iki gatvės, gyventojų skaičius, remonto būklė, baldų būklė, grindų dangos rūšis.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	B	A	Temp	RH	CO2	Benzenas	Toluenas	Etilbenzenas	Ksilenas	Formaldehidas	NO2	Radonas	Viryklė	Ventiliacija
2	B1	A1	21,9	33,2	698	1,00	12,00	3,00	10,00	34,00	6,29		1	0
3		A2	20,4	44,6	658	2,00	5,00	2,00	7,00	33,00	13,05		1	0
4		A3	21,8	37,7	488	1,00	10,00	2,00	7,00	46,00	6,74	17,4	1	0
5		A4	19,6	50,5	2239	1,00	9,00	1,00	7,00	40,00	23,36	17,7	1	0
6		A5	20,8	36,6	760	1,00	4,00	0,50	1,50	37,00	16,78	4,5	1	0
7	B2	A1	18,5	52,0	1095	2,00	8,00	1,00	2,00	34,00	9,31	30,2	1	2
8		A2	18,2	49,5		1,00	4,00	1,00	2,00	43,00	8,28	16,9	1	2
9		A3	18,7	54,3	1473	1,00	31,00	2,00	6,00	30,00	8,16	25,7	1	2
10		A4	19,2	56,2	1509	2,00	27,00	4,00	20,00	16,00	13,31		1	2
11		A5	19,1	55,0	1551	1,00	17,00	2,00	6,00	20,00	8,36		1	1
12		A6			540									
13	B3	A1	18,1	51,5	1027	1,29	6,55	0,00	0,00	24,45	18,93	17,8	1	0
14		A2	20,3	43,4	1017	15,87	4,02	0,00	0,00	13,59	21,10	11,5	1	0
15		A3	19,2	33,0	743	0,98	1,59	0,00	0,00	16,01	8,63	12,4	1	2
16		A4	21,5	42,8	1458	3,07	1,40	0,00	0,00	14,64	15,68		1	2

3.1 pav. Pradinių duomenų fragmentas

Patalpų oro teršalai matuojami santykių skalėje, kitos buto buitinės charakteristikos matuojamos vardų skalėje. (3.1 lentelė) pateikiami kintamųjų apibūdinimai.

3.1 lentelė. Kintamųjų apibūdinimas

Kintamasis	**Aprašymas**
Namas	Namo numeris
Butas	Buto numeris
Temperatūra	Oro temperatūros vertė Celsijais
Drėgmė (RH)	Oro drėgmės santykinė vertė
Anglies dioksidas	Tolydus kintamasis apibūdinantis CO_2 kiekį
Benzenas	Tolydus kintamasis apibūdinantis benzeno kiekį
Toluenas	Tolydus kintamasis apibūdinantis tolueno kiekį
Etilbenzenas	Tolydus kintamasis apibūdinantis etilbenzeno kiekį
Ksilenas	Tolydus kintamasis apibūdinantis ksileno kiekį
Formaldehidas	Tolydus kintamasis apibūdinantis formaldehido kiekį
Azoto dioksidas	Tolydus kintamasis apibūdinantis NO_2 kiekį
Radonas	Tolydus kintamasis apibūdinantis radono kiekį
Virykle	0- Kita 1 - Dujinė
Ventiliacija	0 - Natūrali 1 - Mechaninė 2 - Natūrali ir mechaninė
Aukstas	0 - Aukščiau nei trečias 1 - Pirmieji trys aukštai
Atstumas	0 – Toliau nuo gatvės (>50m) 1 – Arčiau gatvės (<50m)
Gyventojai	0 - Du arba mažiau 1 - Daugiau nei du
Remontas	0 – Nėra 1 - <5 metų senumo
Baldai	0 - Seni 1 - Nauji (iki 5 metų)
Grindys	0 – Medžio parketas 1 - Laminatas arba linoleumas 2 - Laminatas ir kilimas 3 - Parketas ir kilimas 4 - Kiliminė danga

Pirmiausia bus atliekamas tyrimas su duomenimis matuotais prieš atliekant renovaciją Lietuvoje esantiems daugiabučiams. Nustačius oro taršos priežastis, ištirsime oro kokybę po atliktos renovacijos. Tuomet apžvelgsime analogišką situaciją Suomijoje ir palyginsime rezultatus.

Tyrimo metu taip pat bus reikalinga informacija apie nustatytas mikroklimato higienos normas (3.2 lentelė) ir didžiausią teršalų koncentraciją (3.3 lentelė) leidžiamą gyvenamosiose patalpose.

3.2 lentelė. Gyvenamųjų patalpų mikroklimato higienos normos

Eil. Nr.	Mikroklimato parametrai	Ribinės vertės
1.	Oro temperatūra, ^oC	18-22
2.	Santykinė oro drėgmė, %	35-60

3.3 lentelė. Didžiausia leidžiama teršalų koncentracija gyvenamosios aplinkos ore

Medžiagos pavadinimas	Didžiausia leidžiama koncentracija
CO_2	60 $\mu g/m^3$
Benzenas	100 $\mu g/m^3$
Toluenas	600 $\mu g/m^3$
Etilbenzenas	20 $\mu g/m^3$
Ksilenas	200 $\mu g/m^3$
Formaldehidas	10 $\mu g/m^3$
NO_2	1000 $\mu g/m^3$
Radonas	200 Bq/m^3

SAS programinio įrankio pagalba importuoti duomenys konvertuojami į .sas7bdat formatą.

3.2. Aprašomosios statistikos analizė

Butuose prieš renovaciją buvo atlikta patalpų oro teršalų analizė. Matuoti tokie dujiniai oro teršalai kaip: anglies dioksidas, benzenas, toluenas, etilbenzenas, ksilenas, formaldehidas, azoto dioksidas ir radonas. Vertinant oro kokybę taip pat buvo išmatuota oro temperatūra ir drėgmė. Vertinant oro kokybę galima pamatyti, jog nors ir patalpų oro temperatūros vidutinė reikšmė atitinka reikalaujamas normas, tačiau minimali bute užfiksuota temperatūra siekė vos 12.06 ^oC. Santykinės oro drėgmės vidurkis taip pat neviršija kritinių ribų, bet butuose buvo užfiksuotas ir 15.54% bei 71.74% drėgmės lygis. Atsižvelgus į dujinių oro teršalų rodiklius, visi neviršija didžiausios leidžiamos teršalų koncentracijos gyvenamosios aplinkos ore. Duomenys pateikti 3.4 lentelėje.

3.4 lentelė. Skaitinės kintamųjų charakteristikos (prieš renovaciją Lietuvoje)

Kintamasis	Imties plotis	Vidurkis	Standartinis nuokrypis	Mažiausia reikšmė	Didžiausia reikšmė
Temperatūra	92	19.59806	2.07320	12.06250	23.68024
Drėgmė	92	45.25144	10.90075	15.54260	71.74503
CO_2	89	1010	378.47318	427.00000	2239
Benzenas	95	6.84209	12.61135	0.54866	101.76803
Toluenas	95	10.11971	13.75644	0	90.28284
Etilbenzenas	95	1.87929	4.50830	0	38.23195

Ksilenas	95	3.92556	6.83339	0	45.82500
Formaldehidas	95	23.16019	10.47222	7.17292	51.37720
NO_2	93	13.98742	7.81951	2.44000	43.77000
Radonas	43	27.53891	16.36935	4.46000	70.20000

3.2.2. Duomenų paruošimas klasterizavimui

Naudojant *SAS* programinį paketą procedūros *proc corr* pagalba įvertinama patalpų oro teršalų tarpusavio priklausomybė. Ryšio stiprumui įvertinti naudojamas Pirsono koreliacijos koeficientas (3.5 lentelė). Tyrimas bus atliekamas nevertinant radono poveikio, kadangi ne visuose stebiniuose buvo matuojamas jo kiekis ore. Įtraukus į analizę radoną prarasime daug duomenų ir rizikuosime gauti klaidingas išvadas.

Pirsono koreliacijos koeficientas (3.5 lentelė) rodo stiprų teigiamą ryšį tik tarp dviejų kintamųjų - etilbenzeno ir ksileno, tarp kitų kintamųjų reikšmingas tarpusavio ryšys nenustatytas. Tai reiškia, kad nėra informacijos dubliavimo ir negalime atmesti nei vieno kintamojo.

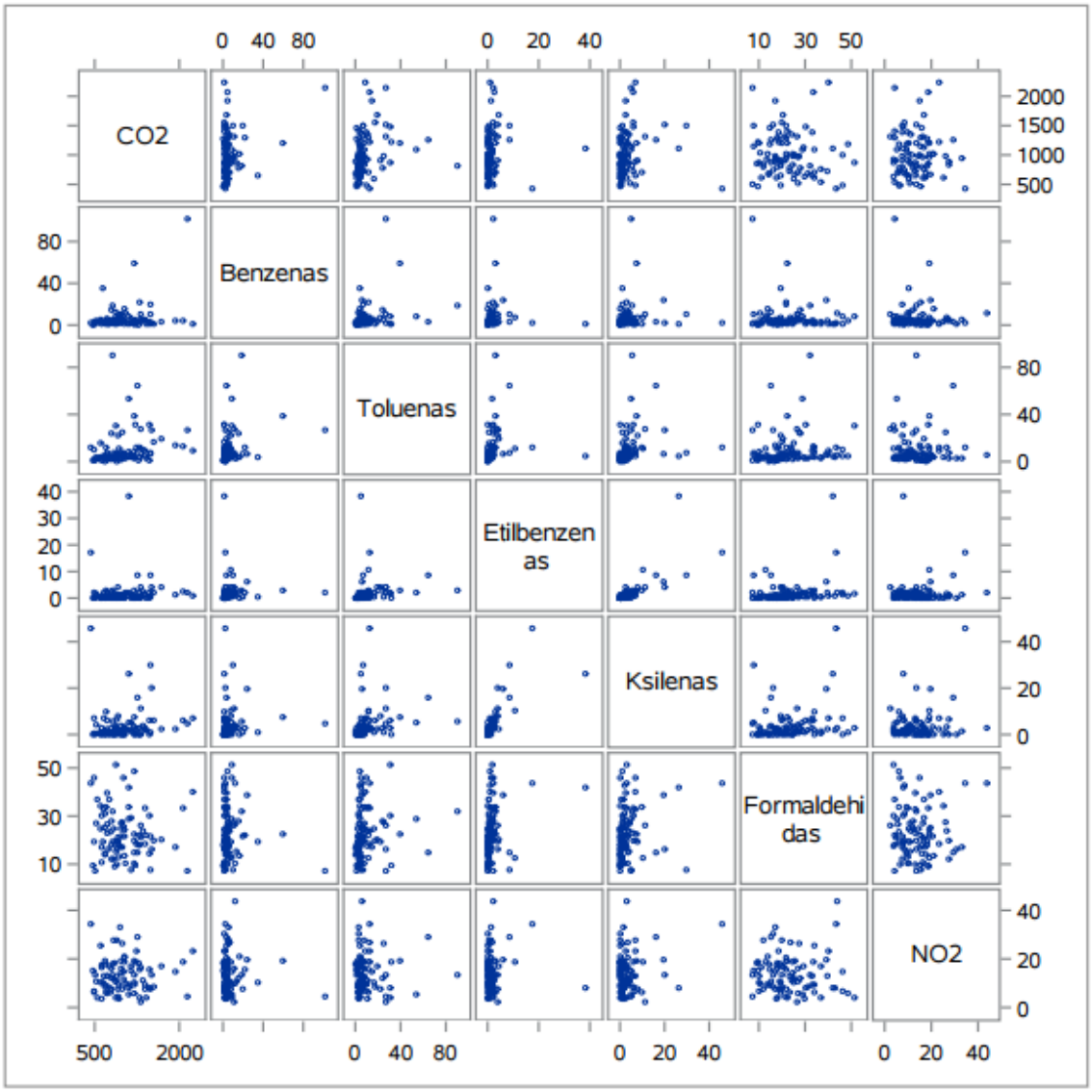
3.5 lentelė. Pirsono koreliacijos tarp kintamųjų matavimas (prieš renovaciją Lietuvoje)

	CO_2	**Benzenas**	**Toluenas**	**Etilbenzenas**	**Ksilenas**	**Formaldehidas**	**NO_2**
CO_2	1.00000	0.28746	0.22148	0.06296	0.09515	-0.09311	0.00384
Benzenas	0.28746	1.00000	0.28022	0.02062	0.06101	-0.10015	-0.06945
Toluenas	0.22148	0.28022	1.00000	0.14764	0.24284	0.10656	0.01508
Etilbenzenas	0.06296	0.02062	0.14764	1.00000	**0.74754**	0.25536	0.07687
Ksilenas	0.09515	0.06101	0.24284	**0.74754**	1.00000	0.21959	0.19232
Formaldehidas	-0.09311	-0.10015	0.10656	0.25536	0.21959	1.00000	-0.00601
NO_2	0.00384	-0.06945	0.01508	0.07687	0.19232	-0.00601	1.00000

Toliau pateikiama duomenų tarpusavio santykį apibūdinanti sklaidos diagrama, tačiau iš pradinių duomenų išsidėstymo negalėsime prognozuoti klasterių skaičiaus (3.2 pav.). Matome, kad duomenys yra prastai atsiskiriantys į klasterius, neaiškus klasterių dydis bei forma, o klasteriai galimai sutampa. Todėl šiame tyrime labai svarbus ir kitas žingsnis - kintamųjų standartizavimas, kadangi kintamųjų reikšmės yra išsidėsčiusios labai plačiame diapazone, kas stipriai trukdo kintamuosius tinkamai lyginti tarpusavyje. Naudojant procedūrą *proc stdize* standartizuojame pradinius duomenis. Standartizuojama naudojant *method = range* parinktį, kas reiškia imties ribų metodą.

Svarbiausias tyrimo požymis pagal kurį galima ne tik skirstyti duomenis į klasterius, bet ir susieti su kitomis buto charakteristikomis yra patalpų oro teršalai, pagal ką ir bus grupuojame tiriami objektai.

Kaip ir minėjau tyrimas bus atliekamas nevertinant radono poveikio, kadangi ne visuose stebiniuose buvo matuojamas jo kiekis ore.



3.2 pav. Pradinių duomenų sklaidos diagrama (prieš renovaciją Lietuvoje)

Tyrime naudojamas Euklido atstumo matas, kadangi jo rezultatų beveik neįtakoja išskirtys ir jis dažniausiai naudojamas praktikoje. *SAS* programinis paketas Euklido atstumo matą naudoja kaip numatytąjį, todėl programos kode nereikia nieko nurodyti.

3.2.3. Klasterizavimas naudojant hierarchinius metodus

Atliekant hierarchinio klasterizavimo tyrimą naudojami jungimo metodai. Nustatyti panašius klasterius naudoti Vordo atstumo matai. Realizacijai naudota procedūrą *proc cluster*. Vordo metodo realizavimui *proc cluster* procedūroje nurodoma *method = ward* parinktis.

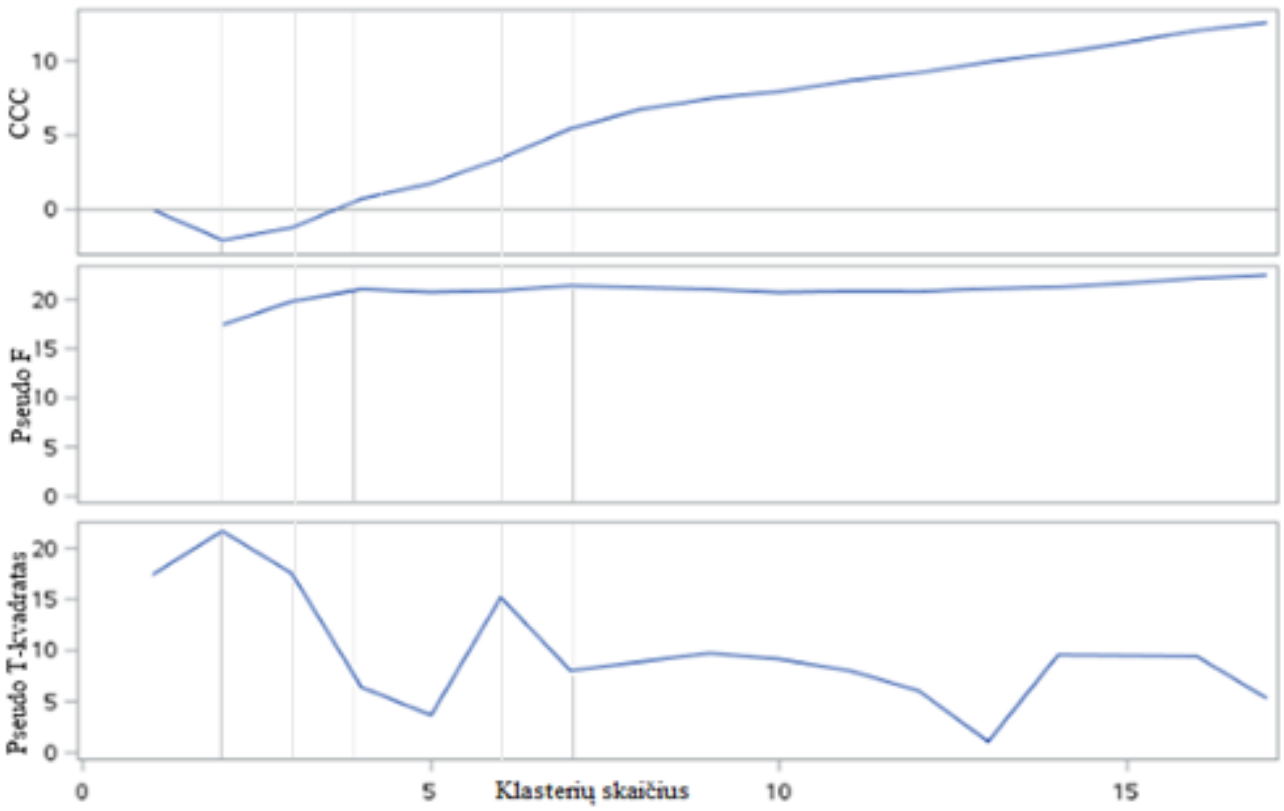
Galima klasterizavimo schema ar dendogramos pakaitalas gali būti klasterių istorijos lentelė, kurios fragmentas pateikiamas žemiau (3.6 lentelė). Vykdant klasterių jungimo procesą, pirmiausiai suformuojami 85 klasteriai, tuomet stebiniai jungiami kol sujungiami į vieną klasterį. Lentelėje galima matyti kada ir koks stebinys susijungė su kitu stebiniu ar klasteriu skiltyje „*Sujungti klasteriai"*. Lentelėje pateikiamas stebinių kiekis klasteryje bei klasterių skaičiaus nustatymui naudojamos statistikos, tokios kaip kubinis klasterizavimo kriterijus, pseudo F statistika ir pseudo T^2 kriterijus. (3 Priedas) pateiktas 3P.1 lentel4je pateiktas pilnas klasterių jungimo protokolas.

3.6 lentelė. Klasterių jungimo protokolas (prieš renovaciją Lietuvoje)

Klasterių skaičius	**Sujungti klasteriai**		**Dažnis**	**Kubinis klasterizavimo kriterijus**	**Pseudo F**	**Pseudo T^2**
85	OB14	OB16	2	.	125	.
84	OB18	OB35	2	.	99.2	.
...	...	...	...	...	...	...
10	CL16	CL18	21	3.14	20.7	9.1
9	CL19	CL14	32	2.74	21.0	9.7
8	OB28	OB29	2	2.14	21.2	.
7	CL46	CL12	15	1.44	21.4	8.0
6	CL10	CL9	53	0.02	20.9	15.2
5	CL13	OB90	4	-1.2	20.7	3.7
4	CL7	CL5	19	-2.0	21.0	6.4
3	CL11	CL8	14	-3.1	19.8	17.5
2	CL4	CL6	72	-3.1	17.4	21.7
1	CL3	CL2	86	0.00	.	17.4

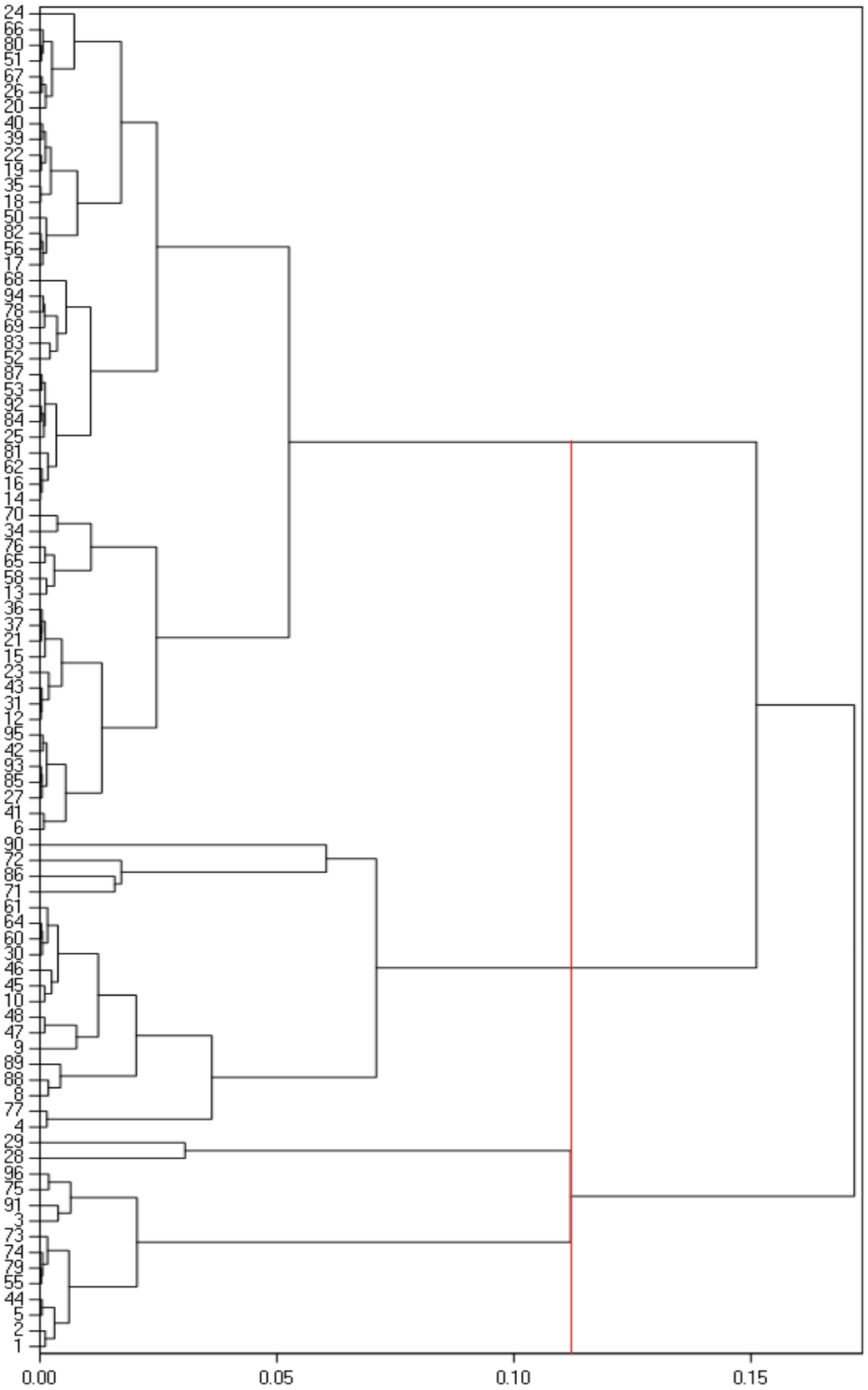
Klasterių skaičiaus nustatymas.

Iš 3.6 lentelės ir 3.3 pav. matyti, jog kubinis klasterizavimo kriterijus siūlo duomenis grupuoti į 8 arba 9 klasterius. Pseudo F statistika rodo, kad duomenis galima skirstyti į 4 arba 7 klasterius. Pseudo T^2 kriterijus rodo, jog duomenis galima skirstyti į 5 klasterius.



3.3 pav. Klasterių skaičiaus nustatymo kriterijai (prieš renovaciją Lietuvoje)

Optimalus klasterių skaičius ekspertiškai nustatomas iš dendrogramos. Dendrogamos ašyse atidedami stebinių numeriai ir atstumai. Nustatytą tyrėjo pjūvį žymi raudona linija, ji išskiria tris klasterius (3.4 pav.).



3.4 pav. Dendograma (prieš renovaciją Lietuvoje)

Nustačius didesnį grupių skaičių atsiranda klasterių su pavieniais elementais.

Procedūroje *proc tree* parenkamas $n = 3$ kriterijus tam, kad atliekant tolimesnius veiksmus tiriamoji imtis būtų skirtoma į tris klasterius.

Norint nustatyti kiek kiekvienam klasteriui priklauso stebinių galima naudojant *proc freq* procedūrą:

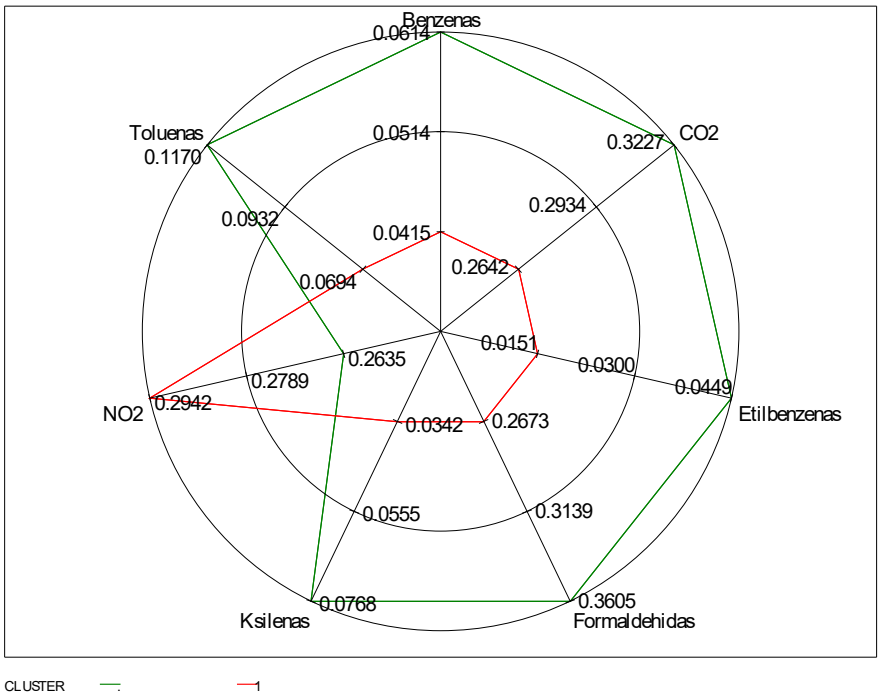
- 53 stebiniai priklauso pirmajam klasteriui;
- 19 stebinių priklauso antrajam klasteriui;
- 14 stebinių priklauso trečiajam klasteriui.

Ne vienam šių suformuotų klasterių nepriklauso dar 10 stebinių.

Sudarytų klasterių požymių tyrimas

Klasterių požymių tyrimui naudojamama procedūra *proc gradar*. Kiekvienam klasteriui atskirai yra nubraižomi žvaigždės formos grafikai naudojant *chart* sakinį. Šių grafikų pagalba galima palyginti konkretaus vieno klasterio matavimų vidurkius su visos imties matavimų vidurkiais ir taip nustatyti kiekvienam klasteriui dominuojančius teršalus. Pavaizduoti duomenys yra standartizuoti.

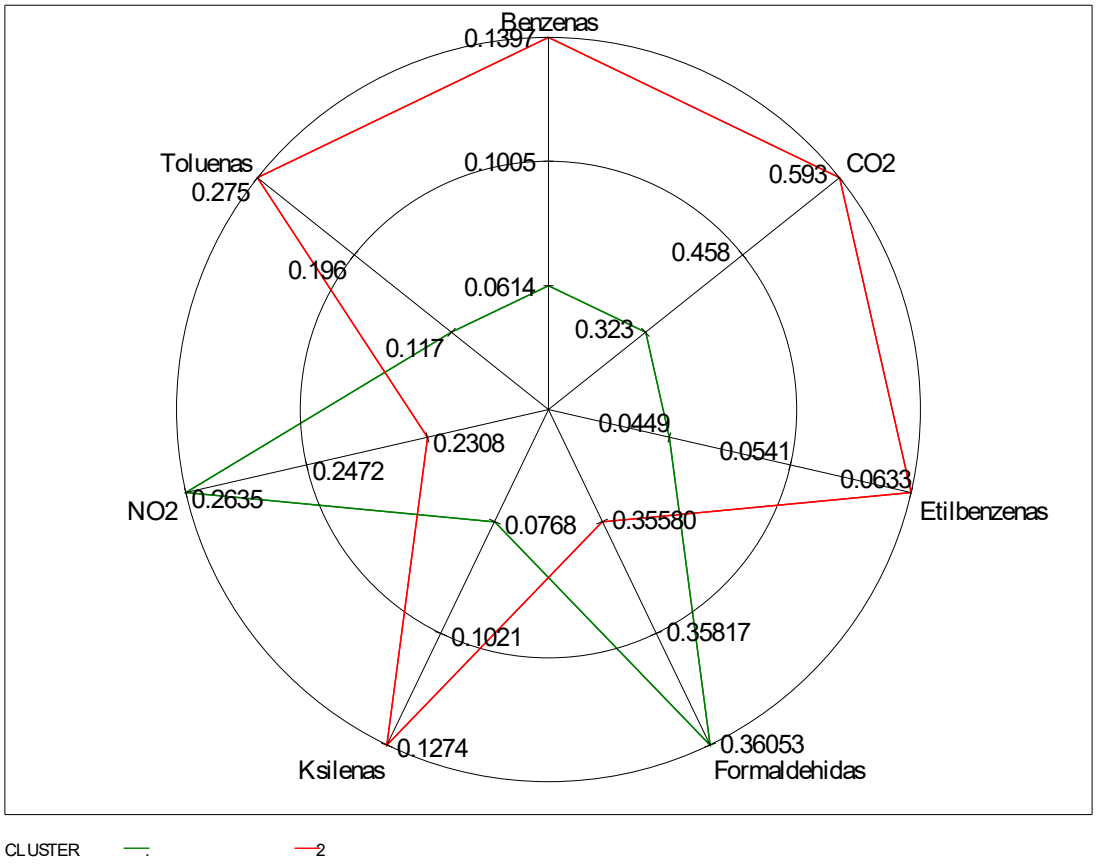
Nagrinėjamam klasteriui (3.5 pav.) būdingas didesnis nei vidutinis azoto dioksido kiekis. Vidutinis NO_2 kiekis tiriamoje imtyje 0.264 $\mu g/m^3$, o pirmame klasteryje šio teršalo vidurkis sudaro net 0.294 $\mu g/m^3$.



3.5 pav. Teršalų būdingų I-am klasteriui nustatymas (prieš renovaciją Lietuvoje)

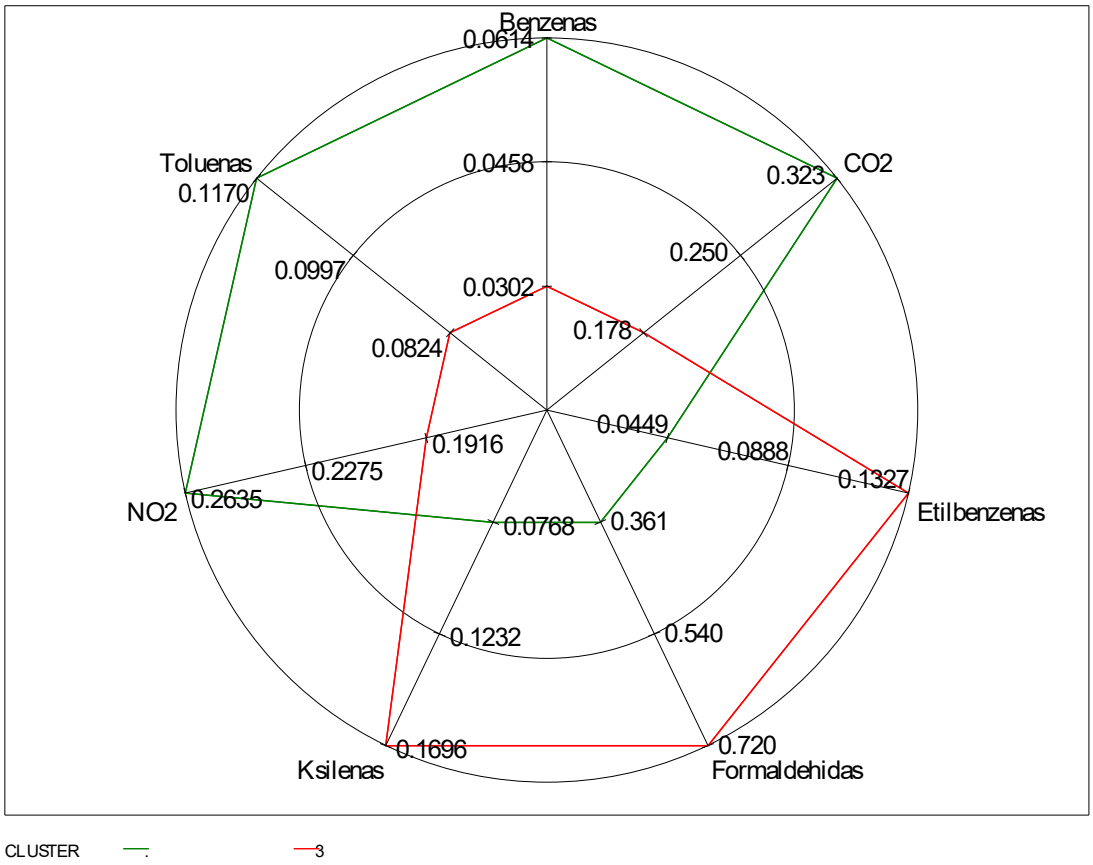
Toliau apžvelgsime antrojo (3.6 pav.) ir trečiojo (3.7 pav.) klasterių matavimų vidurkius ir nustatysime jiems būdingus teršalus.

Iš žvaigždės formos grafikų matoma, kad antrajam klasteriui būdingas didesnis nei vidutinis anglies dioksido, benzeno, tolueno, etilbenzeno bei ksileno kiekis (3.6 pav.). Anglies dioksido vidutinis kiekis tarp visų stebinių yra 0.323 $\mu g/m^3$, tuo tarpu nagrinėjamame klasteryje 0.593 $\mu g/m^3$. Vidutinis benzeno kiekis tiriamoje imtyje 0.061 $\mu g/m^3$, o antrame klasteryje šio teršalo vidurkis yra daugiau nei dvigubai didesnis - sudaro net 0.1397 $\mu g/m^3$. Tolueno vidutinis kiekis tiriamų patalpų ore sudaro 0.117 $\mu g/m^3$ tuo tarpu šiame klasteryje teršalo vidurkis yra taip pat didesnis nei dvigubas - 0.275 $\mu g/m^3$. Etilbenzeno vidutinis kiekis tarp visų stebinių yra 0.045 $\mu g/m^3$, o klasteryje siekia 0.063 $\mu g/m^3$. Ir ksileno kiekis tiriamoje imtyje 0.077 $\mu g/m^3$, o antrame klasteryje jo vidurkis sudaro 0.127 $\mu g/m^3$.



3.6 pav. Teršalų būdingų II-am klasteriui nustatymas (prieš renovaciją Lietuvoje)

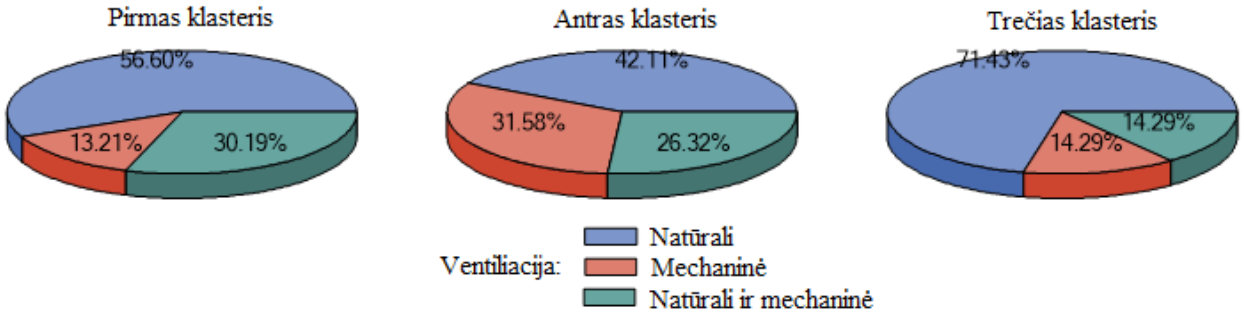
Trečiajam klasteriui būdingi didesni nei vidutiniai etilbenzeno, ksileno ir formaldehido kiekiai (3.7 pav.). Matome, kad vidutinis etilbenzeno kiekis tarp visų stebinių yra 0.045 $\mu g/m^3$, tuo tarpu nagrinėjamame klasteryje 0.133 $\mu g/m^3$. Ksileno kiekis tiriamoje imtyje 0.077 $\mu g/m^3$, o trečiame klasteryje šio teršalo vidurkis yra daugiau nei dvigubai didesni - sudaro net 0.170 $\mu g/m^3$. Formaldehido vidutinis kiekis tiriamų patalpų ore sudaro 0.361 $\mu g/m^3$, o tuo tarpu trečiame klasteryje šio teršalo vidurkis yra 0.720 $\mu g/m^3$.



3.7 pav. Teršalų būdingų III-am klasteriui nustatymas (prieš renovaciją Lietuvoje)

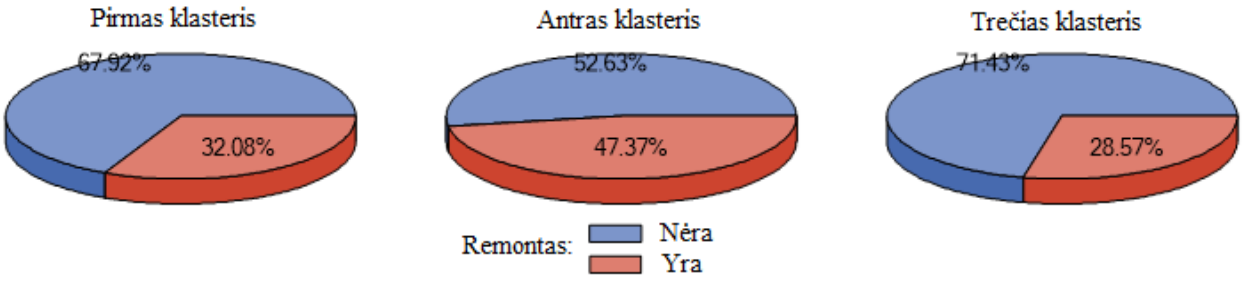
Toliau yra tiriamos pagal į grupę patekusių butų nagrinėtas buitines sąvybes klasterių buitinės specifikacijos. Tyrimui nanudojama *proc gchart* procedūra, o *pie3D* sakinys naudojamas skritulinėms diagramoms realizuoti. Šiuo tyrimu susiejami klasteriui būdingi patalpų oro teršalai su galimais taršos šaltiniais.

Didesnis nei 71% procentas nagrinėjamo trečiojo klasterio butų naudoja natūralią ventiliaciją (3.8 pav.). Formaldehido ir kitų lakiųjų organinių junginių kiekio ore padidėjimo priežastis gali būti būtent prastas patalpų vėdinimas, nes ore lieka medienos konservantų, valymo ir dezinfekavimo priemonių, oro gaiviklių bei aerozolinių purškalų nuodingos liekanų. Todėl reikėtų nepamiršti kuo dažniau vėdinti patalpas, ypač jei jos neseniai remontuotos.

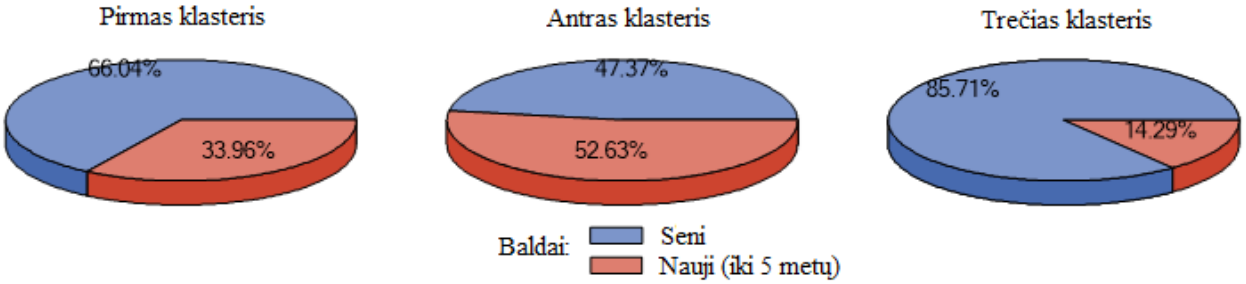


3.8 pav. Ventiliacijos tipas skirtinguose klasteriuose (prieš renovaciją Lietuvoje)

Antrajame klasteryje vyrauja didesnis nei kituose kiekis suremontuotų (3.9 pav.) ir naujai apstatytų baldais butų (3.10 pav.). Būtent tai galėtų būti padidėjusio benzeno kiekio priežastis. Todėl naujai apstačius butą, reikia jį daug vėdinti ir pasistengti jame nebūti bent jau pirmomis dienomis, pirmomis savaitėmis.

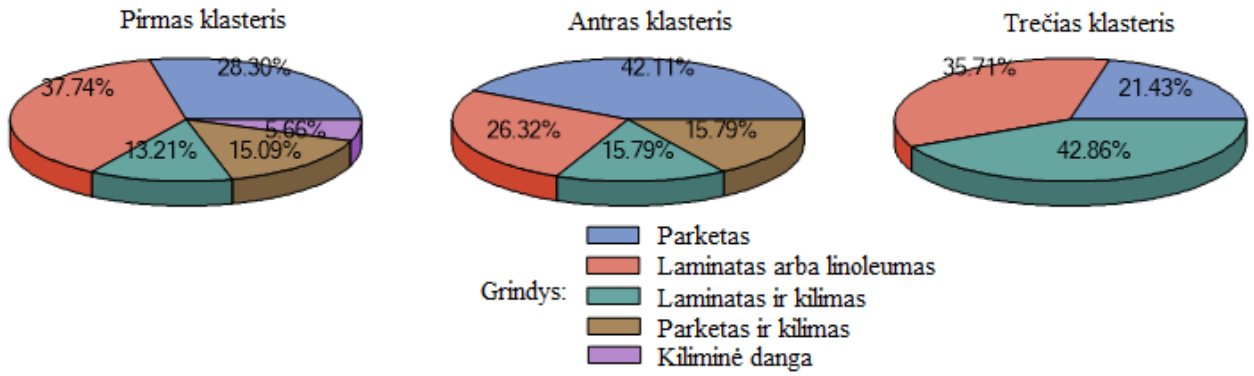


3.9 pav. Remonto vertinimas skirtinguose klasteriuose (prieš renovaciją Lietuvoje)



3.10 pav. Baldų amžius būdingas skirtingiems klasteriams (prieš renovaciją Lietuvoje)

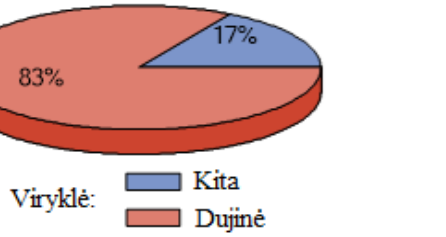
Lyginant dominuojančius grindų tipus (3.11 pav.) skirtinguose tiriamų klasterių butuose, matoma, jog antrame klasteryje dominuoja natūralios parketo grindys ir parketas su kilimu. Šių tipų grindys sudaro 57.90% tiriamo klasterio butų grindų. Trečio klasterio tiriamuose butuose vyrauja laminatu, linoleumu ir kilimais padengtos grindys, kas bendrai sudaro net 78.57%. Atlikus tyrimą, pastebėta, jog trečiajame klasteryje stipriai padidėjęs formaldehido kiekis, kurį, remiantis Jungtinių Amerikos Valstijų aplinkos apsaugos agentūros duomenimis [24], sąlygoja grindys išklotos laminatu ar linoliaumu.



3.11 pav. Grindų tipas skirtinguose klasteriuose (prieš renovaciją Lietuvoje)

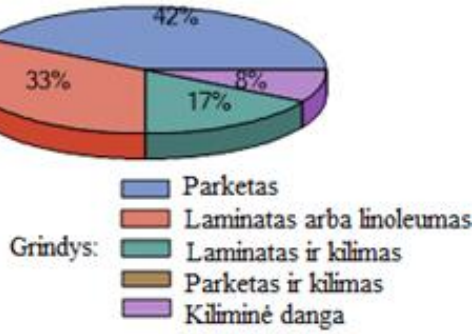
Atlikus patalpų oro taršos tyrimą po daugiabučių renovacijos galima apžvelgti kaip pasikeitė oro kokybė. Išsamūs tyrimo rezultatai po renovacijos pateikiami priede (2 Priedas). Plačiau bus nagrinėjamas tik 2-iasis klasteris, kadangi jam būdingi taršos rodikliai yra ženkliai didesni, nei kituose klasteriuose.

Matome (3.12 pav.), jog 83% tiriamąjį klasterį sudarančių butų naudoja dujines virykles. Tai galėtų būti padidėjusio NO_2 bei CO_2 kiekio priežastis, kadangi dujiniai prietaisai yra šių dujų taršos šaltiniai.



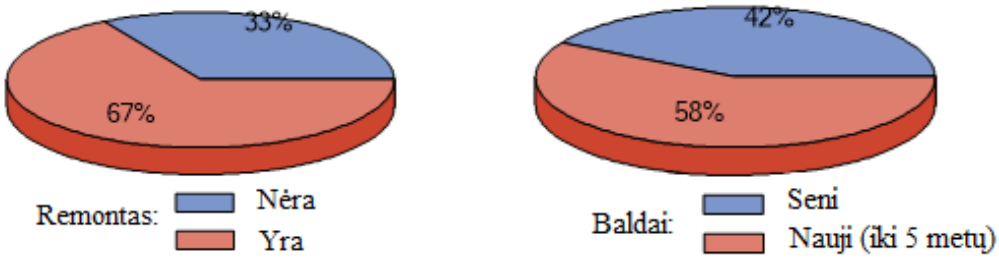
3.12 pav. Viryklės tipas II-ame klasteryje (po renovacijos Lietuvoje)

Formaldehido koncentracija paprastai yra didesnė būstuose, kuriuose yra laminuota grindų danga. Lyginant antrame klasteryje būdingus grindų tipus (3.13 pav.) matoma, jog butuose dominuoja būtent laminatu, linoliaumu ir kilimais dengtos grindys, kas ir sąlygoja ore padidėjusį formaldehido kiekį.



3.13 pav. Grindų tipas II-ame klasteryje

Oro tarša taip pat didesnė būstuose, kuriuose grindų danga keista anksčiau nei prieš metus. Antrame klasteryje vyrauja naujai apstatyti baldais butai, kuriems atliktas remontas (3.14 pav.).



3.14 pav. Remonto ir baldų amžiaus vertinimas II-ame klasteryje

Atliekant kai kuriuos patalpų vidaus apdailos darbus gali išsiskirti dideli formaldehido kiekiai. Formaldehidas naudojamas baldų, plastikinių apdailos medžiagų, presuotos medienos, izoliacinių

medžiagų gamyboje. Tikėtina, kad seniai remontuotame name formaldehido ore bus daug mažiau vien dėl to, kad jis jau bus išgaravęs. Taip pat įtakos turi ir baldų iš medienos plokščių kiekis: kuo naujesnės minėtos medžiagos, tuo didesnė formaldehido koncentracija patalpos ore.

Toliau atliksime analogišką tyrimą, kad palygintume Lietuvos ir Suomijos patalpų oro kokybę.

3.3. Aprašomosios statistikos analizė

Pirmiausiai apžvelgę apskaičiuotas skaitines patalpų oro teršalų charakteristikas, pastebime, jog etilbenzeno kiekis ore buvo matuojamas ne visuose stebiniuose. Dėl šios priežasties atliekant analizę nebus vertinamas jo poveikis, kadangi rizikuojama prarasti nemažai duomenų ir taip gauti klaidingas išvadas. Duomenys pateikti 3.7 lentelėje.

3.7 lentelė. Skaitinės kintamųjų charakteristikos (prieš renovaciją Suomijoje)

Kintamasis	Imties plotis	Vidurkis	Standartinis nuokrypis	Mažiausia reikšmė	Didžiausia reikšmė
Temperatūra	167	22.69786	1.04679	19.26136	24.92524
Drėgmė	167	27.65964	6.85406	14.16200	49.85143
CO_2	164	713.06653	202.99105	385.89012	1686
Benzenas	156	2.77516	4.27896	0	47.66004
Toluenas	157	3.79787	4.92877	0.14561	43.40517
Etilbenzenas	146	0.63612	1.18691	0	10.26848
Ksilenas	157	2.02117	3.95462	0	33.22224
Formaldehidas	164	21.10163	12.55204	4.08561	88.90861
NO_2	162	6.87395	3.89347	0	32.54000
Radonas	168	70.05952	67.05523	20.00000	350.00000

Matoma, kad santykinės oro drėgmės vidurkis siekia tik 27.66%, o mažiausias užfiksuotas dydis vos 14.16%.

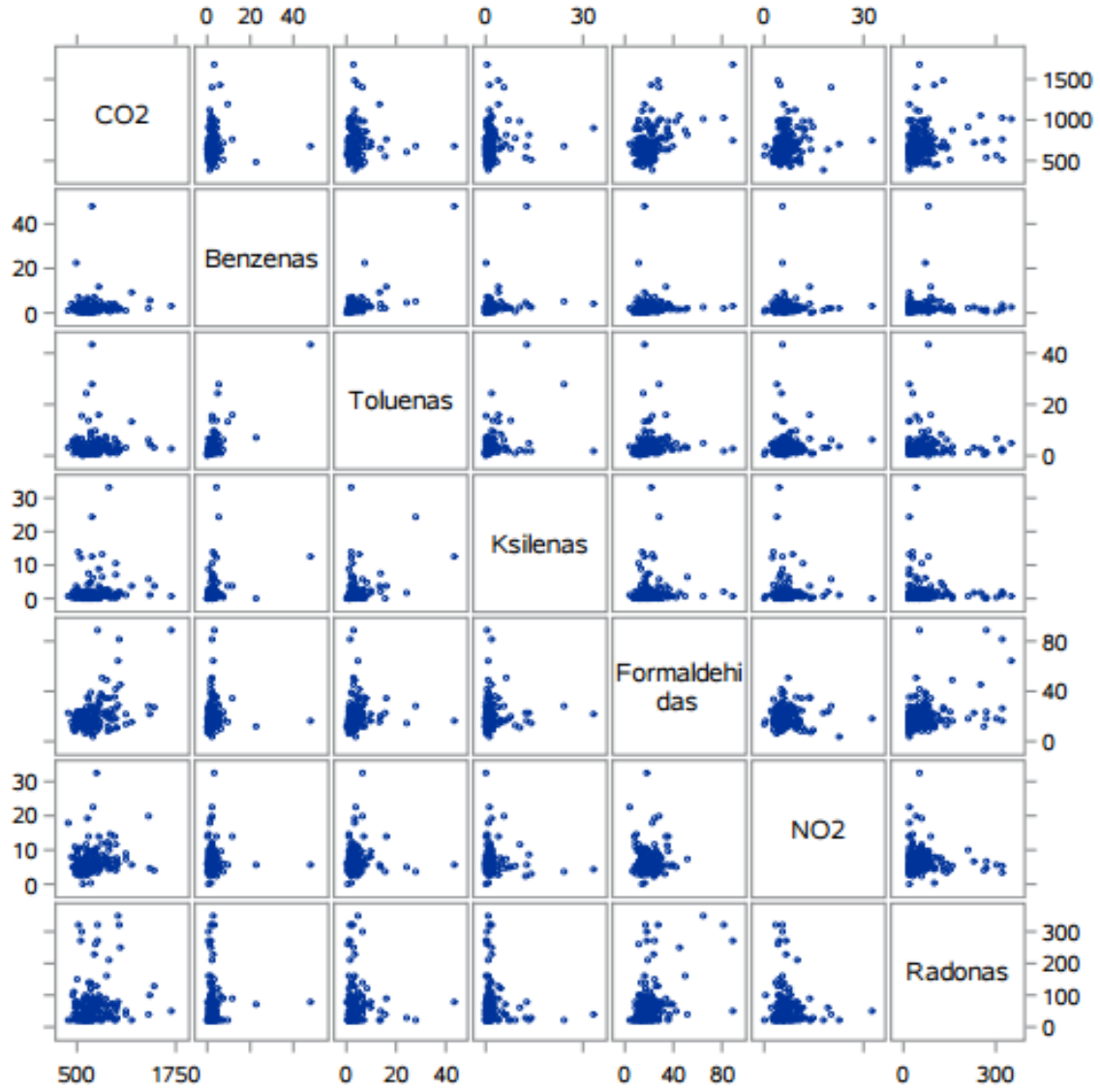
3.3.2. Duomenų paruošimas klasterizavimui

Tarp kintamųjų reikšmingas tarpusavio ryšys nenustatytas, todėl nėra informacijos dubliavimo ir negalime atmesti nei vieno kintamojo (3.8 lentelė).

3.8 lentelė. Pirsono koreliacijos tarp kintamųjų matavimas (prieš renovaciją Suomijoje)

	CO_2	Benzenas	Toluenas	Ksilenas	Formaldehidas	NO_2	Radonas
CO_2	1.00000	0.02116	0.02153	0.10733	0.43805 <.0001	0.15154	0.13056
Benzenas	0.02116	1.00000	0.69914 <.0001	0.25345	-0.00835	-0.00921	-0.01700
Toluenas	0.02153	0.69914 <.0001	1.00000	0.35836 <.0001	0.06841	0.05263	-0.04364
Ksilenas	0.10733	0.25345	0.35836 <.0001	1.00000	0.03425	-0.07116	-0.12345
Formaldehidas	0.43805 <.0001	-0.00835 0.9182	0.06841 0.3976	0.03425	1.00000	0.01597	0.38693 <.0001
NO_2	0.15154	-0.00921	0.05263	-0.07116	0.01597	1.00000	-0.13191
Radonas	0.13056	-0.01700	-0.04364	-0.12345	0.38693 <.0001	-0.13191	1.00000

Pateikiama pradinių duomenų išsidėstymą ir tarpusavio santykį apibūdinanti sklaidos diagrama (3.15 pav.).



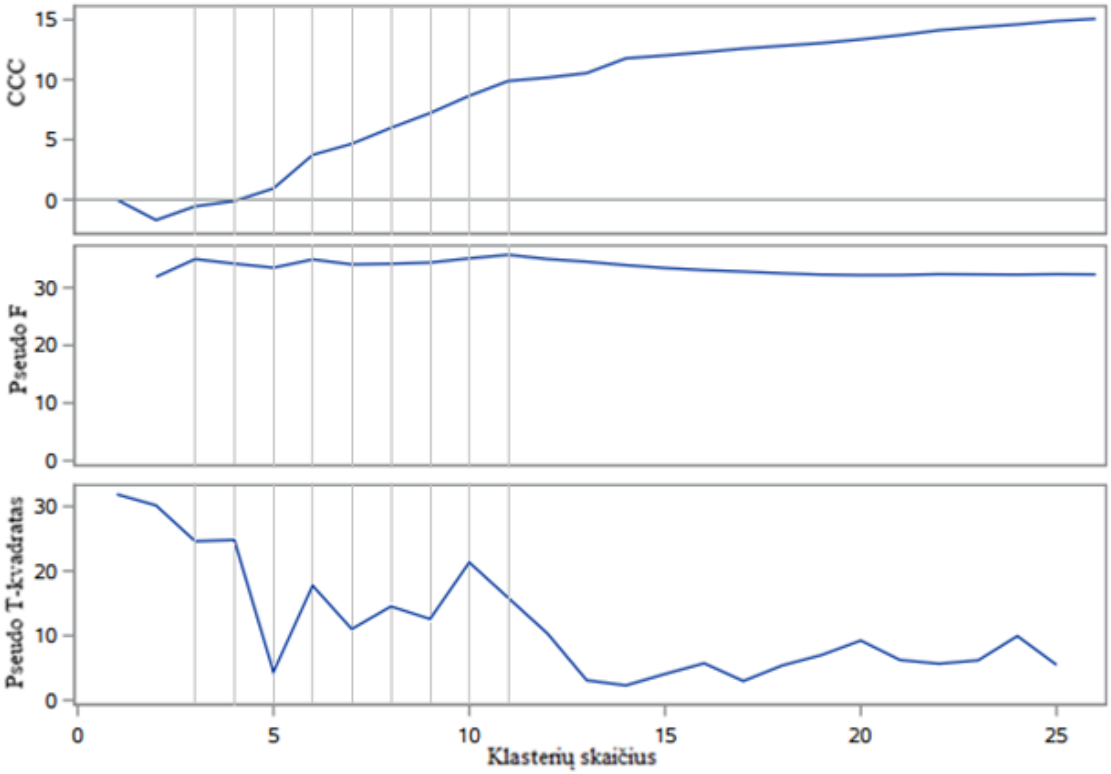
3.15 pav. Pradinių duomenų sklaidos diagrama (prieš renovaciją Suomijoje)

3.3.3. Klasterizavimas naudojant hierarchinius metodus

Atliekant hierarchinį klasterizavimą naudojami jungimo metodai taikant Vordo matą. Pateikiamas klasterių jungimo protokolo fragmentas (3.9 lentelė) ir diagrama (3.16 pav.).

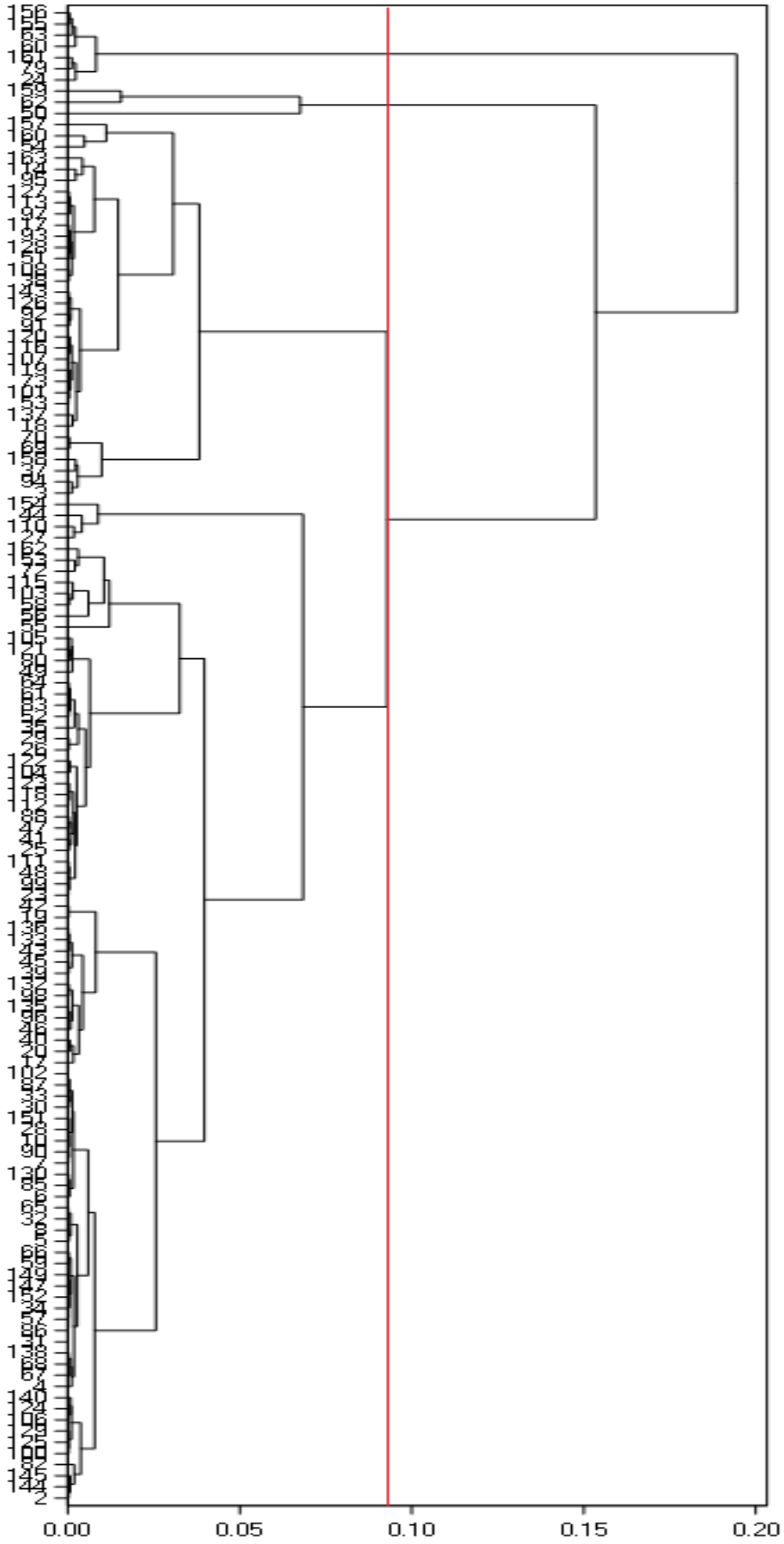
3.9 lentelė. Klasterių jungimo protokolas (prieš renovaciją Suomijoje)

Klasterių skaičius	Sujungti klasteriai		Dažnis	Kubinis klasterizavimo kriterijus	Pseudo F	Pseudo T^2
133	OB31	OB86	2	.	215	.
132	OB10	OB28	2	.	202	.
...	...	...	...	...	...	...
11	OB62	OB159	2	9.89	35.7	.
10	CL20	CL19	54	8.66	35.1	21.3
9	CL12	CL14	28	7.22	34.4	12.6
8	CL22	CL13	32	5.98	34.1	14.5
7	CL16	CL9	34	4.66	34.0	11.0
6	CL10	CL8	86	3.73	34.9	17.8
5	OB50	CL11	3	0.93	33.5	4.4
4	CL6	CL17	90	-.12	34.2	24.8
3	CL4	CL7	124	-.55	35.0	24.6
2	CL3	CL5	127	-1.7	31.9	30.2
1	CL2	CL18	134	0.00	.	31.9



3.16 pav. Kriterijai klasterių skaičiaus nustatymui (prieš renovaciją Suomijoje)

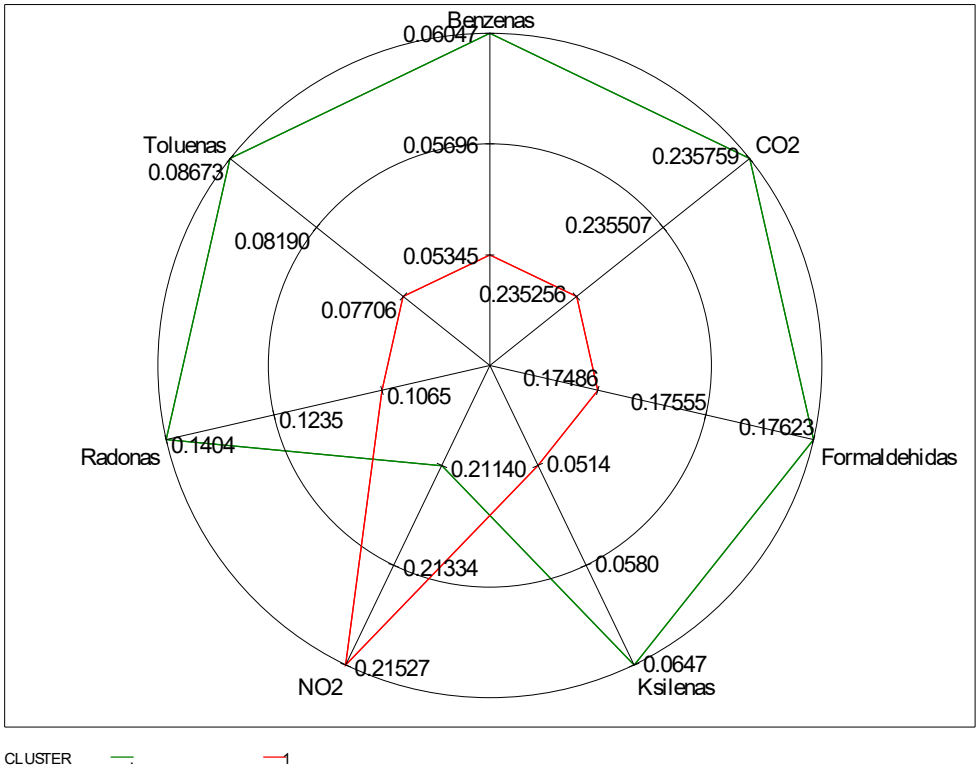
Nustatytas optimalus klasterių skaičius trys (3.17 pav.).



3.17 pav. Dendograma (prieš renovaciją Suomijoje)

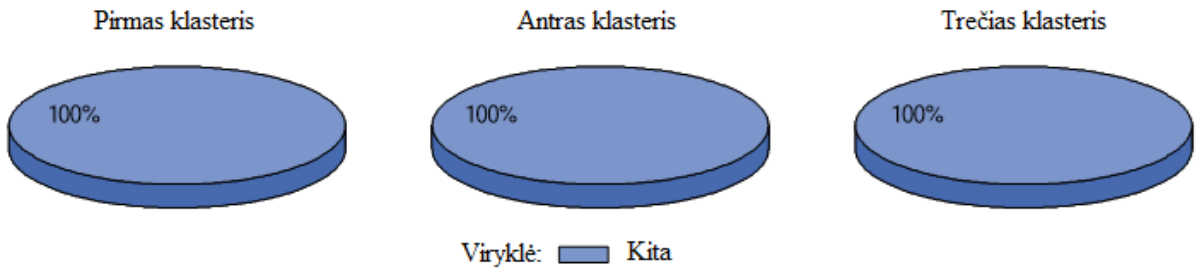
Sudarytų klasterių požymių tyrimas.

Pateikti žvaigždės formos grafikai, kurie rodo kiekvienam klasteriui būdingus teršalus.

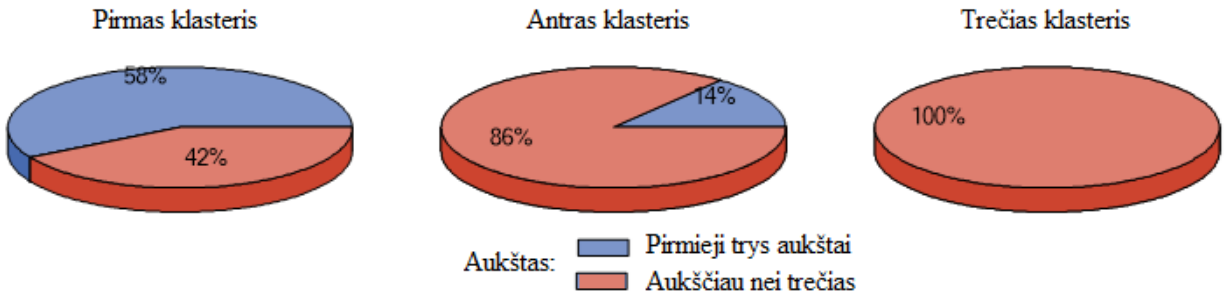


3.18 pav. Teršalų būdingų I-am klasteriui nustatymas (prieš renovaciją Suomijoje)

Pirmajam nagrinėjamam klasteriui (3.18 pav.) būdingas nors ir nežymus, tačiau didesnis nei vidutinis azoto dioksido kiekis, ko pagrindinė priežastis gali būti naudojamos dujinės viryklės. Tačiau pastebima, kad Suomijoje tokios viryklės nenaudojamos ne viename tirtame bute (3.19 pav.). Todėl šiuo atveju dujų kiekis gali skirtis dėl buto aukšto (3.20 pav.). Kuo aukščiau butas yra, tuo mažesnė tikimybė, jog NO_2 bus padidėjęs.

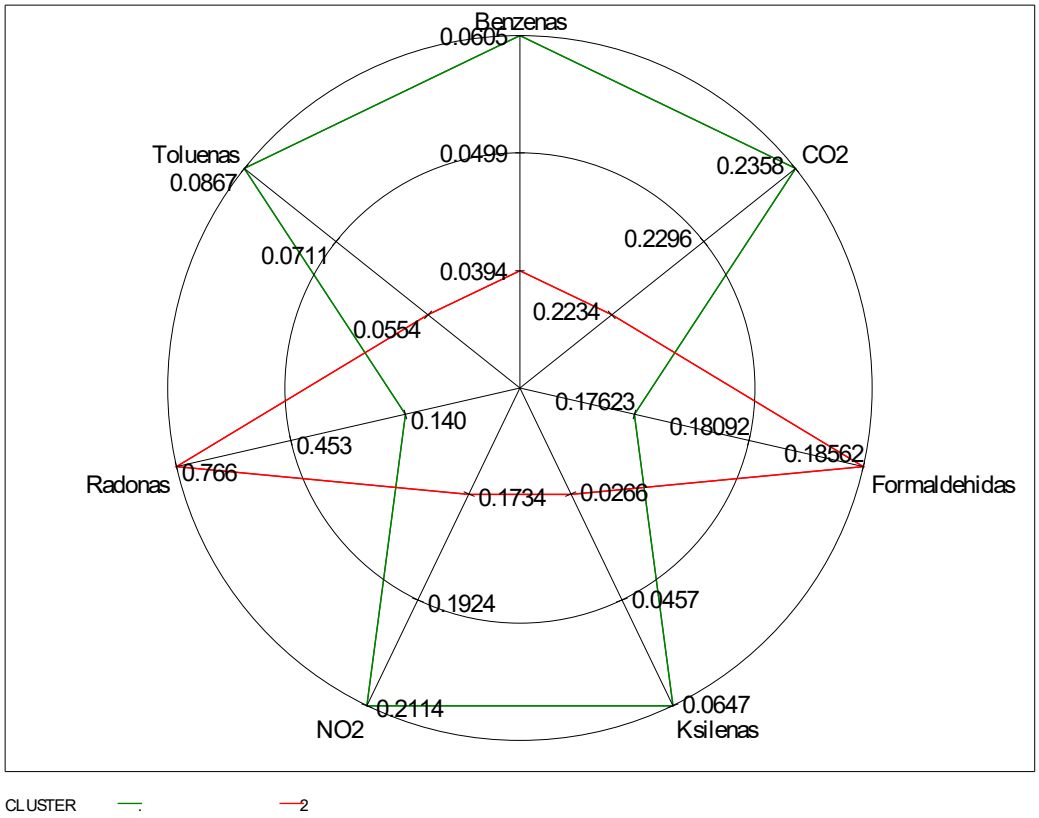


3.19 pav. Viryklės tipas skirtinguose klasteriuose (prieš renovaciją Suomijoje)



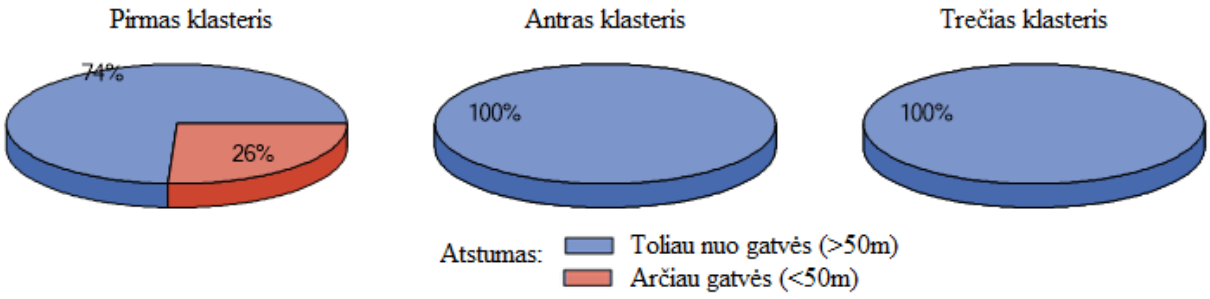
3.20 pav. Aukšto įtaka skirtinguose klasteriuose (prieš renovaciją Suomijoje)

Antrojo klasterio butuose padidėjęs radono ir formaldehido kiekis (3.21 pav.).



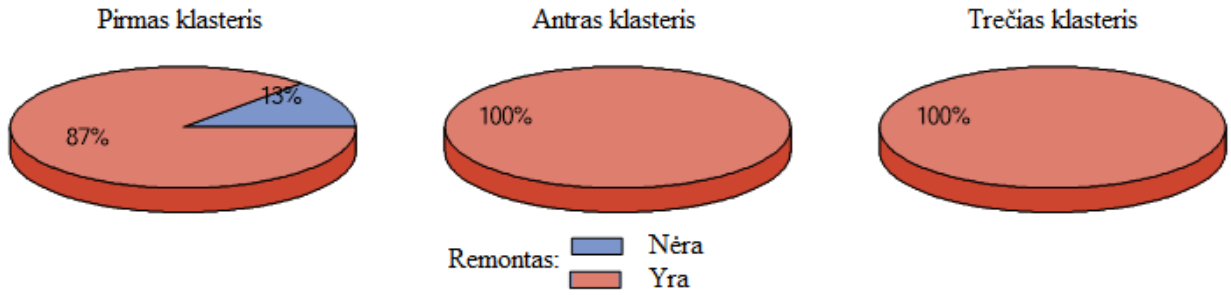
3.21 pav. Teršalų būdingų II-am klasteriui nustatymas (prieš renovaciją Suomijoje)

Atitinkamus vidutinius anglies dioksido ir azoto dioksido kiekius patalpose antrame klasteryje lemia butų išsidėstymas atokiau nuo gatvės (3.22 pav.).



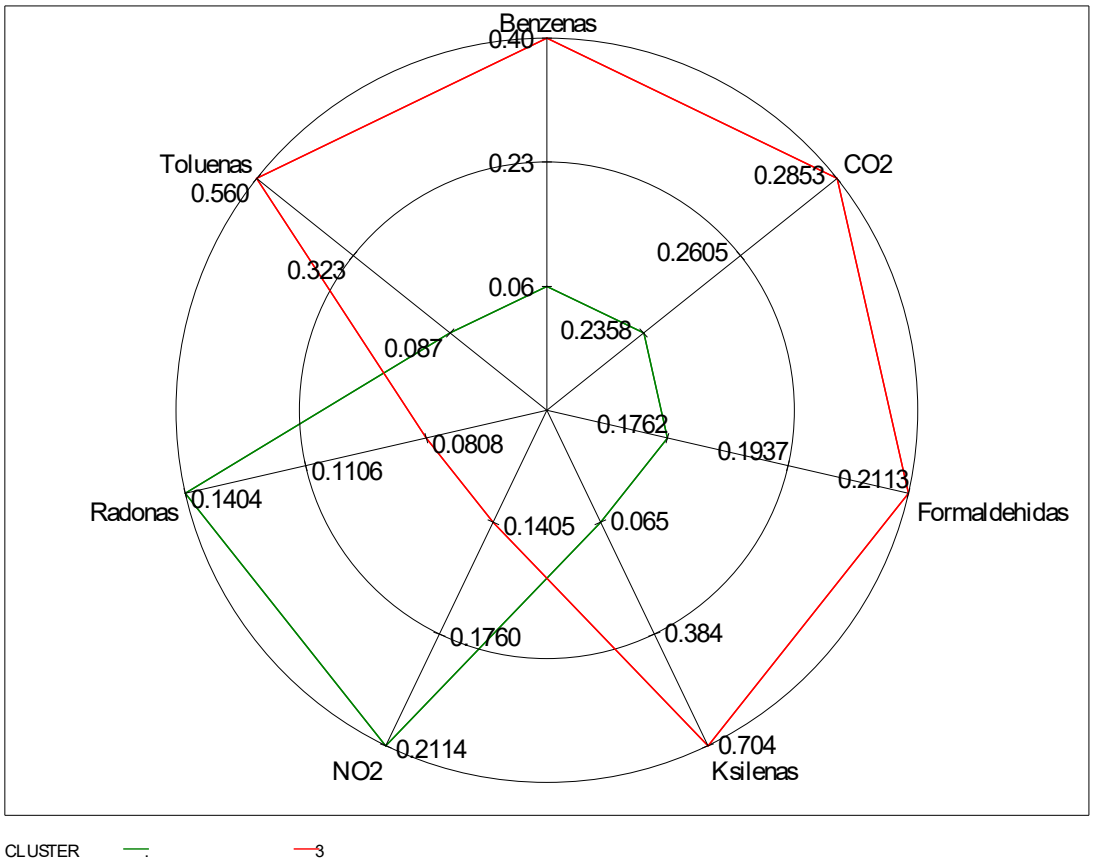
3.22 pav. Butų atstumas iki gatvės skirtinguose klasteriuose (prieš renovaciją Suomijoje)

Tačiau nuo kitų dujinių teršalų tai neapsaugo. Matome, jog visuose tiek antro, tiek trečio klasterio butuose buvo atliktas remontas, kurio metu išsiskiria didelis formaldehido kiekis. Jis naudojamas baldų, plastikinių apdailos medžiagų, presuotos medienos, izoliacinių medžiagų gamyboje. Tikėtina, kad seniai remontuotame name formaldehido ore bus daug mažiau vien dėl to, kad jis jau bus išgaravęs. Būtent tai ir paaiškina padidėjusio radono ir formaldehido kiekio priežastį antrame klasteryje (3.23 pav.).



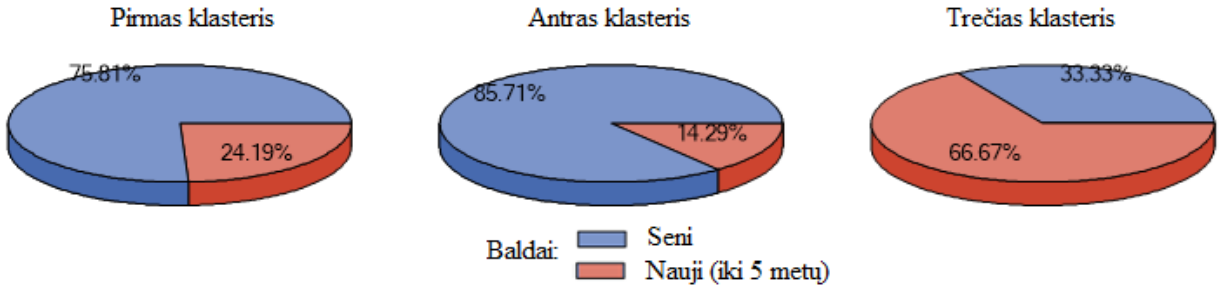
3.23 pav. Remonto vertinimas skirtinguose klasteriuose (prieš renovaciją Suomijoje)

Trečiajame klasteryje stipriai viršinami vidutiniai benzeno, tolueno, ksileno, anglies dioksido ir taip pat formaldehido kiekiai (3.24 pav.).



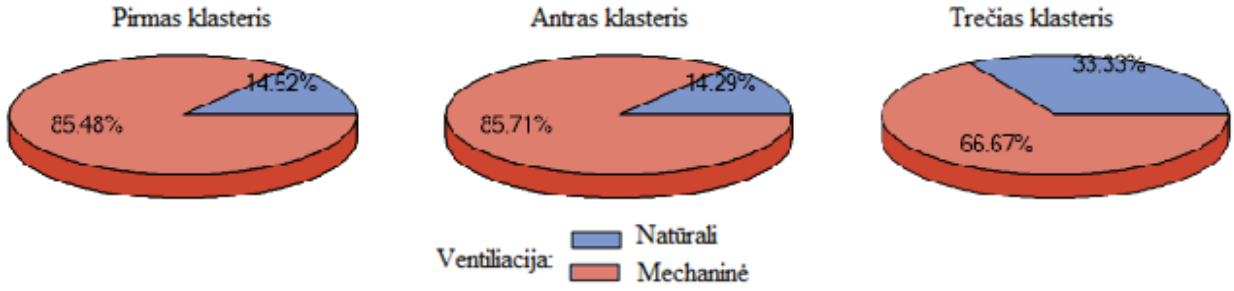
3.24 pav. Teršalų būdingų trečiam klasteriui nustatymas (prieš renovaciją Suomijoje)

Tai yra ne tik jau minėtų butų remontų priežastis. Šiuose butuose buvo naujai pakeista ir didžioji dalis baldų (3.25 pav.). Formaldehido koncentracijai patalpų ore taip pat įtakos turi ir baldų iš medienos plokščių kiekis: kuo naujesnės minėtos medžiagos, tuo didesnis išskiriamų dujų kiekis.



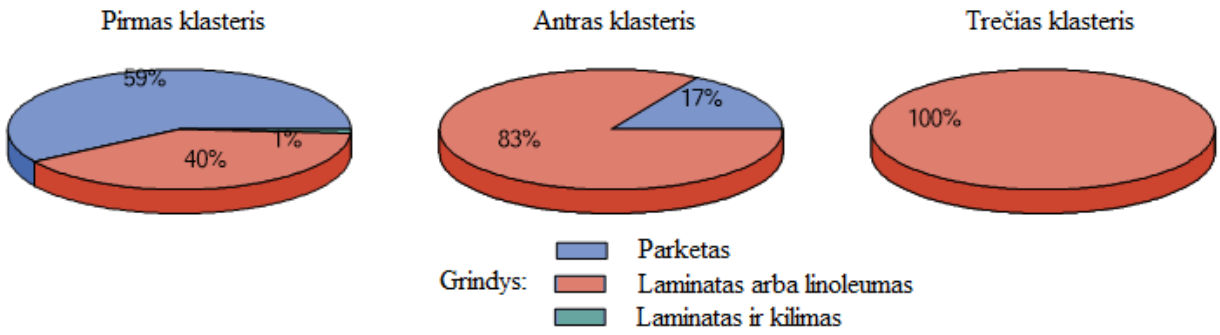
3.25 pav. Baldų amžius būdingas skirtingiems klasteriams (prieš renovaciją Suomijoje)

Todėl po remonto ir baldų atnaujinimo, dar svarbesnį vaidmenį atlieka namų ventiliacija. Tačiau būtent šiame klasteryje net trečdalio butų patalpos vėdinamos natūraliu būdu, kuris neužtikrina kokybiško oro namuose (3.26 pav.). Kaip ir buvo minėta, būtent prastas patalpų vėdinimas ir įtakoja tiek formaldehido, tiek ir kitų lakiųjų organinių junginių kiekio ore padidėjimą.



3.26 pav. Ventiliacijos tipas skirtinguose klasteriuose (prieš renovaciją Suomijoje)

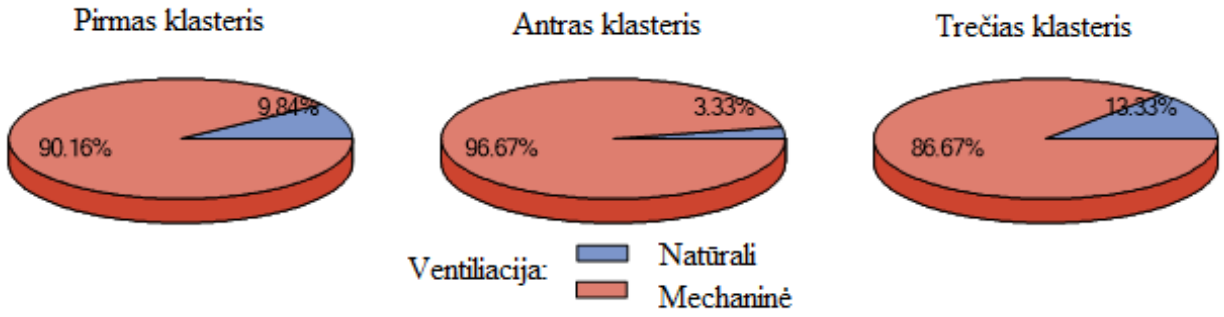
Formaldehido kiekį ore lemia ir laminatu arba linoliaumu dengtos grindys.



3.27 pav. Grindų tipas skirtinguose klasteriuose (prieš renovaciją Suomijoje)

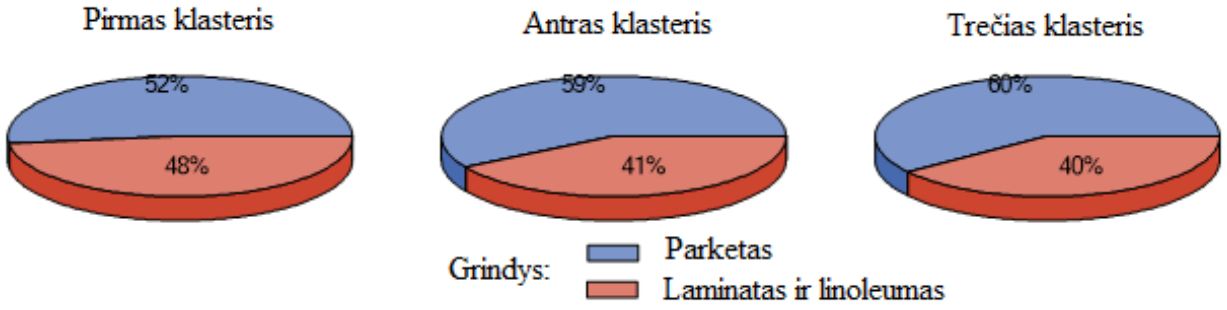
Atlikus patalpų oro taršos tyrimą po daugiabučių renovacijos Suomijoje (4 Priedas) galima apžvelgti kaip pasikeitė situacija.

Matome, kad daugelis gyventojų savo butuose įsirengė mechaninę vėdinimo sistemą, kuri užtikrina efektyvų buto vėdinimą bei pastovią oro cirkuliaciją, kas lemia švaresnį orą ir geresnę savijautą (3.28 pav.).



3.28 pav. Ventiliacijos tipas skirtinguose klasteriuose (po renovacijos Suomijoje)

Dalis gyventojų taip pat atsisakė laminatu ir limoleumu dengtų grindų ir pasirinko natūralaus medžio parketą, kuris nedaro įtakos patalpų oro taršai.



3.29 pav. Grindų tipas skirtinguose klasteriuose (po renovacijos Suomijoje)

Atlikus daugiabučių renovaciją pastebima geresnė patalpų oro kokybė. Nors remontui naudojamos medžiagos ir baldai išskiria didesnį lakiųjų medžiagų kiekį, tačiau bendras taršos lygis sumažėjo dėl efektyvesnės vėdinimo sistemos bei natūralaus medžio grindų, kuris neišskiria kenksmingų formaldehido dujų.

3.4. Patalpų oro taršos Lietuvoje ir Suomijoje palyginimas

Atlikus patalpų oro būklės analizę ir identifikavus taršos šaltinius, palyginsime oro kokybę Lietuvoje ir Suomijoje. Tyrimo rezultatai apie patalpų oro būklę prieš atliekant renovaciją ir po jos pateikti 3.10 lentelėje. Matoma, kaip pasikeitė kiekvienas matuotas rodiklis po atliktų remonto darbų skirtingose šalyse.

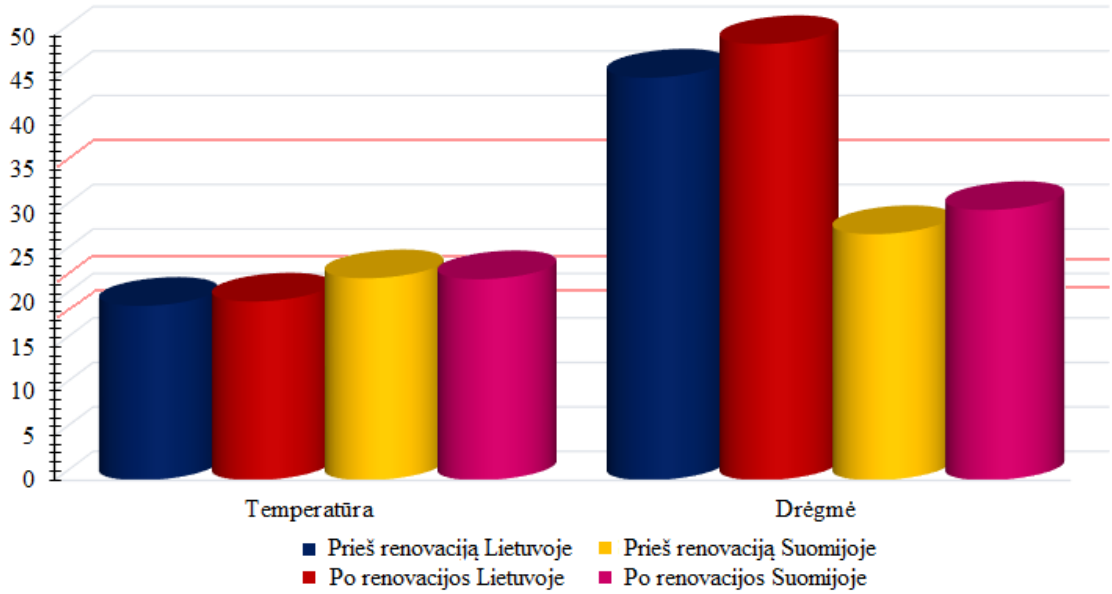
Apžvelgus duomenis, matomas akivaizdus skirtumas tarp patalpų oro kokybės Lietuvoje ir Suomijoje. Žemiau pateiktos diagramos, atspindinčios oro taršos situaciją skirtingose šalyse. Atlikus

daugiabučių renovaciją vidutinė oro temperatūra išliko beveik nepakitusi, tačiau Lietuvoje ji keliais laipsniais žemesnė. Suomijoje santykinė patalpų drėgmė vis dar neatitinka gyvenamųjų patalpų mikroklimato higienos normų, kadangi mažiausia užfiksuota vertė turėtų siekti bent 35%.

3.10 lentelė. Oro būklę lemiančių charakteristikų palyginimas

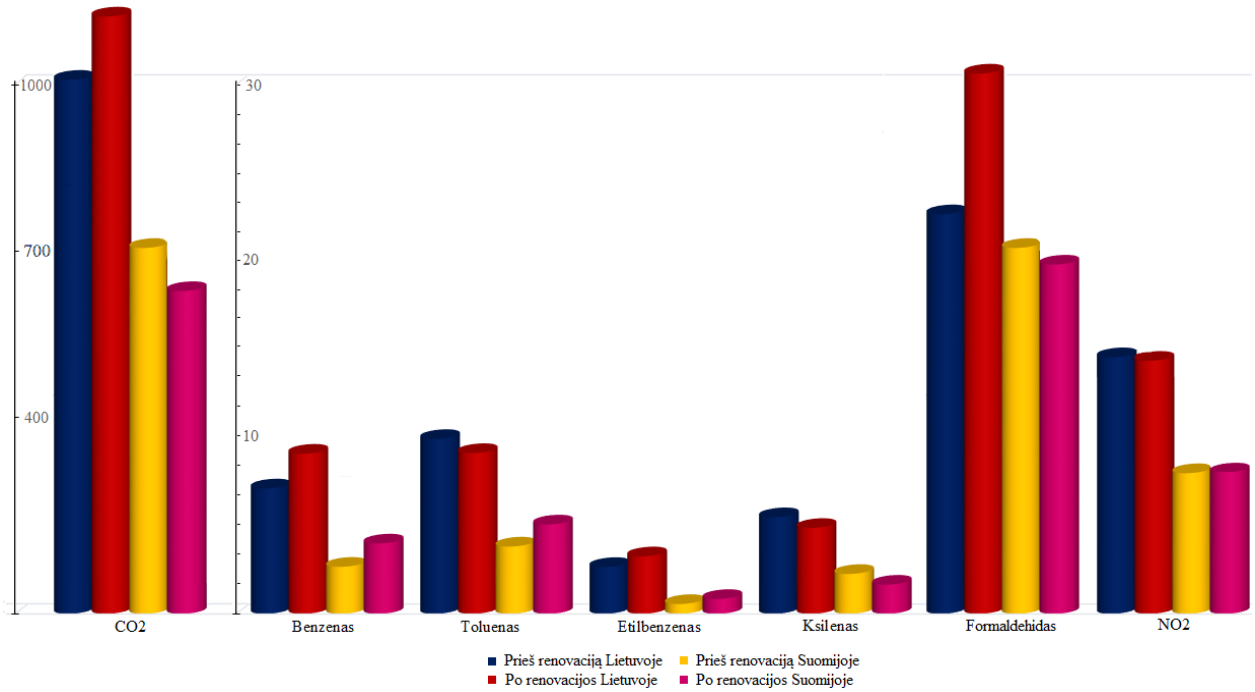
	Lietuva		Suomija	
Kintamasis	**Imties dydis**	**Vidutinė reikšmė**	**Imties dydis**	**Vidutinė reikšmė**
Temperatūra	92 ⇨ 65	19.598 ⇧ 20.086	167 ⇨ 122	22.698 ⇩ 22.592
Drėgmė	92 ⇨ 65	45.251 ⇧ 49.016	167 ⇨ 122	27.660 ⇧ 30.366
CO_2	89 ⇨ 65	1010 ⇧ 1077	164 ⇨ 119	713.067 ⇩ 671.858
Benzenas	95 ⇨ 63	6.842 ⇧ 8.008	156 ⇨ 122	2.775 ⇧ 3.557
Toluenas	95 ⇨ 63	10.120 ⇩ 9.638	157 ⇨ 122	3.798 ⇧ 4.539
Etilbenzenas	95 ⇨ 63	1.879 ⇧ 2.237	146 ⇨ 122	0.636 ⇧ 0.806
Ksilenas	95 ⇨ 63	3.926 ⇩ 3.567	157 ⇨ 122	2.021 ⇩ 1.662
Formaldehidas	95 ⇨ 65	23.160 ⇧ 31.277	164 ⇨ 123	21.102 ⇩ 18.769
NO_2	93 ⇨ 65	13.987 ⇩ 13.738	162 ⇨ 118	6.874 ⇩ 6.822
Radonas	43 ⇨ 35	27.539 ⇧ 41.214	168 ⇨ 103	70.060 ⇩ 65.146

Diagramoje matomas rodiklių kitimas bei ryškus santykinės drėgmės patalpose Lietuvoje ir Suomijoje skirtumas (3.30 pav.).



3.30 pav. Oro temperatūros ir santykinės drėgmės apžvalga

Įvertinus kiekvieną dujinį teršalą individualiai, pastebime, jog po atliktos daugiabučių renovacijos patalpų oro kokybė Lietuvoje nepagerėjo, o štai Suomijoje, nors ir ne ženkliai, tačiau beveik visi tirti rodikliai tapo mažesni.



3.31 pav. Dujinių oro teršalų apžvalga

Atlikus patalpų oro taršos tyrimą, nustatėme, jog norint pagerinti oro kokybę gyvenamojoje aplinkoje gyventojai turi atidžiai rinktis medžiagas, baldus, valymo ar kitas buitines priemones ir įvertinti kokį poveikį, patalpų oro taršai, gali turėti jų veikla. Žmogus būdamas užterštose ir nevėdinamose patalpose pradeda jausti galvos, akių, nosies skausmus, gerklės dirginimą, jį ima kamuoti sausas kosulys, sloga, odos išsausėjimas ir niežėjimas, pykinimas, taip pat sutrinka miegas, jaučiamas nuovargis, žmogus negeba susikoncentruoti, sutelkti dėmesio. Išvengti minėtų nemalonių sveikatos sutrikimų, galima:

- Efektyviai vėdinant patalpas;
- Pastoviu drėgmės lygio palaikymu;
- Reguliariu grindų plovimu;
- Vengiant sintetinių oro gaiviklių ir valiklių;
- Atidžiai renkantis patalpų vidaus apdailos medžiagas.

3.5. Klasterizavimas naudojant nehierarchinius metodus

Kaip jau ir buvo minėta, nehierarchinio klasterizavimo pagrindinis skirtumas nuo hierarchinių metodų yra tas, kad dar prieš pradedant analizę, tyrėjas jau turi žinoti norimą klasterių skaičių. Tyrime atliekant nehierarchini klasterizavimą naudojami k-vidurkių bei k-artimiausių kaimynų metodai.

Radono įtaka nėra vertinama kadangi prarandama daug stebinių, kas darytų įtaką išvadoms. Paprastai nehierarchiniai metodai yra skirti didelės imties tyrimams.

K-vidurkių metodas

K-vidurkių metodui realizuoti yra naudojama procedūra *proc fastclus*. Tačiau pirmiausia naudojant kubinio klasterinzavimo kriterijų, psiaudo F statistiką bei ekspertinį vertinima nustatomas grupių skaičius, į kurias bus suskirtyta imtis. Klasterių skaičius nurodomas sakiniu *maxclusters*.

Kubinis klasterizavimo kriterijus rekomenduoja stebinius grupuoti į 8-9 klasteriu. Psiaudo F kriterijus rodo, kad optimalus klasterių skaičius turėtų būti arba 2, arba 5, arba 7. Šiuo atveju optimalus klasterių skaičius turėtų būti du, kadangi didžiausia F kriterijaus reikšmė yra 20.43. Ši prielaida sutampa ir su ekspertiniu vertinimu.

Įmtis grupuojama į du klasterius:

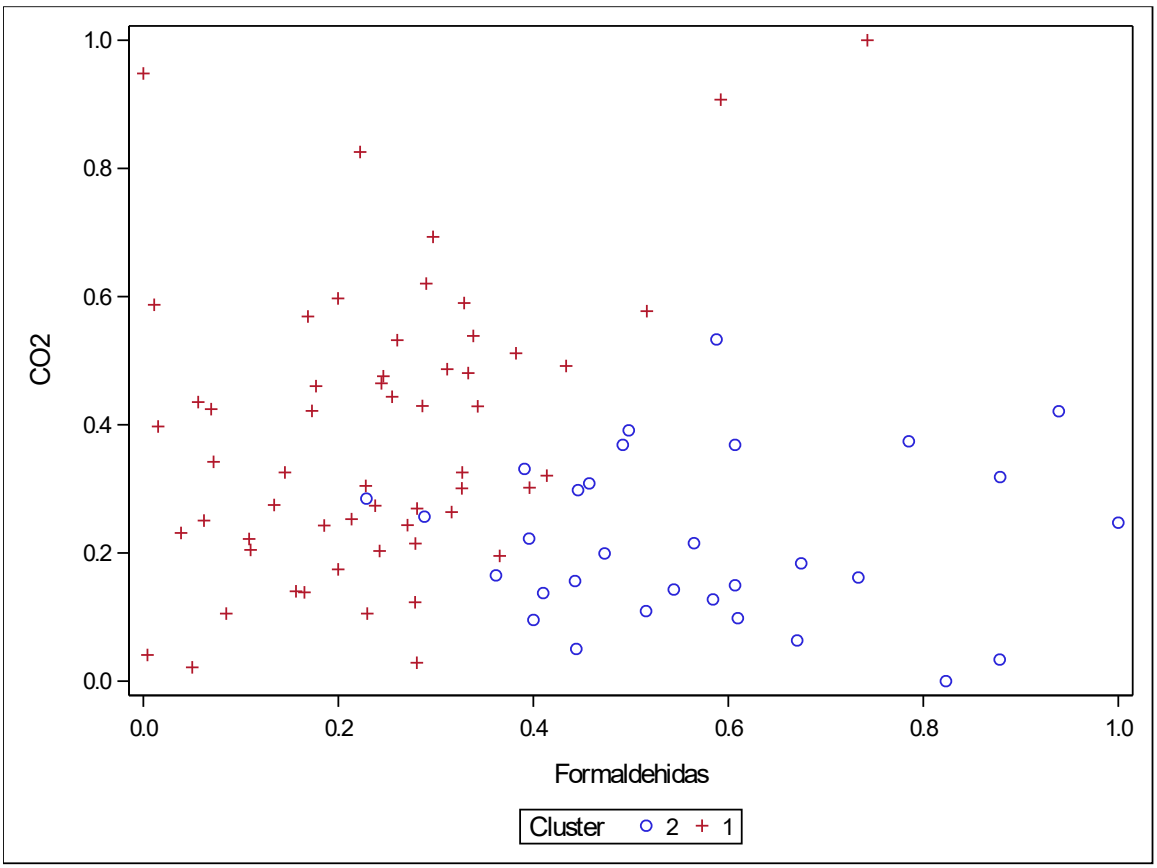
- 60 stebinių priklauso I-ajam klasteriui;
- 36 stebiniai priklauso II-jam klasteriui.

Klasterius sudarančių požymių charakterizuojantys vidurkiai pateikiami (3.11 lentelė).

3.11 lentelė. Klasterių vidurkiai

Klasteris	CO_2	Benzenas	Toluenas	Etilbenzenas	Ksilenas	Formaldehidas	NO_2
1	0.38209209	0.069424468	0.100820766	0.033288444	0.06943430	0.231568204	0.26580842
2	0.21485237	0.049750371	0.131405989	0.076354607	0.11348683	0.584695949	0.30191144

Toliau braižomi požymių tarpusavio priklausomybės sklaidos grafikai, kurie skirstomi pagal grafikus. Tam naudojama funkcijos *proc sgplot* komanda *scatter*. Kadangi duomenų pasiskirtymas į grupes nėra gerai matomas iš visų perpektyvų, pateikiama sklaidos diagrama apie CO_2 ir formaldehido priklausomybę (3.32 pav.). Nors ir tiriami duomenys prastai atsiskiria į klasterius, neaiški jų forma ir dydis, tačiau galima pamatyti duomenų grupavimą į būtent du klasterius.



3.32 pav. CO_2 priklausomybės nuo formaldehido sklaidos diagrama

Sudarytų klasterių požymių tyrimas

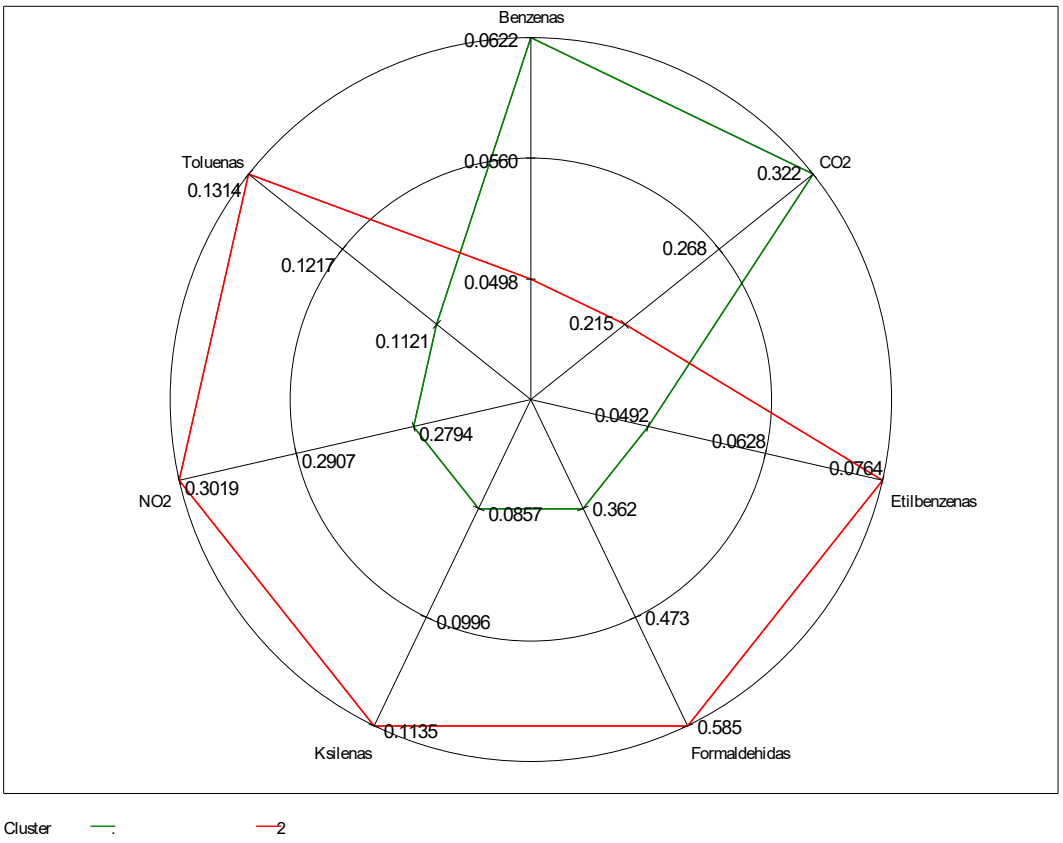
Klasterių požymių tyrimui naudojama procedūra *proc gradar*. Žvaigždės formos grafikai buvo braižomi atskirai kiekvienam klasteriui naudojant *chart* sakinį (3.33 pav.). Šių grafikų tikslas, kaip ir jau buvo minėta, yra palyginti konkretaus klasterio matavimų vidurkius su visos imties matavimų vidurkiais ir taip nustatyti teršalus būdingus būtent tiriamam klasteriui. Grafikams braižyti buvo naudojami standartizuoti duomenys.

Toliau plačiau nagrinėjamas tik antrasis klasteris, o pirmojo klasterio požymių tyrimo rezultatai pateikiami (5 priedas).

(3.33 pav.) matoma, kad antrajam klasteriui būdingas didesnis nei vidutinis tolueno, etilbenzeno, ksileno, formaldehido ir NO_2 kiekis patalpų ore.

Tolueno vidutinis kiekis tiriamų patalpų ore sudaro 0.1121 μg/m3 tuo tarpu šiame klasteryje teršalo vidurkis yra 0.1314 μg/m3. Etilbenzeno vidutinis kiekis klasteryje siekia 0.0764 μg/m3, o tarp visų stebinių yra 0.0492 μg/m3. Vidutinis ksileno kiekis tiriamoje imtyje 0.0857 μg/m3, o pirmame klasteryje šio teršalo vidurkis sudaro 0.1135 μg/m3. Formaldehido vidutinis kiekis tiriamų patalpų

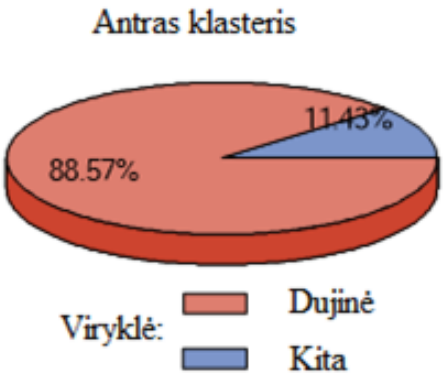
ore sudaro 0.362 μg/m3, o tuo tarpu klasteryje šio teršalo vidurkis yra 0.585 μg/m3. Azoto dioksido vidutinis kiekis tarp visų stebinių yra 0.2794 μg/m3, tuo tarpu nagrinėjamame klasteryje 0.3019 μg/m3.



3.33 pav. Antrojo klasterio požymių tyrimas

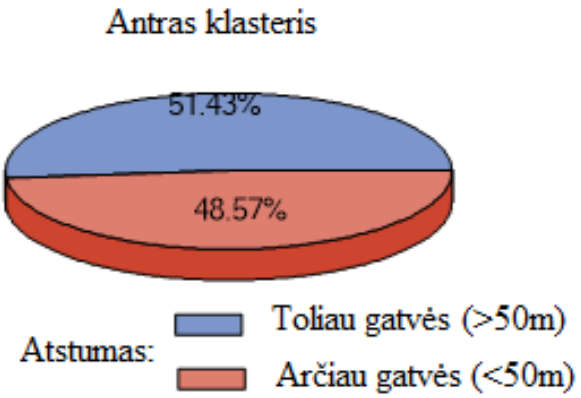
Toliau nagrinėjamos antrojo klasterio buitinės sąlygos. Jos galimai ir yra oro taršos priežastys. Pirmojo klasterio buitinių specifikacijų tyrimo rezultatai taip pat pateikiami (5 priede).

(3.34 pav.) matoma, kad net 85.57% tiriamąjį klasterį sudarančių butų maistą gamina naudojant dujines virykles, kas galėtų būti padidėjusio NO_2 kiekio priežastis, kadangi būtent dujiniai prietaisai ir yra pagrindiniai NO_2 taršos šaltiniai.



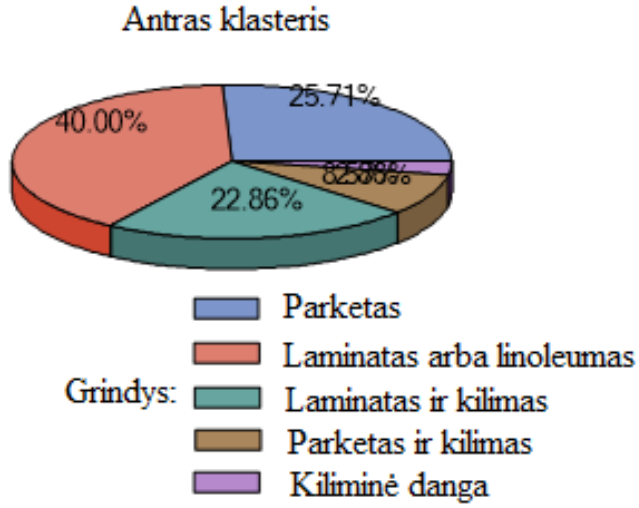
3.34 pav. Viryklės tipas II-ame klasteryje

CO_2 bei NO_2 kiekio patalpų ore svyravimo priežastis taip pat gali būti astumas iki gatvės. (3.35 pav.) duomenimis pusė butų yra toliau nei 50 metrų nuo gatvės (51.43%), o kita pusė arčiau nei 50 metrų nuo gatvės (48.57%).



3.35 pav. Butų atstumas iki gatvės II-ame klasteryje

Vienas iš pagrindinių rodiklių, sąlygojančių padidėjusį formaldehido bei kitų lakiųjų organinių junginių kiekį ore, gali būti grindų tipas. 22.86% tiriamo klasterio butų turi natūralias parketo grindis, likusią dalį sudaro laminatu, linoliaumu ir kilimais padengtos grindys (3.36 pav.).



3.36 pav. Grindų tipas II-ame klasteryje

K-artimiausių kaimynų metodas

K-artimiausių kaimynų metodas realizuojamas procedūra *proc modeclus*, parinktimi *k* nurodome norimą kaimynų skaičių. Tiriamasis stebinys gali priklausyti bet kuriam klasteriui, taigi randami jo k-artimiausi kaimynai. Šis stebinys priskiriamas tam klasteriui, su kuriuo jį sujungus tankis būna didžiausias ir ne mažesnis nei sujungus su bet kuriuo kitu kaimynu.

Klasterių skaičiaus nustatymas

Nagrinėti rezultatai su dviem, trim, keturiais, penkiais, šešiais, septyniais, aštuoniais ir devyniais kaimynais. Klasterių tikslinimas baigiamas, kai klasteriai nusistovi, o jų struktūra nebekinta.

Optimalus klasterių skaičius yra keturi. Tolimesnis tyrimas vykdomas kai artimiausių kaimynų yra penki, kadangi šiuo atveju klasterių tankiai yra didžiausi, be to vertinant ekspertiškai parinkus penkis kaimynus sklaidos diagramoje matosi aiškios klasterių stuktūros.

Kitus skaičius kaimynų būtų galima atmesti, nes vertinant pagal klasterius suformuotas sklaidos diagramas matoma, kad grupavimas yra netinkamas. Klasteriai taip pat suformuojami su mažesniu tankiu.

Žemiau pateikiami klasterius sudarančių objektų skaičius ir tankiai (3.12 lentelė).

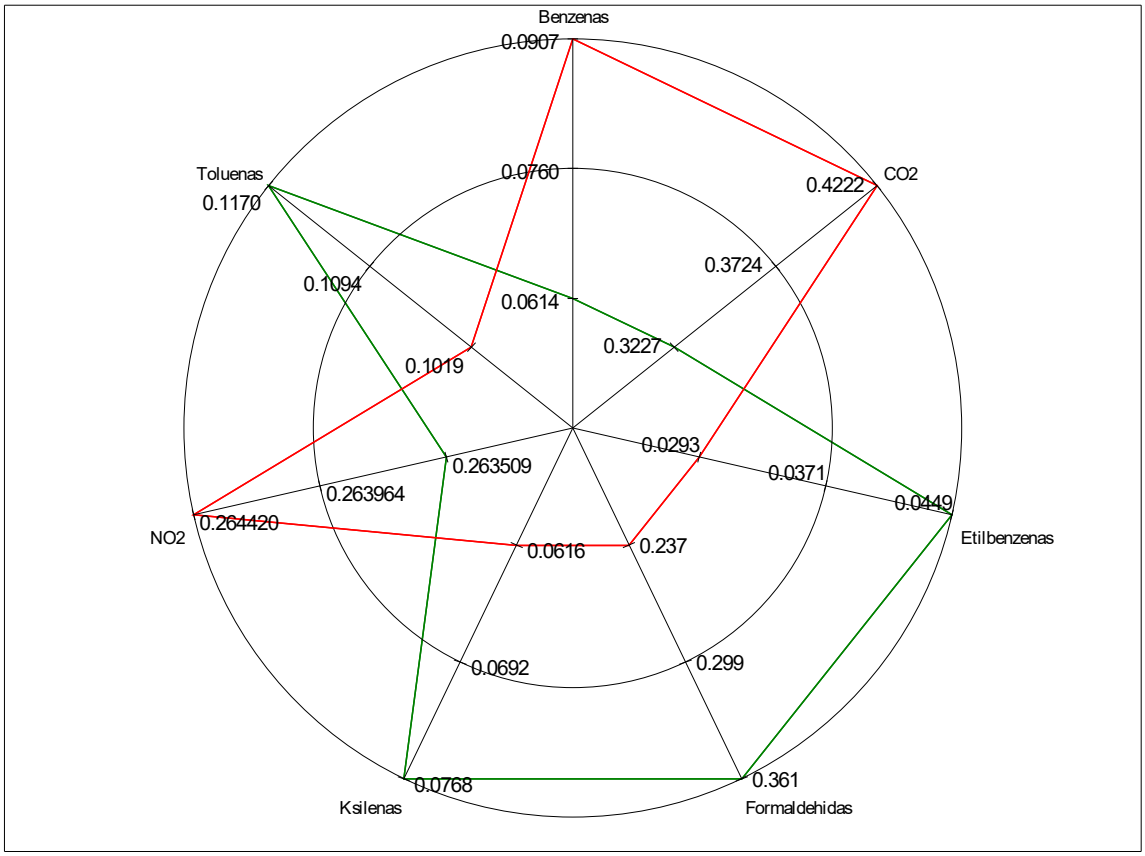
3.12 lentelė. Klasterių apibūdinimas kai K=5

Klasteris	Objektų skaičius	Maksimalus numatomas tankis
1	39	23159.9631
2	26	14401.585
3	5	9106.66571
4	16	3156.12933

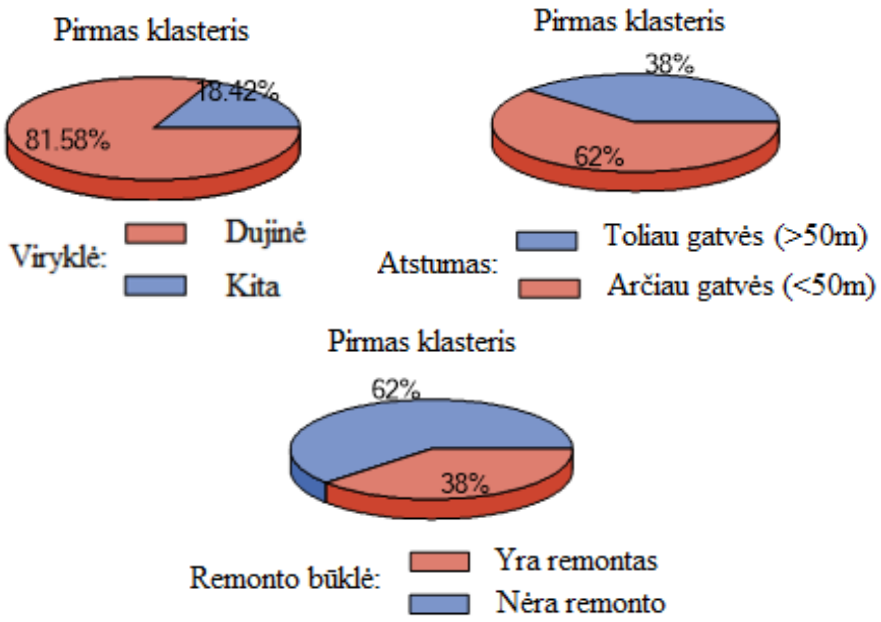
Sudarytų klasterių požymių tyrimas

Tyrimui nustatyti klasterių požymius naudojama procedūra *proc gradar*. Taip pat braižomi žvaigždės formos grafikai atskirai kiekvienam klasteriui. Plačiau nagrinėjamas tik pirmasis klasteris. Kitų klasterių rezultatai pateikiami (5 priedas).

(3.37 pav.) matoma, jog pimajam klasteriui būdingas didesnis nei vidutinis NO_2, benzeno bei CO_2 kiekis. Vidutinis benzeno kiekis tiriamoje imtyje 0.0614μg/m3, o pirmame klasteryje šio teršalo vidurkis sudaro 0.090μg/m3. NO_2 vidutinis kiekis tiriamų patalpų ore sudaro 0.263509μg/m3 tuo tarpu tiriamajame klasteryje šio teršalo vidurkis yra nežymiai didesnis jis sudaro 0.264420μg/m3. Vidutinis CO_2 kiekis tarp visų objektų yra 0.3227μg/m3, o pirmame klasteryje 0.4222μg/m3.



3.37 pav. Pirmojo klasterio požymių tyrimas



3.38 pav. Pirmajam klasteriui būdingos buitinės specifikacijos

(3.38 pav.) matome, kad net 81.58% tiriamąjį klasterį sudarančių butų naudoja dujines virykles ir daugiau nei pusė butų (62%) yra šalia gatvės. Tai galėtų būti padidėjusio NO_2 bei CO_2 kiekio priežastis. 38% tiriamojo klasterio butų kelių metų bėgyje buvo atlikti remonto darbai, kas galėtų būti padidėjusio benzeno kiekio priežastis.

Išvados

Šiame darbe aptariamos technologijos, skirtos didelių duomenų analizei atlikti, pateikiama duomenų analizavimo įrankių apžvalga, apžvelgiami duomenų analizės uždaviniai, tokie kaip klasterizavimas, statistinė ir vizuali analizė, kurie leidžia aptikti, išrinkti ir efektyviai panaudoti naudingą informaciją.

SAS programinio įrankio pagalba atlikta patalpų oro taršos duomenų analizė Lietuvoje ir Suomijoje analizė trimis skirtingais klasterizavimo metodais.

Atlikus hierarchinį klasterizavimą taikant jungimo metodą nustatyta, kad geriausių rezulatų pasiekta naudojant Vordo atstumo matą, Jungimo metodas sprendimą apie rezultatus palieka pačiam tyrėjui, nes tyrėjas sprendžia žiūrėdamas į dendogramą. Priimta, kad paskirstymas į tris klasterius yra optimalus. Grupuojant tiriamąją imtį naudojant nehierarchinius K-vidurkių bei K-artimiausių kaimynų metodus nustatytas optimalus klasterių skaičius yra atitinkamai du ir keturi klasteriai. Taigi, visais trimis metodais nustatytas optimalus klasterių skaičius skiriasi.

Nagrinėjant suformuotų klasterių teršalų kiekius nustatyta, kad skirtingoms objektų grupėms būdingi skirtingų oro teršalų kiekių padidėjimai bei daugeliu atvejų esant kurio nors iš oro teršalų kiekio padidėjimui bute galima rasti taršos šaltinį, literatūroje įvardinamą kaip šio teršalo skleidėju.

Gauti rezultatai atskleidė pagerėjusias gyvenimo sąlygas po pastatų renovacijos. Daugumoje tirtų butų vidaus aplinkos kokybės rodikliai atitiko nustatytas nacionalines rekomendacines normas, tačiau po daugiabučių renovacijos kai kurie vidaus taršos šaltiniai suintensyvėjo.

Vidaus oro kokybė pagėrėjo pastatuose, kuriuose įrengtas mechaninis vėdinimas, o pastatuose su natūraliu vėdinimu tendencija buvo priešinga. Norint užtikrinti švarų patalpų orą ir gerą savijautą po renovacijos vėdinimo sistema turi būti patikrina ir tinkamai subalansuota.

Literatūros sąrašas

1. The Inside Story: A Guide to Indoor Air Quality [interaktyvus]. Consumer Production Safety Commission, 2016 [žiūrėta 2018-02-20]. Prieiga per: http://www.cpsc.gov/en/Safety-Education/SafetyGuides/Home/The-Inside-Story-A-Guide-to-Indoor-Air-Quality/

2. SHAPLEY, Dan. 6 Surprising Sources of Home Air Pollution [interaktyvus]. Popular Mechanics, 2013 [žiūrėta 2018-01-20]. Prieiga per: http://www.popularmechanics.com/home/how-to/g54/indoor-airpollution-47020404/

3. An Introduction to Indoor Air Quality [interaktyvus]. Enviromental Protection Agency [žiūrėta 2018-01-18]. Prieiga per: https://www.epa.gov/indoor-air-quality-iaq/introduction-indoor-air-quality

4. 7 million premature deaths annually linked to air pollution [interaktyvus]. World Health Organisation, 2014 [žiūrėta 2018-01-13]. Prieiga per: http://www.who.int/mediacentre/news/releases/2014/airpollution/en/

5. Lietuvos higienos norma HN 23:2007 „Cheminių medžiagų profesinio poveikio ribiniai dydžiai. Matavimo ir poveikio vertinimo bendrieji reikalavimai“, patvirtinta Lietuvos Respublikos sveikatos apsaugos ministro 2007 m. spalio 15 d. įsakymu Nr. V-827/A1-287. Valstybės žinios. 2007;108-4434. [žiūrėta 2018-02-22]. Prieiga per: http://www3.lrs.lt/pls/inter3/dokpaieska.showdoc_l?p_id=306641&p_query=&p_tr2=

6. ŠEDUIKYTĖ, L. and BLIŪDŽIUS, R. Apdailos Medžiagų Ir Aplinkos Parametrų Įtaka Patalpų Mikroklimato Kokybei: Monografija, 2007. [žiūrėta 2018-01-15]. Prieiga per: http://vddb.laba.lt/fedora/get/LT-eLABa-0001:B.03~2007~ISBN_978-9955-25-268-9/DS.001.0.01.BOOK04.19

7. DZEMYDA, G.; KURASOVA, O.; ŽILINSKAS, J. (2008). Daugiamačių duomenų vizualizavimo metodai. Vilnius: Mokslo aidai. [žiūrėta 2017-12-19] Prieiga per: http://web.vu.lt/mii/j.zilinskas/DzemydaKurasovaZilinskasDDVM.pdf

8. JAGADISH, H. V. (2015). Big Data and Science: Myths and Reality. Big Data Research, vol. 2(2), p. 49–52. [žiūrėta 2017-12-19] Prieiga per: http://iranarze.ir/wp-content/uploads/2016/10/E414.pdf

9. SHERMAN, C. (2014). What's the Big Deal About Big Data? In: Online Searcher 38.2. ProQuest Central, p. 10–17. [žiūrėta 2017-12-19] Prieiga per: http://connection.ebscohost.com/c/articles/95273135/whats-big-deal-about-big-data

10. KURASOVA, O., MARCINKEVIČIUS, V., MEDVEDEV, V., RAPEČKA, A., STEFANOVIČ, P. (2014). Strategies for big data clustering. In: Proceedings of IEEE 26th International Conference on Tools with Artificial Intelligence, ICTAI 2014, p. 740–747. [žiūrėta 2017-12-19] Prieiga per: https://www.mii.lt/paslaugu_internetas/rodikliai/5veikla/5.3.pdf

11. KAZAKEVIČIŪTĖ, J., et al. Europos Valstybių Švietimo Duomenų Statistinės Analizės Modeliai Ir Programinė Įranga, 2011. [žiūrėta 2017-12-19] Prieiga per: http://vddb.library.lt/fedora/get/LT-eLABa0001:E.02~2011~D_20110902_093229-84256/DS.005.0.02.ETD

12. STAŠKŪNAITĖ, K. Statistinių Paketų Taikymai Mokymui, 2015, Lietuvos Edukologijos universitetas. [žiūrėta 2017-12-21] Prieiga per: http://talpykla.elaba.lt/elabafedora/objects/elaba:8644554/datastreams/MAIN/content

13. SALLAM, R.L., et al. Magic Quadrant for Business Intelligence and Analytics Platforms. Gartner RAS Core Research Notes.Gartner, Stamford, CT, 2014. [žiūrėta 2018-01-30] Prieiga per: https://www.gartner.com/doc/reprints?id=1-2AJGAKH&ct=150225&st=sb

14. STRAVINSKIENĖ, A. and GUDAS, S. Duomenų Gavyba Paremta Veiklos Modeliu. [žiūrėta 2017-12-13] Prieiga per: http://isd.ktu.lt/it2010/material/Research/Artificial_Intelligence_and_Knowledge_Egineering.pdf

15. STRAVINSKIENĖ, A. And ŽUKAUSKAITĖ, A. and GUDAS, S. Duomenų gavybos įrankių pritaikymas mažose įmonės. Vilniaus universitetas. [žiūrėta 2017-12-12] Prieiga per: http://isd.ktu.lt/it2010/material/Research/1_AI_2.pdf

16. Azevedo, Ana Isabel Rojão Lourenço. KDD, SEMMA and CRISP-DM: A Parallel Overview. Iads-Dm, 2008. [žiūrėta 2017-12-18] Prieiga per: http://recipp.ipp.pt/bitstream/10400.22/136/1/KDD-CRISP-SEMMA.pdf

17. JANILIONIS, Vytautas. Klasterinė analizė: paskaitos konspektas. KTU, Taikomosios matematikos katedra, 1999. Prieiga per: https://moodle.ktu.edu/

18. YEO, David. Applied Clustering Techniques Course Notes. USA, SAS Institute, 2003. ISBN 978-162960-004-8

19. LAZDAUSKAITĖ, Sandra. Klasterinės ir diskriminantinės analizės taikymai mokinių pasiekimų tyrimui [interaktyvus]. Vilnius, 2007 [žiūrėta 2017-12-30]. Prieiga per: http://vddb.laba.lt/fedora/get/LT-eLABa-0001:E.02~2007~D_20070816_17111469914/DS.005.0.01.ETD

20. JANILIONIS, Vytautas, Vaidas, MORKEVIČIUS ir Rimantas RAULECKAS. Klasterinė analizė: Statistinė kiekybinių duomenų analizė su spss ir stata. [interaktyvus]. Lidata, 2008 [žiūrėta 2017-12-28]. Prieiga per: http://www.lidata.eu/index.php?file=files/mokymai/stat/stat.html&course_file=stat_III_8_1_1.html

21. RUZGAS, T. Daugiamačio pasiskirstymo tankio neparametrinis įvertinimas naudojant stebėjimų klasterizavimą: Daktaro Disertacija: Fiziniai Mokslai, Matematika (01P) [interaktyvus]. Vilnius: 2007. [žiūrėta 2017-12-30] Prieiga per: http://www.mii.lt/files/imi_dis_07_ruzgas.pdf

22. ČEKANAVIČIUS, Vydas, Gediminas, MURAUSKAS. Statistika ir jos taikymai, II. Vilnius: TEV, 2002. ISBN 995-491-16-7

23. MILERIS, Ričardas. Ekonominių reiškinių daugiamatė statistinė analizė: mokomoji knyga [interaktyvus]. Kaunas: Technologija, 2013 [žiūrėta 2017-12-29]. ISBN 978-609-02-0985-1. Prieiga per: https://www.ebooks.ktu.lt/einfo/1251/ekonominiu-reiskiniu-daugiamate-statistine-analize/

24. KUMAR, A. and KUMAR, K. Indoor Air Pollution [interaktyvus]. Pollution Issues, 2014 [žiūrėta 2018-01-20]. Prieiga per: http://www.pollutionissues.com/Ho-Li/Indoor-Air-Pollution.html

25. KAUFMAN, Leonard, Peter, J. ROUSSEEUW. Finding Groups in Data An Introduction to Cluster Analysis [interaktyvus]. New Jersey: John Wiley & Sons, 2005 [žiūrėta 2018-01-02]. ISBN 0-47173578-7. Prieiga per: https://books.google.lt/books?hl=lt&lr=&id=YeFQHiikNo0C&oi=fnd&pg=PR11&dq=cluster+analysis &ots=5zt9D9MHAH&sig=QQnJIxewGzQNrMqajAx-PH1ayZs&redir_esc=y#v=onepage&q&f=false

Priedai

1 priedas

Programos kodas

Hierarchinis klasterizavimas Lietuvoje prieš renovaciją

```
Proc import datafile="/home/gintare.zuzevici/sasuser.v94/PRE_LT.xlsx"
out=duomenys1 dbms=xlsx replace;
getnames=yes;
run;
proc corr data=duomenys1;
var Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2 Radonas;
run;
ods graphics on;
proc sgscatter data=duomenys1;
matrix CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
run;
ods graphics off;
proc stdize data=duomenys1 method=range out=stand1;
var Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
run;
proc corr data=stand1;
var Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2
Radonas;
run;
proc cluster data=stand1 method=ward pseudo outtree=tree1 ccc plots=all;
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
copy B A Temp RH Radonas Virykle Ventiliacija Aukstas Atstumas Gyventojai
Remontas Baldai Grindys;
run;
proc tree data=tree1 horizontal spaces=2 n=3 out=rez1;
copy B A Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas
NO2 Radonas Virykle Ventiliacija Aukstas Atstumas Gyventojai Remontas
Baldai Grindys;
run;
proc freq data=rez1;
tables cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=Toluenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=Etilbenzenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=Ksilenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=Formaldehidas / group=cluster;
```

```
run;
proc sgplot data=rez1;
scatter y=Benzenas x=NO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=CO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=CO2 x=Toluenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=CO2 x=Etilbenzenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=CO2 x=Ksilenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=CO2 x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=CO2 x=NO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Toluenas x=Etilbenzenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Toluenas x=Ksilenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Toluenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Toluenas x=NO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Etilbenzenas x=Ksilenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Etilbenzenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Etilbenzenas x=NO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Ksilenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Ksilenas x=NO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Formaldehidas x=NO2 / group=cluster;
```

```
run;
proc summary data=rez1;
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
output out=vid_c1 mean()=;
class cluster;
run;
proc transpose data=vid_c1 out=vid_transp1(rename=(col1=MATAVIMAS));
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
by cluster;
run;
axis1 value=(h=1.5) label=(h=1.5);
axis2 value=(h=1.5) label=(h=1.5);
axis3 value=(h=1.5) label=(h=1.5);
axis4 value=(h=1.5) label=(h=1.5);
axis5 value=(h=1.5) label=(h=1.5);
axis6 value=(h=1.5) label=(h=1.5);
axis7 value=(h=1.5) label=(h=1.5);
run;
proc gradar data=vid_transp1;
chart _NAME_ / sumvar=MATAVIMAS
staraxis=(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 1);
run;
quit;
proc gradar data=vid_transp1;
chart _NAME_ / sumvar=MATAVIMAS
staraxis=(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 2);
run;
quit;
proc gradar data=vid_transp1;
chart _NAME_ / sumvar=MATAVIMAS
staraxis=(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 3);
run;
quit;
goptions reset=all border;
proc gchart data=rez1;
pie3D Virykle / group=cluster discrete legend slice=none
percent=inside  across=3 coutline=black noheading;
where cluster ^=.;
run;
```

```
quit;
proc gchart data=rez1;
pie3D Ventiliacija / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3D Aukstas / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3D Atstumas / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3D Gyventojai / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3d Remontas / group=cluster discrete legend slice=none
percent=inside
across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3d Baldai / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3d Grindys / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
```

Hierarchinis klasterizavimas Lietuvoje po renovacijos

```
Proc import datafile="/home/gintare.zuzevici/sasuser.v94/POST_LT.xlsx"
out=duomenys2 dbms=xlsx replace;
getnames=yes;
run;
```

```
proc corr data=duomenys2;
var Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2 Radonas;
run;
ods graphics on;
proc sgscatter data=duomenys2;
matrix CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
run;
ods graphics off;
proc stdize data=duomenys2 method=range out=stand2;
var Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
run;
proc corr data=stand2;
var Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2 Radonas;
run;
proc cluster data=stand2 method=ward pseudo outtree=tree2 ccc plots=all;
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
copy B A Temp RH Radonas Virykle Ventiliacija Aukstas Atstumas
Gyventojai Remontas Baldai Grindys;
run;
proc tree data=tree2 horizontal spaces=2 n=3 out=rez2;
copy B A Temp RH CO2 Benzenas Toluenas Etilbenzenas
Ksilenas Formaldehidas NO2 Radonas Virykle Ventiliacija
Aukstas Atstumas Gyventojai Remontas Baldai Grindys;
run;
proc freq data=rez2;
tables cluster;
run;
proc sgplot data=rez2;
scatter y=Benzenas x=Toluenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Benzenas x=Etilbenzenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Benzenas x=Ksilenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Benzenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Benzenas x=NO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Benzenas x=CO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=CO2 x=Toluenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=CO2 x=Etilbenzenas / group=cluster;
run;
```

```
proc sgplot data=rez2;
scatter y=CO2 x=Ksilenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=CO2 x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=CO2 x=NO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Toluenas x=Etilbenzenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Toluenas x=Ksilenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Toluenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Toluenas x=NO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Etilbenzenas x=Ksilenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Etilbenzenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Etilbenzenas x=NO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Ksilenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Ksilenas x=NO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Formaldehidas x=NO2 / group=cluster;
run;
proc summary data=rez2;
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
output out=vid_c2 mean()=;
class cluster;
run;
proc transpose data=vid_c2 out=vid_transp2(rename=(col1=MATAVIMAS));
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
by cluster;
run;
axis1 value=(h=1.5) label=(h=1.5);
axis2 value=(h=1.5) label=(h=1.5);
axis3 value=(h=1.5) label=(h=1.5);
```

```
axis4 value=(h=1.5) label=(h=1.5);
axis5 value=(h=1.5) label=(h=1.5);
axis6 value=(h=1.5) label=(h=1.5);
axis7 value=(h=1.5) label=(h=1.5);
run;
proc gradar data=vid_transp2;
chart _NAME_ / sumvar=MATAVIMAS
staraxis=(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 1);
run;
quit;
proc gradar data=vid_transp2;
chart _NAME_ / sumvar=MATAVIMAS
staraxis=(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 2);
run;
quit;
proc gradar data=vid_transp2;
chart _NAME_ / sumvar=MATAVIMAS staraxis=
(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 3);
run;
quit;
goptions reset=all border;
proc gchart data=rez2;
pie3D Virykle / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Ventiliacija / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Aukstas / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
```

```
proc gchart data=rez2;
pie3D Atstumas / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Gyventojai / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Remontas / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Baldai / group=cluster discrete legend slice=none
percent=inside  across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Grindys / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
```

Hierarchinis klasterizavimas Suomijoje prieš renovaciją

```
Proc import datafile="/home/gintare.zuzevici/sasuser.v94/PRE_FI.xlsx"
out=duomenys1 dbms=xlsx replace;
getnames=yes;
run;
proc corr data=duomenys1;
var Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2 Radonas;
run;
ods graphics on;
proc sgscatter data=duomenys1;
matrix CO2 Benzenas Toluenas  Ksilenas Formaldehidas NO2 Radonas;
run;
ods graphics off;
proc stdize data=duomenys1 method=range out=stand1;
var Temp RH CO2 Benzenas Toluenas Ksilenas Formaldehidas NO2 Radonas;
run;
proc cluster data=stand1 method=ward pseudo outtree=tree1 ccc plots=all;
var CO2 Benzenas Toluenas Ksilenas Formaldehidas NO2 Radonas;
copy B A Temp RH Virykle Ventiliacija Aukstas Atstumas Gyventojai
```

```
Remontas Baldai Grindys;
run;
proc tree data=tree1 horizontal spaces=2 n=3 out=rez1;
copy B A Temp RH CO2 Benzenas Toluenas Ksilenas Formaldehidas
NO2 Radonas Virykle Ventiliacija Aukstas Atstumas Gyventojai Remontas
Baldai Grindys;
run;
proc freq data=rez1;
tables cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=Toluenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=Ksilenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=NO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=CO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Benzenas x=Radonas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=CO2 x=Toluenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=CO2 x=Ksilenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=CO2 x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=CO2 x=NO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Radonas x=CO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Toluenas x=Ksilenas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Toluenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Toluenas x=NO2 / group=cluster;
```

```
run;
proc sgplot data=rez1;
scatter y=Toluenas x=Radonas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Ksilenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Ksilenas x=NO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Ksilenas x=Radonas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Formaldehidas x=NO2 / group=cluster;
run;
proc sgplot data=rez1;
scatter y=Formaldehidas x=Radonas / group=cluster;
run;
proc sgplot data=rez1;
scatter y=NO2 x=Radonas / group=cluster;
run;
proc summary data=rez1;
var CO2 Benzenas Toluenas Ksilenas Formaldehidas NO2 Radonas;
output out=vid_c1 mean()=;
class cluster;
run;
proc transpose data=vid_c1 out=vid_transp1(rename=(col1=MATAVIMAS));
var CO2 Benzenas Toluenas Ksilenas Formaldehidas NO2 Radonas;
by cluster;
run;
axis1 value=(h=1.5) label=(h=1.5);
axis2 value=(h=1.5) label=(h=1.5);
axis3 value=(h=1.5) label=(h=1.5);
axis4 value=(h=1.5) label=(h=1.5);
axis5 value=(h=1.5) label=(h=1.5);
axis6 value=(h=1.5) label=(h=1.5);
axis7 value=(h=1.5) label=(h=1.5);
run;
proc gradar data=vid_transp1;
chart _NAME_ / sumvar=MATAVIMAS
staraxis=(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 1);
run;
quit;
proc gradar data=vid_transp1;
chart _NAME_ / sumvar=MATAVIMAS
staraxis=(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
```

```
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 2);
run;
quit;
proc gradar data=vid_transp1;
chart _NAME_ / sumvar=MATAVIMAS
staraxis=(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 3);
run;
quit;
goptions reset=all border;
proc gchart data=rez1;
pie3D Virykle / group=cluster discrete legend slice=none
percent=inside  across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3D Ventiliacija / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3D Aukstas / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3D Atstumas / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3D Gyventojai / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3d Remontas / group=cluster discrete legend slice=none
percent=inside
across=3 coutline=black noheading;
where cluster ^=.;
```

```
run;
quit;
proc gchart data=rez1;
pie3d Baldai / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez1;
pie3d Grindys / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
```

Hierarchinis klasterizavimas Suomijoje po renovacijos

```
Proc import datafile="/home/gintare.zuzevici/sasuser.v94/POST_FI.xlsx"
out=duomenys2 dbms=xlsx replace;
getnames=yes;
run;
proc corr data=duomenys2;
var Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2 Radonas;
run;
ods graphics on;
proc sgscatter data=duomenys2;
matrix CO2 Benzenas Toluenas Ksilenas Formaldehidas NO2;
run;
ods graphics off;
proc stdize data=duomenys2 method=range out=stand2;
var Temp RH CO2 Benzenas Toluenas Ksilenas Formaldehidas NO2;
run;
proc cluster data=stand2 method=ward pseudo outtree=tree2 ccc plots=all;
var CO2 Benzenas Toluenas Ksilenas Formaldehidas NO2;
copy B A Temp RH Virykle Ventiliacija Aukstas Atstumas
Gyventojai Remontas Baldai Grindys;
run;
proc tree data=tree2 horizontal spaces=2 n=3 out=rez2;
copy B A Temp RH CO2 Benzenas Toluenas
Ksilenas Formaldehidas NO2 Virykle Ventiliacija
Aukstas Atstumas Gyventojai Remontas Baldai Grindys;
run;
proc freq data=rez2;
tables cluster;
run;
proc sgplot data=rez2;
scatter y=Benzenas x=Toluenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Benzenas x=Ksilenas / group=cluster;
run;
```

```
proc sgplot data=rez2;
scatter y=Benzenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Benzenas x=NO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Benzenas x=CO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=CO2 x=Toluenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=CO2 x=Ksilenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=CO2 x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=CO2 x=NO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Toluenas x=Ksilenas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Toluenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Toluenas x=NO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Ksilenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Ksilenas x=NO2 / group=cluster;
run;
proc sgplot data=rez2;
scatter y=Formaldehidas x=NO2 / group=cluster;
run;
proc summary data=rez2;
var CO2 Benzenas Toluenas Ksilenas Formaldehidas NO2;
output out=vid_c2 mean()=;
class cluster;
run;
proc transpose data=vid_c2 out=vid_transp2(rename=(col1=MATAVIMAS));
var CO2 Benzenas Toluenas Ksilenas Formaldehidas NO2;
by cluster;
run;
axis1 value=(h=1.5) label=(h=1.5);
axis2 value=(h=1.5) label=(h=1.5);
axis3 value=(h=1.5) label=(h=1.5);
```

```
axis4 value=(h=1.5) label=(h=1.5);
axis5 value=(h=1.5) label=(h=1.5);
axis6 value=(h=1.5) label=(h=1.5);
run;
proc gradar data=vid_transp2;
chart _NAME_ / sumvar=MATAVIMAS
staraxis=(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 1);
run;
quit;
proc gradar data=vid_transp2;
chart _NAME_ / sumvar=MATAVIMAS
staraxis=(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 2);
run;
quit;
proc gradar data=vid_transp2;
chart _NAME_ / sumvar=MATAVIMAS staraxis=
(axis1, axis2, axis3, axis4, axis5, axis6, axis7)
overlayvar=cluster
cstars=(green, red) wstars=2 2 lstars=1 1
spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 3);
run;
quit;
goptions reset=all border;
proc gchart data=rez2;
pie3D Virykle / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Ventiliacija / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Aukstas / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
```

```
pie3D Atstumas / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Gyventojai / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Remontas / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Baldai / group=cluster discrete legend slice=none
percent=inside  across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=rez2;
pie3D Grindys / group=cluster discrete legend slice=none
percent=inside across=3 coutline=black noheading;
where cluster ^=.;
run;
quit;
```

Nehierarchinis klasterizavimas K-vidurkių metodas

```
Proc import datafile="/home/gintare.zuzevici/sasuser.v94/PRE_LT.xlsx"
out=duomenys1 dbms=xlsx replace;
getnames=yes;
run;
proc stdize data=duomenys1 method=range out=stand1;
var Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
run;
proc fastclus data=stand1 maxclusters=2 out=kmeans1;
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
run;
proc sgplot data=kmeans1;
scatter y=Benzenas x=Toluenas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Benzenas x=Etilbenzenas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Benzenas x=Ksilenas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Benzenas x=Formaldehidas / group=cluster;
```

```
run;
proc sgplot data=kmeans1;
scatter y=Benzenas x=NO2 / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Benzenas x=CO2 / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=CO2 x=Toluenas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=CO2 x=Etilbenzenas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=CO2 x=Ksilenas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=CO2 x=Formaldehidas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=CO2 x=NO2 / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Toluenas x=Etilbenzenas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Toluenas x=Ksilenas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Toluenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Toluenas x=NO2 / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Etilbenzenas x=Ksilenas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Etilbenzenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Etilbenzenas x=NO2 / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Ksilenas x=Formaldehidas / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Ksilenas x=NO2 / group=cluster;
run;
proc sgplot data=kmeans1;
scatter y=Formaldehidas x=NO2 / group=cluster;
run;
proc summary data=kmeans1;
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
output out=vid_k1 mean()=;
class cluster;
```

```
run;
proc transpose data=vid_k1 out=vid_transposed1(rename=(col1=MATAVIMAS));
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
by cluster;
run;
proc gradar data=vid_transposed1;
chart _NAME_ / sumvar=MATAVIMAS overlayvar=cluster cstars=(green, red)
wstars=2 2 lstars=1 1 spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 1);
run;
quit;
proc gradar data=vid_transposed1;
chart _NAME_ / sumvar=MATAVIMAS overlayvar=cluster cstars=(green, red)
wstars=2 2 lstars=1 1 spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 2);
run;
quit;
goptions reset=all border;
proc gchart data=kmeans1;
pie3D Virykle / group=cluster discrete legend slice=none
percent=inside across=2 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=kmeans1;
pie3D Ventiliacija / group=cluster discrete legend slice=none
percent=inside across=2 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=kmeans1;
pie3D Aukstas / group=cluster discrete legend slice=none
percent=inside across=2 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=kmeans1;
pie3D Atstumas / group=cluster discrete legend slice=none
percent=inside across=2 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=kmeans1;
pie3D Gyventojai / group=cluster discrete legend slice=none
percent=inside across=2 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=kmeans1;
pie3D Remontas / group=cluster discrete legend slice=none
percent=inside across=2 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=kmeans1;
pie3D Baldai / group=cluster discrete legend slice=none
```

```
percent=inside  across=2 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=kmeans1;
pie3D Grindys / group=cluster discrete legend slice=none
percent=inside across=2 coutline=black noheading;
where cluster ^=.;
run;
quit;
```

Nehierarchinis klasterizavimas K-kaimynų metodas

```
Proc import datafile="/home/gintare.zuzevici/sasuser.v94/PRE_LT.xlsx"
out=duomenys1 dbms=xlsx replace;
getnames=yes;
run;
proc stdize data=duomenys1 method=range out=stand1;
var Temp RH CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
run;
proc modeclus data=stand1 k=5 method=1 out=nearest1;
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
run;
proc summary data=nearest1;
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
output out=vid_n1 mean()=;
class CLUSTER;
run;
proc transpose data=vid_n1 out=vid_transpose1(rename=(col1=MATAVIMAS));
var CO2 Benzenas Toluenas Etilbenzenas Ksilenas Formaldehidas NO2;
by cluster;
run;
proc gradar data=vid_transpose1;
chart _NAME_ / sumvar=MATAVIMAS overlayvar=cluster cstars=(green, red)
wstars=2 2 lstars=1 1 spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 1);
run;
quit;
proc gradar data=vid_transpose1;
chart _NAME_ / sumvar=MATAVIMAS overlayvar=cluster cstars=(green, red)
wstars=2 2 lstars=1 1 spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 2);
run;
quit;
proc gradar data=vid_transpose1;
chart _NAME_ / sumvar=MATAVIMAS overlayvar=cluster cstars=(green, red)
wstars=2 2 lstars=1 1 spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 3);
run;
quit;
proc gradar data=vid_transpose1;
chart _NAME_ / sumvar=MATAVIMAS overlayvar=cluster cstars=(green, red)
wstars=2 2 lstars=1 1 spokescale=vertex starcircles=(0.5 1.0);
where cluster in (., 4);
run;
quit;goptions reset=all border;
```

```
proc gchart data=nearest1;
pie3D Virykle / group=cluster discrete legend slice=none
percent=inside across=4 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=nearest1;
pie3D Ventiliacija / group=cluster discrete legend slice=none
percent=inside across=4 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=nearest1;
pie3D Aukstas / group=cluster discrete legend slice=none
percent=inside across=4 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=nearest1;
pie3D Atstumas / group=cluster discrete legend slice=none
percent=inside across=4 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=nearest1;
pie3D Gyventojai / group=cluster discrete legend slice=none
percent=inside across=4 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=nearest1;
pie3D Remontas / group=cluster discrete legend slice=none
percent=inside across=4 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=nearest1;
pie3D Baldai / group=cluster discrete legend slice=none
percent=inside  across=4 coutline=black noheading;
where cluster ^=.;
run;
quit;
proc gchart data=nearest1;
pie3D Grindys / group=cluster discrete legend slice=none
percent=inside across=4 coutline=black noheading;
where cluster ^=.;
run;
quit;
```

Klasterių jungimo protokolas

2P.1 lentelė. Klasterių jungimo protokolas (prieš renovaciją Lietuvoje)

Klasterių skaičius	Sujungti klasteriai		Dažnis	Kubinis klasterizavimo kriterijus	Pseudo F statistika	Pseudo T kvadratas
85	OB14	OB16	2	.	125	.
84	OB18	OB35	2	.	99.2	.
83	OB56	OB82	2	.	88.1	.
82	OB12	OB31	2	.	76.5	.
81	OB84	OB92	2	.	68.6	.
80	OB19	OB22	2	.	63.4	.
79	OB85	OB93	2	.	60.2	.
78	OB51	OB80	2	.	57.6	.
77	OB21	OB37	2	.	55.8	.
76	OB60	OB64	2	.	54.6	.
75	OB55	OB79	2	.	53.8	.
74	CL82	OB43	3	.	51.9	1.8
73	OB53	OB87	2	.	50.5	.
72	CL85	OB62	3	.	49.2	4.5
71	CL77	OB36	3	.	48.1	1.4
70	OB27	CL79	3	.	47.2	1.5
69	OB26	OB67	2	.	46.6	.
68	OB5	OB44	2	.	46.1	.
67	OB30	CL76	3	.	45.4	1.6
66	CL75	OB74	3	.	44.8	1.6
65	OB39	OB40	2	.	44.2	.
64	OB17	CL83	3	.	43.6	3.3
63	OB42	OB95	2	.	42.9	.
62	CL78	OB66	3	.	42.4	2.0
61	OB78	OB94	2	.	41.6	.
60	OB25	CL81	3	.	40.5	3.2
59	OB6	OB41	2	.	39.6	.
58	OB69	CL61	3	.	38.5	1.3
57	OB47	OB48	2	.	37.6	.
56	OB10	OB45	2	.	36.8	.
55	CL60	CL73	5	.	36.0	2.2
54	OB65	OB76	2	.	35.3	.
53	OB15	CL71	4	.	34.8	2.9
52	OB1	OB2	2	.	34.3	.
51	CL80	CL65	4	.	33.8	2.8
50	OB20	CL69	3	.	33.4	2.8
49	CL64	OB50	4	.	32.9	3.7
48	OB13	OB58	2	.	32.5	.
47	CL70	CL63	5	.	32.1	3.2
46	OB4	OB77	2	.	31.8	.
45	CL67	OB61	4	.	31.4	3.9

Klasterių skaičius	Sujungti klasteriai		Dažnis	Kubinis klasterizavimo kriterijus	Pseudo F statistika	Pseudo T kvadratas
44	CL66	OB73	4	.	31.1	3.9
43	CL72	OB81	4	.	30.8	6.6
42	OB8	OB88	2	.	30.6	.
41	CL74	OB23	4	.	30.3	5.9
40	OB75	OB96	2	.	30.1	.
39	OB52	OB83	2	.	29.8	.
38	CL84	CL51	6	.	29.5	4.2
37	CL56	OB46	3	.	29.2	2.4
36	CL50	CL62	6	.	28.9	3.9
35	CL48	CL54	4	.	28.4	2.4
34	CL52	CL68	4	.	28.0	3.9
33	CL43	CL55	9	.	27.6	5.0
32	CL39	CL58	5	.	27.1	2.8
31	OB34	OB70	2	.	26.8	.
30	CL37	CL45	7	.	26.5	3.2
29	OB3	OB91	2	.	26.4	.
28	CL42	OB89	3	.	26.1	2.4
27	CL41	CL53	8	.	25.9	6.4
26	CL59	CL47	7	.	25.5	7.4
25	CL32	OB68	6	.	25.2	3.0
24	CL34	CL44	8	.	24.9	5.1
23	CL29	CL40	4	.	24.6	2.3
22	CL36	OB24	7	.	24.3	7.0
21	OB9	CL57	3	.	24.0	7.8
20	CL49	CL38	10	.	23.9	9.6
19	CL33	CL25	15	.	23.3	6.5
18	CL35	CL31	6	.	22.9	4.7
17	CL21	CL30	10	7.81	22.5	5.3
16	CL26	CL27	15	7.28	22.2	9.4
15	OB71	OB86	2	6.56	21.7	.
14	CL20	CL22	17	5.85	21.2	9.5
13	CL15	OB72	3	5.31	21.1	1.1
12	CL28	CL17	13	4.22	20.8	6.1
11	CL24	CL23	12	3.76	20.8	8.0
10	CL16	CL18	21	3.14	20.7	9.1
9	CL19	CL14	32	2.74	21.0	9.7
8	OB28	OB29	2	2.14	21.2	.
7	CL46	CL12	15	1.44	21.4	8.0
6	CL10	CL9	53	0.02	20.9	15.2
5	CL13	OB90	4	-1.2	20.7	3.7
4	CL7	CL5	19	-2.0	21.0	6.4
3	CL11	CL8	14	-3.1	19.8	17.5
2	CL4	CL6	72	-3.1	17.4	21.7
1	CL3	CL2	86	0.00	.	17.4

Iš 2P.1 lentelės matome, kad CCC kriterijus siūlo duomenis grupuoti į 8 arba 9 klasterius. Pseudo F statistika sufleruoja, kad optimalus klasterių skaičius galėtų būti 4 arba 7. Pseudo T kvadratas siūlo optimaliu klasterių skaičiumi priimti 5 klasterius.

3 priedas

Duomenų analizė atlikus renovaciją Lietuvoje

Aprašomosios statistikos analizė

3P.1 lentelė. Skaitinės kintamųjų charakteristikos (po renovacijos Lietuvoje)

Kintamasis	Imties plotis	Vidurkis	Standartinis nuokrypis	Minimumas	Maksimumas
Temperatūra	65	20.08615	1.20220	16.40000	22.30000
Drėgmė	65	49.01600	8.57316	33.23000	72.70000
CO_2	65	1077	457.22484	447.00000	3003
Benzenas	63	8.00794	9.16338	1.50000	53.50000
Toluenas	63	9.63810	6.30551	2.60000	38.00000
Etilbenzenas	63	2.23657	1.72005	0	8.70000
Ksilenas	63	3.56667	3.04329	0.40000	17.20000
Formaldehidas	65	31.27692	13.07030	9.20000	71.60000
NO_2	65	13.73754	7.63459	3.02000	36.32000
Radonas	35	41.21429	26.45738	8.60000	148.30000

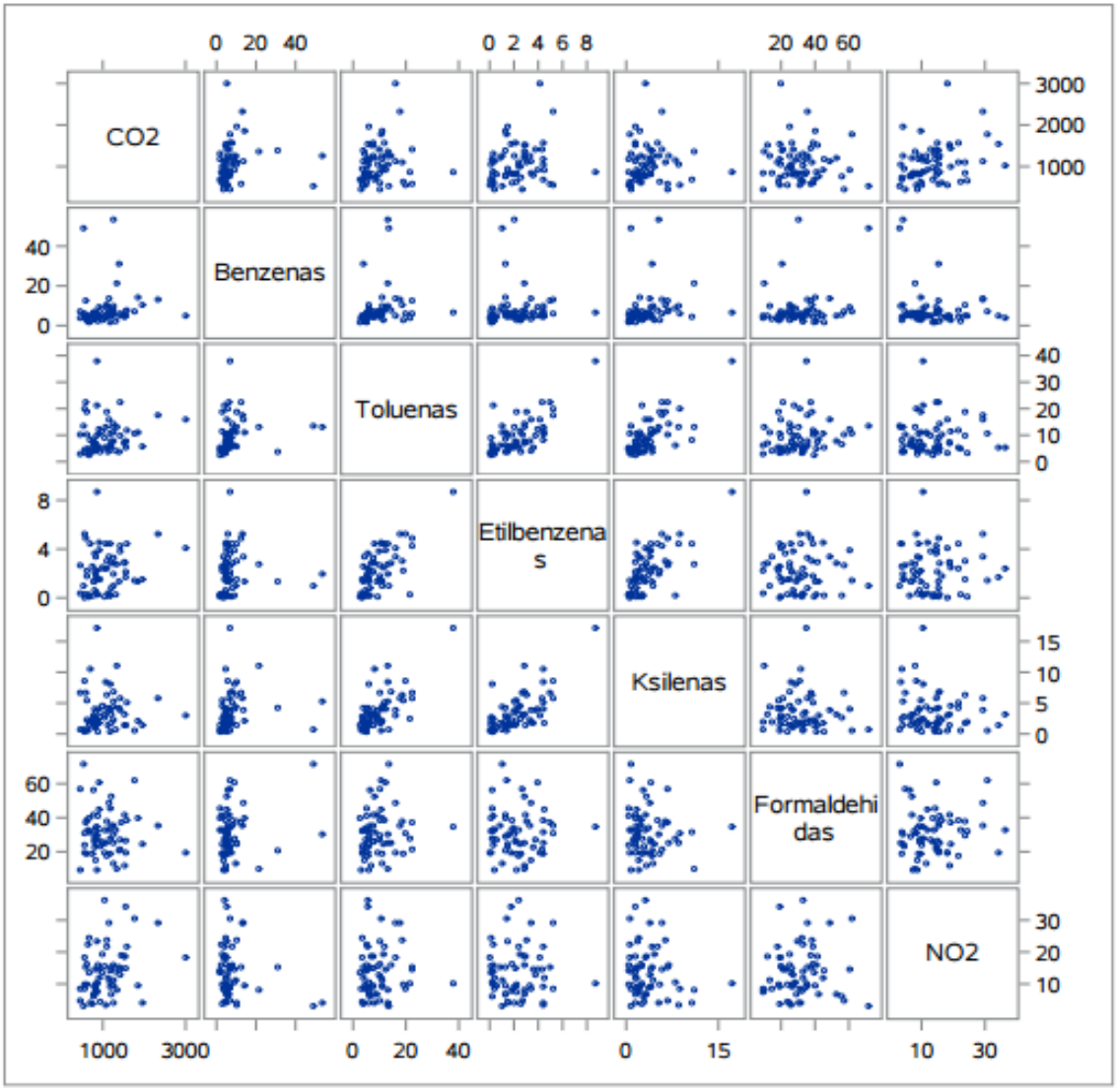
Stebint vidutinę oro temperatūrą, matome, kad ženklaus pasikeitimo nėra, tačiau minimali užfiksuota temperatūra gerokai aukštenė. Santykinis drėgmės kiekis per laikotarpį nuo minimalios 15.54% ribos taip pat padidėjo iki 33.23%, kas lemia ženkliai geresnę oro kokybę namuose. Atlikus daugiabučių renovaciją padidėjo dujinių oro teršalų kiekis, tačiau leistinas normas atitinka.

Duomenų paruošimas klasterizavimui

3P.2 lentelė. Pirsono koreliacijos tarp kintamųjų matavimas (po renovacijos Lietuvoje)

	CO_2	Benzenas	Toluenas	Etilbenzenas	Ksilenas	Formaldehidas	NO_2
CO_2	1.00000	0.09177	0.12355	0.20290	-0.00512	-0.08879	0.33289
Benzenas	0.09177	1.00000	0.18715	0.01878	0.14616	0.20512	-0.18619
Toluenas	0.12355	0.18715	1.00000	0.67811 <.0001	0.69057 <.0001	0.12794	0.02921
Etilbenzenas	0.20290	0.01878	0.67811 <.0001	1.00000	0.68629 <.0001	0.00041	0.01382
Ksilenas	-0.00512	0.14616	0.69057 <.0001	0.68629 <.0001	1.00000	-0.09543	-0.16053
Formaldehidas	-0.08879	0.20512	0.12794	0.00041	-0.09543	1.00000	0.06075
NO_2	0.33289	-0.18619	0.02921	0.01382	-0.16053	0.06075	1.00000

Pateikiama sklaidos diagrama, kuri apibūdina išsidėstymą ir tarpusavio santykį (3P.1 pav.).



3P.1 pav. Pradinių duomenų sklaidos diagrama (po renovacijos Lietuvoje)

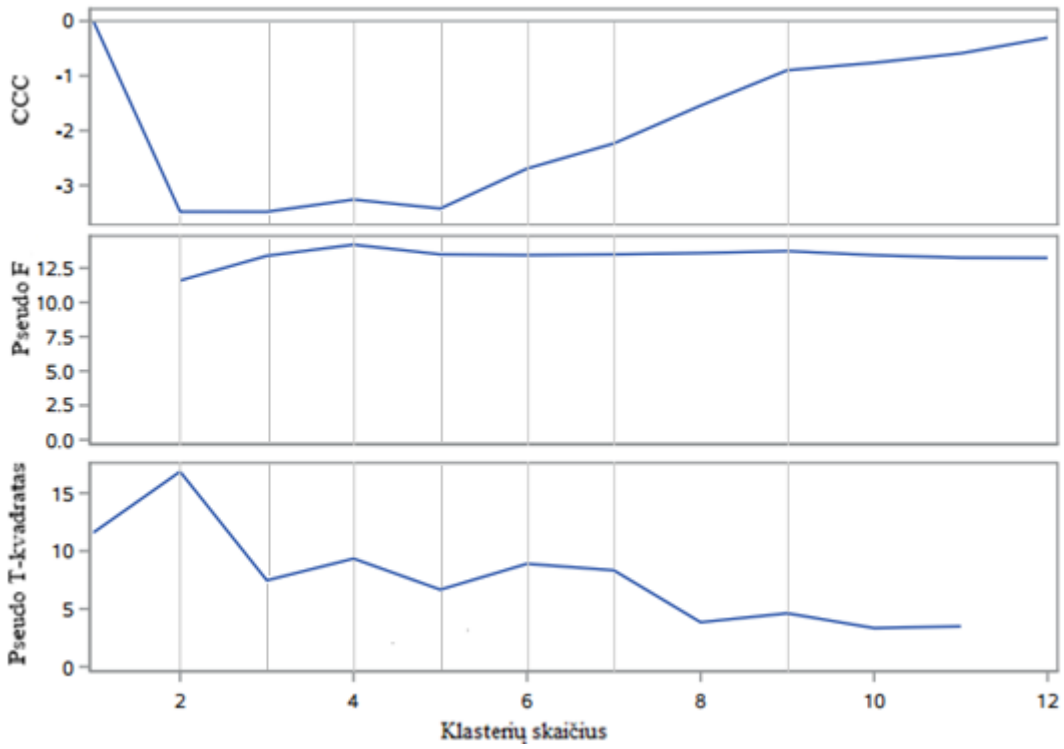
Klasterizavimas naudojant hierarchinius metodus

3P.3 lentelė. Klasterių jungimo protokolas (po renovacijos Lietuvoje)

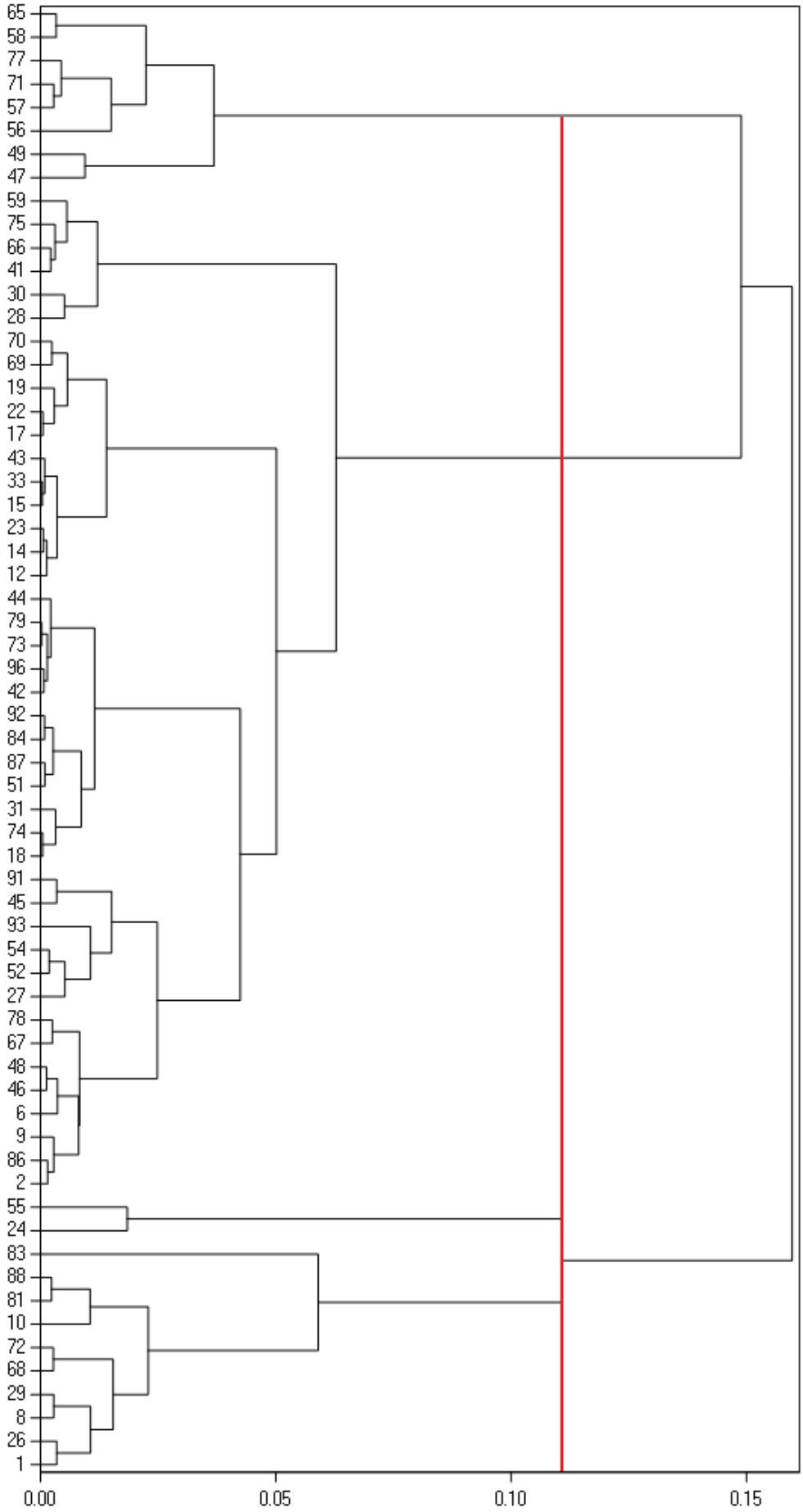
Klasterių skaičius	Sujungti klasteriai		Dažnis	Kubinis klasterizavimo kriterijus	Pseudo F statistika	Pseudo T kvadratas
62	OB73	OB79	2	.	81.1	.
61	OB15	OB33	2	.	52.5	.
...	...	...	...	...	...	...
10	CL13	CL21	9	-.76	13.4	3.3
9	CL24	CL14	14	-.90	13.7	4.6
8	CL22	CL11	8	-1.5	13.6	3.8
7	CL9	CL18	26	-2.2	13.5	8.3
6	CL7	CL16	37	-2.7	13.4	8.9
5	CL10	OB83	10	-3.4	13.5	6.7
4	CL6	CL17	43	-3.3	14.2	9.3
3	CL5	CL12	12	-3.5	13.4	7.5
2	CL4	CL8	51	-3.5	11.6	16.9
1	CL3	CL2	63	0.00	.	11.6

Klasterių skaičiaus nustatymas

Iš 3P.3 lentelės ir 3P.2 pav. matyti, jog kubinis klasterizavimo kriterijus siūlo duomenis grupuoti į 6 arba 7 klasterius. Psiaudo F statistika sufleruoja, kad optimalus klasterių skaičius galėtų būti 4 arba 9. Psiaudo T - kvadratas rodo, jog duomenis galima skirstyti į 3 arba 5 klasterius.



3P.2 pav. Kriterijai klasterių skaičiaus nustatymui (po renovacijos Lietuvoje)



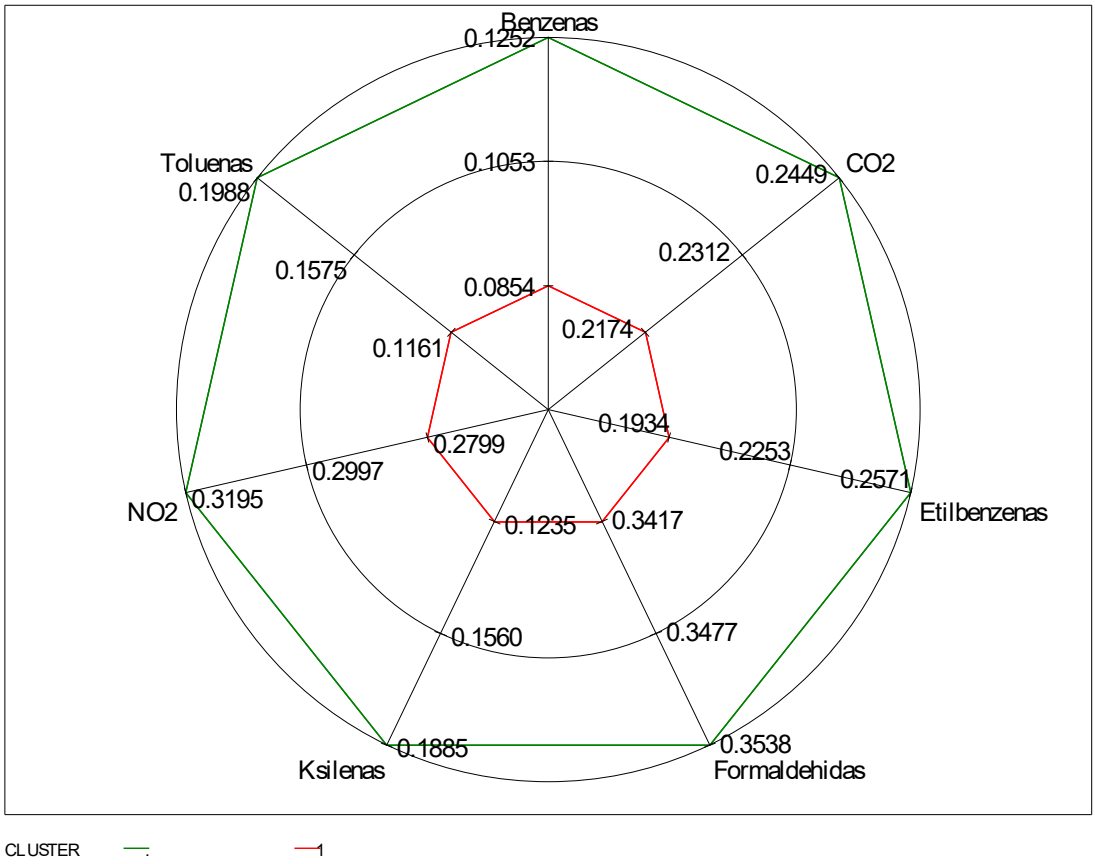
3P.3 pav. Dendograma (po renovacijos Lietuvoje)

Naudojant *proc freq* procedūrą nustatoma kiek stebinių priklauso kiekvienam iš klasterių:

- 43 objektai priklauso pirmajam klasteriui;
- 12 objektų priklauso antrajam klasteriui;
- 8 objektai priklauso trečiajam klasteriui.

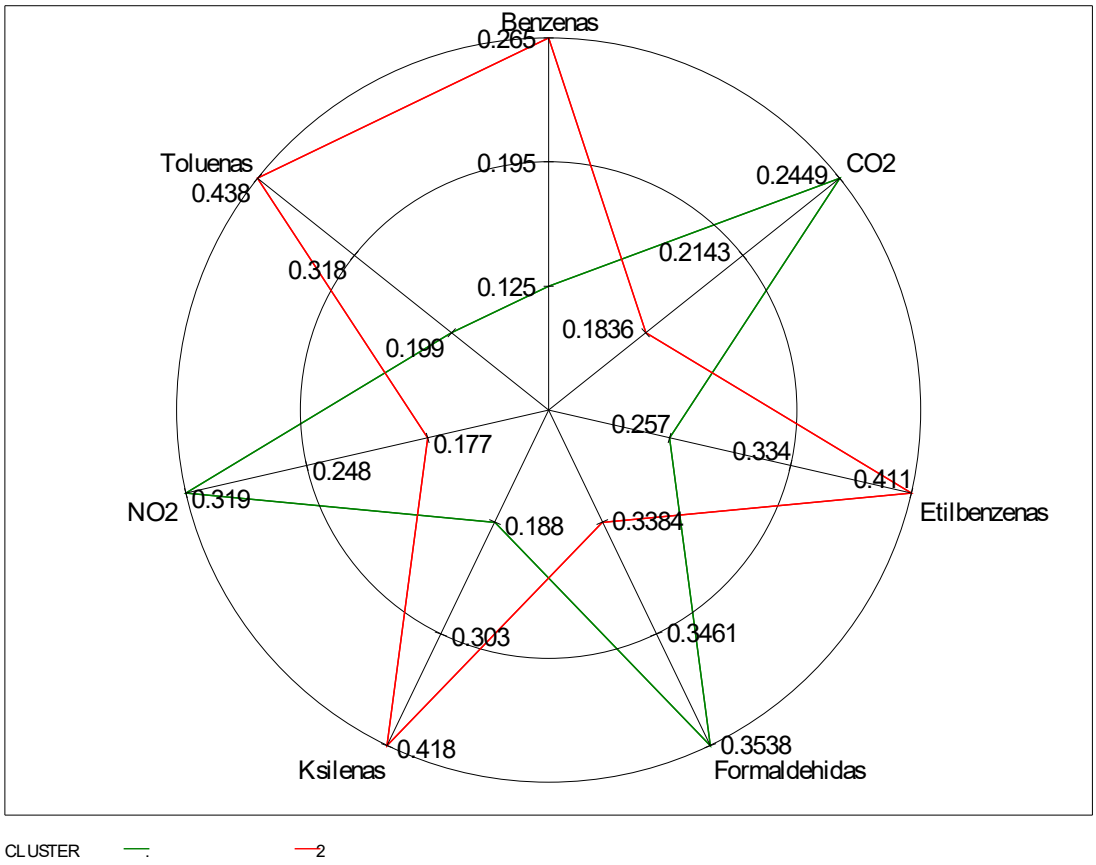
Taip pat nustatyta, kad 33 stebiniai nepriskirti nei vienam iš suformuotų klasterių.

Sudarytų klasterių požymių tyrimas



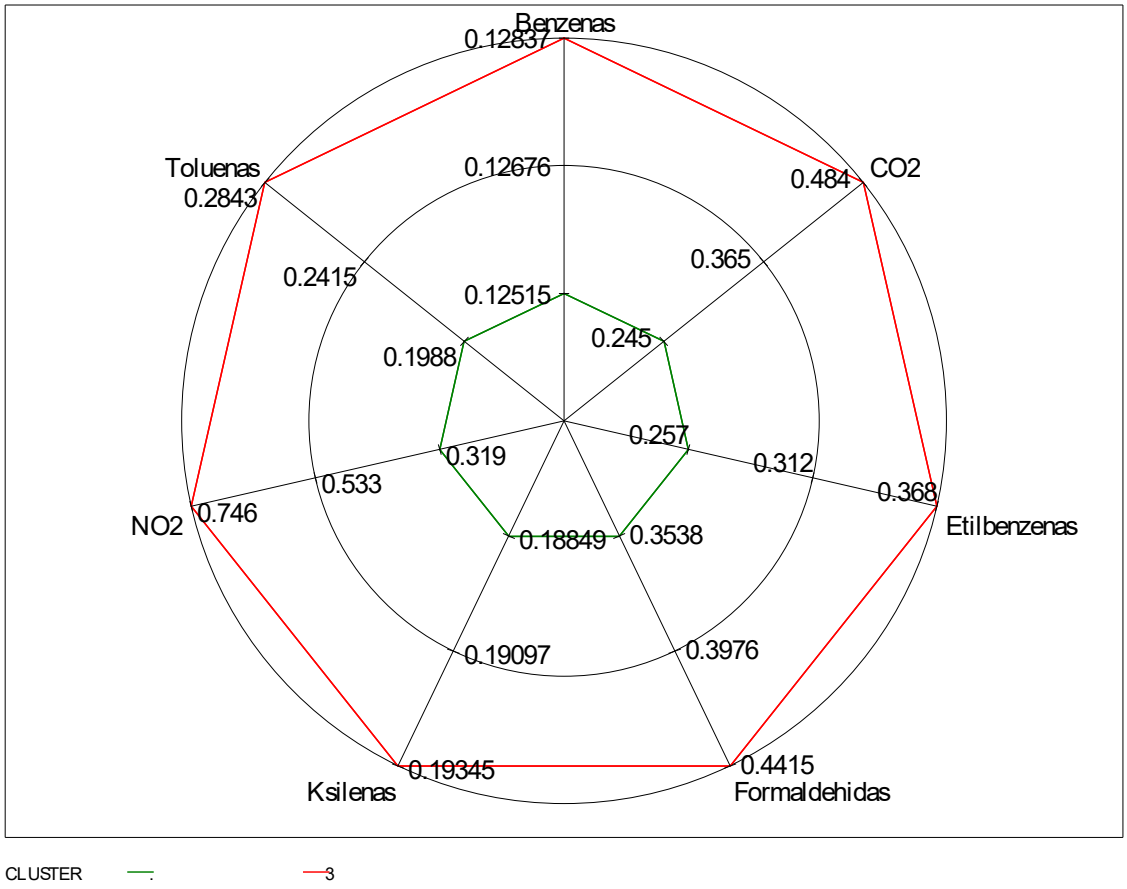
3P.4 pav. Teršalų būdingų I-am klasteriui nustatymas (po renovacijos Lietuvoje)

Nagrinėjamame klasteryje visų tiriamų teršalų kiekis neviršina vidutinių visos imties kiekių (3P.4 pav.). Ko negalima pasakyti apie antram klasteriui būdingus požymius. Iš žvaigždės formos grafiko matoma, kad antrajam klasteriui būdingas didesnis nei vidutinis benzeno, tolueno, etilbenzeno bei ksileno kiekis (3P.5 pav.). Vidutinis benzeno kiekis tiriamoje imtyje 0.125 $\mu g/m^3$, o antrame klasteryje šio teršalo vidurkis yra daugiau nei dvigubai didesnis - sudaro net 0.265 $\mu g/m^3$. Tolueno vidutinis kiekis tiriamų patalpų ore sudaro 0.199 $\mu g/m^3$ tuo tarpu šiame klasteryje teršalo vidurkis yra taip pat didesnis nei dvigubas - 0.438 $\mu g/m^3$. Etilbenzeno vidutinis kiekis tarp visų stebinių yra 0.257 $\mu g/m^3$, o klasteryje siekia 0.411 $\mu g/m^3$. Ir ksileno kiekis tiriamoje imtyje 0.188 $\mu g/m^3$, o antrame klasteryje jo vidurkis sudaro 0.418 $\mu g/m^3$.



3P.5 pav. Teršalų būdingų II-am klasteriui nustatymas (po renovacijos Lietuvoje)

Trečiajam klasteriui būdingi didesni visų tiriamų teršalų kiekiai, tačiau daugelio labai nežymūs (3P.6 pav.). Matome, kad vidutinis etilbenzeno kiekis tarp visų stebinių yra 0.257 $\mu g/m^3$, tuo tarpu nagrinėjamame klasteryje 0.368 $\mu g/m^3$. Ksileno kiekis tiriamoje imtyje 0.188 $\mu g/m^3$, o trečiame klasteryje šio teršalo vidurkis yra neženkliai didesnis - siekia 0.193 $\mu g/m^3$. Formaldehido vidutinis kiekis tiriamų patalpų ore sudaro 0.3538 $\mu g/m^3$, o tuo tarpu trečiame klasteryje šio teršalo vidurkis yra - 0.442 $\mu g/m^3$. Vidutinis benzeno kiekis tiriamoje imtyje 0.125 $\mu g/m^3$, o trečiame klasteryje truputį didesnis ir yra 0.128 $\mu g/m^3$. Tolueno vidutinis kiekis tiriamų patalpų ore sudaro 0.242 $\mu g/m^3$ tuo tarpu šiame klasteryje teršalo vidurkis - 0.284 $\mu g/m^3$. Anglies dioksido vidutinis kiekis tarp visų stebinių yra 0.245 $\mu g/m^3$, tuo tarpu nagrinėjamame klasteryje 0.484 $\mu g/m^3$. Vidutinis NO_2 kiekis tiriamoje imtyje 0.319 $\mu g/m^3$, o trečiame klasteryje šio teršalo vidurkis sudaro net 0.746 $\mu g/m^3$.



3P.6 pav. Teršalų būdingų III-iam klasteriui nustatymas (po renovacijos Lietuvoje)

4 priedas

Duomenų analizė po renovacijos Suomijoje

Aprašomosios statistikos analizė

4P.1 lentelė. Skaitinės kintamųjų charakteristikos (po renovacijos Suomijoje)

Kintamasis	Imties plotis	Vidurkis	Standartinis nuokrypis	Minimumas	Maksimumas
Temperatūra	122	22.59248	1.12542	19.56833	25.27821
Drėgmė	122	30.36603	6.07432	17.34580	52.89767
CO_2	119	671.85813	195.97626	390.25494	1552
Benzenas	122	3.55741	3.36828	0	26.10000
Toluenas	122	4.53880	3.43133	0.14872	25.80000
Etilbenzenas	122	0.80616	0.89794	0	6.80000
Ksilenas	122	1.66228	1.57254	0.09458	11.50000
Formaldehidas	123	18.76908	8.61631	6.07786	52.69788
NO_2	118	6.82153	4.52400	2.16000	39.16000
Radonas	103	65.14563	53.88557	20.00000	300.00000

Tyrimas bus atliekamas nevertinant radono poveikio, kadangi ne visuose stebiniuose buvo matuojamas jo kiekis ore.

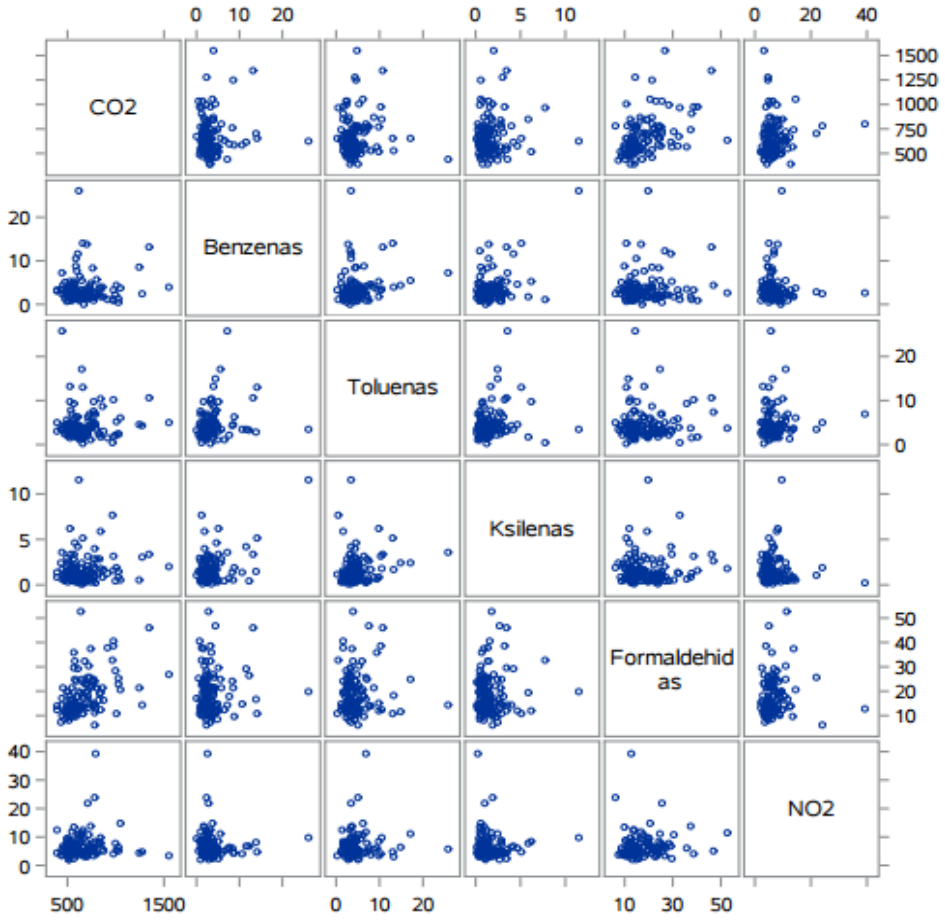
Duomenų paruošimas klasterizavimui

4P.2 lentelė. Pirsono koreliacijos tarp kintamųjų matavimas (po renovacijos Suomijoje)

	CO_2	Benzenas	Toluenas	Etilbenzenas	Ksilenas	Formaldehidas	NO_2
CO_2	1.00000	0.06249	0.02124	0.10601	0.09894	0.43571 <.0001	0.08231
Benzenas	0.06249	1.00000	0.21868	0.38179 <.0001	0.48699 <.0001	0.06721	-0.03032
Toluenas	0.02124	0.21868	1.00000	0.48950 <.0001	0.19719	0.04933	0.06148
Ksilenas	0.09894	0.48699 <.0001	0.19719	0.50763 <.0001	1.00000	0.07533	-0.07119
Formaldehidas	0.43571 <.0001	0.06721	0.04933	0.13651	0.07533	1.00000	0.04329
NO_2	0.08231	-0.03032	0.06148	-0.01241	-0.07119	0.04329	1.00000

Pirsono koreliacijos koeficientas (4P.2 lentelė) rodo stiprų teigiamą ryšį tarp kintamųjų etilbenzeno ir ksileno bei tolueno. Išvengiant gauti neteisingus rezultatus ir pateikti klaidingas išvadas, etilbenzenas bus pašalinamas.

Duomenų išsidėstymą ir tarpusavio santykį apibūdinanti sklaidos diagrama (4P.1 pav.).



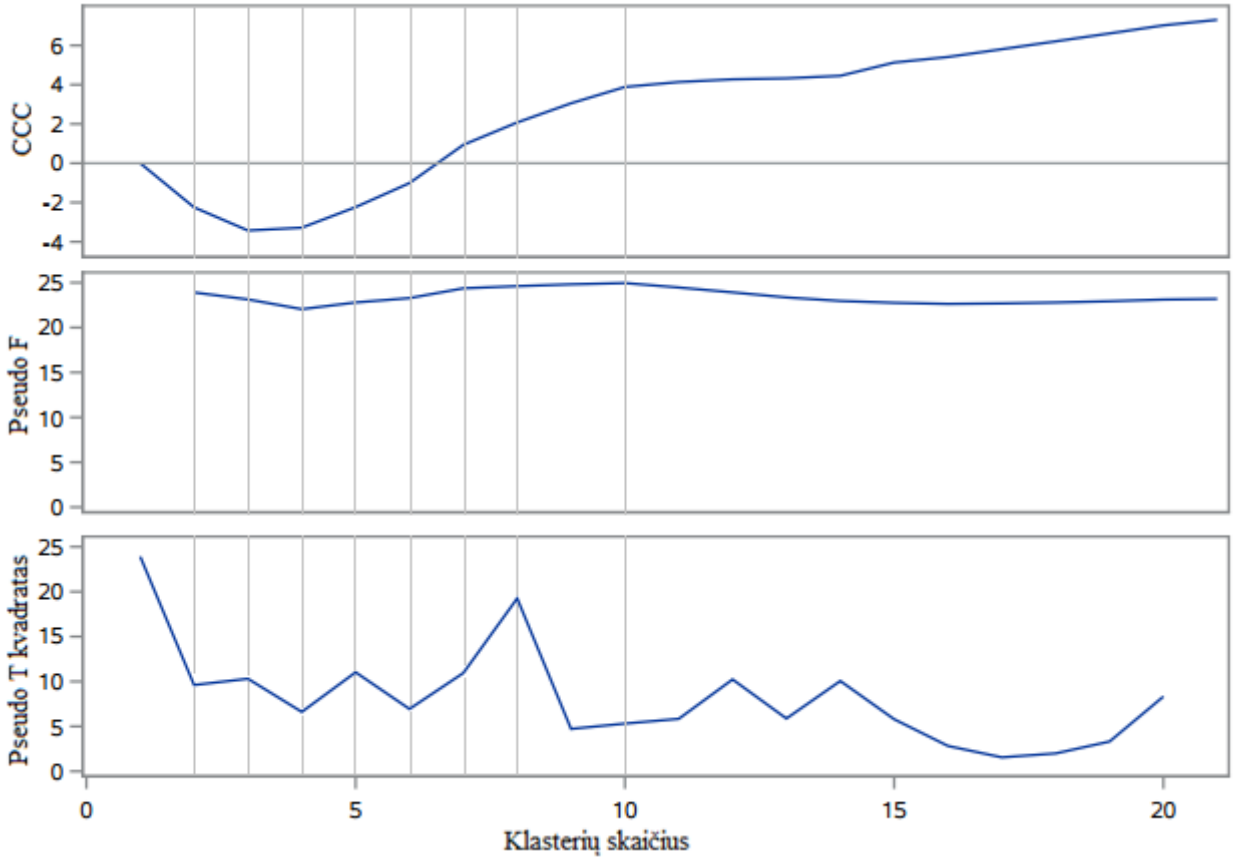
4P.1 pav. Pradinių duomenų po renovacijos sklaidos diagrama (po renovacijos Suomijoje)

Klasterizavimas naudojant hierarchinius metodus

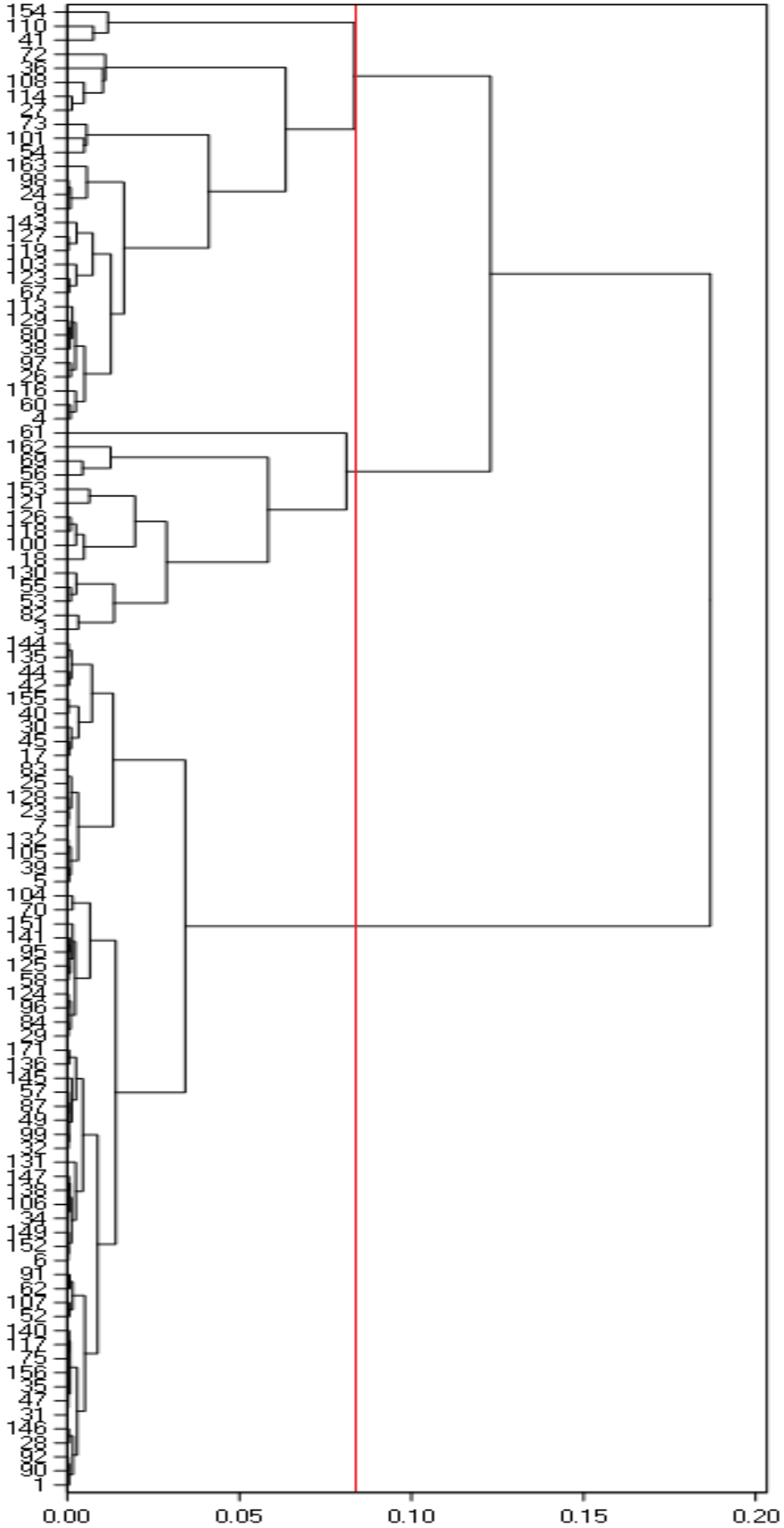
4P.3 lentelė. Klasterių jungimo protokolas (po renovacijos Suomijoje)

Klasterių skaičius	Sujungti klasteriai		Dažnis	Kubinis klasterizavimo kriterijus	Pseudo F statistika	Pseudo T - kvadratas
105	OB25	OB83	2	.	219	.
104	OB31	OB47	2	.	157	.
...						
11	CL15	CL26	19	4.14	24.5	5.8
10	CL32	CL25	6	3.88	25.0	5.3
9	CL13	CL10	11	3.05	24.8	4.7
8	CL12	CL14	61	2.08	24.6	19.3
7	CL11	CL27	22	0.94	24.4	10.9
6	CL9	CL16	14	-1.0	23.3	6.9
5	CL7	CL18	27	-2.2	22.8	11.0
4	CL6	OB61	15	-3.3	22.0	6.6
3	CL5	CL17	30	-3.4	23.1	10.3
2	CL4	CL3	45	-2.2	23.9	9.6
1	CL8	CL2	106	0.00	.	23.9

Klasterių skaičiaus nustatymas



4P.2 pav. Kriterijai klasterių skaičiaus nustatymui (po renovacijos Suomijoje)



4P.3 pav. Dendograma (po renovacijos Suomijoje)

Priimtas optimalus klasterių skaičius 3, kadangi nustačius didesnį grupių skaičių atsiranda klasterių su pavieniais elementais.

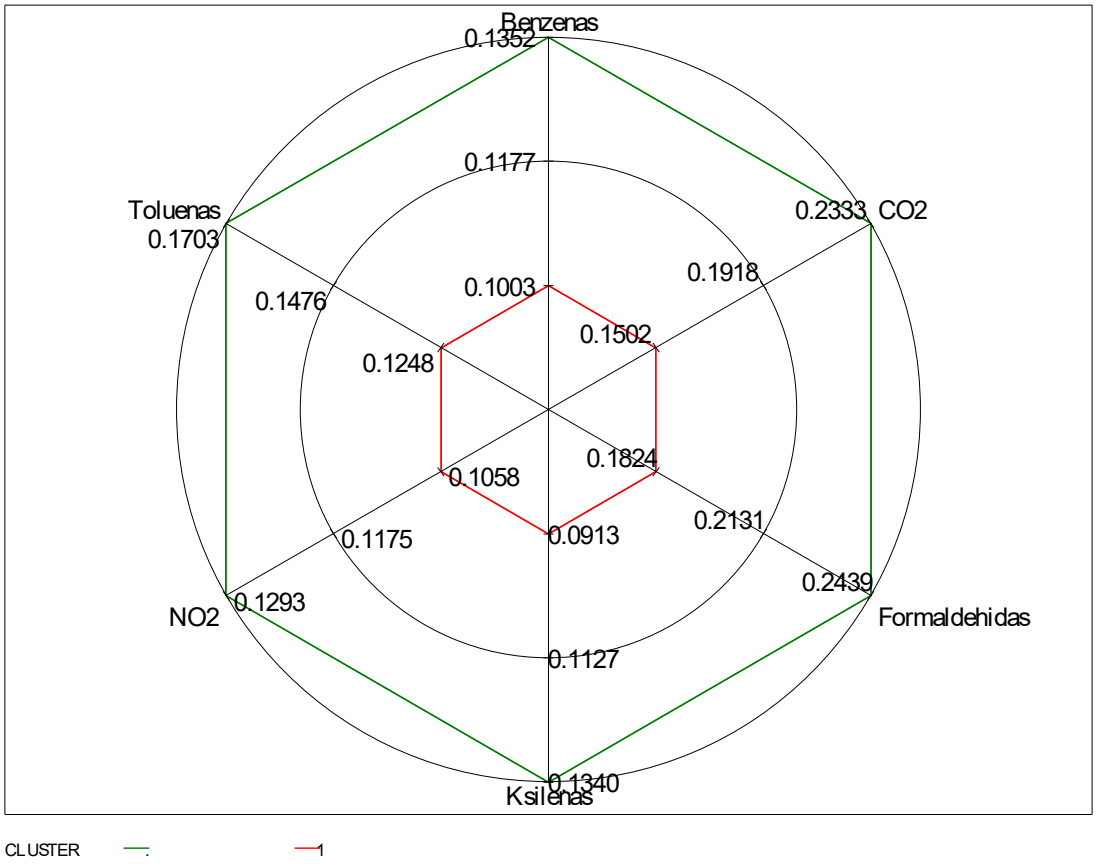
Nustatoma kiek stebinių priklauso kiekvienam iš klasterių:

- 1 klasteriui priklauso 61 objektas;
- 2 klasteriui priklauso 30 objektų;
- 3 klasteriui priklauso 15 objektų.

Taip pat nustatyta, kad 66 stebiniai nepriskirti nei vienam iš suformuotų klasterių.

Sudarytų klasterių požymių tyrimas

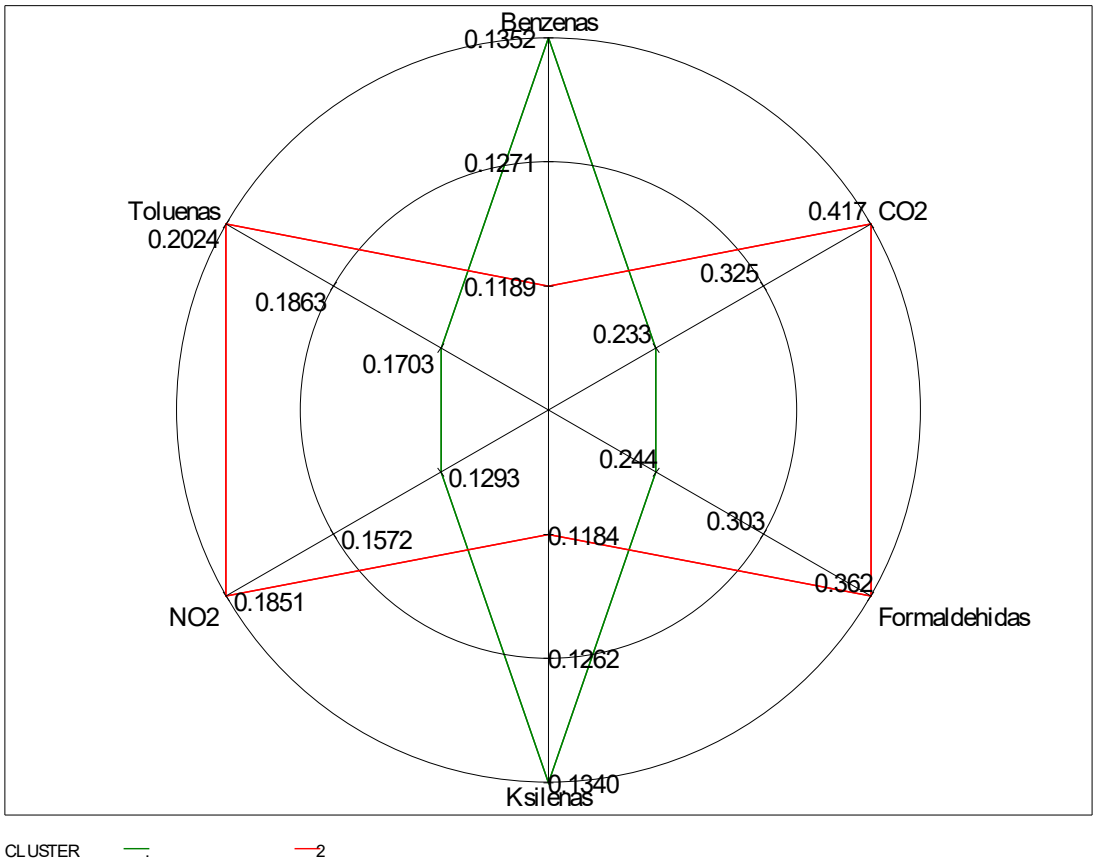
Nagrinėjamame klasteryje visų tiriamų teršalų kiekis neviršina vidutinių visos imties kiekių (4P.4 pav.).



4P.4 pav. Teršalų būdingų I-am klasteriui nustatymas (po renovacijos Suomijoje)

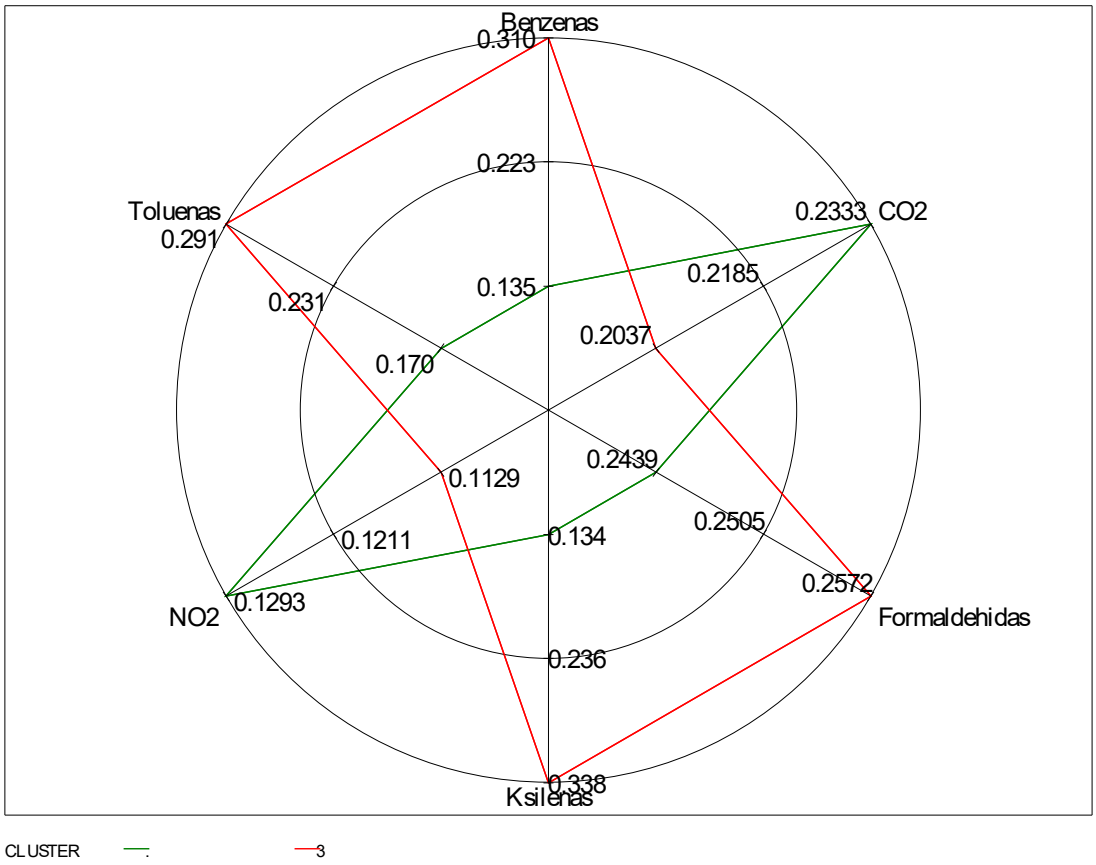
Iš žvaigždės formos grafiko matoma, kad antrajam klasteriui būdingas didesnis nei vidutinis tolueno, formaldehido bei anglies dioksido ir azoto dioksido kiekis (4P.5 pav.). Matome, kad tolueno vidutinis kiekis tiriamų patalpų ore sudaro 0.170 μg/m^3 tuo tarpu šiame klasteryje teršalo vidurkis - 0.202 μg/m^3. Formaldehido vidutinis kiekis tiriamų patalpų ore sudaro 0.244 μg/m^3, o tuo tarpu pirmame klasteryje šio teršalo vidurkis yra - 0.362 μg/m^3. NO_2 vidutinis kiekis tarp visų stebinių yra 0.129

μg/m^3, tuo tarpu nagrinėjamame klasteryje 0.185 μg/m^3. Vidutinis anglies dioksido kiekis tiriamoje imtyje 0.233 μg/m^3, o trečiame klasteryje šio teršalo vidurkis beveik dvigubai didesnis - 0.417 μg/m^3.



4P.5 pav. Teršalų būdingų II-am klasteriui nustatymas (po renovacijos Suomijoje)

Trečiajam klasteriui būdingas didesnis nei vidutinis benzeno, tolueno, radono, ksileno, bei formaldehido kiekis (4P.6 pav.). Vidutinis benzeno kiekis tarp visų stebinių yra 0.135 μg/m^3, tuo tarpu nagrinėjamame klasteryje 0.310 μg/m^3. Tolueno vidutinis kiekis tiriamų patalpų ore sudaro 0.170 μg/m^3 tuo tarpu šiame klasteryje teršalo vidurkis - 0.291 μg/m^3. Ksileno kiekis tiriamoje imtyje 0.134 μg/m^3, o klasteryje šio teršalo vidurkis daugiau nei du kartus didesnis - siekia 0.338 μg/m^3. Formaldehido vidutinis kiekis tiriamų patalpų ore sudaro 0.244 μg/m^3, o tuo tarpu pirmame klasteryje šio teršalo vidurkis yra - 0.257 μg/m^3.



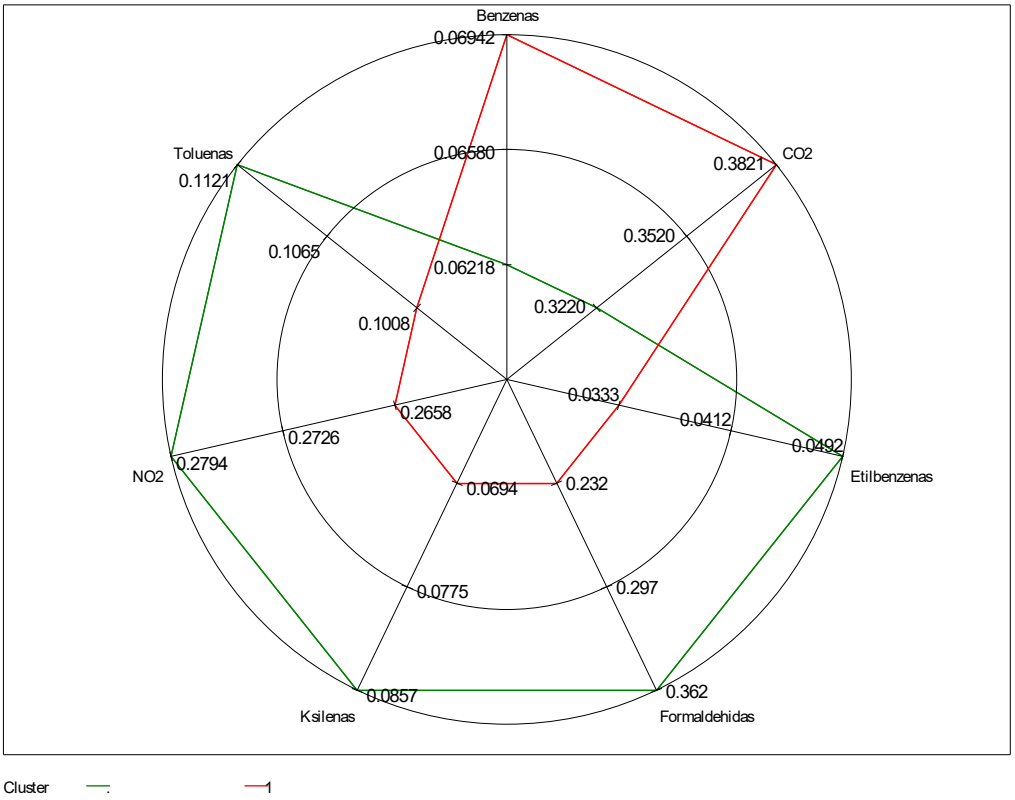
4P.6 pav. Teršalų būdingų III-iam klasteriui nustatymas (po renovacijos Suomijoje)

5 priedas

Klasterizavimas naudojant nehierarchinius metodus

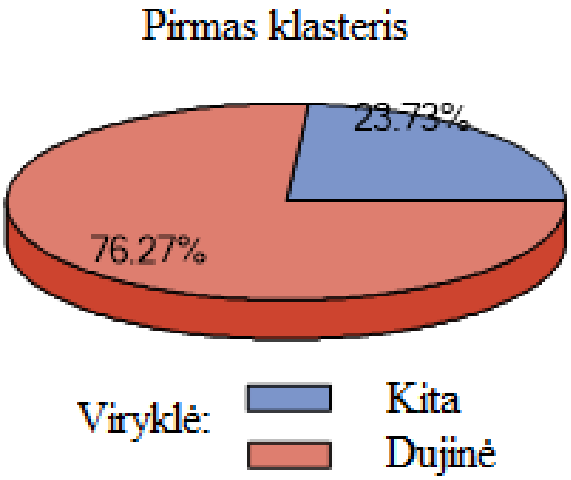
K-vidurkių metodas

Pirmajam klasteriui būdingi didesni nei vidutiniai benzeno ir anglies dioksido kiekiai (5P.1 pav.). Matoma, kad vidutinis benzeno kiekis tiriamoje imtyje 0.06218 μg/m3, o pirmame klasteryje šio teršalo vidurkis kiek didesnis – sudaro 0.06942 μg/m3. Vidutinis anglies dioksido kiekis tiriamoje imtyje 0.3220 μg/m3, o pirmame klasteryje šio teršalo vidurkis sudaro 0.3821 μg/m3.

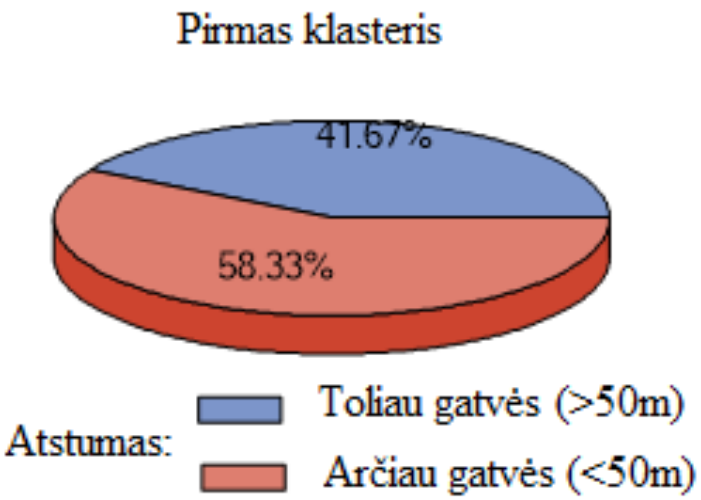


5P.1 pav. Pirmojo klasterio požymių tyrimas

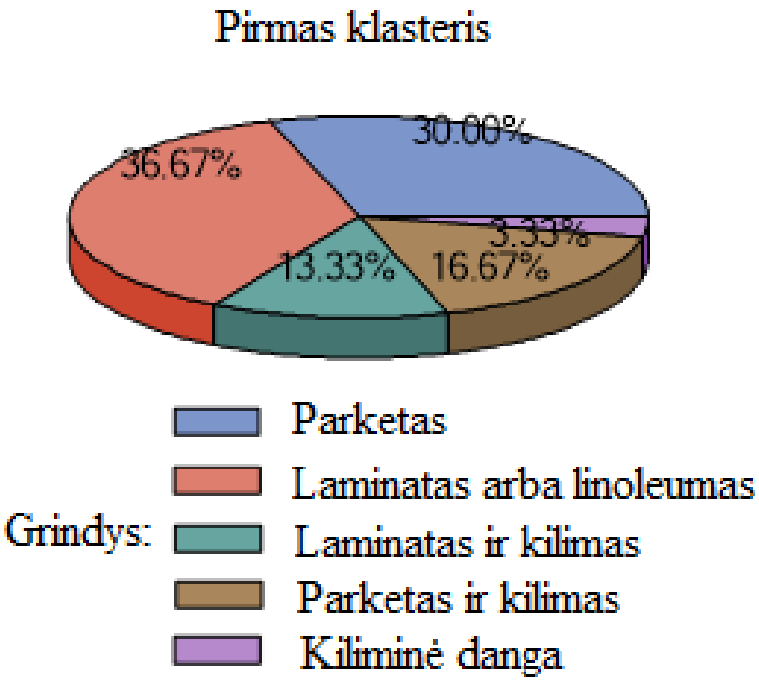
(5P.2 pav.), (5P.3 pav.), (5P.4 pav.) diagramose pateiktos pirmajam klasteriui priklausančių butų buitinių faktorių dažnumas procentais. Atitinkamai: viryklės tipas, atstumas iki gatvės bei grindų tipas.



5P.2 pav. Viryklės tipas I-ame klasteryje



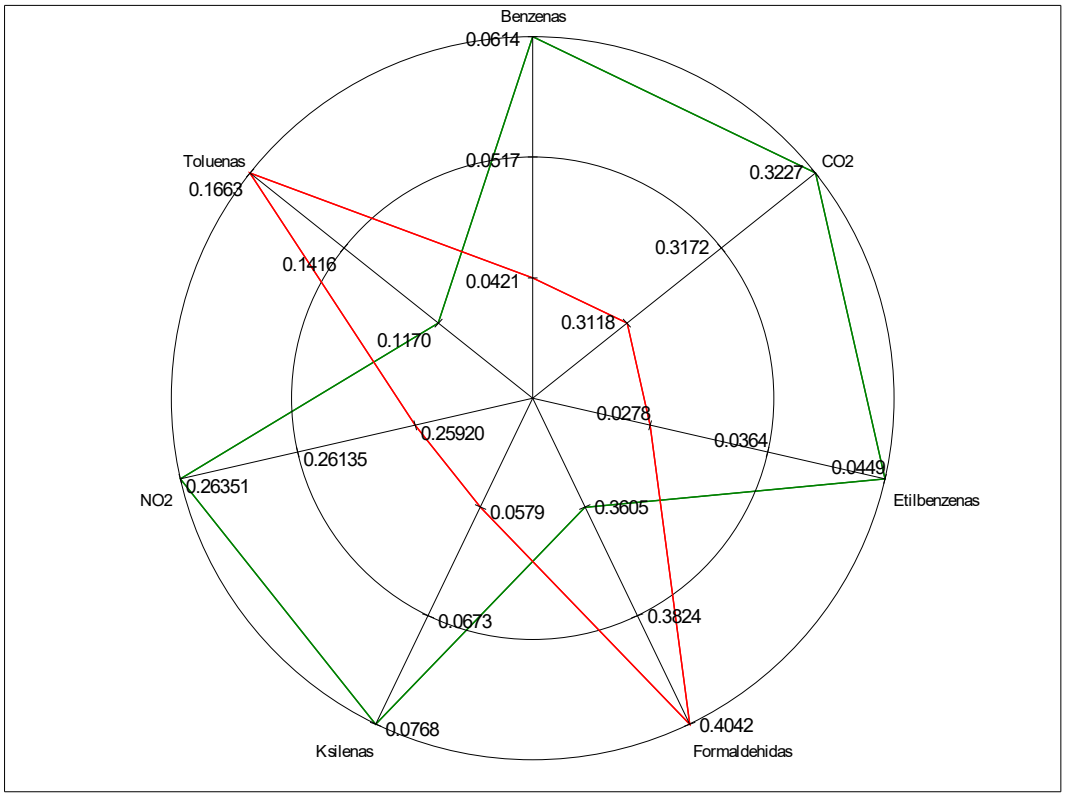
5P.3 pav. Butų atstumas iki gatvės I-ame klasteryje



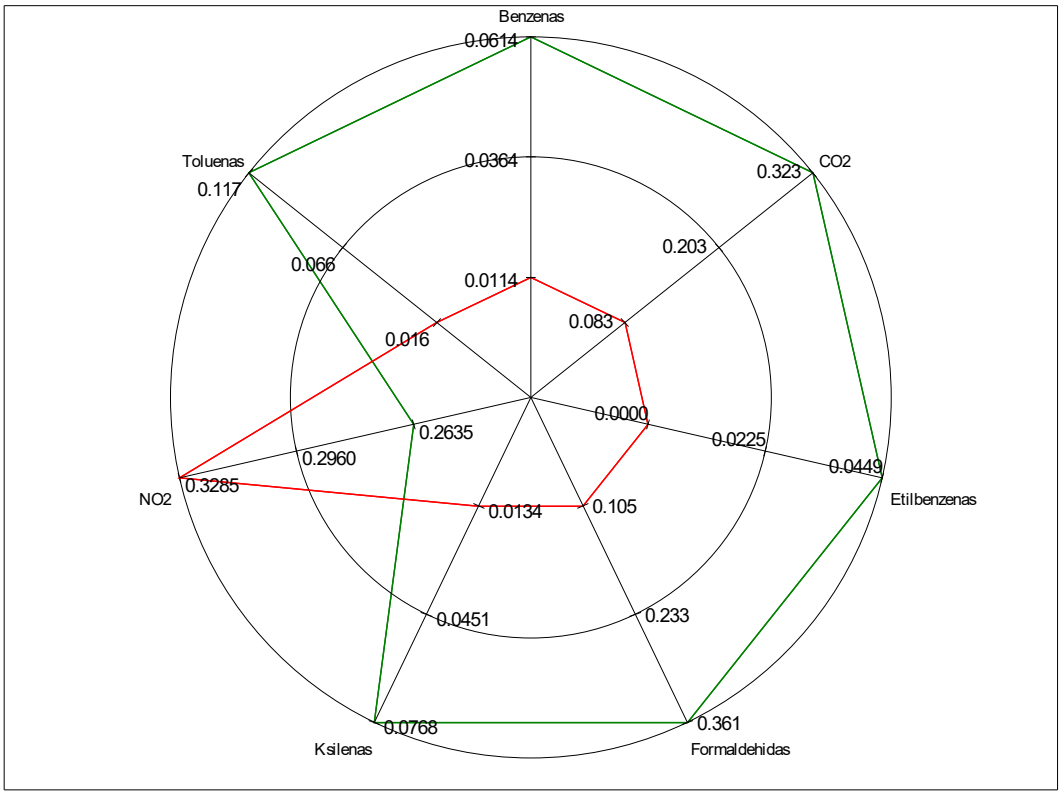
5P.4 pav. Grindų tipas I-ame klasteryje

K-artimiausių kaimynų metodas

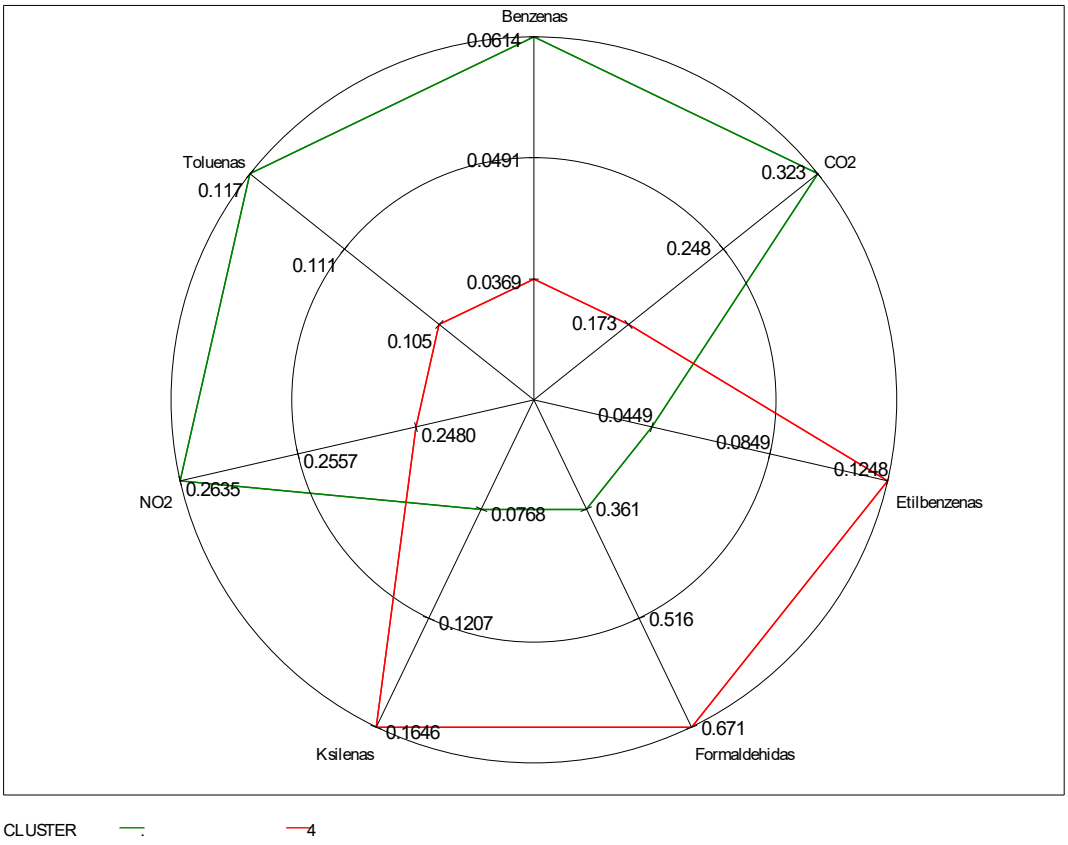
(5P.5 pav.) matoma, kad antrajam klasteriui būdingas didesnis nei vidutinis formaldehido bei tolueno kiekis ore. (5P.6 pav.) matoma, kad trečiajam klasteriui būdingas didesnis nei vidutinis NO_2 kiekis. (5P.7 pav.) matoma, kad pirmajam klasteriui būdinga didesnė nei vidutinė etilbenzeno, ksileno ir formaldehido koncentracija ore.



5P.5 pav. Antrojo klasterio požymių tyrimas

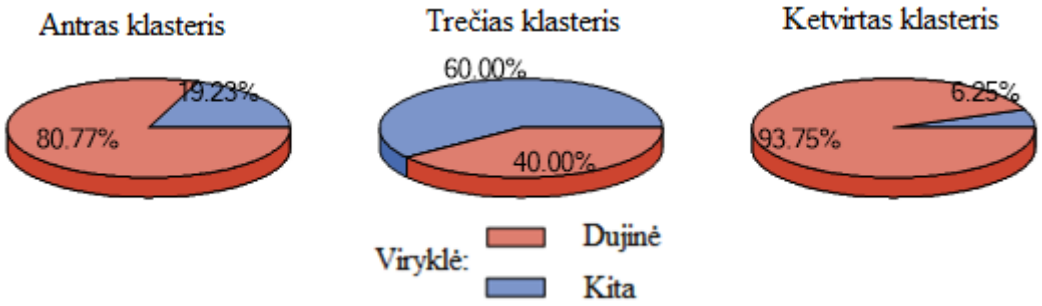


5P.6 pav. Trečiojo klasterio požymių tyrimas



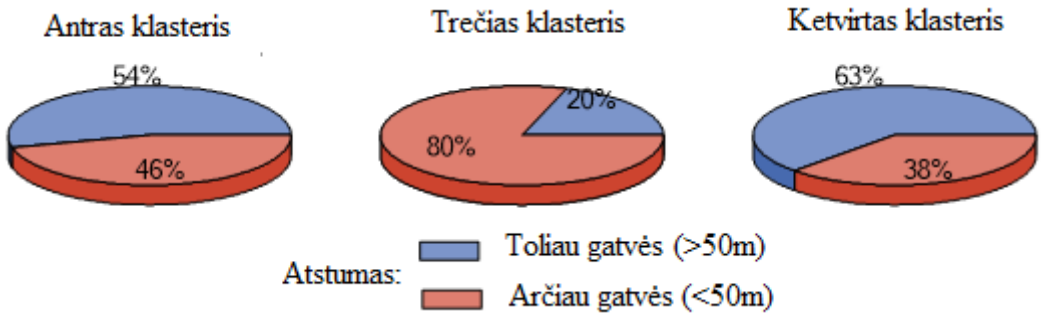
5P.7 pav. Ketvirtojo klasterio požymių tyrimas

(5P.8 pav.) diagramose pateiktos antrajam, trečiajam bei ketvirtajam klasteriams priklausančių butuose naudojamų viryklių tipai.



5P.8 pav. Viryklės tipas skirtinguose klasteriuose

(5P.9 pav.) diagramose taip pat pateiktos antrajam, trečiajam bei ketvirtajam klasteriams priklausančių butų atstumų iki gatvės.



5P.9 pav. Butų atstumas iki gatvės skirtinguose klasteriuose

(5P.10 pav.) diagramose pateiktos antrajam, trečiajam bei ketvirtajam klasteriams priklausančių butų remonto būklė.

5P.10 pav. Remonto būklės dažnumas procentais skirtinguose klasteriuose