TARTU ÜLIKOOL

Arvutiteaduse instituut

Informaatika õppekava

**JOHANNES HORM**

# Hädaabi väljakutsete kategoriseerimine Eesti Päästeameti andmete põhjal 2010–2013

**Bakalaureusetöö (9 EAP)**

Juhendaja: Siim Karus, PhD

Tartu 2016

## Hädaabi väljakutsete kategoriseerimine Eesti Päästeameti andmete põhjal 2010–2013

**Lühikokkuvõte:**

Bakalaureusetöö eesmärgiks on analüüsida väljakutsetelt kogutud andmeid ajavahemikus 2010–2013 automaatsetel andmekaeve meetoditel. Andmeid töödeldakse klasteranalüüsi meetodil, mille abil luuakse omavahel sarnaste väljakutsete kategooriad ehk klastrid. Uuritakse leitud mustreid, erandeid ja trende. Töö aitab kaasa Päästeameti 2015–2025 aastate strateegiale, mille üks sihtidest on tulemuslikkuse tõstmine kasutades väljakutsetelt kogutud äriandmeid. Andmete analüüsimine ärilisel eesmärgil võimaldab teha efektiivsemaid ning targemaid otsuseid ressursside kasutamisel, et tagada eestlastele kõrgem ohutus ja turvalisus.

**Võtmesõnad:**

Andmekaeve, äriteadmus, Päästeamet, Microsoft SQL Server, klasteranalüüs

**CERCS:**

P170 Arvutiteadus, arvanalüüs, süsteemid, kontroll; P175 Informaatika, süsteemiteooria

## Rescue event categorization based on Estonian Rescue Services data from 2010–2013

**Abstract:**

The aim of this Bachelor's thesis is to analyse the data collected from emergency responses between 2010 and 2013 with automatic data mining algorithms. The collected data is processed using cluster analysis methods, in which categories containing similar callouts are grouped into clusters. The focus is on patterns, exceptions and trends. The thesis helps the strategy set by the Estonian Rescue Services from 2015 to 2025. One of the main objectives set the by the strategy is to raise the effectiveness using business data gathered from callouts. Analysing the data for business intelligence helps the Estonian Rescue Services to make more effective and smarter decisions on how to use their limited resources to guarantee the safety and security of Estonians.

**Keywords:**

Data mining, business intelligence, Estonian Rescue Board, Microsoft SQL Server, cluster analysis

**CERCS:**

P170 Computer science, numerical analysis, systems, control; P175 Informatics, systems theory

# Sisukord

# Sissejuhatus

Töö valmib põhinedes Eesti Päästeameti[1] andmetel. Peamiseks eesmärgiks on analüüsida väljakutsetelt kogutud andmeid ajavahemikus 2010–2013 automaatsetel andmekaeve meetoditel. Andmeid grupeeritakse sarnasuste põhjal ja välja tuuakse vähem levinud väljakutsete iseloomuomadused, mis võiks muidu teiste varju jääda. Andmeid töödeldakse klasteranalüüsi meetodil, mille abil luuakse omavahel sarnaste väljakutsete kategooriad ehk klastrid (eestipäraselt ka kobar).

Eesmärgiks on leida väärtuslikku informatsiooni Päästeametile, mille abil oleks võimalik tuvastada trende, mustreid ja erandeid. Lõpptulemuseks on andmetega toetatud vaatlused, loodud kategooriatega, mida on võimalik kasutada järgnevate aastate väljakutsete paremaks mõistmiseks või hüpoteeside testimiseks. Tulemus annab Eesti Päästeametile väärtuslikku teadmust nende poolt reageeritavate väljakutsete kohta. Eesti Päästeameti analüütikutel on suur huvi leida siiani märkamata jäänud mustreid. Tööprotsessi on võimalik korrata koos täiendavate andmetega edasiseks analüüsimiseks.

Käesolev töö aitab kaasa Päästeameti 2015–2025 aastate strateegia [1] elluviimisele, mille üks alapunktidest on tulemuslikkuse tõstmine suurenevat infomahtu kasutades. Eesmärgiks on vähendada päästesündmuste arvu 23% võrra 2025. aastaks võrreldes 2013. aastaga. Andmete analüüsimine ärilisel eesmärgil võimaldab teha Päästeameti ressursside kasutamisel efektiivsemaid ning targemaid otsuseid, et tagada eestlastele kõrgem ohutus ja turvalisus.

Strateegias on seatud mitmeid sihte, kuid käesolevaga tööga kattuvad järgnevad:

- „päästesündmuste arv väheneb“;
- „hoonetulekahjude ja eluhoonete tulekahjude arv väheneb“;
- „varaline kahju hoonetulekahjudest väheneb“;
- „keskkonnaõnnetuste arv ja keskkonnakahju vähenevad“.

Bakalaureusetöö on jaotatud kolmeks: teoreetiline taust, meetod ning tulemused.

Teoreetiline taust annab ülevaate põhilistest kasutatavatest mõistetest, Eesti Päästeametist ning töös kastutatavast tarkvarast. Lugejale esitatakse üldpilt analüüsitavatest andmetest ning nende sisust ja taustast. Kirjeldatakse andmekaeve võimalusi, mida pakub tarkvara Microsoft SQL Server 2014 Analysis Services, ning kuidas neid on kasutatud.

Meetodi osas tuuakse detailselt välja, milliseid tarkvara versioone on kasutatud tulemuste saamiseks. Selgitatakse, kuidas on üles seatud projekt ning kuidas on leitud tulemused Microsoft SQL Server 2014 Analysis Services abiga. Põhjendatakse algoritmite seadistust koos analüüsitavate sisend- ja väljundveergudega ning selgitatakse andmete põhjal juurde loodud andmevälju, lubamaks paremini mõista andmete omavahelist sõltuvust.

Viimases osas tuuakse lisaks andmekaeve algoritmiga loodud kategooriatele välja ka selgitustega toetatud leiud andmete hulgast. Lisaks esitatakse andmetega toetatud vaatlused, mis on Eesti Päästeametile edastatav äriteadmus.

[1] Eesti Päästeamet - http://www.paasteamet.ee/

# 1 Teoreetiline taust

Andmekaeve võimaldab statistikute abita teha kasulikke avastusi andmetes leiduvate seoste kohta. Iga valdkond, mis loob andmeid, võib saada kasu andmekaevest. Töös kasutatavad meetodid, nt klasteranalüüs on vaid osa andmekaeve ning äriteadmuse poolt pakutavatest võimalustest. Peatükis on kirjeldatud töö läbiviimiseks kasutatud tarkvara ja toodud on ülevaade väljakutsete põhjal täidetud andmekogu põhilistest andmeveergudest.

## 1.1 Tähtsamad mõisted

**Andmekaeve (*Data Mining*)**

Andmekaeve tõlgendatuna raamatus „SQL server MVP Deep“ [2]. Andmekaeve võimaldab leida peidetud ehk tundmatuid teadmisi, uurides või treenides andmeid andmekaeve algoritmidega. See on protsess, mis aitab andmestikust leida arengusuundi, mille abil saab luua võimalikke tulevikumudeleid. Suundade avastamiseks kasutatakse ettemääratud tegureid, mida saab kasutada nii trendide ennustamiseks kui avastamiseks.

Algoritmid, millest populaarseimad põhinevad statistikal, aitavad väljendada teadmisi, mis leitakse andmestiku mustrites ning reeglites. Juhtum on andmekaeve vaatlusobjekt, mis võib olla rida, tabel või olem, ning mille atribuute kutsutakse muutujateks. Andmestikust leitud mustreid ning reegleid saab kasutada ennustuste tegemiseks.

**Äriteadmus (*Business Intelligence*)**

Äriteadmuse tähendus põhinedes raamatul „SQL server MVP Deep“ [2]. See on protsess ja infrastruktuur, mis aitab kaasa äriotsuste tegemisele kasutades selleks äriandmeid (*business data*). Äriteadmus põhineb tihti andmekaevest leitud informatsioonil. Teadmuse saamiseks muudetakse suurtes kogustes ärianalüüsi jaoks mõistmatuid andmeid arusaadavaks informatsiooniks, mille abil saab lühema aja jooksul teha äriliselt põhjendatud ning kaalutletud otsuseid. Äriteadmuse lahenduse arendamiseks tuleb mõista äri tuuma, mille abil saaks leida vastuseid, mis aitaks nii olevikus kui ka tulevikus.

**Klasteranalüüs (*Cluster Analysis*)**

Klasteranalüüs toetudes raamatule „Ruumiliste loodusandmete statistiline analüüs“ [3]. Klasteranalüüs on klassifitseerimise liik, milles toimub andmekogu ehk objektide hulga jaotamine alamhulkadesse tunnuste järgi. Tekkinud alamhulkadesse ehk klastritesse kuuluvad lähedased elemendid. Analüüsimeetodit kasutatakse andmetes seoste otsimiseks nende kobaratesse jagunemise järgi. Tegemist on kirjeldava andmeanalüüsi meetodiga, mis ei eelda, et analüüsitavate andmete kohta oleks nendes leiduvate sõltuvuste kohta püstitatud oletused.

Klasteranalüüs üritab leida andmete vahelist klassifikatsiooni ega tõesta klastrite olemasolu, kuigi seda saab kasutada eeldatavate kobarate olemasolu kinnitamiseks. Analüüsi tulemusena tekkinud kobarad ei pruugi olla statistiliselt olulised, kuid need võivad olla objektiivsed, sest otsimine käib kindlate reeglite järgi.

## 1.2 Eesti Päästeamet

Eesti Päästeamet on vastavalt Päästeameti aastaraamatule 2014 [4] asutatud 1992. aasta 25. mail, kui Riiklik Tuletõrjeamet nimetati ümber Päästeametiks. Amet kuulub Siseministeeriumi haldusalasse ning selle eesmärgiks on hoida ja kujundada turvalist elukeskkonda Eestis. Väärtusteks on abivalmidus, julgus ning usaldus. Põhilisteks ülesanneteks on ennetada võimalikke ohte ning aidata õnnetuse korral abivajajaid.

2014. aasta seisuga töötas Päästeametis 2232 inimest, mis teeb sellest suuruselt kolmanda avaliku sektori asutuse. Päästeamet koosneb neljast regionaalsest struktuuriüksusest: Põhja

Päästekeskus, Ida Päästekeskus, Lääne Päästekeskus ja Lõuna Päästekeskus. Bakalaureusetöö aitab kaasa Päästeameti visioonile „Aastaks 2025 on igaühe kaasabil vähenenud õnnetuste arv ja kahju Eestis Põhjamaade tasemele“.

Eesti Päästeamet pakub 16 erinevat avalikku teenust: tulekustutustöö , päästetöö baasteenus, metsatulekahjude kustutustöö, põlevvedelike kustutustöö, keemiapääste, saasteärastus, veepääste, nööripääste, loomapääste, kõrgustest päästetöö, päästetöö juhtimine, naftareostuskorje, üleujutuste pumpamise, logistika ja- transport, logistika sündmuskoha teenindus ja varingupääste . Selle töö valmimisel kasutati eelmainitud kuueteistkümnel teenusel põhinevate väljakutsete andmeid. Lisaks on Päästeamet märkinud väljakutsete alla ka õppuste andmed.

## 1.3 Andmete kirjeldus

Päästeameti poolt väljastatud andmekogumis on igal väljakutsel osaliselt olemas tabelis 1.1 välja toodud informatsioon. Ühel väljakutsel ei saa olla samal ajal kõik väljad täidetud, kuna erinevate sündmuse kirjeldamiseks on kasutatud erisuguseid veerge. See, mis veerud on väljakutsel täidetud, oleneb andmesisestajale teadaolevast informatsioonist ja väljakutse liigist. Seetõttu võib olla täidetud ühel väljakutse liigil „TULEKAHJU“ veerud „Väljakutse number“, „Väljakutse aeg“, „Väljakutse liik SOS“, „Maakond“ ja „Linn/Vald“, kuid teisel sama väljakutse liigiga väljakutsel võib olla lisaks informatsioon veergudes „Päästetöö algus“ ja „C jugade arv“. Mõned väljakutse liigid välistavad teised andmeväljad. Kirje „PT - ABI OSUTAMINE“ tüüpi väljakutse liigil ei saa olla täidetud „Veevõtukoha kaugus“. Autor arvestab seetõttu enimkasutavate andmeväljadega, mis ei ole täidetud ainult ühe teatud väljakutse liigi puhul.

*Tabel 1.1: Eesti Päästeameti andmekogu kirjeldus*

| Nr. | Andmevälja nimi | Andmevälja kirjeldus |
|---|---|---|
| **1.** | Väljakutse number | Väljakutseid eristav ainulaadne tunnus ehk võti |
| **2.** | Väljakutse aeg | Häirekeskusesse helistamise aeg formaadis pp/kk/aaaa tt/mm |
| **3.** | Päästetöö algus | Päästetööde alustamise aeg sündmuskohal formaadis pp/kk/aaaa tt/mm |
| **4.** | Lokaliseerimise aeg | Formaadis pp/kk/aaaa tt/mm |
| **5.** | Likvideerimise aeg | Päästetööde lõpetamise aeg formaadis pp/kk/aaaa tt/mm |
| **6.** | Väljakutse liik SOS | Väljakutsel Päästeameti poolt pakutav teenuseliik |
| **7.** | Väljakutse alamliik SOS | Väljakutsel pakutava teenuseliigi täiendav alamliik |
| **8.** | Sündmuse liik OPIS | Väljakutse sündmuse liik Operatiivinfosüsteemis |
| **9.** | Maakond | Eesti 15 maakonda ning Tallinn |
| **10.** | Linn/vald | Alates 2013. aastast on märgitud Tallinna linnaosad |
| **11.** | Hooneil korruseid | |
| **12.** | ATeS | Automaatse tulekahjuhäire edastamise süsteem |
| **13.** | ATeS Seisund | |
| **14.** | Suitsueemaldus. | |
| **15.** | Eritegevused | Väljakutsel tehtud eritegevused |
| **16.** | Rajatise tulekahju | Rajatise liik |
| **17.** | Sõiduki tulekahju | Sõiduki liik |
| **18.** | Haagise tulekahju | Haagise liik |
| **19.** | Maastiku tulekahju | Maastiku liik |
| **20.** | Paigaldise tulekahju | Paigaldise liik |

| Nr. | Andmevälja nimi | Andmevälja kirjeldus |
|---|---|---|
| **21.** | Muu tulekahju | Muu tulekahju liik |
| **22.** | Hukkunute arv | |
| **23.** | Hukkunud päästjate arv | |
| **24.** | Vigastatute arv | |
| **25.** | Päästetute arv | |
| **26.** | C jugade arv | Harilike tulekustutusvoolikute arv |
| **27.** | B jugade arv | Liitmiku küljes olevate jämedate voolikute arv |
| **28.** | Veevõtukoha kaugus | Meetrites |
| **29.** | Kulunud vesi | |
| **30.** | Kulunud A vahuaine | Kuupmeetrites |
| **31.** | Kulunud B vahuaine | Kuupmeetrites |
| **32.** | SS paaride arv | Suitsusukeldumise paaride arv (Suitsu sukeldutakse ainult paaris) |
| **33.** | SS aeg | Suitsusukeldumise aeg |
| **34.** | Maastiku põlemise pindala | Hektarites |
| **35.** | EVHK SOS | Häirekeskuse infosüsteemi automaatse aadressi sisestuse veerg. Kasutuses alates 2013. aastast |

Aastate 2010–2013 kohta on Eesti Päästeamet edastanud käesoleva töö jaoks 81 426 väljakutse kirjet: 2010. aastal 20 792, 2011. aastal 21132, 2012. aastal 19 237 ja 2013. aastal 20 265 väljakutset. Väljakutsete arvud aastatel 2010 ning 2011 ei kattu arvudega, mis on toodud välja Eesti Päästeameti 2014. aastaraamatus [4], kus on kirjas, et 2010. aastal toimus 23 164 ning 2011. aastal 22 124 väljakutset, kuna andmekogust on eemaldatud on väljakutsed liigiga „EKSLIKUD VÄLJAKUTSED TULEKAHJUDELE“.

## 1.4 Andmete taust

Käesolevas töös pärinevad andmed Eesti Päästeameti infosüsteemist OPIS[2], millest on tehtud väljavõte Microsoft Exceli failitüüpi xlsx[3]. OPIS oli kasutuses Päästeametis vahemikus 2010 kuni 2014 aprill, seejärel võeti kasutusele päästetöö andmestik PÄVIS[4]. Uus süsteem sisaldab peale sündmuste ning väljasõitude ka ressursside valmisolekut ja valvegraafikut.

Päästeameti poolt esitatud failis on viis erinevat lehte, millest igaüks esindab aastat vahemikust 2010–2014. Töös ei kasutata 2014. aasta andmeid, kuna need on poolikult täidetud uuele süsteemile liikumise tõttu, ja nende kasutamine äriteadmuse saamiseks on raskendatud. Aastatel 2010 kuni 2012 on tulekahjusündmused andmetes märgitud osaliselt topelt ning Päästeameti esindaja sõnul ei ole võimalik duplikaate tagantjärgi leida.

Töö autoril ei ole võimalik teha erinevate andmeväljade puhul täiendavat kontrolli, välja arvatud juhtudel kui need pole kirjavead või on sisestatud vale andmetüüp. Mõned väljakutse liigid võivad aastas esineda vaid ühel korral, mistõttu on nende põhjal raske teha järeldusi. Järgnevas andmevälju kirjeldavas alampeatükis on selgitatud nende sisu kasutades enim esinevaid kirjeid.

2013. aasta andmetest on kustutatud 111 real PT algus veeru lahtri kirjed, kuna formaat oli ebatäpne ning parandamine võimatu. Samal aastal on muudetud ühte väljakutset, mille

[2] OPIS - Operatiivinfosüsteem
[3] xlsx - Office Open XML Workbook formaat
[4] PÄVIS - Pääste valdkonna infosüsteem (https://www.riigiteataja.ee/akt/127032015012)

maakonnaks oli märgitud „endine LÄÄNE-VIRUMAA“ , uueks väärtuseks „LÄÄNE-VIRUMAA“.

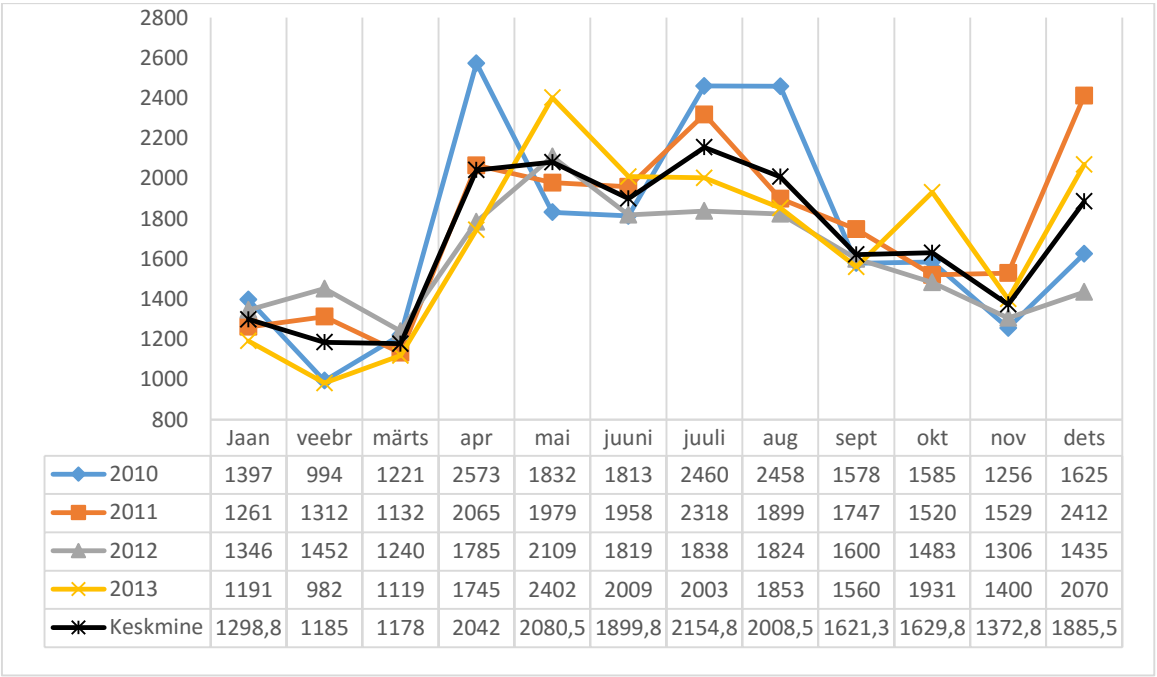
### 1.4.1 Tähtsamate andmeväljade ülevaade

Tähtsamate andmeväljade ülevaade põhineb andmetel, mis on saadud dokumendina Eesti Päästeametilt. Alampeatükis esitatud statistika ning esinemissagedus väljakutse liikide kohta erineb sellest, mis on esitatud Päästeameti aastaraamatus 2014 [4]. Autorile ei ole teada, millised väljakutsed võisid olla ekslikud või kas väljakutse liigid vastavad tegelikult sellele, mis tegevus väljakutsel teostati.

**Väljakutse aeg**

Veerg „Väljakutse aeg“ väljendab aega, millal toimus kõne häirekeskusesse. See on esitatud igal väljakutsel.

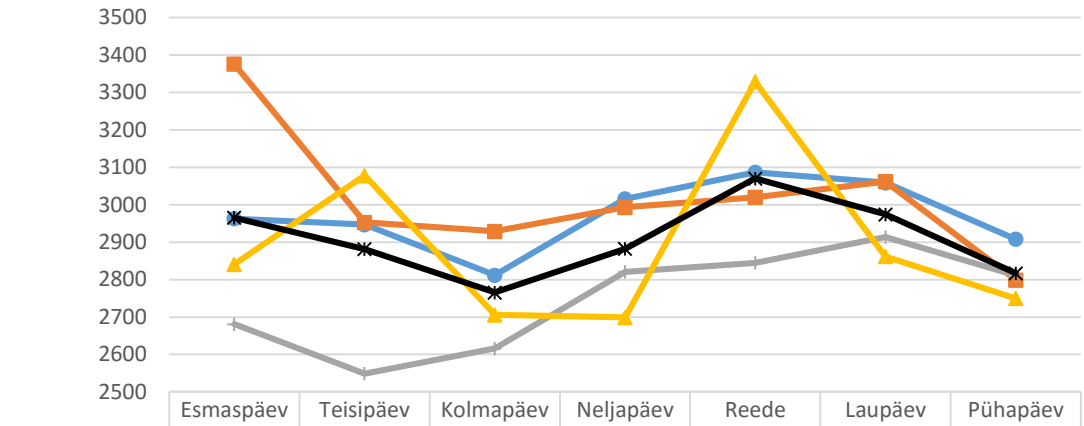
Ülevaade vahemiku 2010–2013 kõikide kuude kohta on välja toodud joonisel 1.1. Aastate iseloomujooned on omavahel võrreldes sarnased: väljakutsete arv püsib madal kuni märtsini, tõustes seejärel kõrgele aprillis, edaspidi langedes kuni novembrini, kuid suurenedes taas detsembris. Viie kõige kõrgema väljakutsete arvuga kuude hulgas kattuvad igal aastal 3 kuud: mai, juuni ja juuli.




| | Jaan | veebr | märts | apr | mai | juuni | juuli | aug | sept | okt | nov | dets |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2010 | 1397 | 994 | 1221 | 2573 | 1832 | 1813 | 2460 | 2458 | 1578 | 1585 | 1256 | 1625 |
| 2011 | 1261 | 1312 | 1132 | 2065 | 1979 | 1958 | 2318 | 1899 | 1747 | 1520 | 1529 | 2412 |
| 2012 | 1346 | 1452 | 1240 | 1785 | 2109 | 1819 | 1838 | 1824 | 1600 | 1483 | 1306 | 1435 |
| 2013 | 1191 | 982 | 1119 | 1745 | 2402 | 2009 | 2003 | 1853 | 1560 | 1931 | 1400 | 2070 |
| Keskmine | 1298,8 | 1185 | 1178 | 2042 | 2080,5 | 1899,8 | 2154,8 | 2008,5 | 1621,3 | 1629,8 | 1372,8 | 1885,5 |

*Joonis 1.1: Väljakutsete arv kuus*

Erinevalt kuude jagunemisest ei ole joonisel 1.2 toodud nädalapäevadel aastate lõikes ühtset trendi tekkinud. Väljakutsete maksimaalne amplituud nädalapäevade suhtes on alla 900, samal ajal kui suurim amplituud esmaspäeval 695 väljakutsega.




| | Esmaspäev | Teisipäev | Kolmapäev | Neljapäev | Reede | Laupäev | Pühapäev |
|---|---|---|---|---|---|---|---|
| 2010 | 2963 | 2947 | 2812 | 3016 | 3087 | 3059 | 2908 |
| 2011 | 3376 | 2953 | 2929 | 2993 | 3020 | 3062 | 2799 |
| 2012 | 2681 | 2549 | 2616 | 2821 | 2845 | 2914 | 2811 |
| 2013 | 2841 | 3078 | 2706 | 2699 | 3329 | 2862 | 2750 |
| Keskmine | 2965,25 | 2881,75 | 2765,75 | 2882,25 | 3070,25 | 2974,25 | 2817 |

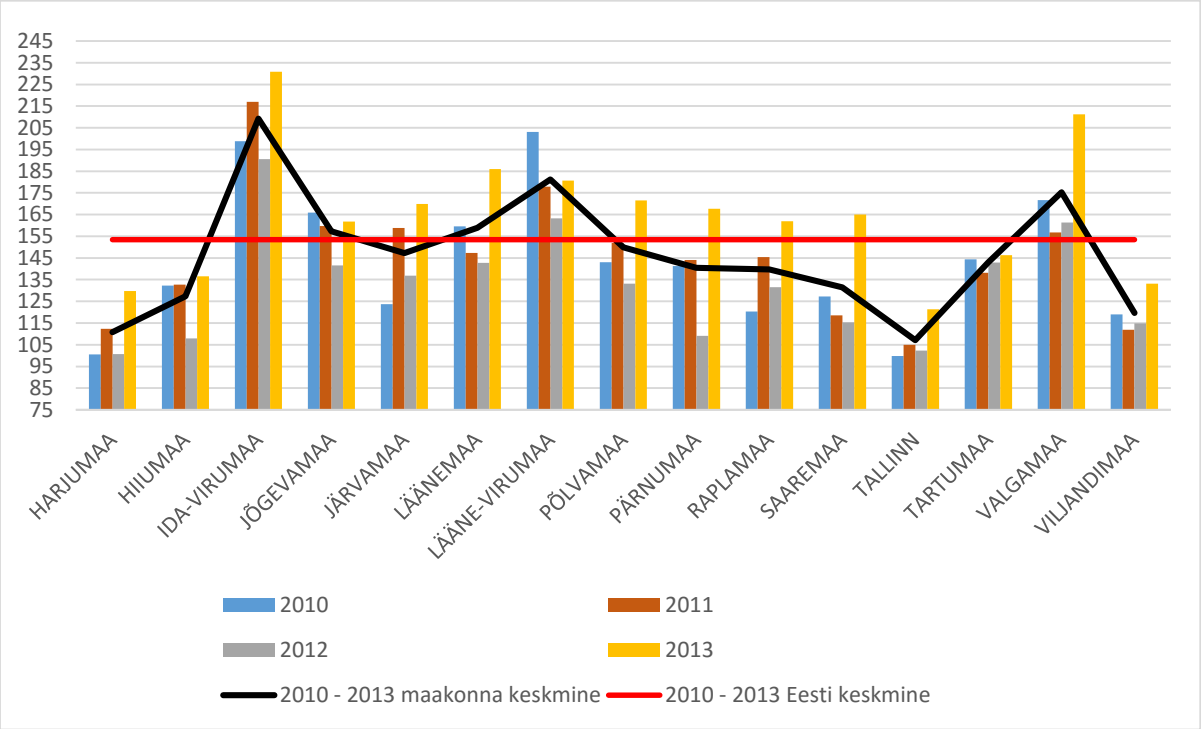
*Joonis 1.2: Väljakutsete arv nädalapäevas*

Võrreldes ülejäänud aastatega, erinevad 2010. aastal nädalapäevad omavahel väljakutsete arvu suhtes kõige vähem, maksimaalne amplituud on 247. Samal ajal on suurim kõikumine 2013. aastal, kus olenevalt nädalapäevast erineb arv maksimaalselt 630 väljakutse võrra.

## Maakond

Väljakutsed on märgitud kõikide Eesti maakondade kohta, lisaks on eraldi maakonna kirjena välja toodud Tallinn, mistõttu on veerus kokku 16 unikaalset kirjet. Andmeväli „Maakond“ on märgitud vahemikus 2010–2012 suuremale osale väljakutsetele ehk vastavalt 85,84; 85,6 ning 87,78 protsendile ja 2013. aastal kõikidele väljakutsetele.

Joonis 1.3 annab ülevaate andmeväljast „Maakond“, kus on võimalik näha aastatevahelisi sarnasusi väljakutsete arvust 10 000 elaniku kohta. Lisaks on toodud maakonna keskmine väljakutsete arv ning Eesti keskmine väljakutsete arv 10 000 elaniku kohta aastatel 2010–2013. Kõigi aastate mood on Tallinn, esinemissageduste arvult järgnevad Ida-Virumaa ja Tartumaa. Joonisel 1.3 kasutatavad maakondade elanike arvud pärinevad Eesti Statistikaameti andmebaasist [5].

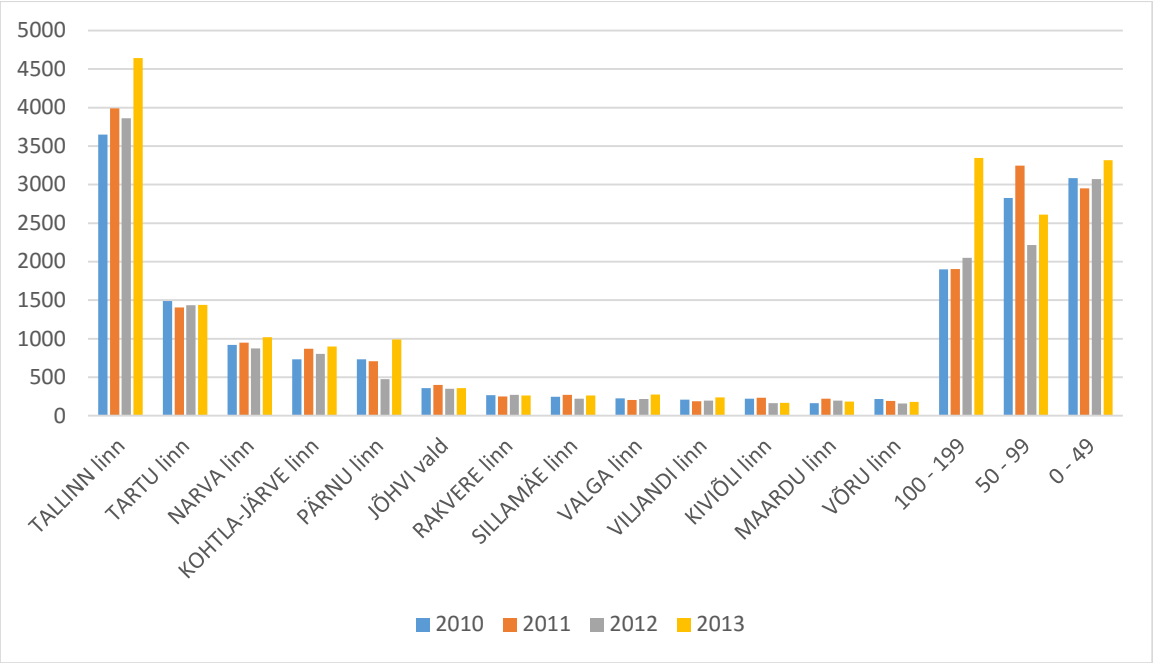



*Joonis 1.3: Väljakutsete arv 10 000 elaniku kohta*

## Linn/Vald

Andmevälja „Linn/Vald“ alla on märgitud Eesti linnad ning vallad. Alates 2013. aastast on eraldi märgitud juurde Tallinna suuremad linnaosad: Kristiine, Mustamäe, Lasnamäe, Haabersti ja Kesklinn. Joonisel 1.4 on välja toodud „Linn/Vald“ andmevälja esinemissagedused, millest selgub, et aastad on omavahel võrreldes sarnased.

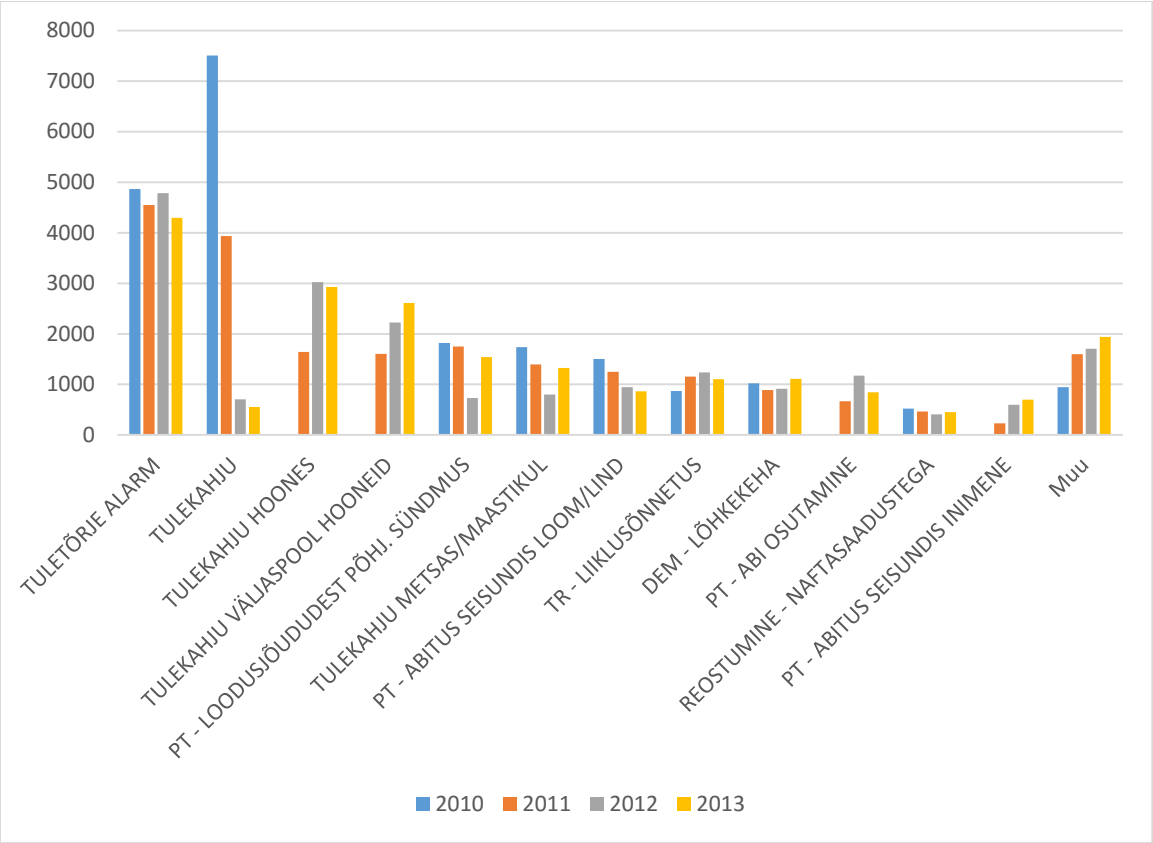
Andmete ühtlustamiseks on 2013. aastal eraldi esitatud linnaosade kirjed lisatud Tallinna väljakutsetele. Joonisel 1.4 on detailselt märgitud kõik kirjed, mille esinemissagedus ületas aastatel 2010–2013 vähemalt korra 200 piiri. Eraldi on grupeeritud „Linn/Vald“ veeru andmete väärtused, mis jäid vahemikku 100–199, 50–99 ning 0–49. 2010. aastast kuni 2012. aastani oli igal aastal 223 unikaalset kirjet, kuid 2013. aastal 234. Väärtused on järjestatud aastate keskmise kirjete arvu järgi.




*Joonis 1.4: Suurima esinemissagedusega „Linn/Vald" kirjed*

## Väljakutse liik SOS

Väljendab Eesti Päästeameti poolt pakutavat kuutteist erinevat teenust ning õppuseid. Mitmed kirjapandud väljakutsete liikidest on suuremate teenuste täpsustused. Näiteks väljakutse liik „REOSTUMINE" omab täpsustusi (andmed 2013. aastast): „REOSTUMINE - GAASILINE", „REOSTUMINE - KEEMILINE", „REOSTUMINE - NAFTASAADUSTEGA" ning „REOSTUMINE - RADIOAKTIIVNE". Igal väljakutsel on täidetud veerg „Väljakutse liik SOS". Joonisel 1.5 on välja toodud ülevaade andmeväljast.




*Joonis 1.5: Suurima esinemissagedusega väljakutse liigid*

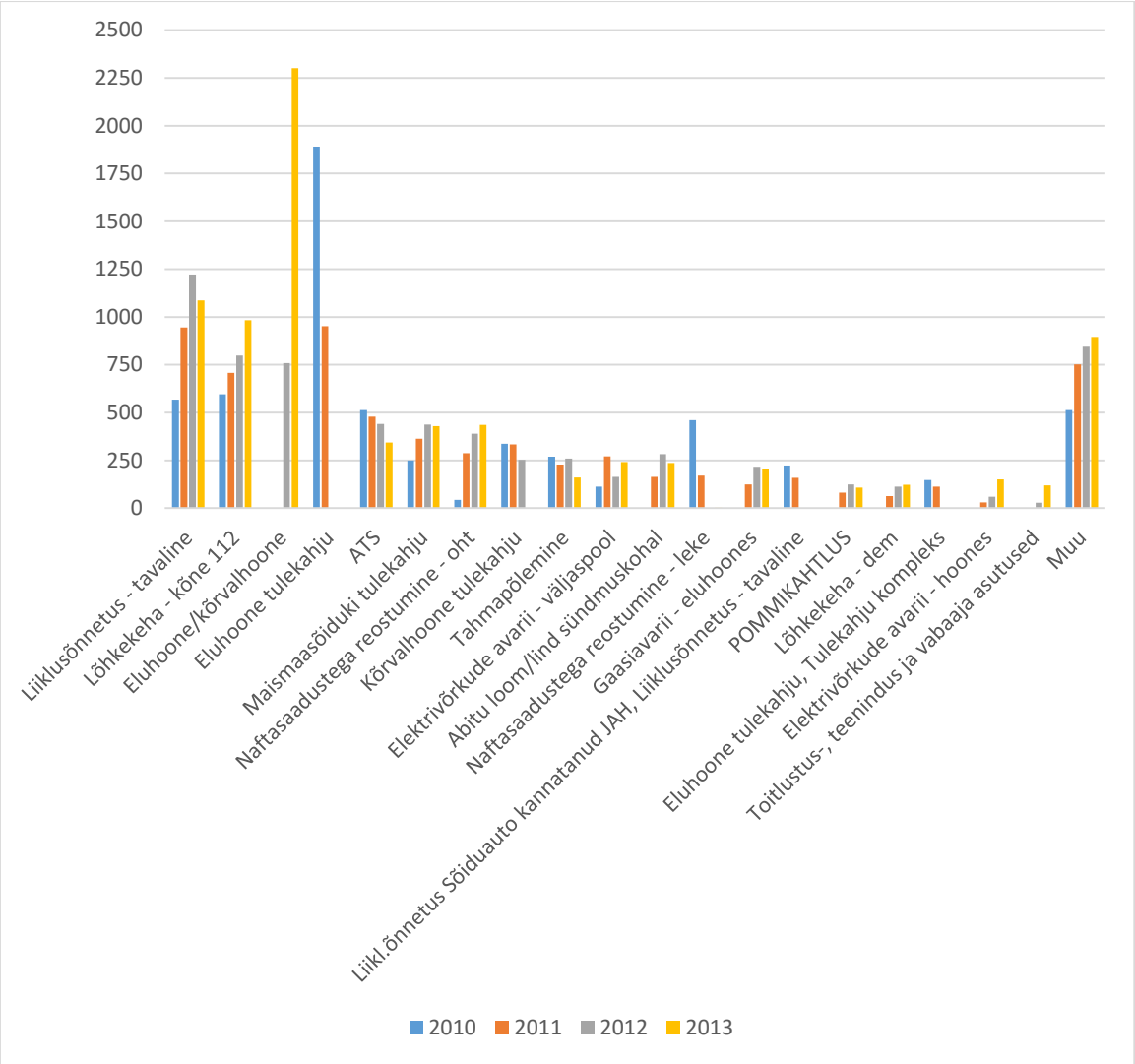
Joonisel 1.5 on kõik kirjed, mille esinemissagedus ületas 500 kordust vahemikus 2010–2013. 2010. aastal on kokku märgitud 18 unikaalset kirjet, kuid aastatel 2011–2013 on neid vastavalt 32, 32 ning 31. Väärtused on järjestatud aastate keskmise kirjete arvu järgi. Peale 2010. aastat lisati rohkem kirjeid väljakutsete väljendamiseks, mistõttu hakati eraldama liiki „TULEKAHJU“ teiste ja täpsemate kirjetega nagu „TULEKAHJU HOONES“ ning „TULEKAHJU VÄLJASPOOL HOONEID“. Lisatud on täiendavaid liike, mis on seotud näiteks päästetöödega „PT - ABI OSUTAMINE“ ning „PT - ABITUS SEISUNDIS INIMENE“.

## Väljakutse alamliik SOS

„Väljakutse alamliik SOS“ täiendab veeru „Väljakutse liik SOS“ kirjeid, millega sarnaselt leiduvad alamliikidel täpsustused. Näiteks olid alamliigil „ATS“[5] 2013. aastal järgnevad täpsustused: „ATS, elektrivõrkude avarii - hoones“; „ATS, Eluhoone/kõrvalhoone“; „ATS, Majutus ja hooldekandeasutused“; „ATS, Meditsiiniasutused“ ja „ATS, Toitlustus-, teenindus ja vabaaja asutused“.

Joonisel 1.6 on toodud ülevaade andmeväljast „Väljakutse alamliik SOS“, kuhu on lisatud kõik kirjed vahemikust 2010–2013, kus vähemalt ühel aastal ületas kirje esinemissagedus 100 piiri. Joonisel 1.6 on liike, mida ei esine igal aastal. Aastatel 2010 kuni 2013 oli märgitud vastavalt 48, 76, 99 ning 77 unikaalset liiki. Väärtused on järjestatud aastate keskmise kirjete arvu järgi.

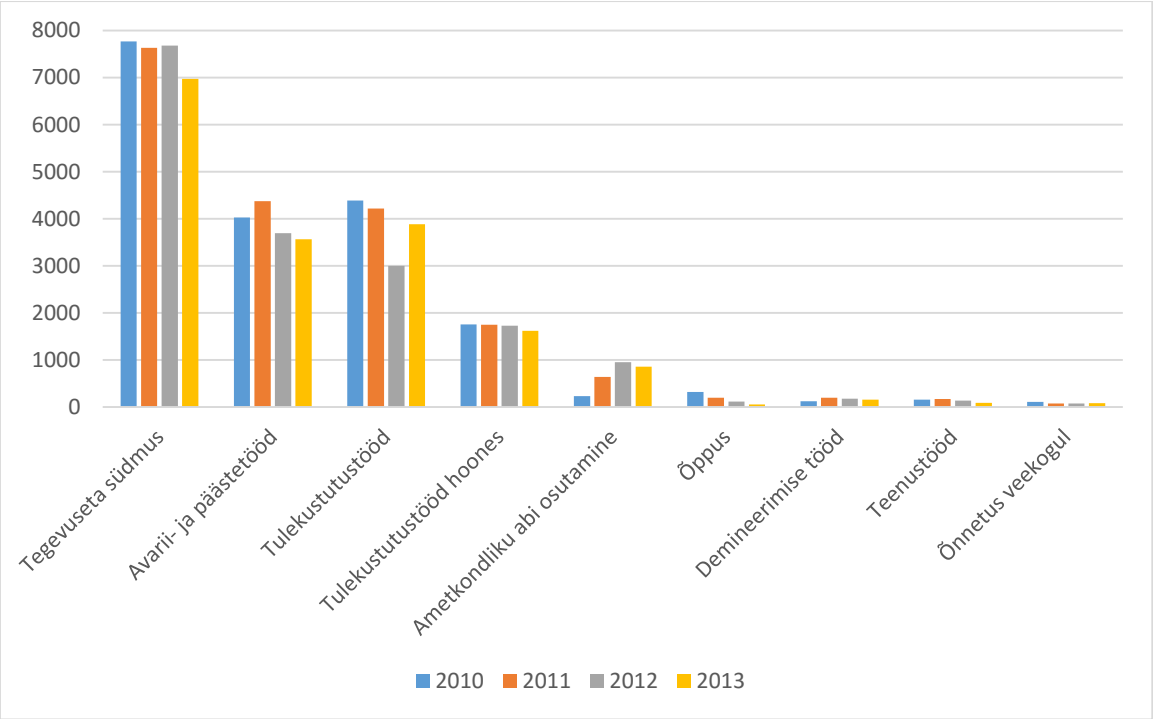
[5] ATS - Automaatne tulekahjusignalisatsioonisüsteem




*Joonis 1.6: Suurimate esinemissagedustega väljakutse alamliigid*

## Sündmuse liik OPIS

„Sündmuse liik OPIS“ sisaldab kirjeid, mida kasutatakse hädaabiteadete menetlemise programmis. OPIS oli Päästeameti peamine infosüsteem ning selle sündmuse liigi abil liigitati väljakutseid. Ülevaade andmeväljast on esitatud joonisel 1.7, kust on välja jäetud kaks sündmuse liiki nende väikese esinemissageduse tõttu: „Ekslik“ esineb vaid ühe korra ning „Teenistuslik väljasõit“ 43 korda nelja aasta jooksul. Väärtused on järjestatud aastate keskmise kirjete arvu järgi.




*Joonis 1.7: Sündmuse liik OPIS esinemissagedus*

## 1.5 Microsoft SQL Server 2014 Analysis Services

Microsoft SQL Server 2014 Analysis Services-i ülevaade põhineb Microsoft Developer Networkil [6]. Analüüsi teenuspakett (*analysis services*) on võrgupõhine analüütiline andmemootor (*data engine*), mida kasutatakse äriteadmuse saamiseks ning selle põhjal otsuste tegemiseks. Teenuspakett on võimeline edastama andmeid äriraportite ning kliendirakenduste (*client application*) jaoks. Peatükis tutvustatakse lähemalt programmi võimalusi ning teooriat.

Käesolevas töös kasutatakse teenuspaketti mitmemõõtmelises (*multidimensional*) konfiguratsioonis, sest sellel on iseseadistuvate andmebaasipäringute, mis käivad äriandmete vastu, korral hea sooritusvõime. Andmebaasipäringute tegemiseks kasutatakse analüüsi teenuspaketis relatsioonilist andmebaasikeelt MDX[6], mis on kaudselt sarnane SQL[7] keelega, kuid ei ole selle täiendus. MDX põhineb XMLA[8] spetsifikatsioonil [7]. Mitmemõõtmeline mudel koosneb kuupidest ning dimensioonidest, mida saab annoteerida ning täiendada, et luua keerulisi konstruktsioone andmebaasipäringuks [8].

### 1.5.1 Andmekaeve algoritmid

Tarkvarasse Microsoft SQL Server 2014 Analysis Services on sisse ehitatud üheksa [9] erinevat andmekaeve algoritmi, mis kuuluvad järgnevate tüüpide alla:

- Klassifikatsiooni algoritmid (*Classification algorithms*),
- Regressiooni algoritmid (*Regression algorithms*),
- Segmentimise ehk lõikudeks jaotamise algoritmid (*Segmentation algorithms*),
- Assotsiatsiooni algoritmid (*Association algorithms*),
- Järjendalgoritmid (*Sequence analysis algorithms*).

Töös kasutatakse andmekaeveks äriteadmuse eesmärgil kaht algoritmi kahest erinevast tüübist.

[6] MDX–Multidimensional Expressions
[7] SQL–Structured Query Language
[8] XMLA–Extensible Markup Language for Analysis

**Microsofti klasterdamise algoritm (Microsoft Clustering Algorithm)**

Alljärgnev selgitus algoritmi kohta pärit Microsoft Developer Networki [10] lehelt. Microsofti klasterdamise algoritm on segmentimise algoritm, mis grupeerib sarnaste tunnustega andmestikku iteratiivselt. Eesmärgiks on leida suhteid andmestiku sees, mida tavapärasel ülevaatusel tuletada ei suudetaks. Mudel treenitakse põhinedes seostel, mis eksisteerivad andmetes, ning klastrites, mis suudetakse leida.

Peale klastrite esmast määratlemist arvutatakse, kui hästi esindab klaster punktide rühmitust ning seejärel proovitakse rühmitust uuesti defineerida, et andmeid paremini esitada. Algoritm kordab sammu kuni ei suuda enam tulemusi parandada. Täpsem protsess oleneb kasutaja valitud klasterdamise meetodist. Võimalik on valida 4 erineva grupeerimisalgoritmi vahel, mida on täpsemalt kirjeldatud algoritmi seadistamise võimaluste juures.

**Microsofti naiivne Bayesi algoritm (Microsoft Naive Bayes Algorithm)**

Microsofti naiivne Bayesi algoritm, tuginedes Microsoft Developer Networki [11] kirjeldusele, on klassifikatsiooni algoritm, mis põhineb Suurbritannia matemaatiku Thomas Bayesi teoreemidel, ning mida kasutatakse ennustavas modelleerimises. Tegemist on naiivse algoritmiga, kuna see kasutab bayeslikke tehnikaid, kuid ei võta arvesse puuduvaid sõltuvusi. See on arvutuslikult vähenõudlikum kui teised SQL Serveri andmekaeve algoritmid. Sisendparameetrid peavad olema diskretiseeritud (*discretized*), ehk üle viidud pidevast diskreetseks, leidmaks lõpliku arvu üksteisest iseseisvaid väärtuseid. Algoritm arvutab sisendveeru iga võimaliku oleku tõenäosuse ennustatava veeru suhtes.

### 1.5.2 Andmekaeve algoritmide vaikeparameetrid

Igal algoritmil on erinevad vaikeparameetrid, mille väärtusi on võimalik muuta vastavalt etteantud väärtuste vahemikule. Parameetrid on võimalik ka ise juurde lisada. Parameetrite vaikeväärtuste muutmine mõjutab sealjuures andmekaeve algoritmi tulemust. Nende abil on võimalik täpsustada lõpptulemust seda laiendades või kitsendades, mistõttu tuleb olla teadlik parameetrite muutmiste tagajärgedest, et algoritme mitte üle treenida.

**Microsofti klasterdamise algoritm**

Microsoft SQL Server 2014 Analysis Services-i klasterdamise algoritm kasutab seadistuseks vaikimisi üheksat erinevat parameetrit. Sisendparameetrite muutmisega saab mõjutada algoritmide lõpptulemust, mistõttu on oluline täita parameetrid vastavalt sisendandmetele, et saada võimalikud täpsed tulemused. Järgnevalt on selgitatud lühidalt lahti kõik vaikeparameetrid algoritmi tehnilisele selgitusele põhinedes [12].

**Klastrite arv (*Cluster count*)** väljendab eeldatavat klastrite arvu, mida algoritm võib luua. Kui algoritm ei suuda luua võrdselt klastreid sellega, mitu on sisendiks antud, luuakse neid võimalikult palju. Sisendväärtuse 0 korral otsustab algoritm ise andmete põhjal, mis on loodavate klastrite optimaalne arv.

**Klastri seemnete arv (*Cluster seed*)** määrab seemnete arvu, mida kasutatakse juhuslike klastrite tekitamisel mudelite ehitamise algses staadiumis.

**Klasterdamise meetod (*Clustering method*)** määrab algoritmis kasutatava klasterdamise meetodi. Kokku on 4 võimalikku meetodit:

1. Skaleeritav ootuste maksimeerimine (*Scalable expectation maximization*),
2. Mitteskaleeritav ootuste maksimeerimine (*Non-scalable expectation maximization*),
3. Skaleeritavad K-vahendid (*Scalabe K-means*),
4. Mitteskaleeritavad K-vahendid (*Non-scalabe K-means*).

**Ootuste maksimeerimine** on pehme klasterdamise meetod, mis lubab ühel andmepunktil kuuluda igasse klastrisse ning iga andmepunkti ja klastri kombinatsiooni jaoks arvutatakse eraldi tõenäosus. Algoritm lõpetab töö, kui tõenäosusmudel sobitub andmetega. Skaleeruvas meetodis kasutatakse vaikimisi 50 000 esimest kirjet algse uurimise seemendamiseks, mille õnnestumise puhul rohkem andmeid ei kasutata. Kui mudel on suurem kui 50 000 kirjet, siis loetakse lisaks sisse veel samaväärselt andmeid kuni kõik andmekogu on sisse loetud. Mitteskaleeruvas meetodis kasutatakse kogu andmehulk suurusest olenemata korraga ära.

**K-vahendid** on tugev klasterdamise meetod, mis lubab andmepunktil kuuluda rangelt vaid ühte klastrisse, seetõttu arvutatakse klastrisse kuulumise tõenäosus vaid korra. Eesmärgiks on siduda andmed klastriga nii, et minimaliseeritakse klastri elementide erinevus, maksimeerides samaaegselt klastrite omavaheline kaugus. „Vahendid" meetodi nimes väljendab klastri raskuskeset, milleks on juhuslikult valitud andmepunkt ning mida rafineeritakse seni, kuni see väljendab kõikide andmepunktide keskmist klastris. „K" viitab juhuslikule punktide arvule, mida kasutatakse klasterdamise protsessis seemendamiseks. Vahe skaleeruva ning mitteskaleeruva meetodi põhimõttes on sama, mis ootuste maksimeerimise meetodis.

**Maksimaalne sisendatribuutide arv *(Maximum input attributes*)** seab maksimaalse sisendatribuutide arvu, mida algoritm suudab töödelda enne tunnuste valiku rakendamist.

**Maksimaalne olekute arv (*Maximum states*)** määrab maksimaalse atribuudi olekute arvu, mida algoritm toetab. Arvu ületamisel kasutatakse vaid kõige populaarsemaid olekuid ning ülejäänud eemaldatakse.

**Minimaalse toetuse (*Minimum support*)** parameeter seab minimaalse juhtumite arvu, mis on vajalik klastri ehitamiseks. Kui arv on väiksem, käsitletakse klastrit tühjana ja see eemaldatakse algoritmist.

**Modelleerimise kardinaalsus (*Modelling cardinality*)** määrab ära ehitatavate proovimudelite arvu klastri ehitamise ajal.

**Näidise suurus (*Sample size*)** määrab ära andmete hulga, mida algoritm iteratsiooni ajal kasutab. Seda kasutatakse juhul kui klasterdamise meetodiks on valitud skaleeruv meetod. Kui seada parameetri väärtuseks 0, loetakse korraga sisse kõik lähteandmed. Kõikide andmete korraga analüüsimine võib tekitada arvutile sooritusprobleeme.

**Peatumistolerants (*Stopping tolerance*)** määrab ära väärtuse, milleni jõudes algoritm otsustab, et koondumine sisendandmetes on toimunud ning mudeli ehitamine lõpetatakse. See juhtub siis, kui üleüldine muutumine klastrite tõenäosustes on väiksem kui suhe peatumistolerantsi jagamisel mudeli suurusega.

**Microsofti naiivne Bayesi algoritm**

Microsofti naiivne Bayesi algoritm kasutab vaikimisi nelja parameetrit [13]. Maksimaalsete sisendatribuutide arvud ning maksimaalsete olekute arvu parameetrid kattuvad Microsofti klasterdamise algoritmi omadega.

**Maksimaalne väljundatribuutide arv (*Maximum output attributes*)** määrab ära maksimaalse arvu väljundatribuute, mida algoritm suudab käsitleda enne atribuutide vahel valiku tegemist (valitakse kõige levinumad). Kui väärtus seada nulliks, siis ei tehta kunagi valikut atribuutide vahel.

**Minimaalne sõltuvuste tõenäosus (*Minimum dependency probability*) s**eab minimaalse sõltuvuse tõenäosuse sisend- ning väljundatribuutide vahel. Selle väärtuse abil saab piirata algoritmist saadavat tulemuse hulka – mida suurem on väärtus, seda väiksem on atribuutide arv lõppmudelis.

### 1.5.3 Sisutüübid (*Content Types*)

Sisutüüpide alampeatükk põhineb Microsoft Developer Networki veebilehel [14]. Igal andmeväljal on enda sisutüüp, mida on Microsoft SQL Server 2014 Analysis Services programmis 10 erinevat. Sisutüübid määravad ära selle, kuidas algoritmid neid mõistavad ja kasutavad. Valesti määramine võib põhjustada algoritmide kasutamisel probleeme: sisendandmeid töödeldakse valesti või algoritmi ei ole võimalik rakendada. Käesolevas töös kasutatakse viit erinevat sisutüüpi.

**Diskreetne (*Discrete*):**

Sisutüüp on diskreetne, kui veerus leidub vaid lõplik arv väärtusi, mille vahel ei leidu kontiinumi[9]. Lubatud on kõik andmekaeve andmetüübid Microsoft SQL Serveris.

**Pidev (*Continiuous*):**

Väärtus on pidev, kui veerus oleval numbrilisel väärtusel on lubatud ka vahepealsed väärtused. Pidev sisutüüp võib omada lõpmatult palju murdosaväärtuseid. Lubatud andmetüübid on kuupäev, pikk täisarv ning ujukomaarv.

**Võti (*Key*):**

Võti on unikaalne identifikaator, mille abil eristatakse andmeridu üksteisest. Toetatud andmetüübid on kuupäev, pikk täisarv, tekst ning ujukomaarv.

**Diskretiseeritud (*Discretized*):**

Diskretiseerida saab ainult numbrilisi väärtusi. Diskretiseerimine on protsess, mille puhul sisestatakse pidevaid väärtusi kogumikesse. See tekitab olukorra, kus kogumikes on ainult piiratud hulk võimalikke väärtusi. Seega on diskretiseeritud väärtused ammutatud pideva veeru andmetest. Toetatud andmetüübid on kuupäev, ujukomaarv, pikk täisarv ning tekst.

**Tsükliline (*Cyclical*):**

Väärtus on tsükliline, kui see on järjestatud ning korduv. Tsükliliseks väärtuseks on näiteks tunnid päevas, kus numbrid 1 kuni 24 on järjestatud ning korduvad. Seda veergu peetakse sisutüübina nii diskreetseks kui ka järjestatuks, ning seda toetab tarkvara Microsoft SQL Server 2014 Analysis Services iga andmekaeve algoritmi puhul. Mitmed algoritmid käsitlevad seda sisutüüpi diskreetsena ning töötlevad seda ka vastavalt.

[9] Kontiinum - punktide või arvude pidev ja ühtlane hulk

# 2 Meetod

Päästeameti väljakutsete uurimise jaoks on vaja üles seada keskkond, milles saab läbi viia andmekaevet. Selleks on vaja valida tarkvara, mis on võimalikult dünaamiline ning mitmekülgne, võimaldades leida andmetest täpset ja väärtuslikku informatsiooni.

Väljakutsete põhjalikumaks analüüsiks on vaja lisada andmekogusse metaandmeid, kuna kogutud andmed sisaldavad omakorda veel andmeid. Andmeveergudest äriteadmuse saamiseks on loodud mudelid, et näha, kas mitmete veergude kirjetest tekib omavahelisi seoseid. Veergude arvu vähendamisel ja võimalikele seostele põhinedes on loodud täpsustatud mudelid.

Päästeametit huvitavad vähem levinud väljakutsed, mistõttu on tekitatud ainult nendega arvestav andmekogu. Viimasena on vaja seadistada keskkonnas kasutatavad algoritmid, et neid mitte üle treenida ja saada selgeid tulemusi.

## 2.1 Kasutatud tarkvara

Järgnevas loetelus on toodud välja töös kasutatud tarkvara ning nende versioonid:

- Microsoft SQL Server 2014 - 12.0.2269.0 (X64) Business Intelligence Edition - Build 10 586;
- Microsoft Visual Studio Professional 2015 Version 14.0.24720.00 Update 1;
- Microsoft SQL Server Management Studio 12.0.2269.0;
- Microsoft Analysis Services Client Tools 12.0.2000.8;
- Microsoft Data Access Components (MDAC) 10.0.10586.0.

## 2.2 Andmekaeveprojekti ülesseadmine

Autor kasutab töös *Server Data Tools Preview* paketti, mis võimaldab luua Visual Studio *2015* programmis malli nimega „*Analysis Services Multidimensional and Data Mining Project*“. See on ülesehituselt eelseadistatud mall, mis on loodud andmekaeveprojektideks. Analüüsiteenuste projektis on lähteandmete all üks lähteandmete kogum (*data source*) nimega „päästeamet.ds“, mille põhjal on loodud vaated (*data source view*) „2010-2013.dsv“, „koondandmed.dsv“ ning „vahendatud.dsv“.

Kasutatud vaadete abil on loodud erinevad andmekaevestruktuurid, mis sisaldavad vastavalt algoritmile töödeldud andmekogusid. Andmekogude klasteranalüüsimisel ehk andmekaeve algoritmidega uurimise tulemuseks on seoste leidmine, mille abil on võimalik luua äriteadmust.

## 2.3 Lisatud andmeväljad

Päästeameti edastatud andmekogumi andmed sisaldavad metaandmeid, mida saab ära kasutada andmehulga suurendamiseks ning parandamiseks. Lisatud andmed aitavad põhjalikumalt mõista väljakutseid põhjustavad põhjused.

Igale aastale on andmeanalüüsi projektis lisatud neli lisaveergu: „VK kellaaeg“, „Tund“, „Nädalapäev“ ning „Kuu“. Uued veerud on tuletatud andmeveerust „VK aeg“, millest on vastavalt eraldatud kellaaeg minutites, väljakutse toimumise tund, väljakutse nädalapäev ning kuu.

Need on lisatud analüüsi selleks, et leida, kas mingil kellaajal, nädalapäeval, tunnil ja/või kuus esineb mõnda alamliiki, väljakutse liiki, maakonda, linna/valda rohkem või vähem võrreldes teiste ajahetkedega.

Koondandmetele ehk andmehulgale, kus on korraga koos 2010–2013 aastate andmed ja vähendatud väljakutsete andmekogule, on lisatud andmeveerud „Päev aastas“, mis väljendab numbriliselt „VK aeg“ veerust saadud kuupäeva alates aasta algusest. Lisatud on ka „Nädalapäev“, „Tund“ ja „Kuu“. Koondandmetesse ega vähendatud andmekogusse ei ole lisatud andmevälja „VK kellaaeg“.

## 2.4 Vähem levinumate väljakutsete andmekogu

Vähem levinumate väljakutsete omadused ja mustrid võivad jääda märkamata suuremate ning enim levinud väljakutsete seas, mistõttu on loodud andmekogu, kust on eemaldatud kõik väljakutse liigid, mis esinevad uuritavas vahemikus kokku rohkem kui 3000 korda. Eemaldatud kirjed on esitatud tabelis 2.1.

*Tabel 2.1: Koondandmetest eemaldatud väljakutsed ja nende esinemissagedus*

| **Eemaldatud Väljakutse liik SOS** | **Väljakutsete arv** |
|---|---|
| TULETÕRJE ALARM | 18 492 |
| TULEKAHJU | 12 696 |
| TULEKAHJU HOONES | 7592 |
| TULEKAHJU VÄLJASPOOL HOONEID | 6441 |
| PT - LOODUSJÕUDUDEST PÕHJ. SÜNDMUS | 5848 |
| TULEKAHJU METSAS/MAASTIKUL | 5259 |
| PT - ABITUS SEISUNDIS LOOM/LIND | 4558 |
| TR - LIIKLUSÕNNETUS | 4369 |
| DEM - LÕHKEPEA | 3931 |
| **Kokku:** | 69 186 |

Peale tabelis 2.1 välja toodud andmete eemaldamist jäi uude andmekogusse alles 12 240 väljakutset ehk 15,03 protsenti kogu väljakutsete arvust. Selle abil loodi uus lähteandmete kogumi vaade nimega „vahendatud“, mis tuleneb eestikeelsest sõnast „vähendatud“.

## 2.5 Algoritmide poolt kasutatavad andmeveerud

Andmeanalüüsi projektis „Paasteamet2010-2013“ on loodud erinevaid andmekaevestruktuure (*mining structures*), mis sisaldavad ühte lähteandmete kogumi vaadet kasutavaid andmekaeve mudeleid. Mudelite abil saab määrata ära kasutatavad sisend- ning väljundandmeveerud koos sisutüüpide ja andmekaeve algoritmidega. Veergude omavaheline loogiline seostamine vastavalt algoritmi eesmärgile on tähtis selleks, et oleks võimalik saada tulemused, millel on Päästeameti jaoks mingisugune praktiline väärtus. Struktuurides on seatud testkomplekti (*testing set*) suuruse väärtuseks null, kuna töö eesmärgiks ei ole teha ennustusi ning seetõttu kasutatakse ära kogu andmekogu.

### 2.5.1 Andmekaevestruktuur „Koond“

Struktuur „Koond“ koosneb ainult Microsofti klasterdamise algoritmi kasutavatest andmekaevemudelitest, mida on kokku kuus ja kasutab lähteandmete kogumi vaadet „koondandmed“. Mudeliteks on „Koond - Koht ja päev“; „Koond - Tund“; „Koond - Kuu“; „Koond - Maakond“; „Koond - Päev aastas“ ja „Koond - Kõik“. Iga struktuuris olev mudel loob 16 klastrit**.** Tabelis 2.2 on toodud kasutatavate andmeväljade sisutüübid.

*Tabel 2.2: Andmeväljade sisutüübid*

| Andmeväli | Sisutüüp |
|---|---|
| Alamliik SOS | Diskreetne |
| Väljakutse liik SOS | Diskreetne |
| Sündmuse liik SOS | Diskreetne |
| Linn/Vald | Diskreetne |
| Maakond | Diskreetne |
| VK Number | Võti |
| Päev Aastas | Pidev |
| Nädalapäev | Tsükliline |
| Kuu | Tsükliline |
| Tund | Tsükliline |

**Koond - Kõik** kasutab kõiki uuritavaid andmeveerge tabelist 2.2, et omavahel grupeerida sarnaste omadustega väljakutseid.

**Koond - Koht ja päev** mudel loob seoseid asukoha ning nädalapäeva vahel. Uuritud on, kas mõni väljakutse liik esineb teatud nädalapäeval rohkem võrreldes teiste nädalapäevadega. Selleks kasutatakse veerge „Alamliik SOS“, „Maakond“, „Linn/Vald“, „Väljakutse liik SOS“ ning „Nädalapäev“.

**Koond - Tund** mudel toob välja seosed ööpäeva tundide ning väljakutse liigi vahel, kasutades selleks veerge „Alamliik SOS“, „Tund“ ja „Väljakutse liik SOS“.

**Koond - Kuu** kasutab sisend- ja väljundveergudeks „Alamliik SOS“, „Kuu“ ja „Väljakutse liik SOS“. Mudeli eesmärgiks on leida seosed kuu ning väljakutse liigi vahel, nt kas mõni väljakutse esineb vaid teatul kuul või kuudel aastas.

**Koond - Maakond** on mudel, mis kasutab sisend- ja väljundveergudeks „Alamliik SOS“, „Maakond“, „Linn/Vald“ ning „Väljakutse liik SOS“. Eesmärgiks on leida klastrites seosed väljakutse liigi ning asukoha vahel. Uuritud on, kas mõnes maakonnas eristub mõni väljakutse liik teistest või kas mõni maakond eraldub oma väljakutsete poolest teistest oluliselt.

**Koond - Päev aastas** on andmekaevemudel, mis toob välja seosed väljakutse liigi ning aastas olevate päevade vahel. Selleks on kasutatud veerge „Alamliik SOS“, „Väljakutse liik SOS“ ja „Päev aastas“.

### 2.5.2 Andmekaevestruktuur „Koond - Naive Bayes“

Struktuur „Koond - Naive Bayes“ sisaldab üht mudelit, milles on kasutatud koondandmete grupeerimise jaoks ainult naiivset Bayesi algoritmi. Struktuur kasutab lähteandmeteks lähteandmete kogumit „koondandmed“.

Mudel „kõik Naive Bayes“ kasutab sisend- ja väljundveergudeks tabelis 2.2 olevaid veerge, sealjuures numbriliste väljade („Päev aastas“, „Kuu“, „Tund“ ja „Nädalapäev“) sisutüübid on diskretiseeritud väärtused. Naiivne Bayesi algoritm loob suhtevõrgustiku andmeveergude vahel, mille tulemusena on näha ühe veeru kõikide erinevate kirjete jagunemist teiste veergude kirjete suhtes, seda aga välja arvatud numbriliste veergude puhul, kus teineteisele lähedal paiknevad numbrid võivad olla automaatselt grupeeritud.

### 2.5.3 Andmekaevestruktuur „Vahendatud“

Andmekaeve struktuuris „Vahendatud“ on kasutusel Microsofti klasterdamise algoritm**.** Loodud on kuus mudelit, mis on eesliitega “Vahendatud“ ning identsed mudelitega struktuuris

„Koond“. Lähteandmete kogumiks on vaade „Vahendatud“. Lähteandmeteks kasutatakse lähteandmete kogumit „vahendatud“, mis tuleneb sõnast vähendatud.

## 2.6 Algoritmide seadistus

Kõigil üheksal algoritmil paketis Microsoft SQL Server 2014 Analysis Services on unikaalne vaikeseadistus, mida muutes on võimalik tulemusi täpsustada vastavalt vajadusele. Autor on muutnud seadistust vastavalt andmekaeve sisend- ning väljundparameetrite vaikeseadistusega töötlemisel saadud informatsioonile. Alljärgnevalt on toodud välja seadistuse muudatused, mis kehtivad kõikidele vastavat algoritmi kasutavatele struktuuridele.

**Microsofti klasterdamise algoritmi** puhul on muudetud järgnevaid vaikeväärtusi: klasterdamise meetod, klastrite arv ning maksimaalne olekute arv.

Klasterdamise meetod on muudetud skaleeritava ootuste maksimeerimise pealt mitteskaleeritava peale. Päästeameti poolt märgitud väljakutsete arv ei ole piisavalt suur, seetõttu kasutatakse meetodit, mis loeb korraga sisse kõik väljakutsed. Ootuste maksimeerimise tulemused võivad vähesel määral varieeruda, kuna algselt valitud andmepunktid on alati suvalised.

Maksimaalse olekute arvu vaikeväärtus on muudetud arvult 100 arvule 350, kuna ühelgi aastal ei leidu andmevälja, mille unikaalsete kirjete arv ületaks 350. Samuti pole arvu suurendamisest saadav tõusnud ressursivajadus väikeste andmemahtude juures probleem. Arvu liiga madalaks seades ignoreerib algoritm vähem populaarsemaid olekuid, mistõttu võivad väärtuslikud andmed osaliselt kaduma minna [10].

Muudetud on tulemuseks saadavate klastrite arvu 16 või 13 peale vastavalt sellele, millist lähteandmete kogumi vaadet on kasutatud. Arvud näitavad miinimumkriteeriumi klastrite arvuks, mis on saadud võttes täisosa logaritm alusel kaks andmekogu suurusest. Koondandmete puhul on klastrite arv 16 ning vähendatud andmekogu korral 13.

**Naiivse Bayesi algoritmi** puhul on sarnaselt klasterdamise algoritmile muudetud maksimaalse olekute arvu väikeväärtus arvult 100 arvule 350.

# 3 Tulemused

Peatükk koosneb kahe erineva andmekogu analüüsil tekkinud tulemuste kirjeldamisest. Koondandmete uurimisel on kasutatud kahte erinevat andmekaeve algoritmi, vähem levinud väljakutsete uurimisel vaid ühte. Mõlemas andmekogus on kasutatud Microsofti klasterdamise algoritmi sarnaste väljakutsete grupeerimiseks. Ainult koondandmete töötlemiseks on kasutatud Microsofti naiivset Bayesi algoritmi. Peatükis leiduvatel joonistel on andmeveergude kirjed, mille jaotused klastris ületavad 0,1. Ülejäänud kirjed on grupeeritud kokku kategooria „Muud“ alla.

## 3.1 Koondandmed 2010–2013

Analüüsides koondandmeid andmekaeve algoritmidega, kasutades erinevaid sisendparameetreid, on autor leidnud mustreid, trende ja anomaaliaid. Järgnevalt on läbi erinevate andmekaevemudelite välja toodud tulemused kõikide käsitletavate aastate kohta. Andmekaevemudelist tuuakse välja vaid klastrid, mis annavad edasi autori meelest väärtuslikku teavet hädaabi väljakutsete kohta.

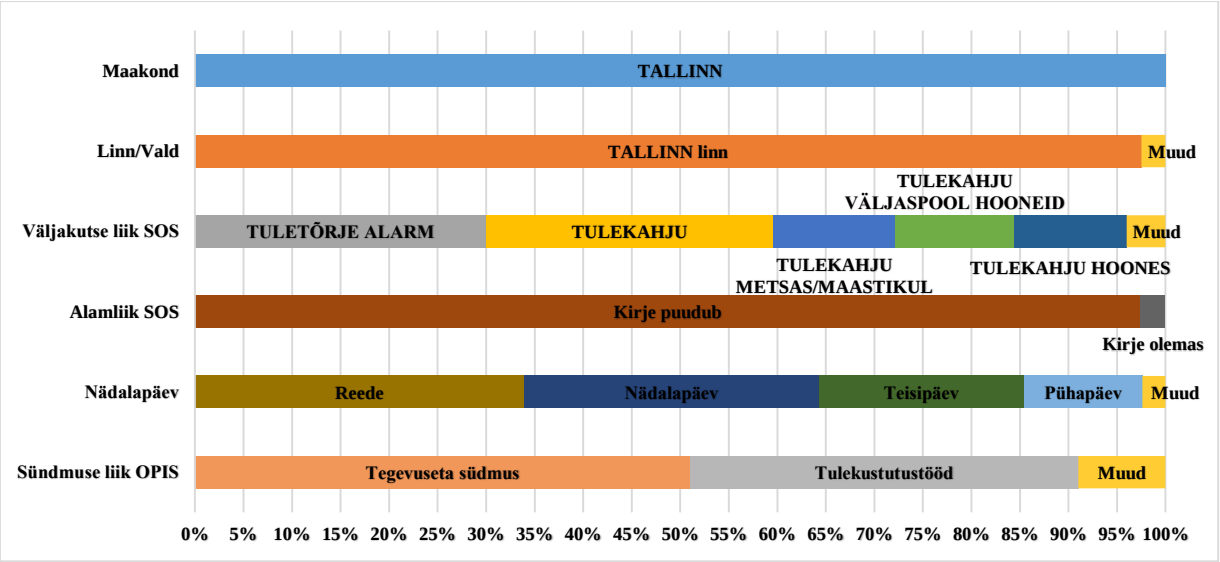
### 3.1.1 Leiud andmekaevestruktuurist „Koond“

Alampeatükk põhineb leidudel andmekaevestruktuurist „Koond“. Klasterdamise algoritm tekitab erinevatele mudelite antud sisend- ja väljundveergude korral klastrid ehk kategooriad. Saadud kategooriad on selles alampeatükis lähemalt kirjeldatud.

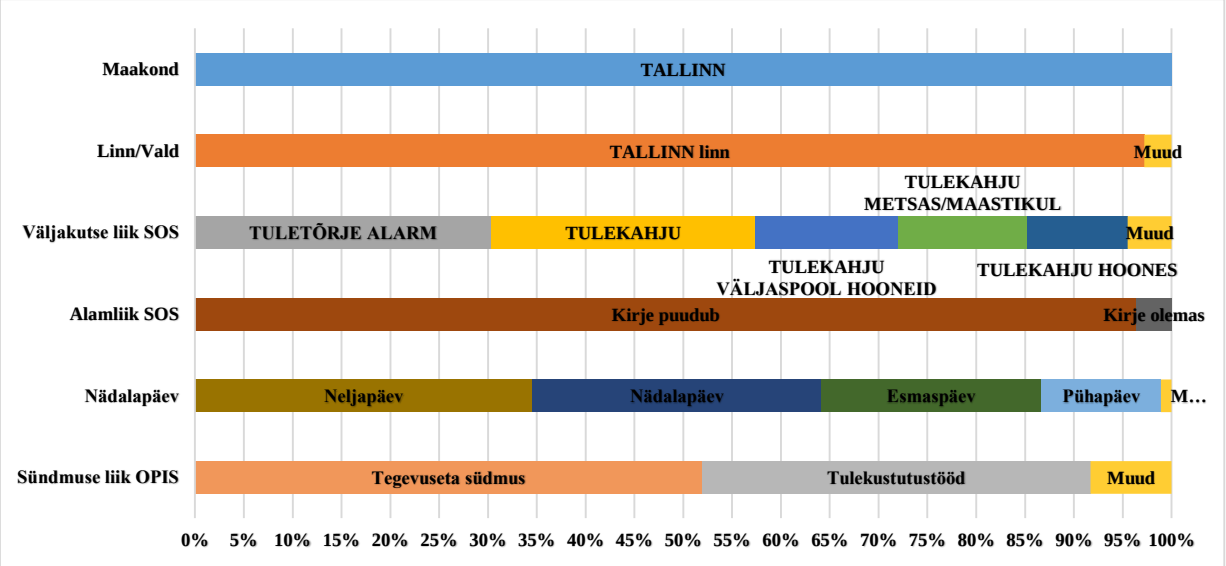
**Tähelepanekud mudelist „Koond - Kõik“**

Mudelis „Koond - Kõik“, kus on kasutatud kõiki töös uuritavaid andmeveerge, eristub mitu klastrit, kus üks andmeveeru väärtus domineerib teiste üle. Järgnevalt on toodud välja klastrid, kus ühe andmeveeru kirje eristus teistest. Kobarate kohta on välja toodud andmeveergude jaotused, mis aitavad kaasa kirjeldamisele, sisaldades üle väheste andmevälja kirjete jaotunud informatsiooni.

Andmeveeru maakond kohta eraldub kolm kobarat, kus kaks väljendavad maakond „TALLINN“ väljakutseid ning üks Ida-Virumaa maakonda. Joonisel 3.1 on ülevaade esimesest, suurusega 4929 väljakutset, ja joonisel 3.2 teisest, Tallinna klastrist suurusega 4724, mis on omavahel kõikide välja toodud veergude suhtes sarnased välja arvatud veerus „Nädalapäev“. Mõlemas klastris on täiesti erinevad nädalapäevad, välja arvatud pühapäev, mis on mõlemas klastris. Mõlema klastrite jooniste 3.1 ja 3.2 nädalapäevade kokku panemisel saab kõik päevad nädalas kaetud, mistõttu võib öelda, et kobarad täiendavad üksteist.

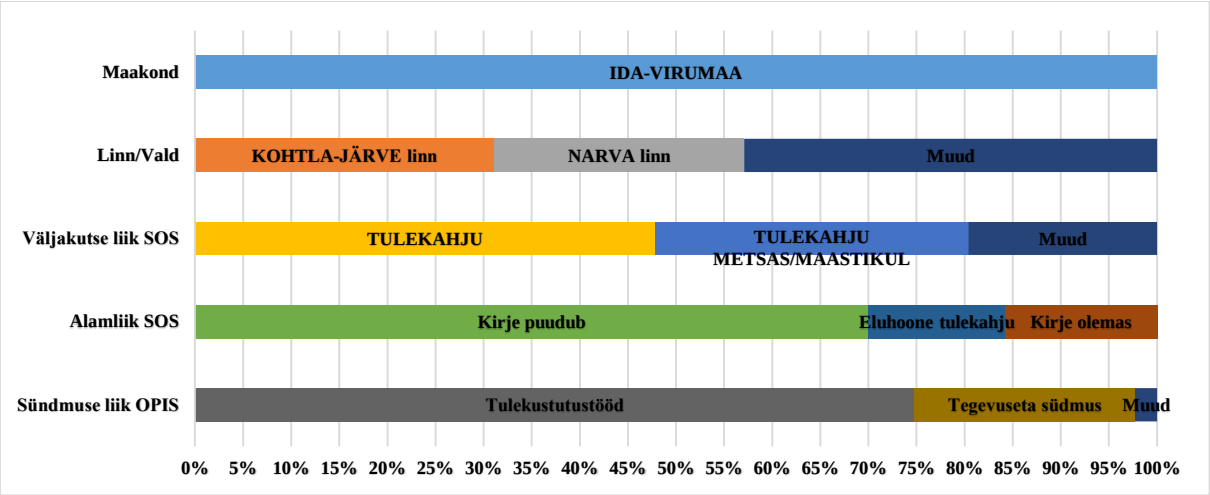



*Joonis 3.1: Esimene eristunud Tallinn maakond klastri andmeveergude jagunemine*




*Joonis 3.2: Teine eristunud Tallinn maakond klastri andmeveergude jagunemine*

Ida-Virumaa klaster joonisel 3.3 erineb Tallinna klastritest, kuna väljakutsed on jagunenud üle maakonna laiali ning nende liigid jaotunud põhiliselt vaid kahe kirje vahel: “TULEKAHJU“ ning „TULEKAHJU METSAS/MAASTIKUL“. Kobara suuruseks on 5314 väljakutset. Suurimaks sündmuse liigiks on „Tulekustutustööd“, mis esineb koos klastrile iseloomulike väljakutse liikidega.

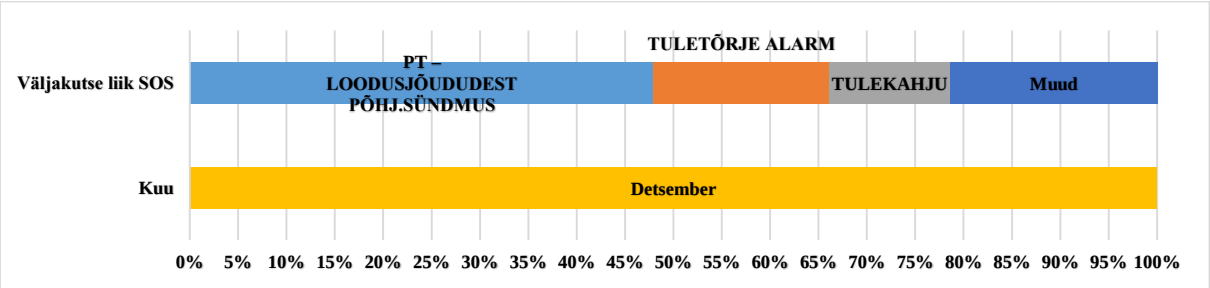



*Joonis 3.3: Eristunud Ida-Virumaa maakond klastri andmeveergude jagunemine*

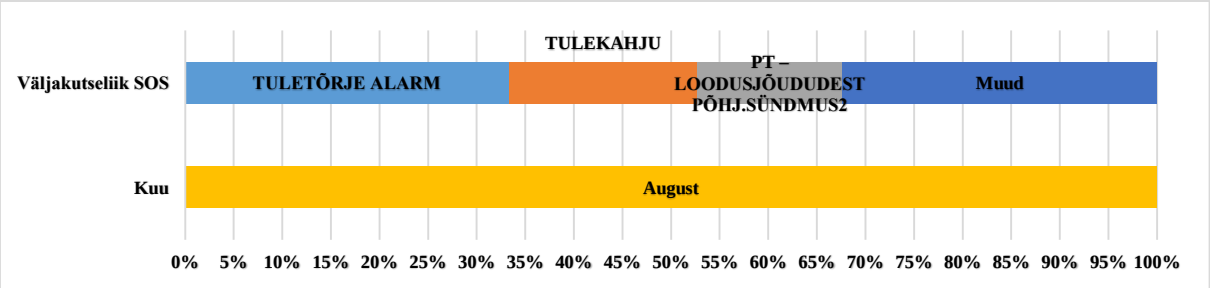
„Päev aastas“ veeru ning „Kuu“ kohta eristuvad kattuvate ajavahemikega kaks klastrit, kus üks väljendab detsembri väljakutseid joonisel 3.4 ning teine augusti väljakutseid joonisel 3.5. Tabelis 3.1 on esitatud klastrite ülevaatlik statistika andmeveerule „Päev aastas“.

Augustikuu klaster kestab vahemikus 29. juuli kuni 31. august ning detsembri kobar 30. novembrist kuni 31. detsembrini. Kuigi mõlema klastri väljakutsete liigid kattuvad, siis loodusjõududest põhjustatud väljakutsed moodustavad detsembri kobarast peaaegu pool kogu klastri kirjetest, sealjuures augusti väljakutsetest vaid üle kümnendiku.

Saadud tulemuste põhjal on loodud mudelid „Koond - Päev aastas“ ning „Koond - Kuu“ struktuuri „Koond“, et luua klastrid, mis annaksid rohkem informatsiooni veergude seoste kohta. Tulemused ja leiud on toodud samuti välja käesolevas peatükis.




*Joonis 3.4: Detsembrikuu klastri andmeveergude jagunemine*

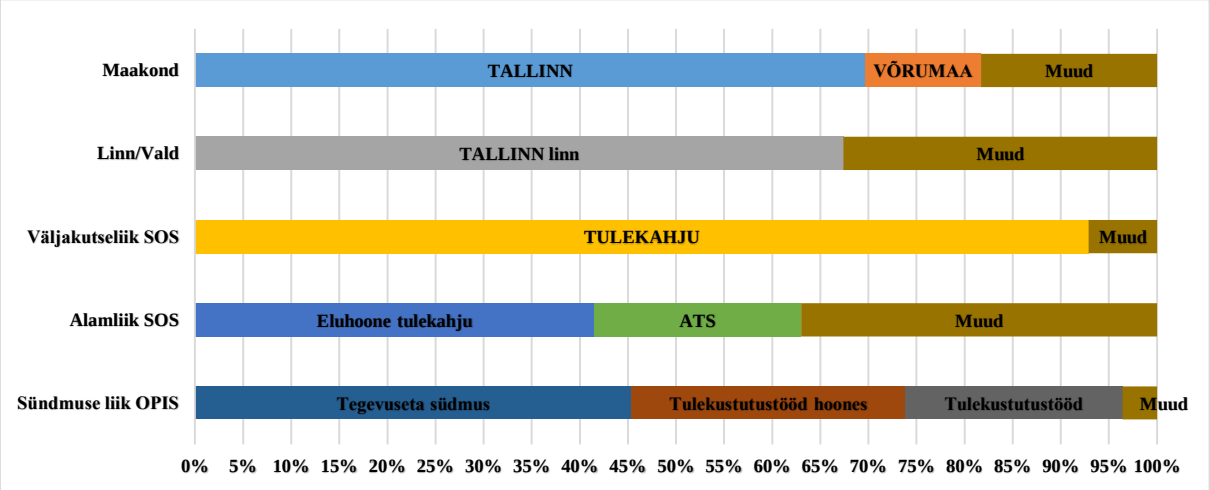



*Joonis 3.5: Augustikuu klastri andmeveergude jagunemine*

*Tabel 3.1: Detsembri ja augusti klastrite statistika*

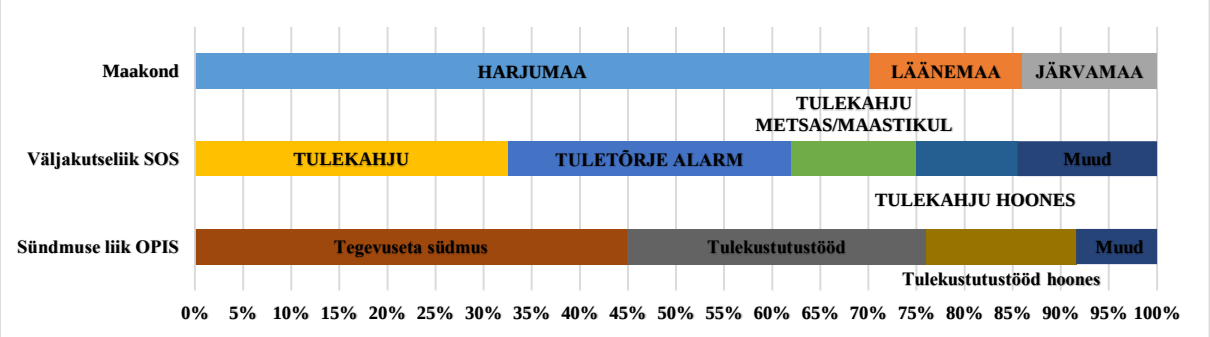
| Detsember | | August | |
|---|---|---|---|
| **Päev aastas** | | | |
| Keskmine | 352,7 | Keskmine | 226,43 |
| Varaseim | 335 | Varaseim | 210 |
| Hiliseim | 366 | Hiliseim | 243 |
| Standardhälve | 8,7 | Standardhälve | 8,16 |
| **Klastri suurus** | | | |
| 5571 | | 4979 | |

Viimasena eristub „Väljakutse liik SOS“ klaster, kus on põhiliseks kirjeks „TULEKAHJU“. Klastri kirjeldusest joonisel 3.6, selgub, et üle poolte 4095 väljakutsest on toimunud Tallinnas. Võrreldes teiste klastritega on paljudele väljakutsetele märgitud ka alamliigid. Nendeks on „Eluhoone tulekahju“ ning „ATS“, mis esinevad koos väljakutse liigiga „TULEKAHJU“. Alamliigi kirje „Eluhoone tulekahju“ viitab sündmuse liigi kirjele „Tulekustutustööd hoones“.




*Joonis 3.6: Eristunud väljakutse liigi Tulekahju andmeveergude jaotused*

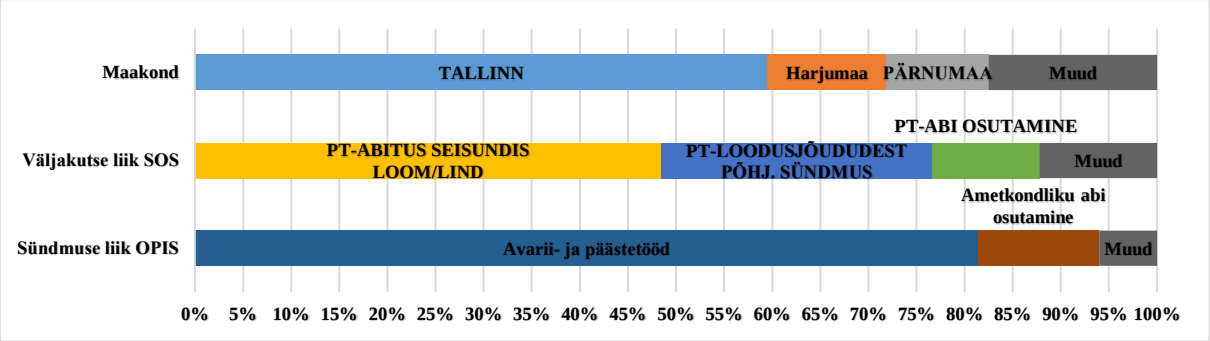
Lisaks eristuvatele klastritele leidub mudelis teisigi kobaraid, kus ükski andmeveerg ei domineeri täielikult, kuid mis sisaldavad endas sellegipoolest väärtuslikku informatsiooni. Tekkis klaster joonisel 3.7, mille väljakutsetest 70% toimuvad Harju maakonnas. Ülejäänud osa moodustavad Läänemaa ning Järvamaa. Põhilisteks väljakutse liikideks on suuruse järjekorras „TULEKAHJU“, „TULETÕRJE ALARM“, „TULEKAHJU METSAS/MAASTIKUL“ ja „TULEKAHJU HOONES“.




*Joonis 3.7: Harjumaa klastri andmeveergude jaotused*

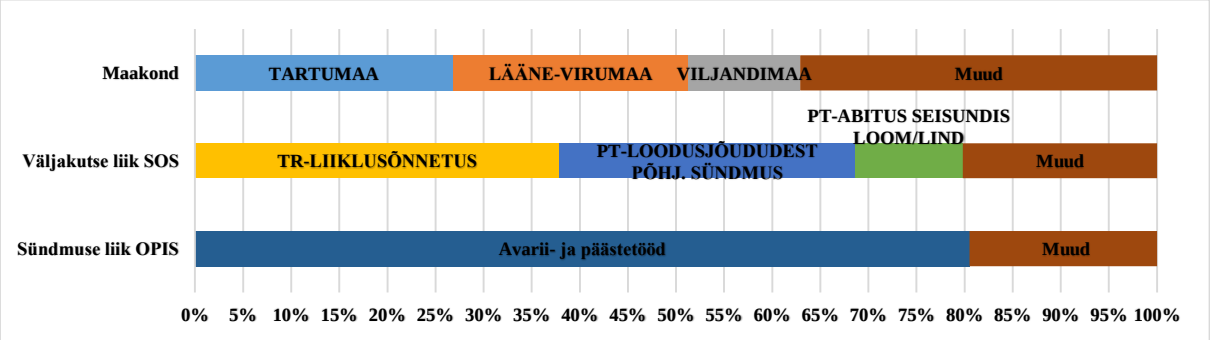
Mudelis on kolm kobarat, milles põhiliseks väljakutsete sündmuse liigiks on üle 75 protsendi „Avarii- ja päästetööd“. Esimeses klastris joonisel 3.8 on levinuimaks väljakutse liigiks „PT - ABITUS SEISUNDIS LOOM/LIND“. Sellele järgnevad liigid „PT - LOODUSJÕUDUDEST

PÕHJ. SÜNDMUS“ ning „PT - ABI OSUTAMINE“. Sündmused on toimunud põhiliselt Tallinnas, kuid ka Harjumaal ning Pärnumaal.



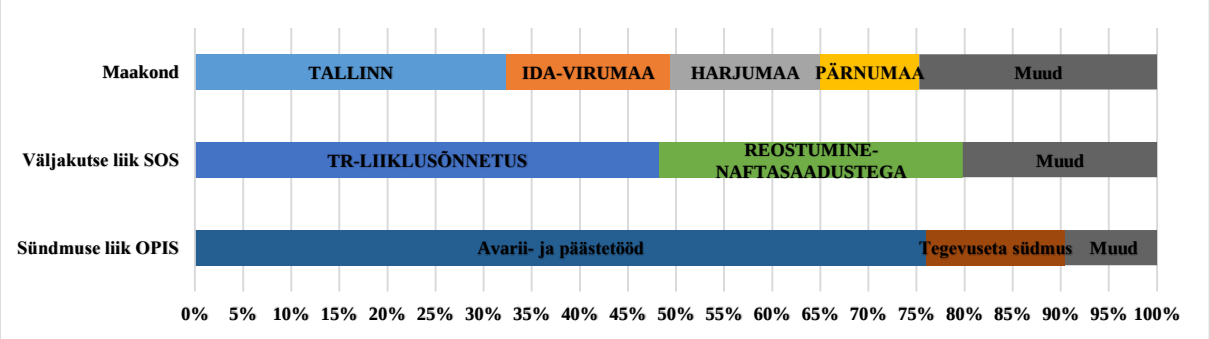

*Joonis 3.8: Esimene „Avarii- ja päästetööd“ sündmusi väljendava klastri andmeveergude kirjete jaotused*

Teisel klastril joonisel 3.9 on suurematekst väljakutse liikideks „TR - LIIKLUSÕNNETUS“ ja sarnaselt eelmisele kirjeldatud klastrile „PT - LOODUSJÕUDUDEST PÕHJ. SÜNDMUS“. Erinevalt eelnevast on kokku pooled väljakutsetest toimunud Tartumaal ja Lääne-Virumaal, väiksemal määral ka Viljandimaal, Raplamaal ning Põlvamaal.




*Joonis 3.9: Teine „Avarii- ja päästetööd“ sündmusi väljendava klastri andmeveergude kirjete jaotused*

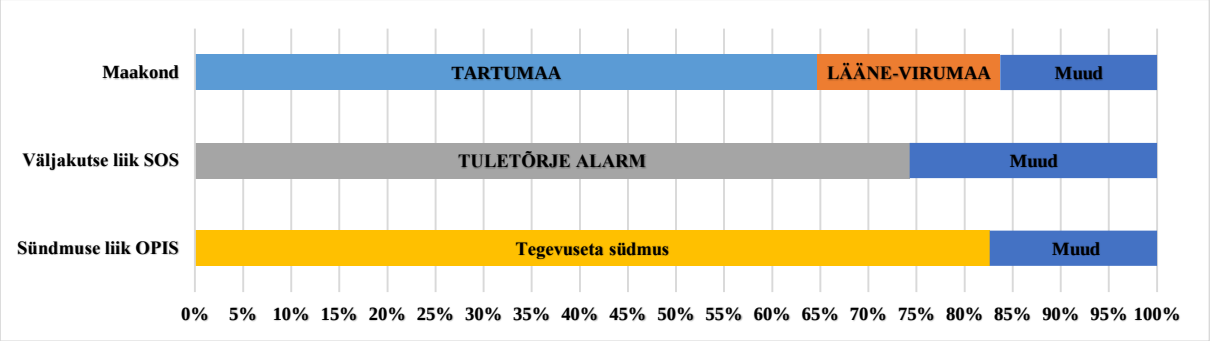
Kolmandas kobaras on peaaegu pooled väljakutsed põhjustanud kirje „TR-LIIKLUSÕNNETUS“, kuid erinevalt kõikidest teistest pääste- ning avariitöödega seotud klastritest on kolmandik väljakutsetest seotud liigiga „REOSTUMINE - NAFTASAADUSTEGA“. Joonisel 3.10 toodud väljakutsed on toimunud Tallinnas, Ida-Virumaal, Harjumaal ning Pärnus. Päästetöödega seotud kobarad on jaotunud ühtlaselt üle terve aasta.




*Joonis 3.10: Kolmas „Avarii- ja päästetööd“ sündmusi väljendava klastri andmeveergude kirjete jaotused*

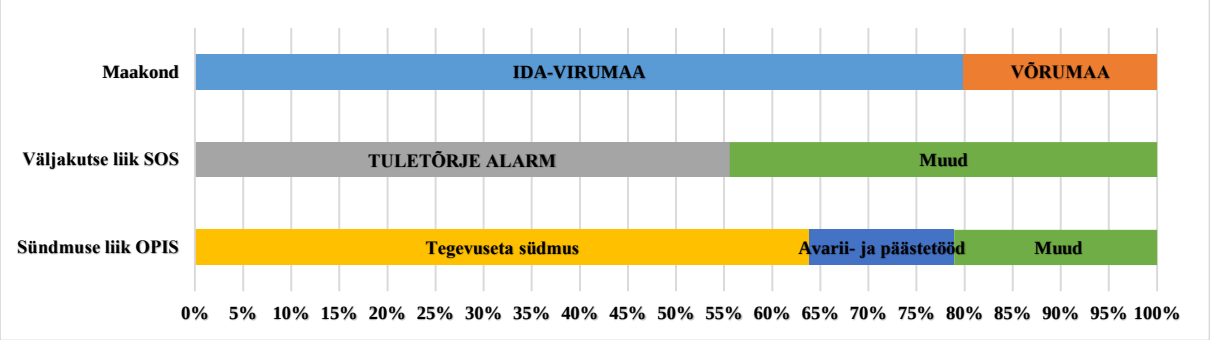
Tartumaa ja Lääne-Virumaa kohta on klaster joonisel 3.11, kus Tartumaaga on seotud 67% väljakutsetest ja Lääne-Virumaaga 19%. Üle 80 protsendil kirjetel on märgitud sündmuse liigiks „Tegevuseta südmus“ (lähteandmete kirjaviga) ning kolmveerandi kõikidest

väljakutsetest on põhjustanud tuletõrjealarm, sealjuures on pooled väljakutsetest toimunud Tartumaal.



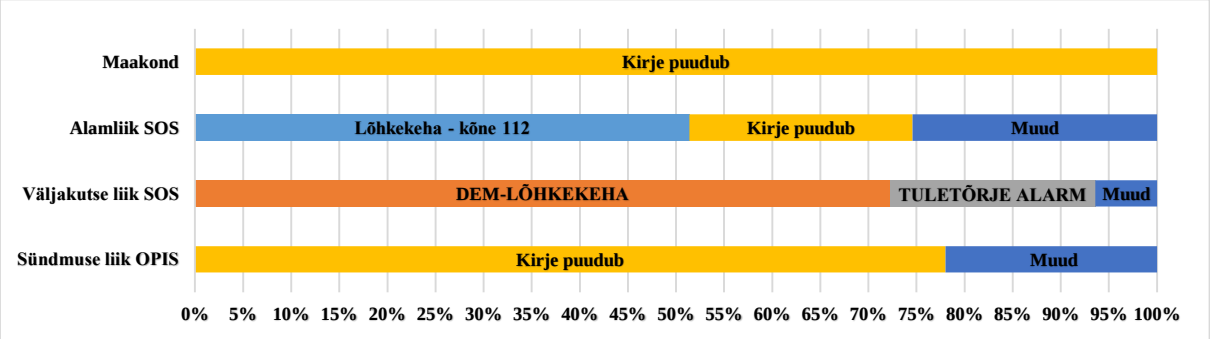

*Joonis 3.11: Esimene tegevuseta sündmusi väljendava klastri andmeveergude kirjete jaotused*

Lisaks joonisel 3.3 välja toodud eristuva Ida-Virumaa klastrile on mudelis veel üks Ida-Virumaa väljakutseid kirjeldav kobar joonisel 3.12, kus eelmainitud maakonna osakaal on 80% ning ülejäänud väljakutsed on toimunud Võrumaal. Üle poolte väljakutsetest on põhjustanud „TULETÕRJE ALARM“ ja suurima osakaaluga on sündmuse liik „Tegevuseta südmus“. Veerand väljakutsetest on toimunud Narvas ja vähem Kohtla-Järvel.




*Joonis 3.12: Ida-Virumaa ja Võrumaa klastri andmeveergude jaotused*

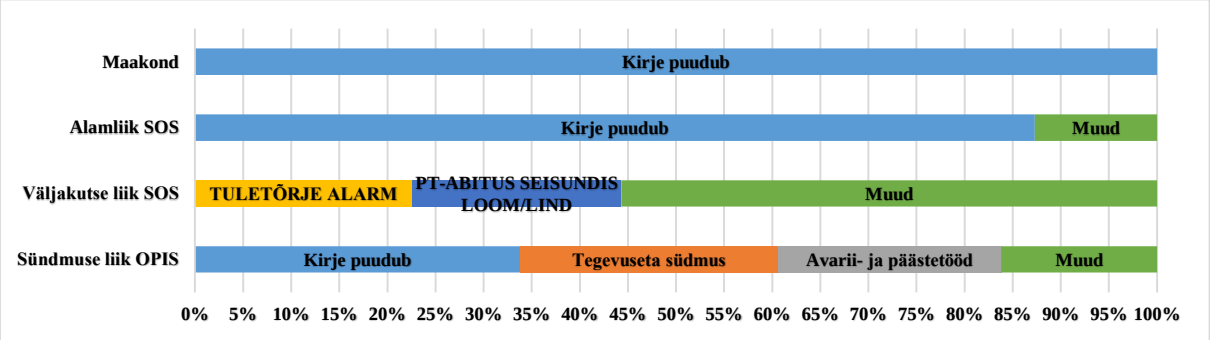
Kokku on kaks klastrit, kus kõikidel väljakutsetel puuduvad asukohad. Esimesel klastril joonisel 3.13 on puudu ka 78% väljakutsetel sündmuse liik OPIS, kuid levinumaks väljakutse liigiks on „DEM - LÕHKEKEHA“, mis on põhjustanud ligikaudu kolmveerand väljakutsetest. Pooled väljakutsetest on alamliigiga „Lõhkekeha - kõne 112“. Ülejäänud väljakutsed on põhjustanud tulekahjud. Tekkinud klastrist saadud infot väljakutse liigi „DEM-LÕHKEKEHA“ kohta on uuritud järgnevas alampeatükis.




*Joonis 3.13: Esimene märkimata asukohaga klastri andmeveergude jaotused*

Teine klaster on jagunenud joonisel 3.14 mitmete erinevate väljakutsete vahel, kuid põhilisteks on „TULETÕRJE ALARM“ ning „PT - ABITUS SEISUNDIS LOOM/LIND“. Erinevalt

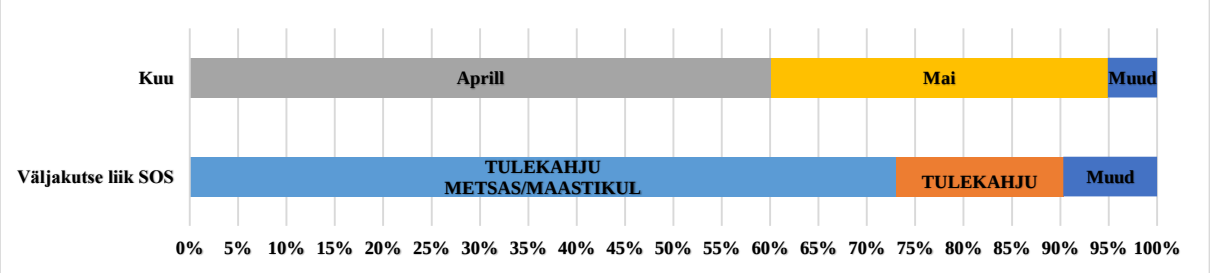
eelmisest klastrist on väljakutsed jagunenud mitmete sündmuste liikide vahel ja 87 protsendil puudub alamliik.




*Joonis 3.14: Teine märkimata asukohaga klastri andmeveergude jaotused*

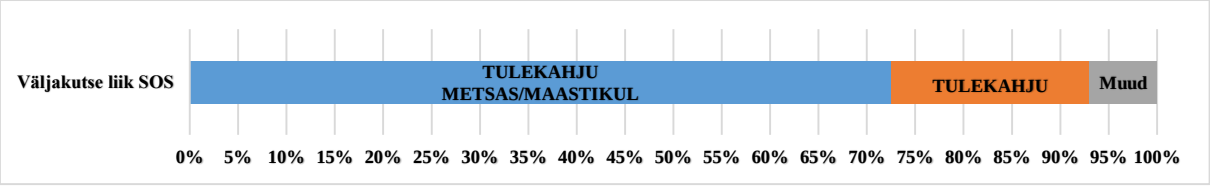
## Metsa- ning maastikutulekahjud kestavad aprilli algusest kuni mai keskpaigani

Mudelist „Koond - Kuu“ eristub klaster suurusega 4384 rida, kus põhiliseks väljakutse liigiks on kirje „TULEKAHJU METSAS/MAASTIKUL“. Joonisel 3.15 on toodud välja „Väljakutse liik SOS“ ja „Kuu“ veergude jaotused, kus suurimate jaotustega on aprill ning mai, mille väljakutse liigiks on tulekahju metsas või maastikul ning väiksemal määral tulekahju.




*Joonis 3.15: Metsatulekahjud aprillis ja mais klastri andmeveergude jaotused*

Kaevates struktuuriga „Koond - Päev aastas“ on leitud sarnaste omadustega klaster, mille väljakutse liigi jaotus on välja toodud joonisel 3.16 ning „Päev aastas“ veeru kirjete kirjeldavstatistika tabelis 3.2. Kobarasse on grupeeritud metsa- ning maastikupõlengu kirjed, millest on leitud klastri maksimaalne vahemik 5. aprill kuni 16. mai ning mediaan 27. aprill. Klastritest pärinev info viitab aprilli algusest kuni maini keskpaigani kestvast suurenenud metsa- ning maastikutulekahjude arvust Eestis.



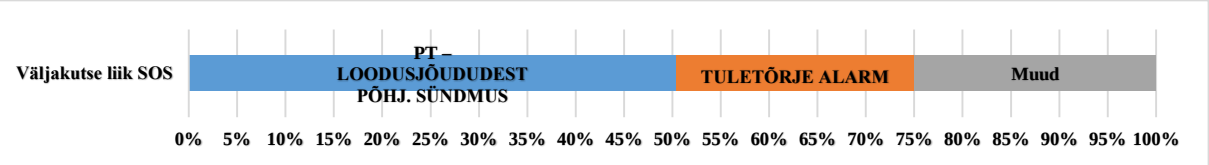

*Joonis 3.16: Metsatulekahjude klastri väljakutse liigi jaotus*

*Tabel 3.2: Metsatulekahjude klastri ülevaatlik statistika*

| Klastri suurus | | | | |
|---|---|---|---|---|
| 5277 | | | | |
| **Keskmine** | **Mediaan** | **Varaseim** | **Hiliseim** | **Standardhälve** |
| 115,29 | 117 | 95 | 136 | 10,2 |

## Loodusjõududest põhjustatud väljakutsed toimuvad novembri lõpust aasta lõpuni

Mudelis „Koond - Päev aastas“ on eristunud klaster, milles on kokku grupeeritud metsa- ning maastikupõlengu kirjed joonisel 3.17. Kobara maksimaalne kuupäevade vahemik 25. november kuni 31. detsember tabelist 3.3. Perioodi mediaaniks on 16. detsember.

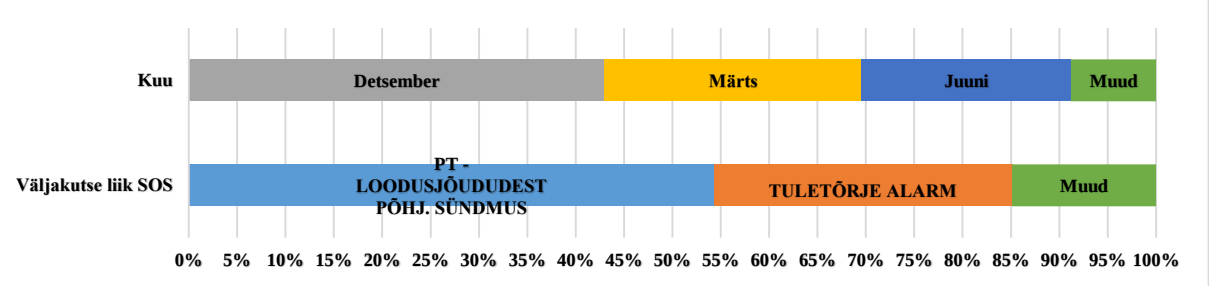



*Joonis 3.17: Loodusjõududest põhjustatud väljakutsete klastri väljakutse liigi jaotus*

*Tabel 3.3: Loodusjõududest põhjustatud klastri ülevaatlik statistika*

| Klastri suurus | | | | |
|---|---|---|---|---|
| 5584 | | | | |
| **Keskmine** | **Mediaan** | **Varaseim** | **Hiliseim** | **Standardhälve** |
| 351,03 | 350 | 329 | 366 | 10,98 |

Klastrit toetab mudelis „Koond - Kuu“ kobar joonisel 3.18, kus on samuti suurima jaotusega väljakutsed, mis on seotud loodusjõududega detsembris. Kuigi klaster on jaotunud ka märtsi ning juuni vahel, on suurima kirjete arvuga kuu selles kobaras detsember. Kobara täielik suurus on 6552 väljakutset. Klastrid viitavad loodusjõududest põhjustatud suurenenud väljakutsete hulgale novembri lõpust, kuid eelkõige detsembrist kuni aasta lõpuni.




*Joonis 3.18: loodusjõududest põhjustatud väljakutsete klastri andmeveergude jaotus*

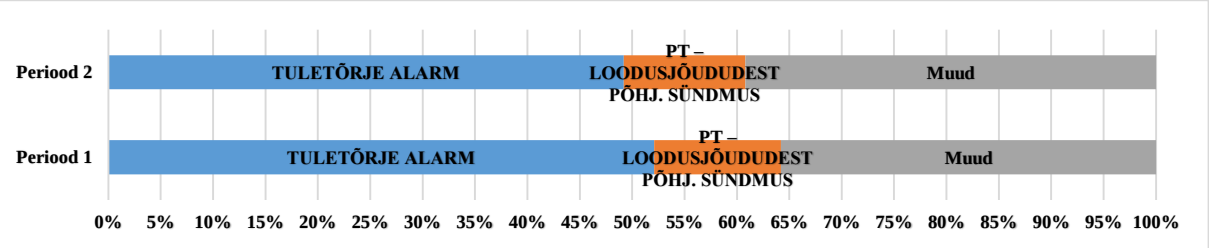
## Tuletõrjealarmidest põhjustatud väljakutsete hooajad

Mudel „Koond - Päev aastas“ leidis aasta lõikes mitu tuletõrjealarmi perioodi, milles eralduvad täpsete ajavahemikega kaks klastrit. Mõlema perioodi statistiline ülevaade on tabelis 3.4.

Esimene periood on vahemikus 1. jaanuar kuni 3. märts, mediaan on 28. jaanuar. Teine periood, kus eristub väljakutse liik „TULETÕRJE ALARM“ on vahemikus 7. september kuni 18. detsember, mediaan 29. oktoober. Perioodide väljakutsed on joonisel 3.19, kus on näha sarnast jaotumist väljakutse liikide vahel.

*Tabel 3.4: Kahe perioodi klastrite ülevaatlik statistika*

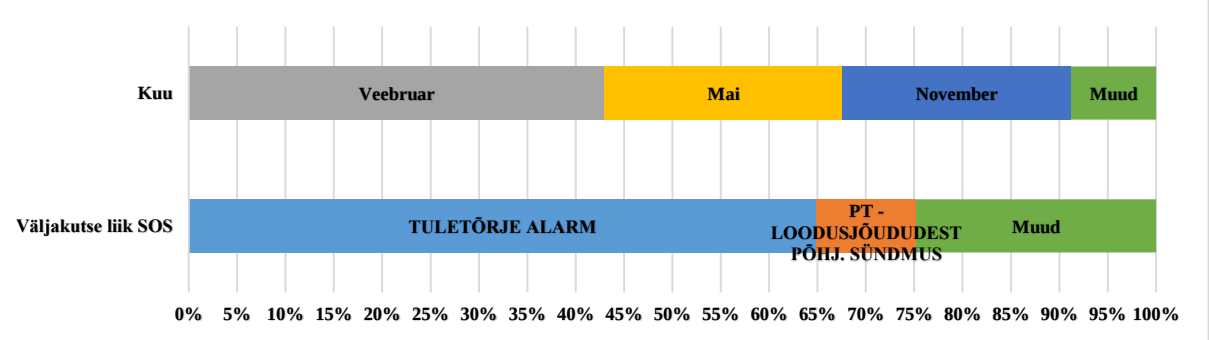
| Periood 1 | | Periood 2 | |
|---|---|---|---|
| **Päev aastas** | | | |
| Keskmine | 27,68 | Keskmine | 301 |
| Mediaan | 28 | Mediaan | 302 |
| Varaseim | 1 | Varaseim | 250 |
| Hiliseim | 62 | Hiliseim | 352 |
| Standardhälve | 17,2 | Standardhälve | 25,31 |
| **Klastri suurus** | | | |
| 5331 | | 6907 | |



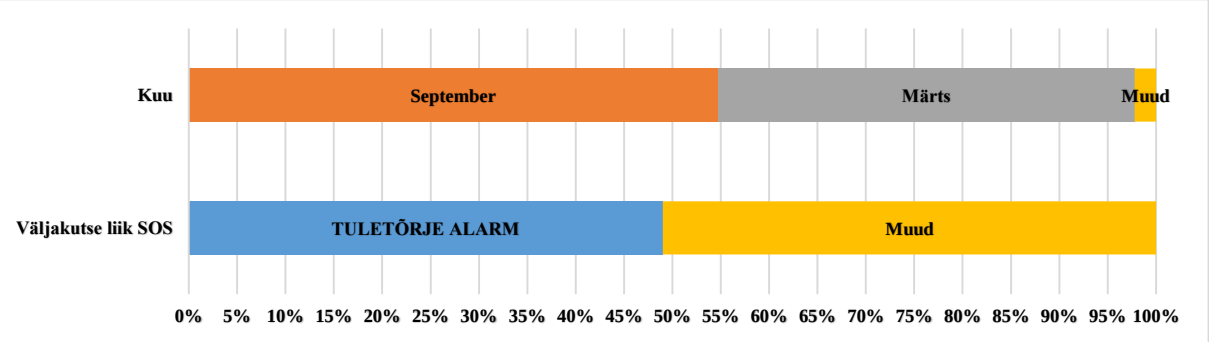

*Joonis 3.19: Väljakutse liikide jagunemine tuletõrjealarmi klastrites kahes perioodis*

Esimest vahemikku toetab mudelist „Koond - Kuu“ tuletõrjealarmiga seotud klaster suurusega 7290 joonisel 3.20, kus suurimat jaotust omab väljakutsetel veebruari, mis kattub tabelis 3.4 leitud ajaperioodiga ning märtsi toetab kuude jaotus joonisel 3.21.

Võrreldes esimese perioodiga, mis kestab 62 päeva, on teine kestvusega 102 päeva pikem. Mudelis „Koond - Kuu“ on kobar joonisel 3.21 suurusega 6196 väljakutset mille andmeveeru „Kuu“ väärtused koos joonisel 3.20 oleva klastri „Kuu“ kirjetega, toetavad „TULETÕRJE ALARM“ kirjega seotud väljakutseid septembris ning novembris.




*Joonis 3.20: Tuletõrjealarmi klastri jagunemine veebruari, mai ja novembri vahel*



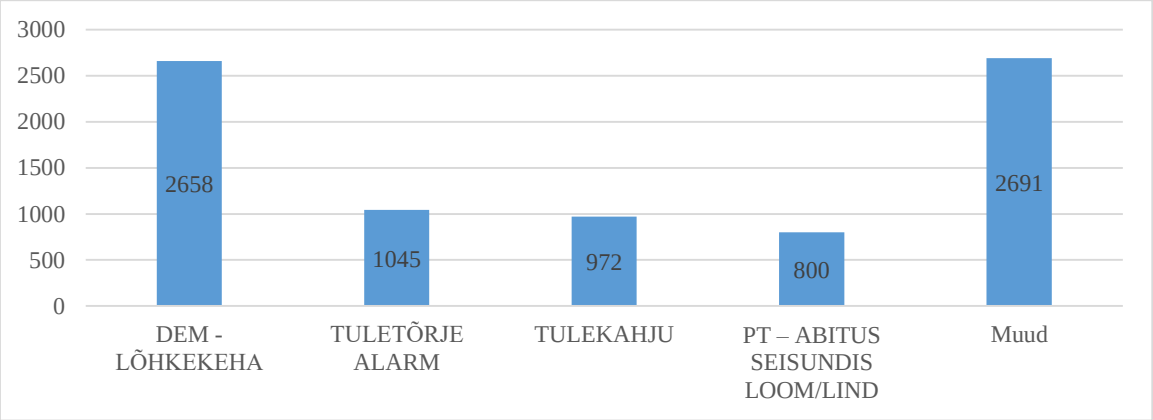

*Joonis 3.21: Tuletõrjealarmi klastri jagunemine septembri ja märtsi vahel*

### 3.1.2 Leiud andmekaevestruktuurist „Koond - Naive Bayes“

Andmekaeve algoritm Microsoft naiivne Bayes aitab leida seoseid andmetes, luues seoseid andmevälja kõikide veergude vahel võrreldes teiste mudelis olevate veergudega. Seetõttu on seda algoritmi kasutades võimalik näha kõiki võimalikke loodud seoseid, mis võisid muidu loogiliselt mõeldes leidmata jääda. Kõik alampeatükis välja toodud leiud pärinevad struktuurist „Koond - Naive Bayes“ mudelist „kõik Naive Bayes“.

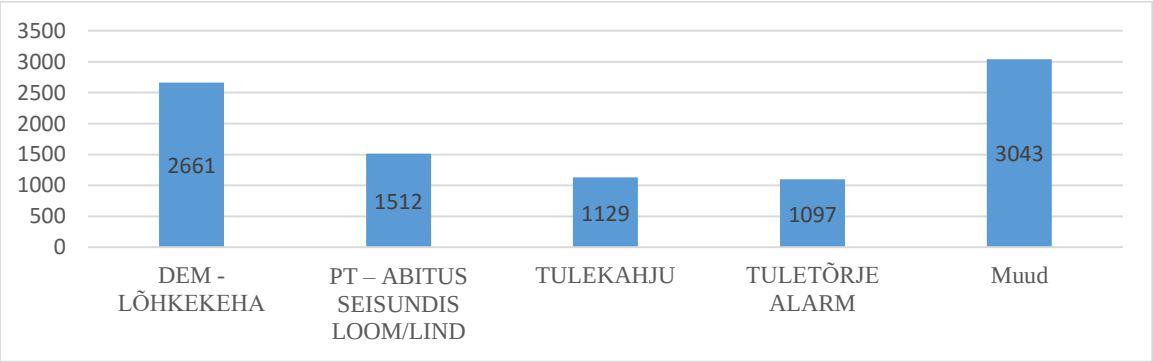
**Lõhkekeha demineerimisel puudub asukoht**

Mudelist on näha, et aastatel 2010–2012, kus väljakutsele on märkimata maakond, olid ligikaudu kolmandik väljakutsetest liigiga „DEM - LÕHKEKEHA“. Teised väljakutse liigid nii suurt osa väljakutsetest, millest puudub maakond ei moodusta. Joonisel 3.22 on esitatud väljakutsed, millel puudub veerg „Maakond“.




*Joonis 3.22: Suurimate esinemissagedustega väljakutse liigid, kus puudub veerg „Maakond“*

Samuti moodustab liik suurima osakaalu kirjetest, millel puudub andmeveerg linn või vald. Puuduvate „Linn/Vald“ veeru kirjete jaotused on toodud välja joonisel 3.23**.** Kokku on nelja aasta jooksul toimunud 3931 väljakutset, mille väljakutse liigiks on märgitud „DEM - LÕHKEKEHA“. 67,62 protsendil puudub maakond ning 67,69 protsendil puudub „Linn/Vald“ kirje.

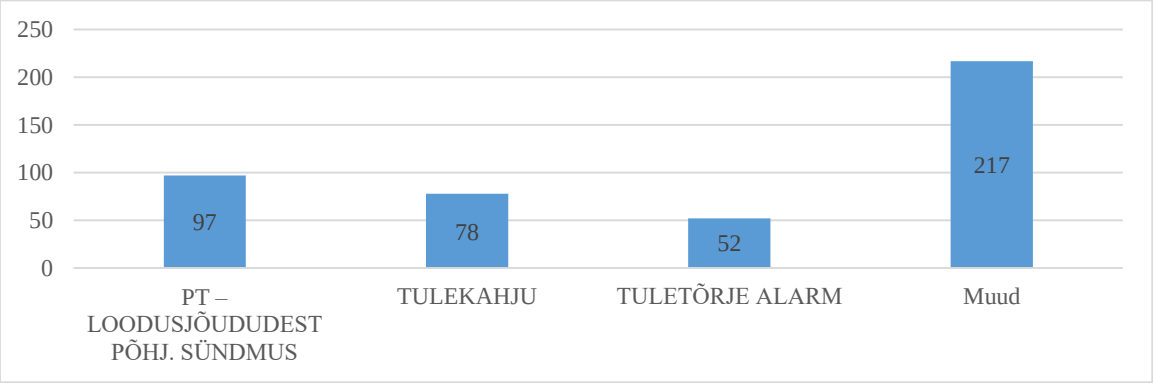



*Joonis 3.23: Suurimate esinemissagedustega väljakutse liigid, kus puudub linn või vald*

Eesti Päästeameti esindaja sõnul täheldati üles ainult läbi häirekeskuse tehtud demineerimisega seotud väljakutsed, mis moodustavad vaid osa kõikidest demineerimissündmustest üle Eesti. Demineerijatel on oma andmebaas nimega DEMIS, kuhu märgitakse väljakutsed, mistõttu on osaliselt OPIS süsteemi märkimata väljakutse asukoht.

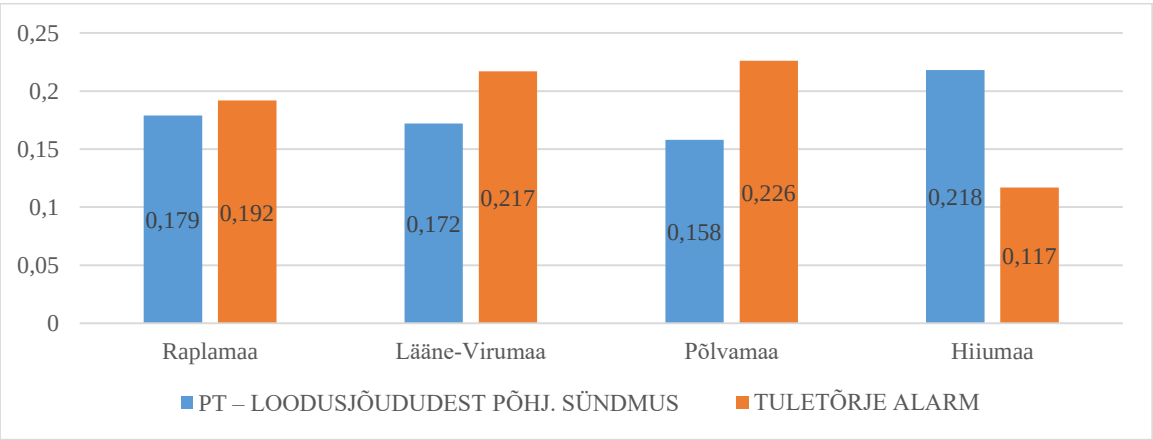
## Hiiumaal on viiendik väljakutsetest põhjustatud loodusjõududest

Teistest maakondadest eristub Hiiumaa, kus põhjustasid loodusjõud üle viiendiku väljakutsetest. Selleks on väljakutsed, mille „Väljakutse liik SOS“ väärtuseks on „PT LOODUSJÕUDUDEST PÕHJ. SÜNDMUS“. Liikide esinemissagedus on joonisel 3.24.




*Joonis 3.24: Hiiumaal toimunud väljakutsete esinemissagedused*

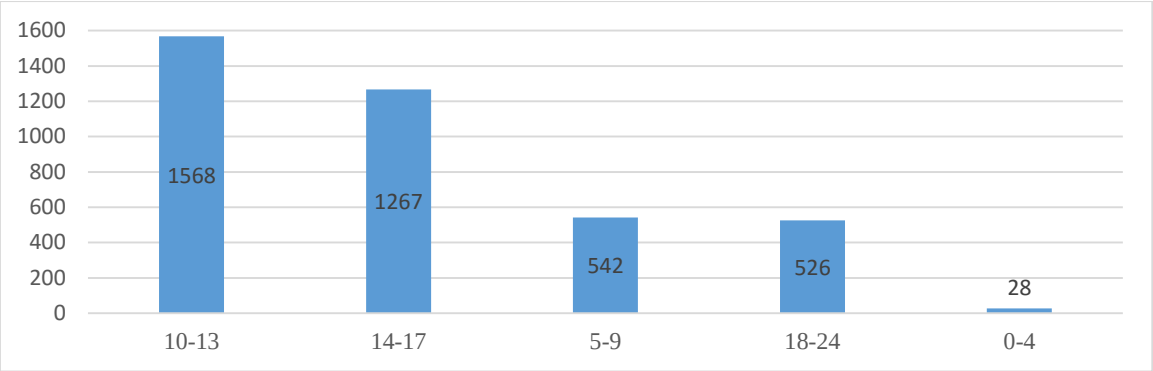
Samuti on Hiiumaa ainuke maakond, kus pole levinuimateks väljakutse liikideks „TULETÕRJE ALARM“ või „TULEKAHJU“. Kuigi Põlvamaal, Raplamaal ja Lääne-Virumaal põhjustavad loodusjõud palju väljakutseid, siis on nendes maakondades liigil „TULETÕRJE ALARM“ suurem osakaal. Jaotused on märgitud joonisel 3.25.




*Joonis 3.25: Maakonnad, kus on põhjustanud väljakutseid loodusjõud*

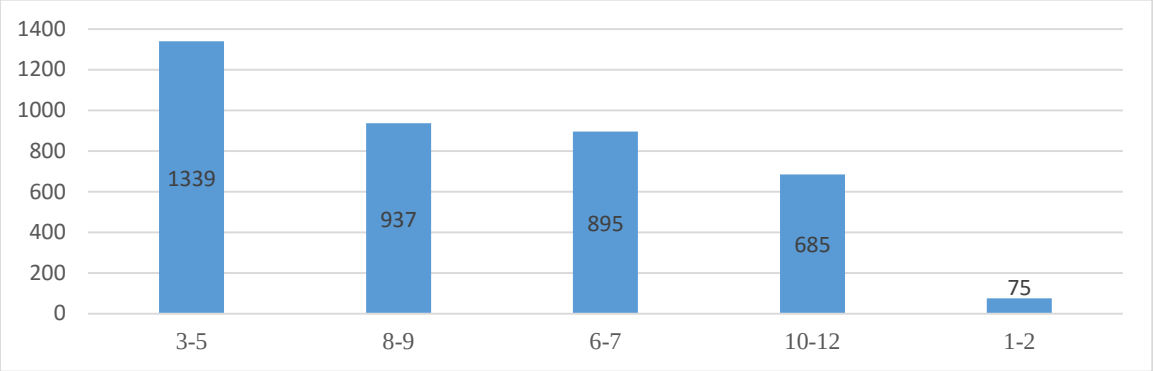
## Lõhkekehade demineerimisi ei toimu öisel ajal ning aasta esimestel kuudel

Mudelist on leitav eripära lõhkekeha demineerimisega. Kirje „DEM - LÕHKEKEHA“, mida vastavalt joonisel 3.26 toodud jaotustele on toimunud öisel ajal, täpsemalt südaööst kuni kella neljani, vaid 28 korda. See moodustab alla 0,01 protsendi kõikidest lõhkekeha demineerimisega seotud väljakutsetest.




*Joonis 3.26: Lõhkekehade demineerimise kellaaegade vahemikud*

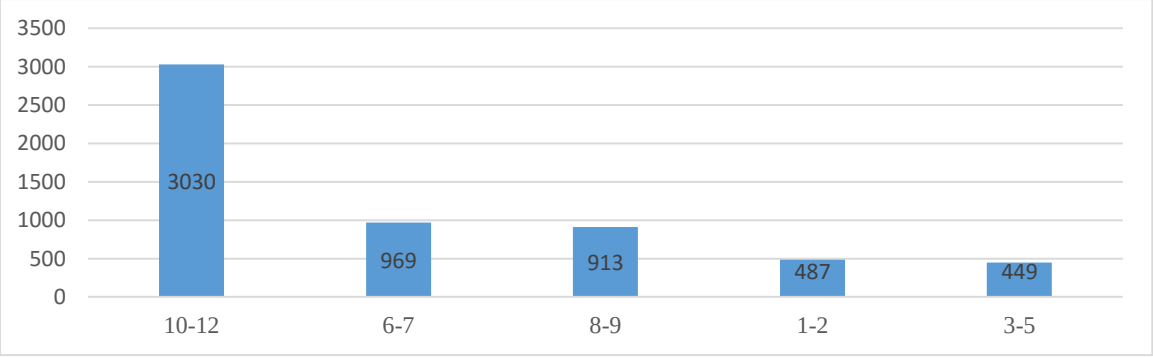
Lisaks selgub joonisel 3.27, et aastatel 2010 kuni 2013 toimus lõhkekeha demineerimisega seotud väljakutseid aasta esimese kahe kuu jooksul minimaalselt ehk ainult 75 korral, sealjuures kui vahemikus märtsist maini toimus 1339 väljakutset. Ülejäänud kuude vahemikes ei kõigu väljakutsete arv suurelt.




*Joonis 3.27: Lõhkekehade demineerimise väljakutsete esinemine kuude vahemikes aastas*

## Loodusjõududest põhjustatud väljakutsed toimuvad aasta lõpus

Automaatselt loodud kuude jaotusest selgus, et üle poolte väljakutsete kirjega „PT - LOODUSJÕUDUDEST PÕHJ. SÜNDMUS“ toimusid vahemikus oktoober kuni detsember. Joonisel 3.28 on loodusjõududest põhjustatud väljakutsete kirjete jaotused kuude lõikes. Väide täiendab eelnevas alampeatükis struktuurist „Koond“ välja toodud erineva algoritmiga leitud sama nimega oletust.

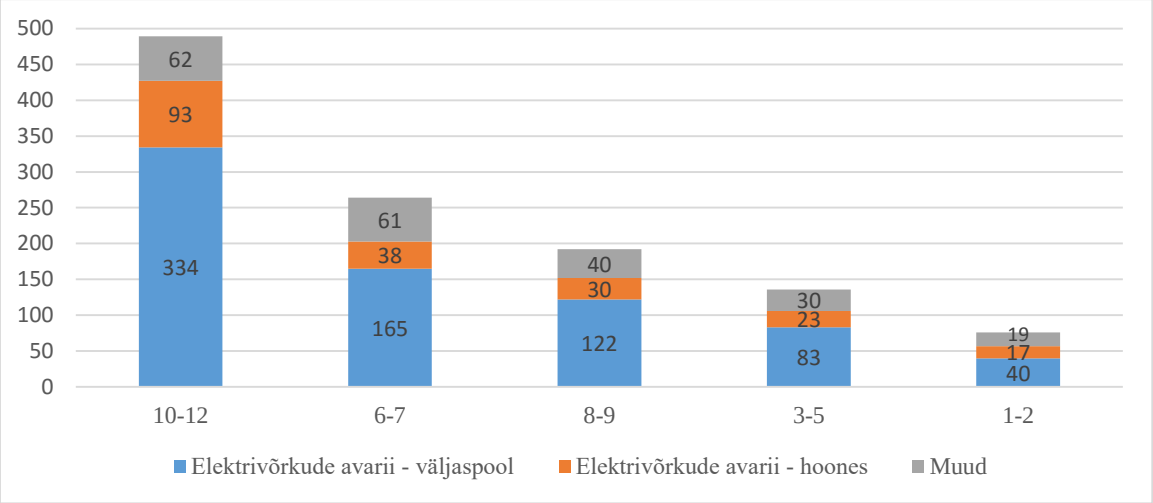



*Joonis 3.28: Loodusjõududest põhjustatud väljakutsete esinemine kuude vahemikes aastas*

## Elektrivõrkude avariidest põhjustatud väljakutsed toimuvad aasta lõpus, mitte alguses

Üle kolmandiku elektrivõrkude avariidest ehk väljakutsetest, mille liigiks on kirje „INFRA - ELEKTRIVÕRKUDE AVARII“ toimuvad aasta lõpus. Väljakutse liigi ning selle alamliigi

jagunemine kuude vahemiku suhtes on toodud välja joonisel 3.29. Kõige rohkem toimub väljakutseid aasta lõpus ning kõige vähem alguses.




*Joonis 3.29: Elektrivõrkude avarii väljakutsete ja alamliikide esinemine kuude vahemikes aastas*

Kõikides kuude vahemikes on levinuimaks alamliik „Elektrivõrkude avarii - väljaspool“, mis on märgitud 64,3 protsendil kõikidest väljakutsetest. Seega toimus väljaspool hooneid üle poolte kõikidest elektrivõrkude avariidest vahemikus 2010 kuni 2013.

Aasta alguses, jaanuaris ja veebruaris toimub minimaalselt ehk alla 10% kõikidest avariidest. Samas toimus 42,26 protsenti väljakutsetest liigiga „INFRA - ELEKTRIVÕRKUDE AVARII“ oktoobrist detsembrini.

## 3.2 2010–2013 vähem levinumad väljakutsed

Erinevate väljakutse liikide esinemissagedus vahemikus 2010 kuni 2013 erineb mitmekümnekordselt. Näiteks väljakutse liiki „TULETÕRJE ALARM“ esineb nelja aasta jooksul 18 942 korral, mis on 6652 väljakutse võrra rohkem, kui kogu peatükis uuritav andmekogu. Selleks, et leida vähem levinumate väljakutsete omadused ning väljakutse liikide sarnasused, on eemaldatud tugevalt domineerivate väljakutse liikidega väljakutsed.

### 3.2.1 Leiud andmekaevestruktuurist „Vahendatud“

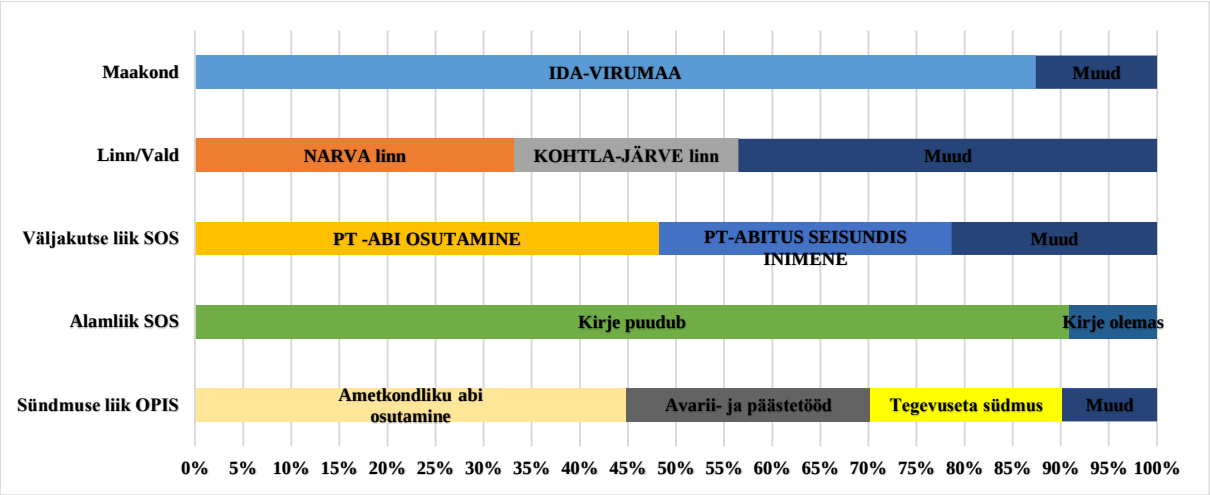
Alampeatükk põhineb andmekaevestruktuuril „Vahendatud“, milles klasteranalüüsitakse vähem levinud väljakutseid.

**Tähelepanekud mudelist „Vahendatud - Kõik“**

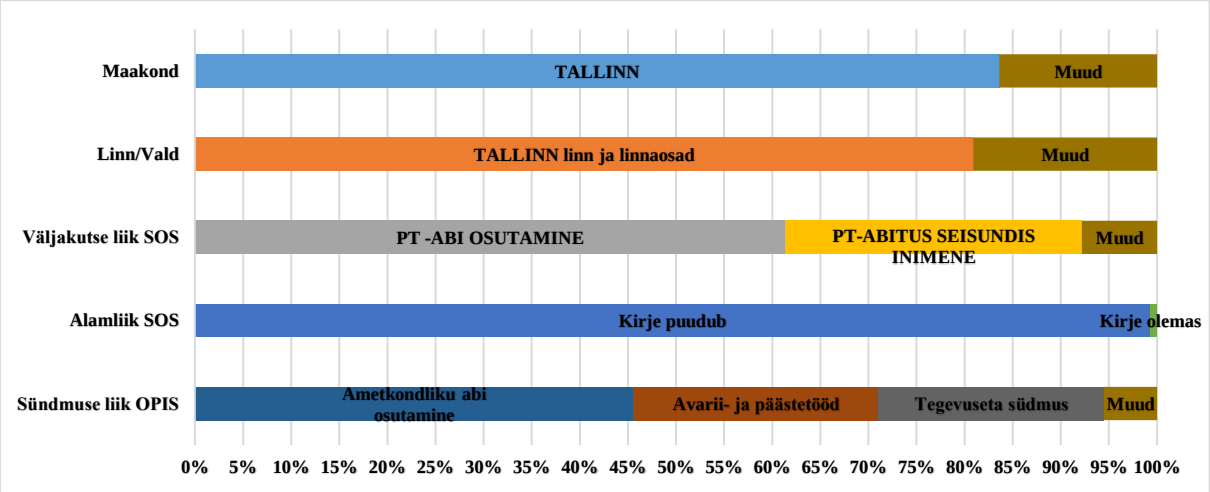
Klastrite kohta on esitatud andmeveergude jaotused, mis aitavad kaasa kirjeldamisele ehk sisaldavad informatsiooni, mis on jaotunud väheste andmevälja kirjete vahel.

Kasutades mudelit „Vahendatud - Kõik“, kus on sisend- ja väljundveergudeks kasutatud kõiki töös uuritavaid andmeveerge, eristub kolm maakonda: Tallinn, Ida-Virumaa ja Tartumaa. Ida-Virumaa klaster on suurusega 1401 ja joonisel 3.30, Tallinna klaster suurusega 1413 ja joonisel 3.31 ning Tartumaa kobar on suurusega 834 ja esitatud 3.32.

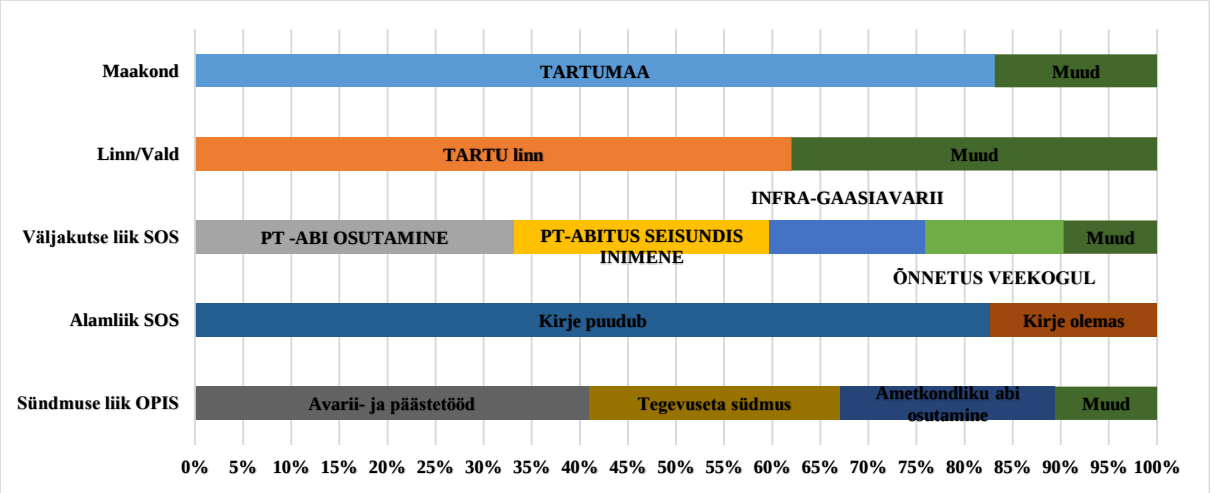
Kuigi kõigil kolmel kobaral kattuvad suurimad väljakutse liigid, siis Tartumaa klaster on erinev, kuna jaotub rohkemate väljakutse liikide vahel. Samuti on kõikidel klastritel suuresti märkimata väljakutsete alamliigid. Kobarad jagunevad ühtlaselt üle aasta laiali ning ükski kuu ei domineeri.




*Joonis 3.30: Ida-Virumaa vähendatud klastri andmeveergude jaotus*




*Joonis 3.31: Tallinn maakonnana vähendatud klastri andmeveergude jaotus*



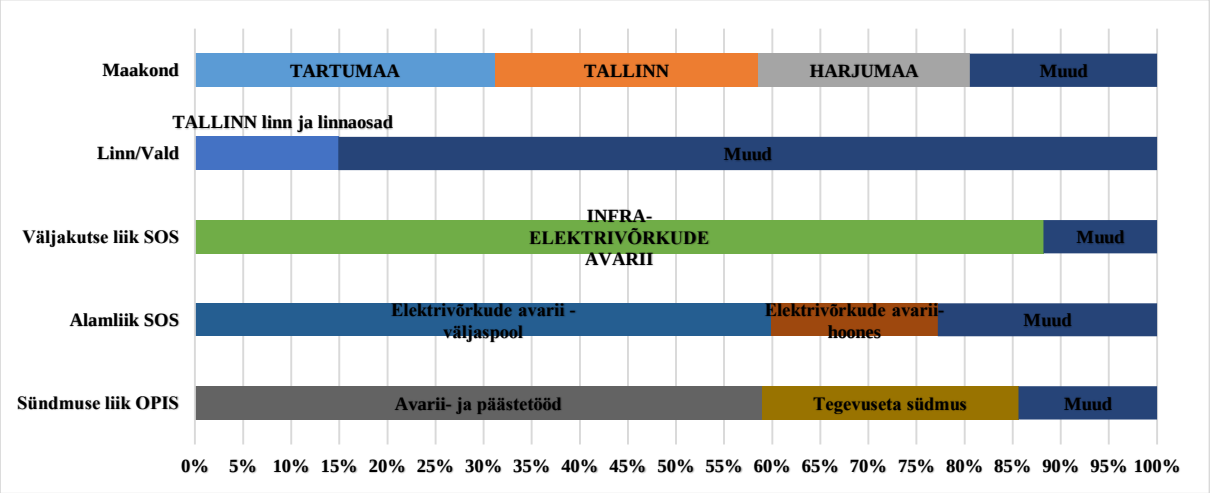

*Joonis 3.32: Tartumaa vähendatud klastri andmeveergude jaotus*

Peale maakondade on eristunud kolm erinevat väljakutse liiki: „INFRA - ELEKTRIVÕRKUDE AVARII“, „REOSTUMINE - NAFTASAADUSTEGA“ ja viimasena „TULEKAHJU TRANSPORDIVAHENDIS“.

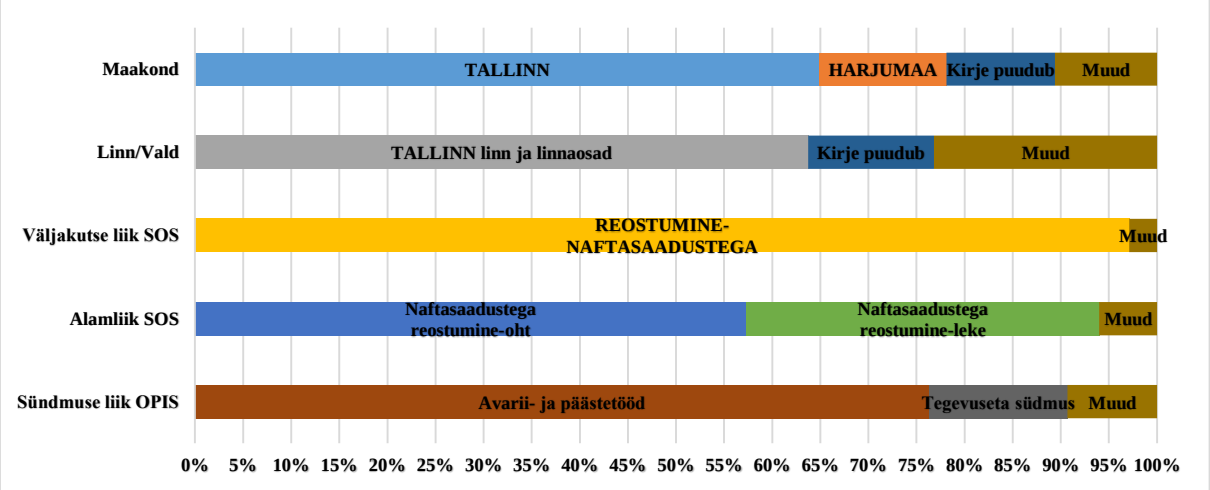
Klastrid on jagunenud erinevate maakondade vahel, kuid suurimate jaotuste hulgas on kõigil Harjumaa ning Tallinna kirjed. Elektrivõrkude avarii klastri suurus on 1045 väljakutset ning

kujutatud joonisel 3.33. Naftasaadustega reostumise kobaral on 1025 väljakutset ja esitatud joonisel 3.34. Viimasena eristuvatest väljakutsetest on joonisel 3.35 esitatud klaster „TULEKAHJU TRANSPORDIVAHENDIS“, mis esineb 834 korda.

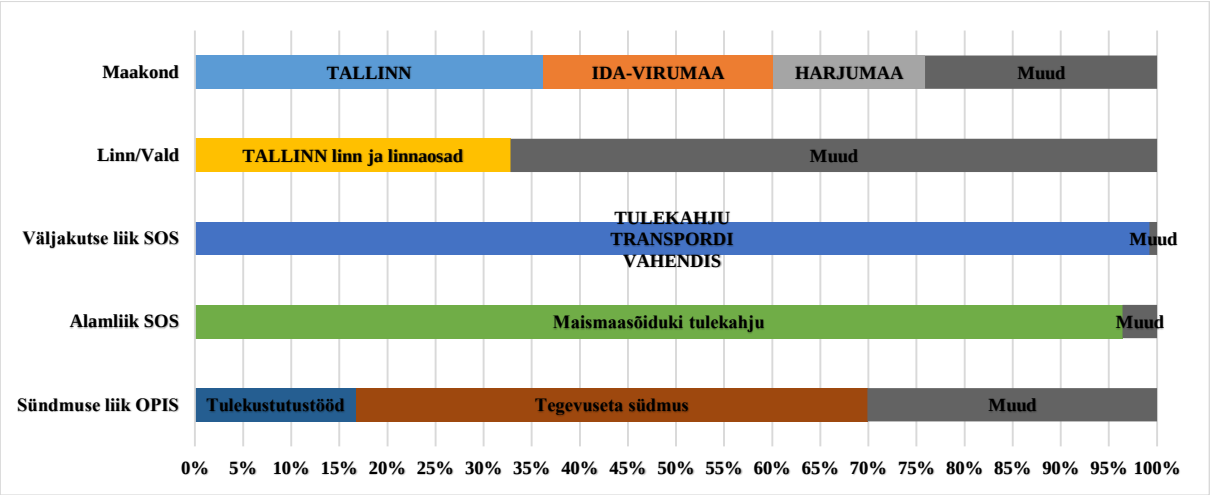
Klastrite väljakutsete liikidel on märgitud täpsustavad alamliigid. Elektrivõrkude avariil on alamliigid „Elektrivõrkude avarii - väljaspool“ ning „Elektrivõrkude avarii hoones“. Naftasaadustega reostumise korral on alamliikideks „Naftasaadustega reostumine - oht“ ja „ Naftasaadustega reostumine - leke“. Viimasena on tulekahju transpordivahendis alamliigina eristunud kirje „Maismaasõiduki tulekahju“.




*Joonis 3.33: Elektrivõrkude avarii vähendatud klastri andmeveergude jaotus*




*Joonis 3.34: Naftasaadusetega reostumine vähendatud klastri andmeveergude jaotus*

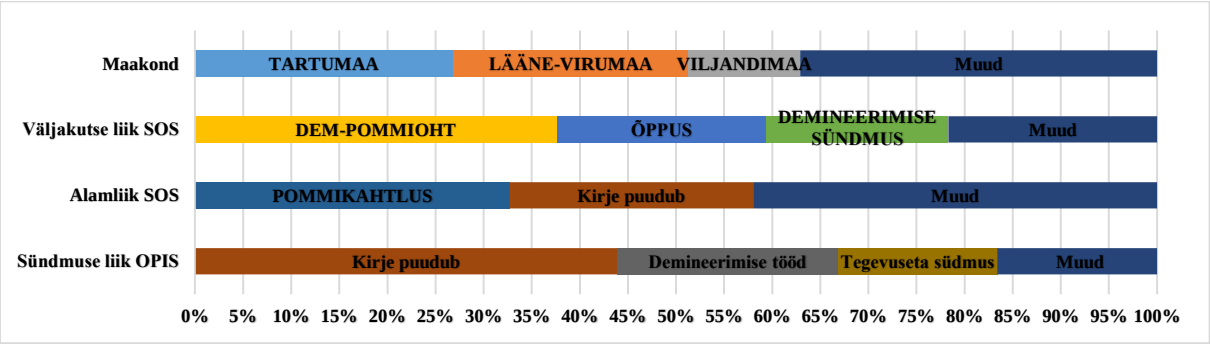



*Joonis 3.35: Tulekahju transpordivahendis vähendatud klastri andmeveergude jaotus*

Lisaks eristunud klastritele leidub mudelis teisi kobaraid, kus ükski andmeveeru kirje ei domineeri täielikult, kuid mis sellegipoolest sisaldavad endas väärtuslikku informatsiooni.

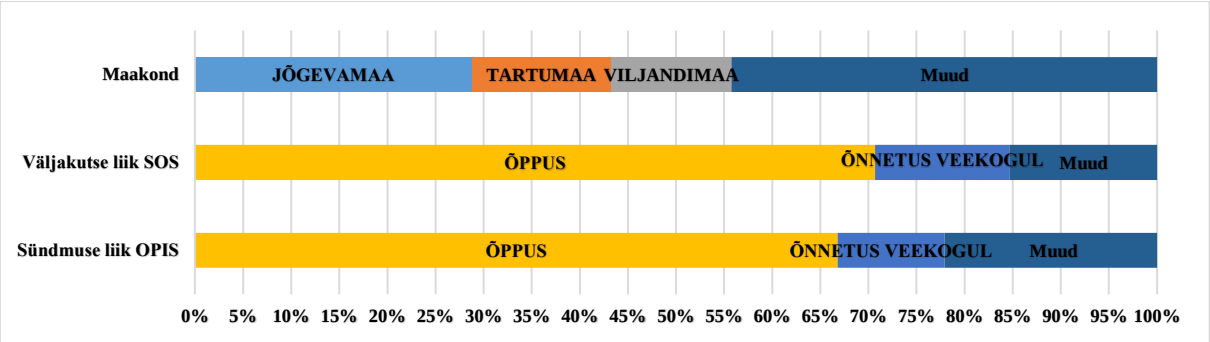
Demineerimissündmustega seotud klaster on esitatud joonisel 3.36, mille suuruseks on 735 väljakutset ning suurimate jaotustega on väljakutse liigid „DEM - POMMIOHT“, „DEMINEERIMISE SÜNDMUS“ ja „DEM- POMMIÄHVARDUS“, mis moodustavad kokku 65,8 protsenti kogu jaotusest. Viiendik jaotusest on liigi „ÕPPUS“ all. Väljakutse alamliigid toetavad demineerimissündmusi, kuna suurima jaotusega on liik „POMMIKAHTLUS“ ja „Sündmuse liik OPIS“ andmevälja puhul on teise jaotusega kirje „Demineerimise tööd“.

Klaster on jagunenud ühtlaselt kuude vahel ning väljakutsed on toimunud kas maakonnas „TALLINN“ või on maakond hoopiski märkimata.



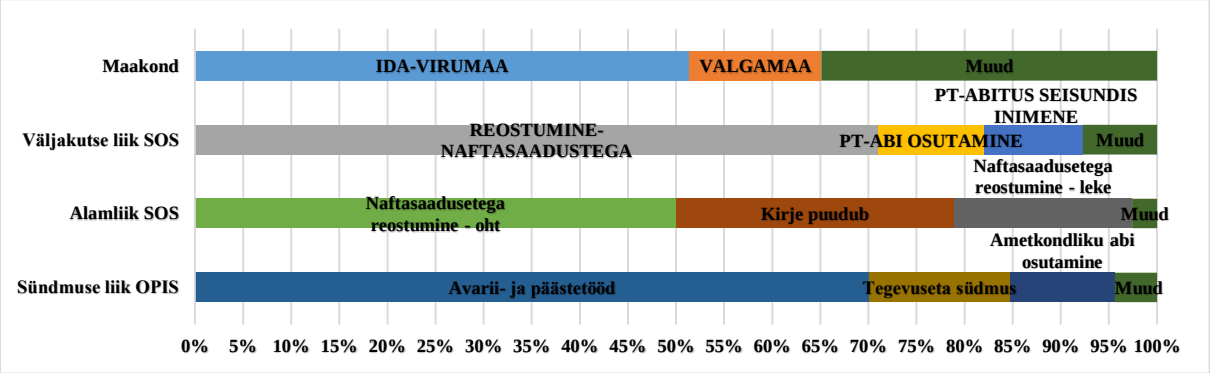

*Joonis 3.36: Demineerimisega seotud klastri andmeveergude jaotumine*

Mudelis on klaster joonisel 3.37, kus väljakutsete sündmused on seotud õppustega erinevates maakondades. Põhiliselt on õppused toimunud Jõgevamaal, Tartumaal ja Viljandimaal. Õppustega on seotud ka õnnetused veekogudel. Kirje „ÕPPUS“ moodustab pool väljakutse ja sündmuse liikidest.



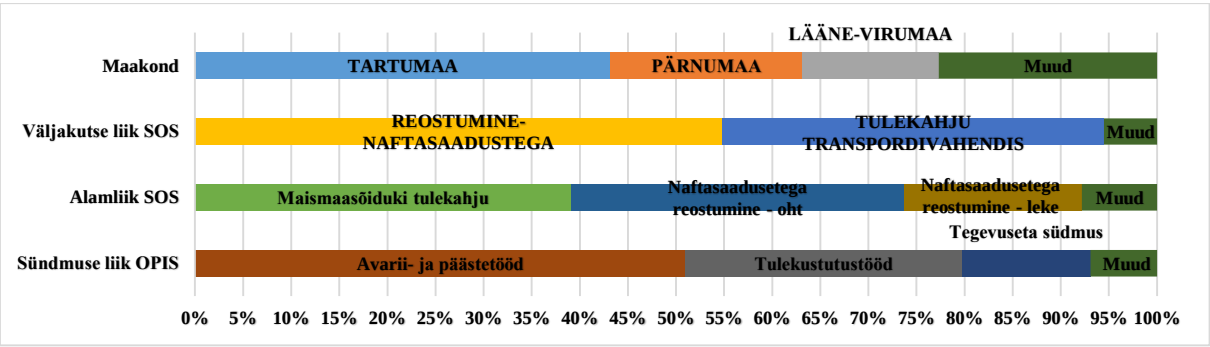

*Joonis 3.37: Õppustega seotud klastri andmeveergude jaotumine*

Lisaks joonisel 3.33 esitatud väljakutse liigiga naftasaadustega reostumine, mis on eelkõige seotud Tallinnaga, esineb seda väljakutse liiki joonisel 3.38 ka Ida-Virumaal ja vähem Valgamaal. Osa klastrist on seotud päästetöödega, kus on aidatud abitus olukorras inimesi. Pooled väljakutse alamliikidest on seotud naftasaadustega reostumine ohuga, väiksem osa aga lekkega.




*Joonis 3.38: Naftasaadustega reostumise Ida-Virumaal klastri andmeveergude jaotumine*

Klaster joonisel 3.39, mille põhilisteks väljakutse liikideks on „REOSTUMINE NAFTASAADUSTEGA“ ja „TULEKAHJU TRANSPORDIVAHENDIS“, on seotud põhiliselt Tartumaaga, kuid ka Pärnumaa ning Lääne-Virumaaga. Väljakutsed toimusid üle pooltel kordadel sündmuse liigiga „Avarii- ja päästetööd“. Suurimaks alamliigiks on „Maismaasõiduki tulekahju“, millele järgnevad väljakutse liigi „REOSTUMINE - NAFTASAADUSTEGA“ täpsustused „Naftasaadustega reostumine - oht“ ning „Naftasaadustega reostumine - leke“.




*Joonis 3.39: Naftasaadustega reostumise ja tulekahju transpordivahendis klastri andmeveergude jaotumine*

# Kokkuvõte

Päästeameti püstitatud strateegia eesmärgiks on 2025. aastaks vähendada kodanike kaasabil õnnetuste arvu ja kahjusid Põhjamaade tasemini [4]. Strateegia üheks osaks on soov panustada tehnoloogia arengusse, et tõsta Päästeameti efektiivsust, mistõttu on oluline kasutada ära töö käigus leitud informatsiooni ning andmeid, et kasutada piiratud ressursse võimalikult arukalt.

Äriteadmus ehk äritegevuse käigus tekkinud andmete mõistmine äri kasuteguri suurendamiseks on tänapäeval ettevõtte konkurentsivõimelisena hoidmise jaoks tähtis. Eesti Päästeamet pole traditsioonilises mõttes küll äri, kuid ka neile on oluline kasutada olemasolevaid vahendeid parima tulemuse saavutamiseks. Bakalaureusetöö eesmärgiks on toetada Päästeameti poolt ettevõetud initsiatiivi.

Uuritud on andmevälju, mis ei olnud väljakutsele vastavalt spetsiifilised ega minimaalselt täidetud. Väljakutsete uurimine klasteranalüüsi meetodi abil grupeeris sarnased juhtumid kokku, mille abil on leitud iseloomustavaid kirjeid maakondadele, väljakutsetele ning kuudele.

Erinevates struktuurides eristusid kobarate seast Eesti rahvaarvult suurimad maakonnad: Ida-Virumaa, eraldi maakonnana Tallinn ning Tartumaa. Sarnaselt eristus koondandmetest üks suurima esinemissagedusega väljakutse liik „TULEKAHJU“, vähendatud andmetest eristus kolm kobarat: „INFRA - ELEKTRIVÕRKUDE AVARII“, „REOSTUMINE - NAFTASAADUSTEGA“ ning „MAISMAASÕIDUKI TULEKAHJU“. Kuudest tekkisid eraldi klastrid augusti ning detsembri kohta. Ülejäänuid on kirjeldatud neile iseloomulike andmeveergude põhjal.

Töö raames leiti mustreid, trende ja erandeid. Aastatel 2010 kuni 2013 toimunud väljakutsete vaatlused on kirjeldatud koos klastritega, et neid oleks võimalik võrrelda järgnevate aastate andmetega. Vaatlustelt selgunud mustrid on „Metsa- ning maastikutulekahjud kestavad aprilli algusest kuni mai keskpaigani“; „Loodusjõududest põhjustatud väljakutsed toimuvad novembri lõpust aasta lõpuni“; „Lõhkekeha demineerimisi ei toimu öisel ajal ning aasta esimestel kuudel“ ja „Elektrivõrkude avariidest põhjustatud väljakutsed toimuvad aasta lõpus, mitte alguses“.

Leitud on erandid „Lõhkekeha demineerimisel puudub asukoht“ ja „Hiiumaal on viiendik väljakutsetest põhjustatud loodusjõududest“ ning trend „Tuletõrjealarmidest põhjustatud väljakutsete hooajad“. Töö tulemused pakuvad informatsiooni sissejuhatuses välja toodud Päästeameti sihtide saavutamiseks.

Autori arvates leidub andmetes veel palju informatsiooni Päästeameti väljakutsete kohta, mis siiski kahjuks ei mahu töö skoopi. Tulevikus on võimalik kasutada rohkem erinevaid andmekaeve algoritme, et töödelda andmeid võimalikult mitmekülgselt. Mõeldav oleks täiendada olemasolevat andmekogu lisaandmetega. Andmete mitmekesistamiseks saab kasutada andmeveergu „Väljakutse aeg“, mis sisaldab endas palju metaandmeid: kas tegu on nädalavahetusega, milline on ilm ja kas on tegu riigipühaga.

Täiendavat informatsiooni saaks tulevikus kasutada väljakutsete põhjuste ning aja paremaks mõistmiseks. Näiteks saaks uurida kobarate muutumist ajas, et näha, kas seosed andmeveergude kirjete vahel jäävad samaks või muutuvad. Samuti on tähtis uurida väljakutse liikide sõltuvust ilmast.

Bakalaureusetöö algne eesmärk oli kasutada lisaks andmeveerge „Sündmuse kirjeldus“ ning „Sündmuse täpsustav kirjeldus“, milles oli väljakutsel viibinud päästja poolt vabas vormis kirjutatud sündmuse kirjeldus. Lisaks on nendel veergudel märgitud ajaline logi ja probleemi kirjeldus ning lahendus. Kahjuks ei soovinud Päästeameti juristid avaldada informatsiooni, mis oleks võinud minna vastuollu isikuandmete kaitsmise nõudega. Lisaks on kaitstud andmeväljad „Pääste x“ ja „Pääste y“, mis märkisid väljakutse koordinaate, „Helistaja nimi“ ning „Helistaja

telefon“ ja „Pääste aadress“ koos „SOS aadress“ veeruga  märkisid väljakutse täpset aadressi. Seetõttu on kasutatud töös vähendatud andmete väljavõtet süsteemist OPIS.

# Kasutatud kirjandus


[1] Eesti Päästeamet, „Eesti Päästeameti strateegia,“ [Võrgumaterjal]. Available: www.paasteamet.ee/et/paasteamet/organisatsioon/strateegia.html. [Kasutatud 10 mai 2016].

[2] P. Nielsen, K. Delaney, G. Low, A. Machanic, P. S. Randal ja K. L. Tripp, Server MVP Deep Dives, Greenwich: Manning Publications Co, 2010.

[3] K. Reim, J. Remm ja A. Kaasik, „Klasteranalüüs,“ %1 *Ruumiliste loodusandmete statistiline analüüs*, Eestikeelsete digitaalsete õpikute hoidla, 2012, pp. 75-76.

[4] Eesti Päästeamet, „Eesti Päästeameti aastaraamat 2014,“ [Võrgumaterjal]. Available: http://www.paasteamet.ee/et/paasteamet/organisatsioon/aastaraamat2014.html. [Kasutatud 10 mai 2016].

[5] Eesti Statistikaamet, „RV022: RAHVASTIK SOO, VANUSERÜHMA JA MAAKONNA JÄRGI, 1. JAANUAR,“ [Võrgumaterjal]. Available: http://www.stat.ee/andmebaas. [Kasutatud 9 mai 2016].

[6] Microsoft Corporation, „Analysis Services,“ [Võrgumaterjal]. Available: https://msdn.microsoft.com/en-us/library/bb522607(v=sql.120).aspx. [Kasutatud 10 mai 2016].

[7] Microsoft Corporation, „Querying Multidimensional Data with MDX,“ [Võrgumaterjal]. Available: https://msdn.microsoft.com/et-ee/library/bb500184(v=sql.120).aspx. [Kasutatud 11 mai 2016].

[8] Microsoft Corporation, „Multidimensional Modeling (SSAS),“ [Võrgumaterjal]. Available: https://msdn.microsoft.com/en-us/library/hh230904(v=sql.120).aspx. [Kasutatud 9 mai 2016].

[9] Microsoft Corporation, „Data Mining Algorithms (Analysis Services - Data Mining),“ [Võrgumaterjal]. Available: https://msdn.microsoft.com/en-us/library/ms175595(v=sql.120).aspx. [Kasutatud 10 mai 2016].

[10] Microsoft Corporation, „Microsoft Clustering Algorithm,“ [Võrgumaterjal]. Available: https://msdn.microsoft.com/en-us/library/ms174879(v=sql.120).aspx. [Kasutatud 10 mai 2016].

[11] Microsoft Corporation, „Microsoft Naive Bayes Algorithm,“ [Võrgumaterjal]. Available: https://msdn.microsoft.com/en-us/library/ms174806(v=sql.120).aspx. [Kasutatud 9 mai 2016].

[12] Microsoft Corporation, „Microsoft Clustering Algorithm Technical Reference,“ [Võrgumaterjal]. Available: https://msdn.microsoft.com/en-us/library/cc280445(v=sql.120).aspx. [Kasutatud 8 mai 2016].

[13] Microsoft Corporation, „Microsoft Naive Bayes Algorithm Technical Reference,“ [Võrgumaterjal]. Available: https://msdn.microsoft.com/en-us/library/cc645902(v=sql.120).aspx. [Kasutatud 10 mai 2016].

[14] Microsoft Corporation, „Content Types (Data Mining),“ [Võrgumaterjal]. Available: https://msdn.microsoft.com/en-us/library/ms174572(v=sql.120).aspx. [Kasutatud 10 mai 2016].

# Lisad

## I. Välised materjalid

Tabelis 1 on kirjas bakalaureusetööle lisatud materjalid.

*Tabel 1: Tööle lisatud materjal*

| Faili nimi | Kommentaar |
|---|---|
| paasteamet2010-2013.zip | Fail sisaldab endas töö tulemuste saamiseks loodud Visual Studio 2015 analüüsipaketi andmekaeveprojekti koos kasutusjuhendiga |

## II. Litsents

**Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks**

Mina, Johannes Horm,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose **Hädaabi väljakutsete kategoriseerimine Eesti Päästeameti andmete põhjal 2010–2013** mille juhendaja on Siim Karus,
   1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 12.05.2016