TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Ethel Maasing

Eesti alaliste elanike määratlemine registripõhises loenduses

Magistritöö (30 EAP)

Matemaatilise statistika eriala

Juhendaja: Mare Vähi, MSc

Tartu 2015

Eesti alaliste elanike määratlemine registripõhises loenduses

Järgmine Eesti rahva ja eluruumide loendus planeeritakse läbi viia registripõhiselt. Üks olulisemaid teemasid loenduste puhul on üldkogumite määratlemine, et kõik isikud ja eluruumid saaksid igas riigis ühekordselt loendatud. Käesoleva magistritöö eemärk on registrite andmete põhjal määratleda loenduse isikute üldkogum 2014. aasta lõpu seisuga. Eesti alalisi elanikke prognoositi koostöös teiste Tartu Ülikooli magistrantidega erinevatel eeldustel logistilise- ja lineaarse regressiooniga ning diskriminantanalüüsiga. Töö autori poolt läbi viidud logistilise regressioonanalüüsiga eristusid registri andmete põhjal kõige paremini residendid mitte residentidest 7–16-aastaste hulgas, kuid kõige raskemini 23–62-aastaste meeste hulgas. Tööd üldkogumi määratlemisega jätkatakse, kasutades analüüsist tehnilistel põhjustel välja jäänud registreid ning kontrollrühmasid täpsustatakse antud magistritöö ettepanekute põhjal.

Märksõnad: loendused, alaline rahvastik, andmeregistrid, regressioonanalüüs.

Permanent Residency Status Determination in Register-Based Census

The next population and housing census in Estonia is intended to be register-based. One of the most important topics in the census in every country is the determination of total population so that all the persons and dwellings are counted once. The aim of this thesis is to determine the total population of individuals using data registers as of the end of 2014. Permanent residents of Estonia were predicted in collaboration with other Tartu University postgraduate students using different assumptions of logistical and linear regression and discriminatory analysis. The work carried out by the author using logistic regression analysis showed that it was easiest to differentiate residents from non-residents using data registers among 7–16-year-olds, but most difficult among men aged 23–62. Work with the determination of total population will continue using data registers that were left out of the analysis because of technical reasons and the control groups will be adjusted using the proposals of this thesis.

Keywords: censuses, resident population, data register, regression analysis.

Sisukord

SISSEJUHATUS

Eesti järgmine, kaheteistkümnes, rahva- ja eluruumide loendus, mis kuulub 2020/2021 loendusvooru, planeeritakse esmakordselt läbi viia registripõhiselt. Rahva- ja eluruumide loenduse eesmärk on koguda andmeid kõigi riigi alaliste elanike ja eluruumide kohta. Varasematel loendustel on Eestis kasutatud traditsioonilist andmete kogumise meetodit ehk loendamist. Eelmisel, 2011. aasta lõpus toimunud rahva ja eluruumide loendusel (REL2011) kasutati lisaks küsitlejatele internetis loendamise võimalust ning andmete parandamiseks ja täiendamiseks registrite andmeid [1].

Viis aastat tagasi alustati registripõhise rahva ja eluruumide loenduse (REGREL) metoodika projektiga. Selle eesmärgiks oli analüüsida Eesti registrite taset ja uurida mida on veel vaja teha, et registrite andmed oleksid piisavalt täielikud ja kvaliteetsed, et läbi viia registripõhine loendus vastavalt Euroopa määrustele: [2] ja [3]. REGREL metoodikaprojekti tulemused ja soovitused on fikseeritud lõppraportis [4]. Pärast projekti lõppu 2013. aasta sügisel on Statistikaameti metoodikud jätkanud registriandmete analüüsidega, sest lahendamist vajavaid probleeme on mitmeid: isikute ja eluruumide üldkogumite määratlemine, erinevate tunnuste moodustamiseks vajalike algoritmide täiendamine ning seaduste ja andmekogude uuendustest tulenevate muudatustega arvestamine.

Järgmine etapp REGRELis on prooviloenduse läbiviimine 2015. aasta lõpu seisuga, mille üheks eesmärgiks on moodustada registri andmete põhjal loenduse isikute üldkogum ehk määrata kõik Eesti alalised elanikud ehk residendid. Antud töö eesmärgiks on moodustada loenduse isikute üldkogum 2014. aasta lõpu seisuga kasutades erinevate Eesti registrite andmeid.

Käesolev töö on jagatud kolmeks osaks. Esimeses peatükis on toodud ülevaade vajaminevatest mõistetest ja andmetest, mida antud töös kasutatakse. Kokku on kasutatud 11 Eesti registri andmeid. Kõige põhjalikumaks andmeallikaks inimeste kohta on Eestis Rahvastikuregister (RR). Kuna andmekogu eesmärgid erinevad loenduse reeglitest, siis esineb RRis loenduse mõttes nii üle- kui alakaetust. Käesoleva magistritöö eesmärk on kindlaks teha RRi ülekaetus statistilisi meetodeid kasutades. Kuna alakaetuse suurus on võrreldes ülekaetusega väike ja vajab teistsuguseid hindamismeetodeid, siis antud töös seda ei ole uuritud. Samuti on selles peatükis kirjeldatud ja põhjendatud eeldusi:

kontrollgruppide ja soo- ning vanusrühmade valik, mis on vajalikud enne statistilise meetodi kasutamist.

Magistritöö autor kasutas Eesti alaliste elanike hindamiseks logistilist regressiooni. Teises peatükis on toodud antud meetodi kirjeldus allikate [8] ja [9] põhjal. Samuti on allikate [10] ja [11] põhjal antud ülevaade kahest mudeli headuse näitajast ning kirjeldatud, kuidas valida arvutatud prognoosidest optimaalne lävend määramaks isik residendiks või mitteresidendiks.

Viimane ehk kolmas peatükk on jagatud neljaks osaks. Esimeses alapeatükis on toodud ülevaade töö autori regressioonanalüüsi tulemustest. 2015. aasta kevadel analüüsisid töö autori poolt moodustatud koondandmeid Tartu Ülikooli magistrandid aine Andmetöötlusmeetodid raames, mille läbiviijad olid Mare Vähi ja Ene-Margit Tiit. Magistrantide eesmärk oli samuti moodustada loenduse üldkogum regressioon-, diskriminant- ja klasteranalüüsi kasutades, kuid erinevatele eeldustele tuginedes. Täpsemalt on tulemusi kirjeldatud teises alapunktis. Järgmises alapunktis on võrreldud antud töö autori ja teiste magistrantide poolt regressioon- ning diskriminantanalüüsiga saadud tulemusi Statistikaameti (SA) avaldatud rahvaarvuga. Viimase punktina on esitatud järeldused tehtud tööst ja ettepanekud edaspidiseks.

Magistritöö lisades on toodud moodustatud koondandmete kirjeldus ja ülevaated inimeste esinemisest registrites sünniaastati.

Töös on kasutatud anonüümitud andmeid ehk isikud ei ole tuvastatavad, sest on eemaldatud ees-ja perekonnanimed ning isikukoodid.

Töö kirjutamiseks on kasutatus tekstitöötlusprogrammi *MS Word*. Analüüsiks on kasutatud statistikaprogrammi *SAS Enterprise Guide* ja tabelite ning jooniste kujundamiseks *MS Excel*'it.

Autor tänab oma lõputöö juhendajat Mare Vähit ja kolleege SA metoodika- ja analüüsiosakonnast Kristi Lehtot ja Ene-Margit Tiitu sisukate märkuste, soovituste ja konsultatsioonide eest. Samuti tänab autor kaasüliõpilasi Maia Arget, Kristi Tüli, Kaidi Jõge ja Hindrek Tederit sisukate arutelude eest aine Andmetöötlusmeetodid raames.

1. ÜLEVAADE MÕISTETEST JA ANDMETEST

Käesolev peatükk on jagatud viieks. Esimeses osas on kirjeldatud erinevaid vajalikke mõisteid ning püstitatud töö eesmärk. Järgnevates alapunktides on vaadeldud, milliseid andmeid on võimalik kasutada ja täpsustatud eesmärgini jõudmist.

1.1. Mõisted ja ülesande püstitus

Rahva- ja eluruumide loenduse eesmärk on koguda andmeid kogu riigi (piirkonna) rahvastiku, leibkondade ja eluruumide kohta fikseeritud ajamomendil (loendusmomendil). Tänu rahvastikusündmustele: sündimus, suremus ja ränne, on rahvastik pidevalt muutuv ja seda infot kasutatakse loenduste vahelistel perioodidel rahvastiku suuruse arvutamiseks. Käesoleva töö eesmärk on määratleda kogu Eesti riigi alaline rahvastik ehk isikute üldkogum ehk residendid seisuga 01.01.2015.

Säilitamaks riikide võrreldavust ja tagamaks iga inimese ühekordset loendamist, on vajalik alalise rahvastiku väga täpne määratlus. Euroopa Liidu liikmesriikide loendused peavad vastama Euroopa Nõukogu ja Parlamendi [2] ning Euroopa Komisjoni määrustele [3]. Antud töös on kasutatud määruseid, mis olid kehtivad eelmises rahva- ja eluruumide loenduse voorus (Eestis toimus viimane loendus momendiga 31.12.2011). Loendusvooruks (2020/2021) ei ole uusi määrusi veel vastu võetud. Praegusel hetkel ei ole ükski riik teinud ettepanekut, et oleks vaja muuta alaliste elanike ehk rahvastiku üldkogumi definitsiooni.

Euroopa Nõukogu ja Parlamendi määruse [2] põhjal:

„Alaliste elanikena käsitletakse ainult järgmisi kõnealuses geograafilises piirkonnas elavaid isikuid:

- neid, kes elasid enne võrdluskuupäeva oma alalises elukohas pidevalt vähemalt kaheteistkümne kuu jooksul, või
- neid, kes saabusid oma alalisse elukohta viimase kaheteistkümne kuu jooksul enne võrdluskuupäeva eesmärgiga elada seal vähemalt üks aasta.

Kui eelmistes punktides kirjeldatud tingimusi ei ole võimalik kindlaks teha, tähendab „alaline elukoht" seaduslikku või registreeritud elukohta."

Eelmistel Eestis korraldatud loendustel on inimesed saanud ise vastata, kas nad on Eestis vähemalt aasta elanud või kavatsevad seda teha. Tulevikus seda teha ei saa ning kogu vajalik info tuleb koguda registrite andmetest.

Kõige põhjalikumat infot Eesti inimeste kohta saab RRist, mis kogub andmeid Eesti kodanike, Eestis elukoha registreerinud Euroopa Liidu, Euroopa Majanduspiirkonna liikmesriigi ja Šveitsi Konföderatsiooni kodanike ning Eestis elamisloa või elamisõiguse saanud välismaalaste kohta. [5]

Kuigi RRi kantakse kõik rahvastikusündmused, sh sünd, surm ja registreeritud elukohavahetus, siis leidub ka mõningaid suuri probleeme loenduse mõttes, mida on kajastatud eelmise loenduse järel tehtud analüüsis:

„Suurimaks puuduseks Eesti põhilises registris – rahvastikuregistris – on erinevus registreeritud ja tegeliku elukoha vahel. Kuni viiendikul juhtudest elavad inimesed registreeritud elukohast erinevas kohas. Sellel nähtusel on terve rida põhjuseid, mis said alguse sellest, kui Riigikogu 90. aastate algul tühistas nõukogude ajal kehtinud sissekirjutuse kohustuse kui igandi. Kuigi praeguseks on elukoha registreerimine taas kohustuslikuks tehtud, pole paljud inimesed seda teadvustanud ja usuvad jätkuvalt, et see on vabatahtlik. Elukoha valesti registreerimist soodustavad (inimliku laiskuse kõrval) mitmesugused paikkondlikud soodustused (koolide ja lasteaedade valimine, pensionilisa, sõidusoodustused). Kõik kohalikud omavalitsused, sealhulgas Tallinn, on huvitatud võimalikult suurest registreeritud elanike arvust. Kõik kirjeldatud probleemid tähendavad seda, et alalise rahvastiku paiknemine riigis võib märgatavalt erineda registreeritust.

Elukoha registreerimise nõude eiramine põhjustab vea ka tegeliku rahvaarvu (üldkogumi) hindamisel. Inimesed, kes ei pea elukoha registreerimist oluliseks ja eiravad seda nõuet, ei pea ka vajalikuks registreerida enda riigist lahkumist. Seega elavad nad vormiliselt riigis edasi, kuigi on siit aastate eest lahkunud. Ka sellisel käitumisel võib olla mõistuspärane (omakasupüüdlik) põhjus: säilitades vormiliselt Eesti elukoha, säilitatakse õigus mõnede teenuste saamisele Eesti riigilt. Teisest küljest võib sellist käitumist vaadelda ka kui soovi säilitada side Eestiga, pidades silmas kavatsust tulevikus kodumaale naasta.“ [6]

Käesolevas töös ei keskenduta täpsele alalisele elukohale, vaid piirdutakse riigi tasandiga: kas isik on Eestis alaline elanik või mitte. Eesmärk on tuvastada RRi väljavõttest seisuga 01.01.2015, millised inimesed kuuluvad Eesti alaliste elanike hulka ja kes on tegelikult Eestist lahkunud. Kokku jagunevad inimesed nelja rühma:

1. Registreeritud elukoht Eesti – tegelik elukoht Eesti
   Siia gruppi kuulub suurem hulk RRis olevaid isikuid, kes kuuluvad loenduse üldkogumisse.
2. Registreeritud elukoht Eesti – tegelik elukoht välismaal
   Isikud, kes on Eestist lahkunud, kuid ei ole RRile sellest teada andnud. Ei kuulu loenduse üldkogumisse.
3. Registreeritud elukoht muu riik või puudu – tegelik elukoht Eesti
   Isikud, kes on Eestisse (tagasi) tulnud ja ei ole sellest teada andnud. Kuuluvad loenduse üldkogumisse.
4. Registreeritud elukoht muu riik või puudu – tegelik elukoht välismaa
   Isikud, kes on ka tegelikult Eestist lahkunud. Ei kuulu loenduse üldkogumisse.

Kasutades Ene-Margit Tiidu poolt soovitatud rühmitamist RRi elukoha, eelmises rahvaloenduses kogutud info ja 2012. aastal hinnatud loenduse alakaetuse põhjal, on võimalik suuremas osas määrata eespool toodud rühmadesse 1 ja 4 kuuluvad isikud, mida saab kasutada kui õpperühmi, et moodustada eeskiri, mille alusel hinnata rühmadesse 2 ja 3 kuuluvate isikute Eestis elamist. Rühmadesse jagamisest annab täpsema ülevaate selle peatüki alapunkt 1.2.

Lisaks RRile on Eestis veel mitmeid registreid. Näiteks, riik kogub kõikide õppijate kohta info Eesti Haridus Infosüsteemi, kõik juhilubade vahetused ja seaduslikud sõidukite müügi ja ostuga seotud omanike vahetused on kantud Liiklusregistrisse jne. Magistritöö eesmärk on määrata inimeste residentsus ehk teha kindlaks, kas inimene registrite andmete põhjal elab alaliselt Eestis või mitte.

Antud töö valmimise ajaks oli lisaks RRile võimalik kasutada 10 registri andmeid. Andmete hõivamisel tekkinud probleemide tõttu ei saa analüüsis kasutada kõiki Eesti isikuandmeid sisaldavaid registreid, sh Isikut tõendavate dokumentide andmekogu, Vangide ja kriminaalhooldusaluste registri ning Kohustusliku kogumispensioni registri andmeid. Täpsem ülevaade andmekogudest, mida kasutatakse, on toodud alapunktis 1.3.

Käesolevas töös ei ole analüüsitud neid inimesi, kes kuuluvad teistesse Eesti registritesse, kuid ei ole esindatud RRis ehk ei ole uuritud RRi alakaetust.

1.2. Kontrollgruppide valimine

Enne residentsuse eeskirja loomist on vajalik määrata inimeste grupid, kes kindlalt elavad Eestis ja kes on siit lahkunud. Üheks võimalikuks variandiks on kasutada RRi elukoha infot ja eelmises loenduses saadud tulemusi. Antud töös on kasutatud Ene-Margit Tiidu poolt soovitatud grupeerimist, kus jagati kõik RR 01.01.2015 väljavõttes olevad isikud järgmiste tunnuste alusel:

- Elukoha riik 01.01.2015 RR-s

  on Eesti: „kuulub RRi“

  on muu riik või puudub: „ei kuulu RRi“;

- Kas rahuldas 2012.a residentsuse kriteeriumit [6]: rahuldas RK12/ ei rahuldanud RK12;
- Tulemus eelmisel rahvaloendusel: loendati Eesti elanikuna/ loendamata/ loendati välismaalasena;
- Pärast viimast rahvaloendust (aastatel 2012-2014) sündinud ja sisse rännanud isikud.

Kolme esimese tunnuse kõikvõimalikud kombinatsioonid moodustavad 12 rühma, mis on toodud tabelis 1. Lisaks on veel 2 eraldi rühma aastail 2012 – 2014 sündinud ja sisserännanud isikutest. Seega kokku moodustub 14 erinevat rühma.

Küllalt kindlalt võib väita, et Eestis elavad alaliselt tabeli 1 järgi rühmadesse 12 – 14 kuuluvad isikud ehk püsielanikud. Rühma number 12 kuuluvad isikud, kes on 01.01.2015 seisuga RRis Eesti elukohaga, nad loendati eelmisel loendused 2012. aasta alguses ning nad vastasid eelmise loenduse alakaetuse hindamisel moodustatud residentsuse kriteeriumile registrite info põhjal. Rühmadesse 13 ja 14 kuuluvad need isikud, kes sündisid või saabusid Eestisse 2012 – 2014 aasta jooksul. Viimasesse rühma võib kuuluda ka neid, kes on saabunud ajutiselt Eestisse, kuid 01.01.2015 momendiks on juba lahkunud ja nad ei ole registreerinud enda lahkumist. Aga antud töö autori hinnangul ei esine selliseid olukordi rohkem kui rühmades 12 ja 13. Edaspidi on seda õpperühma nimetatud kui kindlalt Eesti residendid.

Eestist lahkunuteks võib pidada rühmadesse 1 ja 2 kuuluvaid isikuid, sest neil ei ole RRis elukohaks Eesti, nad ei rahuldanud loenduse järgset alakaetuse hindamiseks loodud residentsuse kriteeriumit ega saanud loendatud või loendati välismaalasena (inimene ise

andis teada, et elab välismaal või perekonnaliige vastas, et see inimene ei ela Eestis). Edaspidi on seda rühma nimetatud kui kindlad Eesti mitteresidendid.

Rühmadesse 3 – 11 jäävad need isikud, kelle kohta võib olla kõige suurema tõenäosusega RRis vale registreeritud elukoha riik ning nende residentsuse staatust antud töö praktilises osas hinnataksegi.

Tabel 1. Residentsuse rühmade jaotus, 01.01.2015

Nr.	Rühma iseloomustus	Staatus	Arvukus
1.	Ei kuulu RR15, ei rahulda RK12, loendati välismaalasena	mitteresident	16 894
2.	Ei kuulu RR15, ei rahulda RK12, loendamata	mitteresident	59 081
3.	Ei kuulu RR15, rahuldas RK12, loendati välismaalasena	uuritav	892
4.	Ei kuulu RR15, ei rahulda RK12, loendati Eesti elanikuna	uuritav	8 145
5.	Ei kuulu RR15, rahuldas RK12, loendamata	uuritav	1 520
6.	Ei kuulu RR15, rahuldas RK12, loendati Eesti elanikuna	uuritav	12 712
7.	Kuulub RR15, ei rahulda RK12, loendati välismaalasena	uuritav	16 971
8.	Kuulub RR15, ei rahulda RK12, loendamata	uuritav	31 145
9.	Kuulub RR15, rahuldas RK12, loendati välismaalasena	uuritav	138
10.	Kuulub RR15, ei rahulda RK12, loendati Eesti elanikuna	uuritav	75 726
11.	Kuulub RR15, rahuldas RK12, loendamata	uuritav	25 094
12.	Kuulub RR15, rahuldas RK12, loendati Eesti elanikuna	resident	1 150 795
13.	Kuulub RR15, sündinud 2012 – 2014	resident	43 640
14.	Kuulub RR15, saabunud 2012 – 2014	resident	20 106
Kokku			**1 462 859**

1.3. Andmestiku loomine

Enne praktilise ülesande lahendamise juurde asumist tehti ära suuremahuline eeltöö, mis hõlmas vajalike registrite ja andmete väljaselgitamist, tellimist ja hõivamist SAsse, anonüümimist ja koondandmestiku loomist. Järgnevalt on toodud protsesside kirjeldused ja teostajad.

Kui RRi andmeid on SAs analüüsitud mitmeid aastaid, siis mõne registri andmed said residentsuse analüüsi jaoks tellitud esmakordselt, näiteks E-toimiku süsteemi andmed. Selle jaoks, et andmed jõuaksid SAsse, on vajalik töötada läbi andmekogu põhimäärus ja kogutavate tunnuste loetelu, vajadusel kohtuda andmekogu esindajatega ning koostada leping. Esmaseks allikaks andmekoguga tutvumisel on Riigi infosüsteemi haldussüsteem (RIHA). Antud töö sai tehtud koostöös SA vanemmetoodiku Kristi Lehto ja juhtivspetsialisti Maret Priimaga.

Esimese sammuna andmete jõudmisel SAsse toimub andmete anonüümimise protsess ja andmete kasutajad töötavad anonüümitud andmetega, milles isikud ei ole tuvastatavad. Kasutades isikukoodi ja isikukoodi puudumisel ees- ja perekonnanime ning sugu ja sünniaega, on isikute eristamiseks loodud kindlate reeglite alusel anonüümne kood SA andmeloa osakonna töötaja poolt. Algoritm on loodud nii, et ühele isikukoodile vastab üks anonüümitud kood ja kui Eesti isikukood on puudu, siis on unikaalne anonüümitud kood loodud soo, sünniaja, ees- ja perekonnanime järgi. Pärast anonüümse koodi moodustamist kustutakse andmestikust isikukood ning ees- ja perekonnanimi. Kirjeldatud koodi abil on võimalik erinevate registrite andmed isiku tasandil omavahel ühendada. Antud töös on uuritud ainult neid isikuid, kellel on Eesti isikukood, sest RRi kuuluvad ainult Eesti isikukoodiga inimesed.

Eestis on isikukoodide väljastajaks RR. Erinevatel põhjustel, näiteks lapsendamine, isikukoodi parandamine, topeltisikukoodi olemasolu, on võimalik, et mõnele isikule on väljastatud aja jooksul mitu isikukoodi või isik on vahetanud oma isikukoodi. RRis on kõik isikud kehtiva isikukoodiga, kuid alles on jäetud ka seos kehtetu isikukoodiga. Välistamaks olukorda, kus info läheb teistest registritest kaotsi, sest isik on nendes vana koodiga, on kõikides teistes SAs paiknevates registrite andmete koopiates antud magistritöö autori poolt anonüümitud koodid uuendatud ehk vana kood asendatud kehtivaga.

Analüüsis saab lisaks RRile kasutada 10 registri andmeid (tabel 2). Olenevalt registrite eripärast on EHIS ja RR aastavahetuse seisuga, kuid teised registrid kajastavad andmeid

perioodi kohta. Töös on kasutatud 8 registri 2014. aasta andmeid ning TÖRi andmed on perioodi 01.07.-01.12.14 kohta. Kokku loodi magistritöö autori poolt registrite andmete põhjal (va RR) 21 binaarset tunnust, mis näitavad inimese aktiivsust registris vaadeldaval perioodil ja seega võiksid näidata inimese elamist Eestis. Lihtsuse mõttes on edaspidi käsitletud 21 binaarset tunnust kui erinevaid registreid.

Tabel 2. Kasutatavate registrite loetelu ja loodud tunnuste arv

Registri täisnimi	**Registri lühinimi**	**Väljavõtte aeg / periood**	**Loodud tunnuste arv**
Eesti hariduse infosüsteem	EHIS	01.01.15	2
Töötuna ja tööotsijana arvel olevate isikute ning tööturuteenuste osutamise register	EMPIS	01.01.-31.12.14	1
E-toimiku süsteem	e-toimik	01.01.-31.12.14	1
Riiklik elamis- ja töölubade register	ETR	01.01.-31.12.14	1
Ravikindlustuse andmekogu	KIRST	01.01.-31.12.14	6
Kaitseväekohustuslaste register	KVKR	01.01.-31.12.14	1
Liiklusregister	liiklusregister	01.01.-31.12.14	2
Riiklik pensionikindlustuse register	PKR	01.01.-31.12.14	5
Rahvastikuregister	RR	01.01.15	4
Sotsiaalteenuste ja -toetuste andmeregister	STAR	01.01.-31.12.14	1
Töötamise register	TÖR	01.07.-01.12.14	1
Kokku			25

Suurte andmemahtude ja iga registri erineva ülesehituse tõttu kulus töö autoril koondandmestiku moodustamiseks väga palju aega. Esiteks vaadati iga registri puhul üle, kas ainult registris olemine võiks viidata Eestis elamisele või leidus andmetes lisatunnuseid, mille abil sai välja jätta mingi hulga inimesi, kelle puhul antud registrisse kuulumine isiku residentsust ei kinnita. Näiteks Ravikindluse andmekogu andmetest, mis näitasid, kas inimesel on ravikindlustus, jäeti välja inimesed, kellel olid järgmised kindlustusliigid: isik kuni 19- aastaseks saamiseni, välismaa üliõpilane, Eesti pensionär teises EL liikmesriigis ja EL liikmesriigis elav pereliige, sest need kindlustusliigid ei näita

kindlalt Eestis elamist või viitavad hoopis välismaal elamisele. Loodud tunnuste kirjeldused on toodud lisas 1.

Teiseks kontrolliti ka esmast andmete kvaliteeti. Kuna kõikides registrites on isiku kohta kirjas ka ajaperiood või moment, millal temaga seotud sündmus on toimunud, siis kontrolliti üle, et analüüsis kasutatakse ainult 2014 aasta jooksul toimunud sündmusi.

Kuna üks isik saab olla mitu korda ühes registris, näiteks võib omada poole aasta jooksul mitut töökohta, siis antud töö jaoks on võrdsustatud ühe- ja mitmekordne esinemine registris.

Kuna töö eemärk on uurida ainult RRi ülekaetust, siis otsiti teistest registritest andmeid ainult nende isikute kohta, kes olid RRi aktiivses väljavõttes.

1.4. Ülevaade andmetest

Järgnevalt ülevaade sellest, kui palju RRis olevaid inimesi teistes registrites esineb. Tabelis 3 on toodud kokkuvõtlik loetelu registri andmete põhjal loodud tunnustest ja nende allikatest. Täpsemad kirjeldused tunnuste sisu kohta on toodud lisas 1. Kõige rohkem inimesi on kajastatud Ravikindlustuse andmekogus. Rohkem kui 1,1 miljonile inimesele on koostatud Eestis raviarve, st inimene on külastanud Eestis arsti 2014. aastal. Natuke üle miljonile inimesele on määratud ravikindlustus Eestis. Tegelikult on neid rohkem, sest välja on jäetud mõned kindlustusliigid, mis kindlalt ei näita Eestis elamist (lisa 1). Eesti apteekidest on digiretsepti välja ostnud 0,8 miljonit inimest. Järgmine suurema infoga register on TÖR, kus on kajastatud üle 0,6 miljoni inimese, kes töötavad Eestis.

Tabel 3. RRi inimeste esinemine teistes registrites

Tunnuse nimi	Allikas	Arvukus
etr_on	ETR	10 726
ehis_on	EHIS	289 776
ehis_peda_on	EHIS	35 611
tor_on	TÖR	657 936
star_on	STAR	171 358
kvkr_on	KVKR	2 949
lr_jl_on	liiklusregister	62 663
lr_om_on	liiklusregister	172 901
et_on	e-toimik	165 631
tk_on	EMPIS	82 504
ska_inst_on	PKR	2 795
ska_pension_on	PKR	411 321
pere_toet_on	PKR	431 538
vanemah_on	PKR	54 211
sots_toet_on	PKR	150 322
hambaravi_on	KIRST	91 906
digiretsept_on	KIRST	833 363
raviarve_on	KIRST	1 103 182
lapsvabastus_on	KIRST	9 869
toovoimetus_on	KIRST	188 004
kindlustus_on	KIRST	1 028 929

Lihtne oleks väita, et kui inimene esineb lisaks RRile veel mõnes registris, siis on ta Eesti alaline elanik. Kahjuks see nii ei ole, sest välismaal elav Eesti kodanik võib käia Eestis juhiluba vahetamas või külastada arsti. Seega on residentsuse eeskirja koostamise mõte lähtuda registrites esinemiste koosmõjust.

Kõige raskem on residentsuse analüüsis kindlaks teha kaht erandlikult käituvat inimrühma. Esiteks leidub Eestis alalisi elanikke, kes ei kuulu ühtegi teise registrisse peale RRi. Selline olukord leiti pärast esmast eelmise loenduse andmete analüüsimist: „Püsielanikena loendatud Eesti elanike seas on üle 2% isikuid, keda ei näita ükski register, neile lisandub ligi 4% inimesi, kes kajastuvad vaid ühes registris. Pole võimalik lähtuda eeskirjast, mis kõik niisugused inimesed välismaalasteks liigitaks. Selgub, et ka RRis olevate Eesti

elanike seas on üsna palju inimesi, keda ükski teine register ei kajasta või kes on ühes registris, kusjuures suur osa neist on püsielanikena loendatud ja pole vähimatki alust kahtlustada neid selles, et nad Eestis ei ela.“ [7] Antud isikuid tuleks kontrollida, muid andmeid kasutades, näiteks, millal nad viimati muutsid oma andmeid RRis.

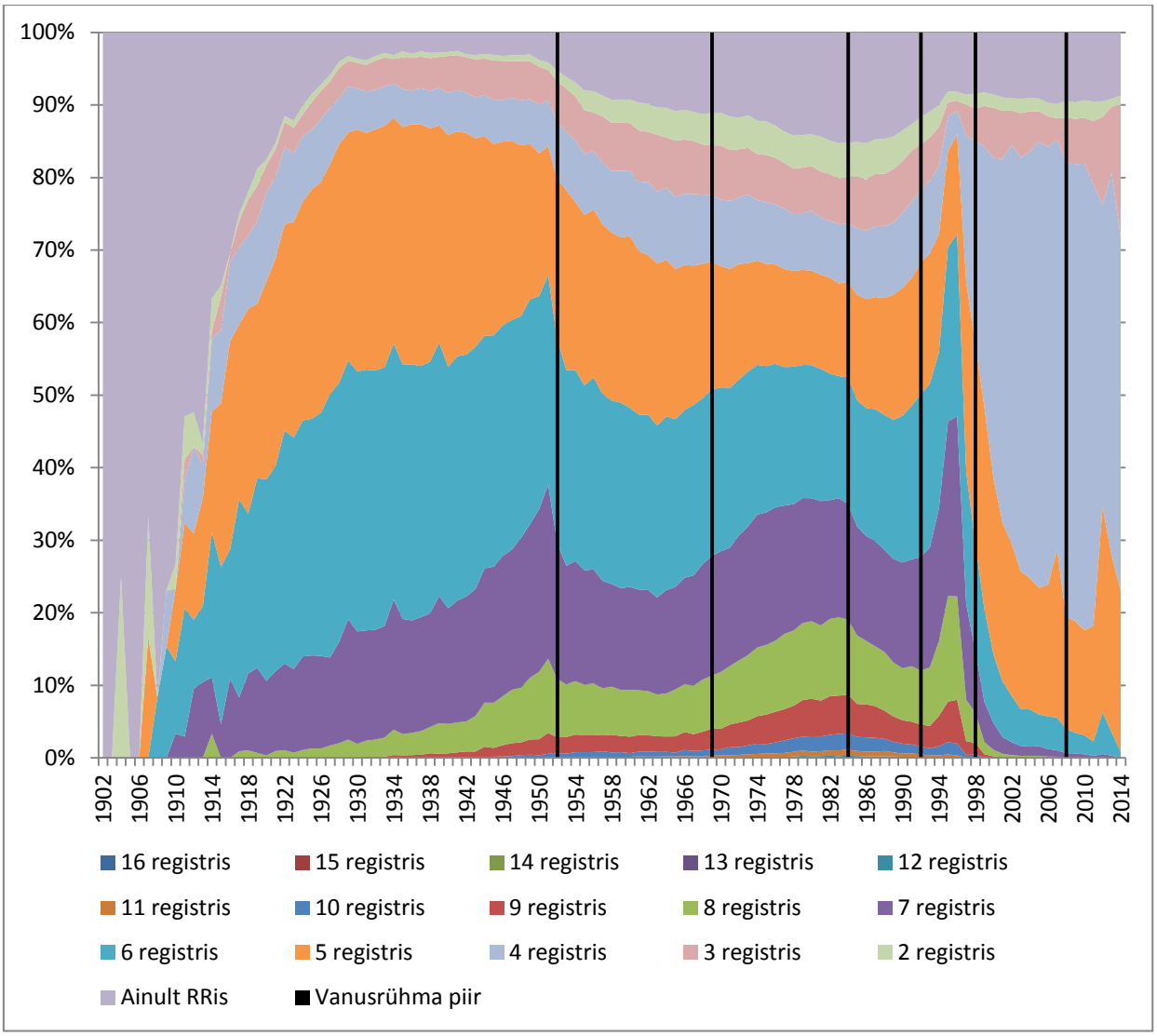
Teine variant on see, et inimene on küll Eestist lahkunud ja ta ei ela enam alaliselt siin, kuid käib aegajalt siiski Eestis, et külastada arsti, osta välja digiretsepti või vahetada juhiluba. Kuid võib arvata, et too Eestist lahkunu ei tööta enam Eestis ning seega ei ole esindatud TÖRis. Selle olukorra peaks suutma lahendada loodud residentsuse eeskiri.

Loomaks vajalikku residentsuse eeskirja on töö autor valinud statistiliseks meetodiks logistilise regressiooni.

1.5. Soo- ja vanusrühmade valimine

Kuna registrites esinemise aktiivsus on sõltuv vanusest ja mõnel juhul ka soost, siis on mõistlik analüüsida erinevaid vanus- ja soorühmi eraldi. Lisas 2 on toodud joondiagrammid, mis näitavad inimeste kuulumist registritesse. Lisas 2 oleval joonisel 1 on näha, et EHISesse kuuluvad põhiliselt noored ning TÖRi tööealised. Joonisel 2 on näha, et peretoetusega on seotud nii laps kui vanem.

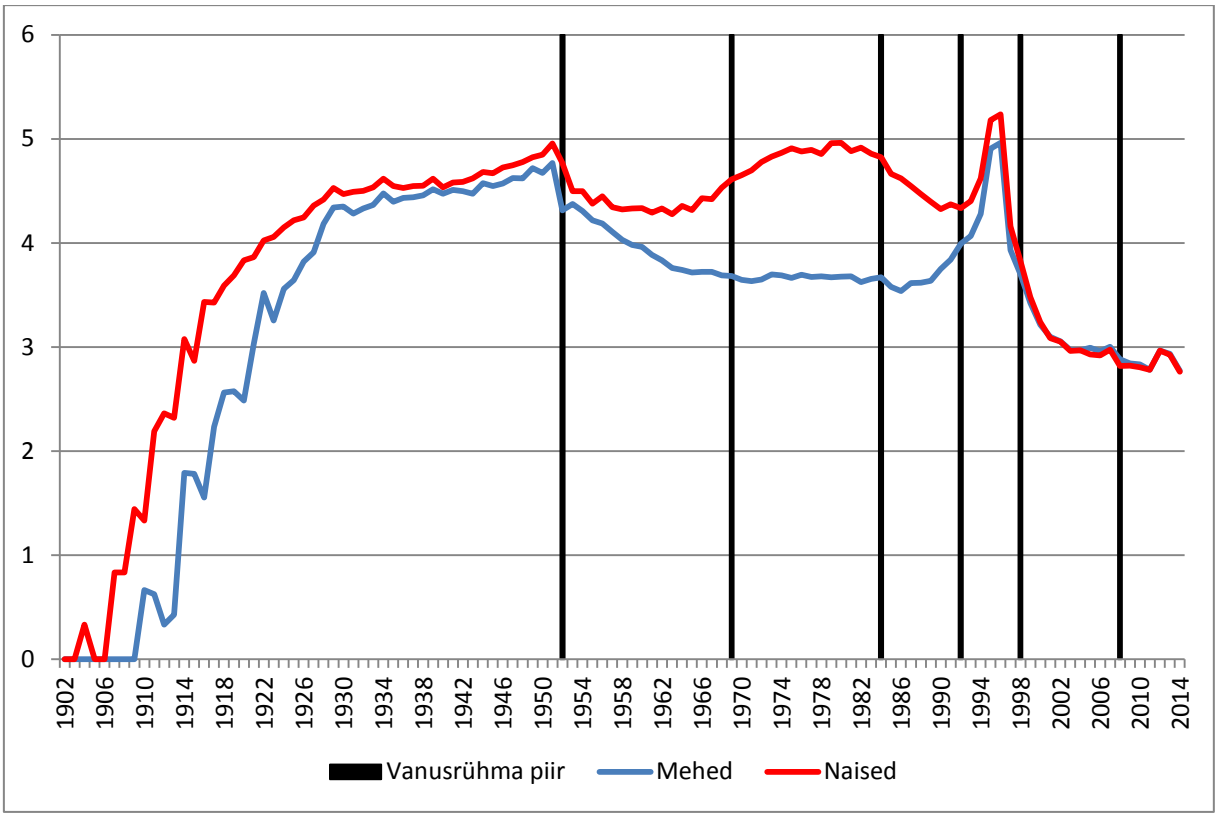
Otsustamaks, milliseid rühmi eraldi uurida, on autor analüüsinud kaht järgmist joonist, millest esimesel on toodud osakaal, mitu protsenti antud aastal sündinutest registritesse kuulub ja teisel meeste ja naiste keskmine registrites esinemiste arv. Joonistele (ka lisas 2) on lisatud mustad jooned, mis tähistavad autori poolt valitud vanusrühmade piire.



Joonis 1. Isikute registrites esinemiste osakaalud sünniaastati

Jooniselt 1 on näha, et igas vanuses leidub inimesi, kes ei kuulu ühessegi teise registrisse, va RR. Joonise järgi tundub, et üle 100-aastaseid esineb teistes registrites vähe, kuid tuleb meeles pidada, et selles vanuserühmas on inimeste aktiivsus on väike. Antud joonisel on info kõikide RRi isikute kohta ja ei ole eraldi vaadeldud, kas nad on alapeatükis 1.2 määratud rühmade järgi residendid, mitteresidendid või uuritavad.

Vaadates joonist 2 on näha, et tööealised inimesed on registrites soo järgi erinevalt esindatud ja seetõttu on neid mõttekas analüüsida eraldi.



Joonis 2. Meeste ja naiste keskmine registrites esinemiste arv sünniaastati

Otstarbekas on moodustada kokku kümme rühma:

1. 0 – 6 aastased
2. 7 – 16 aastased
3. 17 – 22 aastased
4. 23 – 30 aastased mehed
5. 23 – 30 aastased naised
6. 31 – 45 aastased mehed
7. 31 – 45 aastased naised
8. 46 – 62 aastased mehed
9. 46 – 62 aastased naised
10. vähemalt 63 aastased

2. METOODIKA

Järgneva peatüki esimesed kaks alapunkti on koostatud ainult allikate [8] ja [9] põhjal. Kolmandas alapunktides on lisaks eelnevalt toodud allikatele kasutatud ka [10] ja [11]. Viimane alapunkti on allika [9] põhjal.

2.1. Logistilise regressiooni mudel

Logistilise regressioonanalüüsi (nim. ka logistiline mudel või logit-mudel) põhjal saab prognoosida inimese residendiks olemise tõenäosust ja selle muutumist sõltuvalt registrites olemisest. Antud töös ei ole kasutatud lineaarset regressioonanalüüsi, sest see ei garanteeri prognooside jäämist mõistlikku vahemikku nullist üheni. Seevastu logistilise regressiooni abil leitud tõenäosuste hinnangud jäävad alati 0 ja 1 vahele (näidatud järgmisel lehel).

Edaspidi on kasutatud järgnevaid tähistusi:

n – sõltumatute vaatluste (isikute) arv andmestikus ($i = 1, \dots, n$);

y_i – rühmitav tunnus ($0 =$ ei ole resident, $1 =$ resident);

$p_i = P(y_i = 1)$ - residendiks olemise tõenäosus;

k – argumenttunnuste arv;

$\mathbf{x}_i = [1 \quad x_{i1} \quad \dots \quad x_{ik}]^T$ – argumenttunnuste vektor, kus 1 on vabaliikme jaoks;

$\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \dots \quad \beta_k]$ – hinnatavate parameetrite vektor.

Logistilise regressiooni mudel on esitatav kujul

$$logit(p_i) = \ln\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \boldsymbol{\beta}\mathbf{x}_i. \qquad [1]$$

Logistilise regressiooni võrrandi parameetrite tõlgendamine lähtub sellest, et suhe $\frac{p_i}{1-p_i}$ kujutab residendiks olemise **šanssi** – näitab, kui mitu korda tõenäolisem on, et inimene kuulub Eesti alaliste elanike hulka võrreldes sellega, et ta on Eestist lahkunud.

Logistilise regressiooni kordajate $\beta_1, \beta_2, \beta_3, \ldots$ eksponendid e^{β_k} näitavad, mitu korda muutub residendiks olemise šanss *k*-nda argumendi muutumisel ühe ühiku võrra. See tuleneb eelpool toodud logistilise regressiooni võrrandist, sest (lihtsuse mõttes on vaadeldud juhtu $k = 1$)

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_{i1}}.$$

Muutes argumenti x_{i1} ühe ühiku võrra, siis muutub šanss e^{β_1} korda:

$$e^{\beta_0+\beta_1(x_{i1}+1)} = e^{\beta_0}e^{\beta_1 x_{i1}}e^{\beta_1} = e^{\beta_0+\beta_1 x_{i1}}e^{\beta_1} = e^{\beta_1}\frac{p_i}{1-p_i}.$$

Seega kujutavad kordajate eksponendid **šansside suhet**. Positiivse regressioonikordaja korral šansside suhe suureneb ehk kui isik esineb x_1 määravas registris, siis seda tõenäolisem on, et inimene on Eesti alaline elanik võrreldes sellega, et ta ei ela enam Eestis. Negatiivse regressioonkordaja korral šansside suhe väheneb.

Järgnevalt on avaldatud, kuidas leida tõenäosust, kas inimene on resident või mitte:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} + \cdots + \beta_k x_{ik})} = \frac{\exp(\mathbf{\beta x}_i)}{1 + \exp(\mathbf{\beta x}_i)}. \quad [2a]$$

Jagades nii murru lugeja kui nimetaja läbi murru lugejaga lihtsustub tõenäosuse valem järgmiselt:

$$p_i = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_{i1} - \beta_2 x_{i1} - \cdots - \beta_k x_{ik})} = \frac{1}{1 + \exp(-\mathbf{\beta x}_i)}. \quad [2b]$$

Saadud võrrandist näeme, et sõltumata hinnatavate parameetrite ja argumenttunnuste väärtustest jäävad arvutatavad tõenäosused alati 0 ja 1 vahele.

2.2. Parameetrite hindamine

Logistilises regressioonis võib parameetrite hindamiseks kasutada kolme erinevat viisi: vähimruutude meetodit, kaalutud vähimruutude meetodit ja suurima tõepära hinnangut. Kõik kolm on kasutatavad rühmitatud andmete korral, kuid rühmitamata andmetele sobib ainult viimane. Kuna antud töö praktilises osas on kasutatud rühmitamata andmeid, siis on vaadeldud ainult suurima tõepära hinnangut.

Kõigepealt moodustatakse tõepära funktsioon $L = P(y_1, y_2, \dots, y_n)$. Kuna on eeldatud, et vaatlused on sõltumatud, siis tõepära funktsioon avaldub kujul:

$$L = P(y_1)P(y_2) \dots P(y_n) = \prod_{i=1}^{n} P(y_i). \qquad [3]$$

Eelnevalt toodud tähistuste järgi $P(y_i = 1) = p_i$ ja $P(y_i = 0) = 1 - p_i$. Järelikult

$$P(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}.$$

Asendades võrrandisse [3] saadud tulemuse ja lihtsustades:

$$L = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i}\right)^{y_i} (1 - p_i). \qquad [4]$$

Logaritmides võrrandi [4] mõlemaid pooli:

$$\ln L = \sum_i y_i \ln\left(\frac{p_i}{1 - p_i}\right) + \sum_i \ln(1 - p_i).$$

Kasutades võrrandeid [1] ja [2a] saab viimase võrduse kirjutada kujul:

$$\ln L = \sum_i \boldsymbol{\beta}\mathbf{x}_i y_i - \sum_i \ln\left(1 + e^{\boldsymbol{\beta}\mathbf{x}_i}\right).$$

Järgnevalt on vajalik leida $\boldsymbol{\beta}$ nii, et viimane avaldis saaks võimalikult suure väärtuse. Selle jaoks leitakse antud võrrandist tuletis $\boldsymbol{\beta}$ järgi, võrdsustatakse nulliga ja avaldatakse $\boldsymbol{\beta}$:

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_i \mathbf{x}_i y_i - \sum_i \mathbf{x}_i \left(1 + e^{-\boldsymbol{\beta}\mathbf{x}_i}\right)^{-1} = \sum_i \mathbf{x}_i y_i - \sum_i \mathbf{x}_i \hat{y}_i = 0, \qquad [5]$$

kus

$$\hat{y}_i = \frac{1}{1 + e^{-\boldsymbol{\beta}\mathbf{x}_i}}$$

on prognoositud tõenäosus väärtusele y antud argumentväärtuste juures. Kuna $\mathbf{x}_i$ on vektor, siis võrrand [5] on tegelikult $k + 1$ võrrandi süsteem, kus iga võrrand sisaldab tundmatutena kõiki vektori β elemente. Võrrandisüsteemi [5] lahendamiseks ehk $\boldsymbol{\beta}$ leidmiseks kasutatakse kõige laialdasemalt Newton-Raphsoni iteratsioonimeetodit, mis on esitatav kujul:

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j - \mathbf{I}^{-1}(\boldsymbol{\beta}_j)\mathbf{U}(\boldsymbol{\beta}_j), \qquad [6]$$

kus $\mathbf{U}(\boldsymbol{\beta})$ on skoorifunktsioon, mis on leitav kui esimene tuletis tõepärafunktsioonist $\boldsymbol{\beta}$ järgi**:**

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_i \mathbf{x}_i y_i - \sum_i \mathbf{x}_i \hat{y}_i$$

ja $\mathbf{I}^{-1}(\boldsymbol{\beta})$ on pöördmaatriks Hesse maatriksist, mis on leitav kui tõepärafunktsioonist teine tuletis $\boldsymbol{\beta}$ järgi**:**

$$\mathbf{I}(\boldsymbol{\beta}) = \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\sum_i \mathbf{x}_i \mathbf{x}_i{}^{\mathrm{T}} \hat{y}_i (1 - \hat{y}_i).$$

Praktikas on vajalikud algväärtused $\boldsymbol{\beta}_0$. Antud töös on kasutatud protseduuri, mis vaikimisi valib vektori algväärtuseks 0. Algväärtused asendatakse võrduse [6] paremale poole ning tulemuseks on $\boldsymbol{\beta}_1$. Saadud tulemuse asendatakse uuesti võrdusesse [6] ja saadakse järgmine iteratsioonisammu tulemus ehk $\boldsymbol{\beta}_2$. Antud protsess kehtib kuni kahe järjestikuse $\boldsymbol{\beta}_j$ ja $\boldsymbol{\beta}_{j+1}$ vahe on väiksem kui teatud etteantud kriteerium.

SASi vaikimisi koondumise kriteeriumid:

1. Kui $\boldsymbol{\beta}_j \leq |0{,}1|$, siis $\left|\boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j\right| < 0{,}0001$;
2. Kui $\boldsymbol{\beta}_j > |0{,}1|$, siis $\left|\frac{\boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j}{\boldsymbol{\beta}_j}\right| < 0{,}0001$.

2.3. Mudeli headuse näitajad

Kontrollimaks mudeli headust otsitakse vastust kahele küsimusele: kas mudel prognoosib hästi ehk kas mudel suudab eristada residendid mitteresidentidest ja kas mudel sobib andmetega? Esimesele küsimusele annavad vastused erinevad statistikud. Teisele küsimusele on raskem vastata, sest viimasel ajal on simulatsioonide käigus selgunud, et Hosmer-Lemeshovi test ei tööta hästi.

Esmalt on vaadatud erinevaid **prognoosimise headuse** näitajaid. On olemas väga mitmeid viise, kuidas arvutada headust iseloomustavaid statistikud. Tavaliselt on nende väärtused 0 ja 1 vahel ning mida suurem on väärtus, seda paremini mudel prognoosib. Antud töös on kasutatud kaht erinevat statistikut: Cox-Snelli R^2 ja Tjur'i *D*.

Cox-Snelli R^2 on arvutatav järgmiselt

$$R^2_{C\&S} = 1 - \left(\frac{L_O}{L_M}\right)^{\frac{2}{n}},$$

kus L_O on tõepärafunktsioon juhul, kui mudelis ei ole ühtegi hinnatavat parameetrit ehk ainult vabaliikmega mudel, L_M hinnatud mudeli tõepärafunktsioon ja n valimimaht. Toodud statistikut arvutab SASis kasutatav protseduur LOGISTIC.

Tjur'i D leitakse järgmiselt

$$D = |E[P(y = 1)] - E[P(y = 0)]|$$

ehk leitakse absoluutväärtus kahe sündmuse toimumise tõenäosuste keskväärtuse vahel. Tjur'i D statistikut nimetatakse ka diskrimineerimise konstandiks ja see näitab erinevust kahte rühma kuulumise tõenäosuste vahel. [10] ja [11]

Statistik D on ligilähedane ühele siis, kui kindlatele residentidele keskmine prognoositud tõenäosus on ligilähedale ühele ja kindlatele mitteresidentidele keskmine prognoositud tõenäosus on ligilähedane nullile. Kui statistik D on lähedane nullile, siis ei saa väita, et oleks erinevus residentide ja mitteresidentide prognoositud tõenäosuste vahel.

Järgmisena otsitakse vastust küsimusele: kas mudel sobib andmetega hästi? Sellele küsimusele vastamiseks on kasutatud Hosmer ja Lemeshow (HL) testi. Antud testi puhul tuleb meeles pidada, et see otsib vastuseid küsimustele, kas mudel tuleks muuta keerukamaks, näiteks lisades koosmõjusid või kasutada logit funktsiooni asemel mõnd muud funktsiooni, näiteks loglog. Viimasel paaril aastal on arutletud selle üle, et antud testi

tulemused ei ole rühmitamata andmete puhul usaldusväärsed. Antud testi jaoks rühmitamata andmed grupeeritakse, kuid ei ole selge mitmesse gruppi peaks andmed jagama, et saada tõene vastus. On viidud läbi simulatsioone, kus programmides vaikimisi määratud 10 grupi puhul test annab negatiivse vastuse ehk mudelit tuleb muuta paremaks/keerukamaks, kuid 9 või 11 grupi korral annab test positiivse vastuse [10].

2.4. Optimaalne lävend

Olles leidnud mudeli ja saanud teada, et see on piisavalt hea, on vaja veel otsust, alates millisest hinnatud tõenäosusest (lävendist) on tegemist residendiga. Selle vastuse võib leida kasutades Youden'i *J* statistikut. Enne selle defineerimist on vaja teada, mida tähendavad tundlikkus ehk sensitiivsus ja spetsiifilisus.

Kuna uuritav tunnus on binaarne väärtustega – resident ja mitteresident, siis on võimalik isikute arvud jaotada 2×2-tabelisse kujul:

Tabel 4. Tegelike ja prognoositud tulemuste sagedustabel

	$\widehat{\boldsymbol{y}} = \mathbf{0}$	$\widehat{\boldsymbol{y}} = \mathbf{1}$	**Kokku**
$\boldsymbol{y} = \mathbf{0}$	*TN*	*FP*	*TP+FP*
$\boldsymbol{y} = \mathbf{1}$	*FN*	*TP*	*FN+TP*
Kokku	*TN+FN*	*FP+TP*	*TN+FP+FN+TP*

Selgitused:

- *TN* (ingl. *true negative*) – **tõeselt negatiivsete** juhtude arv näitab mitu mitteresidenti prognoositi mitteresidendiks;
- *FP* (ingl. *false positive*) – **valepositiivsete** juhtude arv näitab mitu mitteresidenti prognoositi residentideks;
- *FN* (ingl. *false negative*) – **valenegatiivsete** juhtude arv näitab mitu residenti prognoositi mitteresidentideks;
- *TP* (ingl. *true positive*) – **tõeselt positiivsete** juhtude arv näitab mitu residenti prognoositi residendiks.

Tundlikkus ehk **sensitiivsus** näitab, kui suure osa residentidest ennustab kasutatud mudel õigesti:

$$T(p_j) = \frac{TP}{(TP + FN)}.$$

Spetsiifilisus näitab, kui suure osa mitteresidentidest ennustab kasutatud mudel õigesti:

$$S(p_j) = \frac{TN}{(TN + FP)}.$$

Leidmaks optimaalset lävendit, millisest prognoositud tõenäosusest on tegemist residendiga arvutatakse igale arvutatud tõenäosusele tundlikkus ja spetsiifilisus. Seejärel leitakse nende kahe summa maksimum ehk Youden'i statistik:

$$J = \max_j \{T(p_j) + S(p_j) + 1\}$$

Tõenäosuse väärtus p_j, mis vastab leitud maksimumile, loetakse residentide ja mitteresidentide optimaalseks lävendiks. Seega residentideks loetakse need isikud, kellele prognoositud tõenäosus on suurem kui J ning mitteresidentideks need, kellele prognoositud tõenäousus on väiksem kui J.

3. EESTI ALALISTE ELANIKE MÄÄRATLEMINE

Esimene registripõhine prooviloendus viiakse läbi momendiga 31.12.2015. Üheks eesmärgiks on testida rahvastiku üldkogumi metoodikat ehk antud magistritöö tulemused lähevad kasutusse SA reaalses tegevuses. Eesmärk on moodustada kõik loendustunnused Eesti residentidele. Kui moodustada loendustunnused ka tegelikele mitteresidentidele, kelle kohta on tõenäoliselt raske registritest vajalikku teavet leida, suurendab see nende tunnuste kvaliteediprobleeme. Teine vajadus on ühtlustada loendus- ja rahvastikustatistika. Varasemalt on rahvastikustatistikat täiendatud loenduste andmetega.

Selles peatükis on toodud tulemused praktilisest tööst. Esimeses alapunktis on ülevaade autori tulemustest regressioonanalüüsiga. Järgmises alapunktis on toodud ülevaade, milliseid eeldusi ja meetodeid kasutasid tudengid aine Andmetöötlusmeetodid raames ja võrdlus autori tulemustega. Kolmandas alapunktis on toodud statistiliste meetoditega saadud tulemuste ja SA poolt avaldatud rahvaarvu võrdlus. Viimases alapunktis on toodud järeldused ja ettepanekud järgnevaks tööks.

3.1. Regressioonanalüüs

Mudelid on hinnatud kasutades *SAS*i protseduuri *LOGISTIC* sammuviisilist lähenemist igale vanus- ja/või soorühmale eraldi. Antud meetodi puhul lülitatakse mudelisse järjest olulisi tunnuseid. Vajadusel eemaldatakse varem oluline tunnus, kui ilmneb, et teiste oluliste tunnuste lisamisel varem oluline tunnus muutub statistiliselt mitteoluliseks. Mudelite hindamist alustati 23 tunnuse põhjal:

- 21 binaarset tunnust ehk kuuluvus vastavasse registrisse;
- 2 pidevat tunnust: vanus ja binaarsete tunnuste summa (edaspidi reg).

Saadud **parameetrite hinnangud** on toodud tabelis 5. Vabaliige tähistab teooria osas toodud parameetrit β_0 ning järgmised read vastavalt parameetreid β_1, β_2, β_3, ... (negatiivse väärtuse korral on taust oranži värvi ja positiivse väärtuse korral roheline). Mitte üheski mudelis ei osutunud statistiliselt oluliseks infoks pedagoogina töötamine ja töölt puudumine sünnitamise või lapsendamise tõttu. Samas on seda infot kasutatud tunnuses reg ja seetõttu ei tohi siiski nende registrite info kasutamisest loobuda.

Negatiivsete kordajate esinemise põhjuseks on tunnuste omavaheline sõltuvus – ühe argumenttunnuse suurt mõju kompenseerib teise tunnuse negatiivse kordajaga väärtus.

Kuna tunnus reg sisaldab kõiki ülejäänuid, siis esineb selle ja ülejäänud tunnuste vahel sõltuvus. Lisaks on sõltuvuse allikaid teisigi.

Tabel 5. Mudelite parameetrite hinnangud

Argumenttunnus	0 - 6	7 - 16	17 - 22	23 - 30 mehed	23 - 30 naised	31 - 45 mehed	31 - 45 naised	46 - 62 mehed	46 - 62 naised	üle 63 (ka)
vabaliige	3,94	-1,16	-9,20	3,22	6,06	-0,78	-2,29	0,84	-1,82	1,17
reg	1,96		2,51	1,23	2,24	1,15	1,18	1,18	1,07	0,96
vanus	-1,02	-0,07	0,43	-0,13	-0,26			-0,03		-0,05
etr_on		3,77	1,95	1,26	3,02	1,20	3,24	0,52	2,44	4,20
ehis_on	1,58	5,41	0,87	1,60		0,88				
ehis_peda_on										
tor_on			-0,64	0,39	0,81	0,73	1,55	1,40	2,44	1,57
star_on	-1,41			1,83		1,32	1,70	1,84	1,69	3,00
kvkr_on			-1,68							
lr_jl_on			1,07	1,82	1,28	2,19	2,08	1,73	3,04	2,89
lr_om_on			-1,65	-0,35	-1,86			-0,30		
et_on			-0,66		-1,67					
tk_on							0,89		1,12	
ska_inst_on										-2,32
sots_toet_on							2,62	-0,89		2,29
ska_pens_on	-4,28		-2,83					1,03	1,15	5,67
pere_toet_on	2,48	3,56	1,62		1,04	0,88	3,12	1,65	2,63	
vanemah_on					-2,60		-1,75			
hambaravi_on							-1,70			
digiretsept_on	-2,58		-1,30		-1,42					
raviarve_on		1,82			-0,96				0,44	1,11
lapsvabastus_on										
toovoimetus_on			-2,72	-1,17	-1,20	-0,79	-0,90	-1,45	-1,36	
kindlustus_on				2,14		2,33	2,18	2,16	2,54	

Tulemuste selgitamiseks on vaadatud täpsemalt rühma 17–22-aastased. Silma jääb, et argumenttunnus, mis vastab ajateenistuses või asendusteenistuses olemisele, on saanud negatiivse kordaja. Aga teoreetiliselt peaks just nemad kuuluma Eesti alaliste elanike hulka, sest nad on kindlalt Eestis. Näitamaks pidevate tunnuste olulisust on toodud näide, kus on arvutatud kahe 19-aastase isiku residendiks olemise tõenäosus, kellest üks on ajateenistuses ja teine mitte ning kumbki neist ei kuulu ühessegi ülejäänud registritest.

Arvutustes on kasutatud valemit [2b]:

$$p_i = \frac{1}{1 + \exp\left(-\beta_0 - \beta_1 x_{i1} - \beta_2 x_{i1} - \cdots - \beta_k x_{ik}\right)} = \frac{1}{1 + \exp(-\mathbf{\beta x}_i)}. \quad [2b]$$

Isik, kes ei ole ajateenistuses:

$$p_1 = \frac{1}{1 + \exp\left(9{,}2 - 0{,}43 * vanus\right)} = \frac{1}{1 + \exp\left(9{,}2 - 0{,}43 * 19\right)} \approx 0{,}26.$$

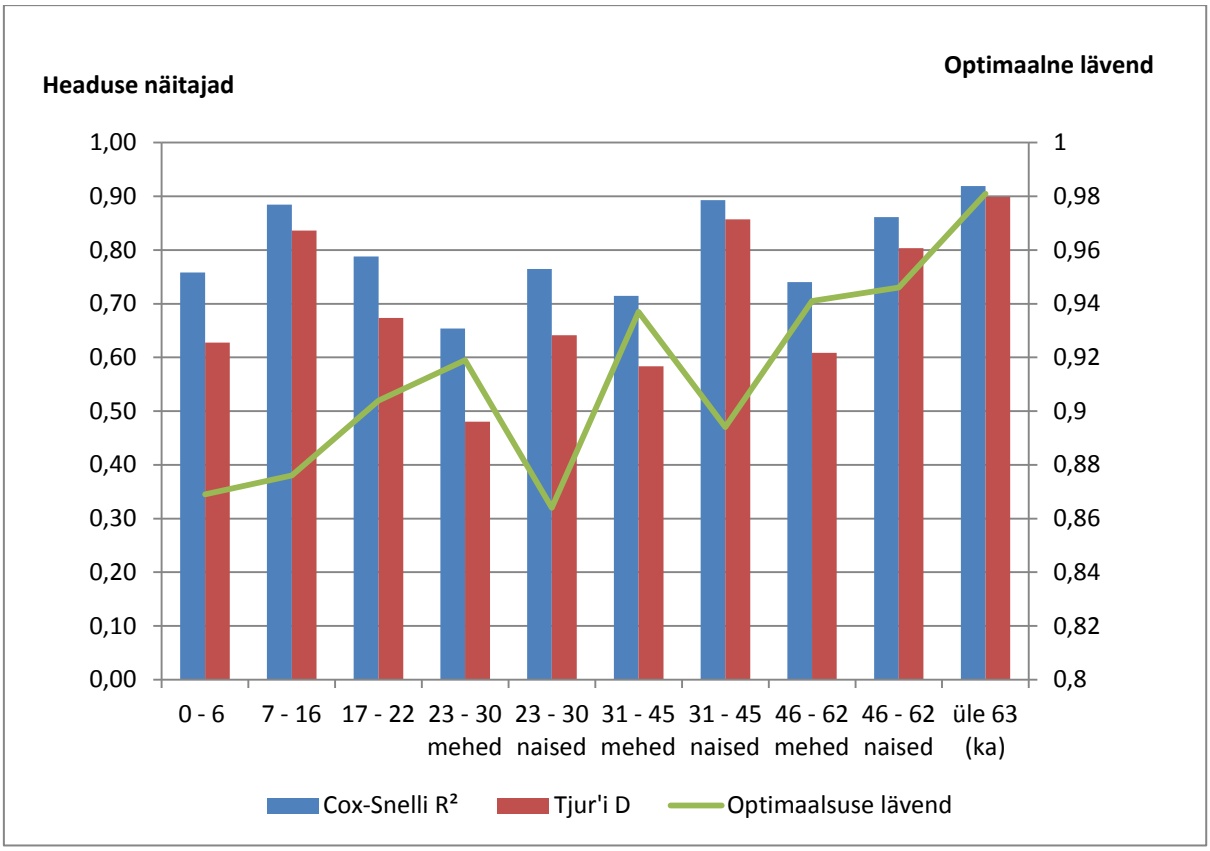
Isik, kes on ajateenistuses:

$$p_2 = \frac{1}{1 + \exp\left(9{,}2 - 2{,}51 * reg - 0{,}43 * vanus + 1{,}68 * kvkr\_on\right)} =$$

$$= \frac{1}{1 + \exp\left(9{,}2 - 2{,}51 * 1 - 0{,}43 * 19 + 1{,}68 * 1\right)} \approx 0{,}45.$$

Seega ajateenistuses oleva isiku tõenäosus olla resident on suurem kui sellel isikul, kes ei ole ajateenijana registris. Kui optimaalse lävendi väärtus on 0,5, siis saavad mõlemad noorukid määratud mitteresidentideks. Kui aga optimaalse lävendi väärtus on 0,4, siis registris olev nooruk määratakse residendiks ja teine jääb mitteresidendiks.

Joonisel 3 on toodud mudelite **headuse näitajad ja optimaalsed lävendid**. Vasakpoolsel teljel on toodud headuse näitajate väärtused ja parempoolsel teljel optimaalse lävendi väärtused. Kõige raskemini eristusid registriandmete põhjal 23–30-aastaste meeste Eestis elamise tõenäosus välismaale lahkunutest ($D = 0{,}48$). Järgnevad 31–45-aastased mehed ja seejärel 46–62-aastased mehed. See on põhjustatud sellest, et antud vanusrühmi on võrreldes teistega kõige vähem registrites kajastatud (joonised 1 ja 2). Kõige paremini eristusid vähemalt 63-aastased, 31–45-aastased naised ja 7–16-aastased. Antud rühmad on vastupidiselt 23–62-aastastele meestele registrites kõige rohkem esindatud.



Joonis 3. Mudelite headuse näitajad ja optimaalsed lävendid

Residendiks kuulumisel kõige madalam lävend on 23–30-aastastel naistel, järgmisena 0–6-aastastel ja 31–45-aastastel naisel (joonis 3). Kõige kõrgem lävend on vähemalt 63-aastastel. Kõige kõrgema ja madalama optimaalse piiri vahe on üle 0,1 ühiku, mis näitab, et vanus- ja soorühmadel vahel on kindlalt erinevused.

Pärast optimaalse lävendi leidmist saab arvutada igale mudelile eraldi tundlikkuse ja spetsiifilisuse (tabel 6). Kõige paremini prognoosib residente vähemalt 63-aastaste mudel – 99,4% residentidest. Teisena prognoosib 7–16-aastastele koostatud mudel 98,4% residentidest. Kõige väiksema osa residentidest prognoosivad mudelid 23–30-aastaste meeste ja 31–45-aastaste meeste kohta, vastavalt 92,2% ja 93,1%. Siin võib samuti põhjuseks pidada, et selles vanuserühmas mehi on raske gruppidesse määrata, sest nende kohta puudub info registrites.

Kõige suurema osa mitteresidentidest prognoosib õigesti 7–16-aastastele noortele ja 31–45 ning 46–62-aastastele naistele koostatud mudelid, vastavalt 99,6%, 99,0% ja 99,0%. Võrreldes teiste vanuserühmadega suudavad kehvemini mitteresidente määrata 23–30-aastastele meestele ja 0–6-aastastele koostatud mudelid, vastavalt 96,3% ja 97,1%.

Tabel 6. Mudelite tundlikkus ja spetsiifilisus

	0-6	7-16	17-22	23-30 mehed	23-30 naised	31-45 mehed	31-45 naised	46-62 mehed	46-62 naised	üle 63 (ka)
Tundlikkus	98,2%	98,4%	96,4%	92,2%	94,9%	93,1%	98,0%	95,9%	98,2%	99,4%
Spetsiifilisus	97,1%	99,6%	98,2%	96,3%	98,2%	97,3%	99,0%	97,3%	99,0%	98,6%

Kokkuvõtlikult võib öelda, et kõige paremini suudab prognoosida 7–16-aastastele koostatud mudel ning kõige keerulisem on olukord 23–30-aastaste meestega. Esimesel juhul on tegemist õppuritega, kes kõik peavad kohustuslikus korras käima üldhariduskoolis ja on seega lisaks muudele registritele kindlasti EHISes kajastatud. Samas 23–30-aastaste meeste puhul on tegemist isikutega, kelle kohta ei saa välja tuua üht kindlat registrit, kus nad võiksid olla või ei ole nad üldse endast registritele märku andnud. Lisaks on 7–16-aastased noored paiksemad kui 23–30-aastased mehed, kellel on suuremad võimalused välismaale elama minna.

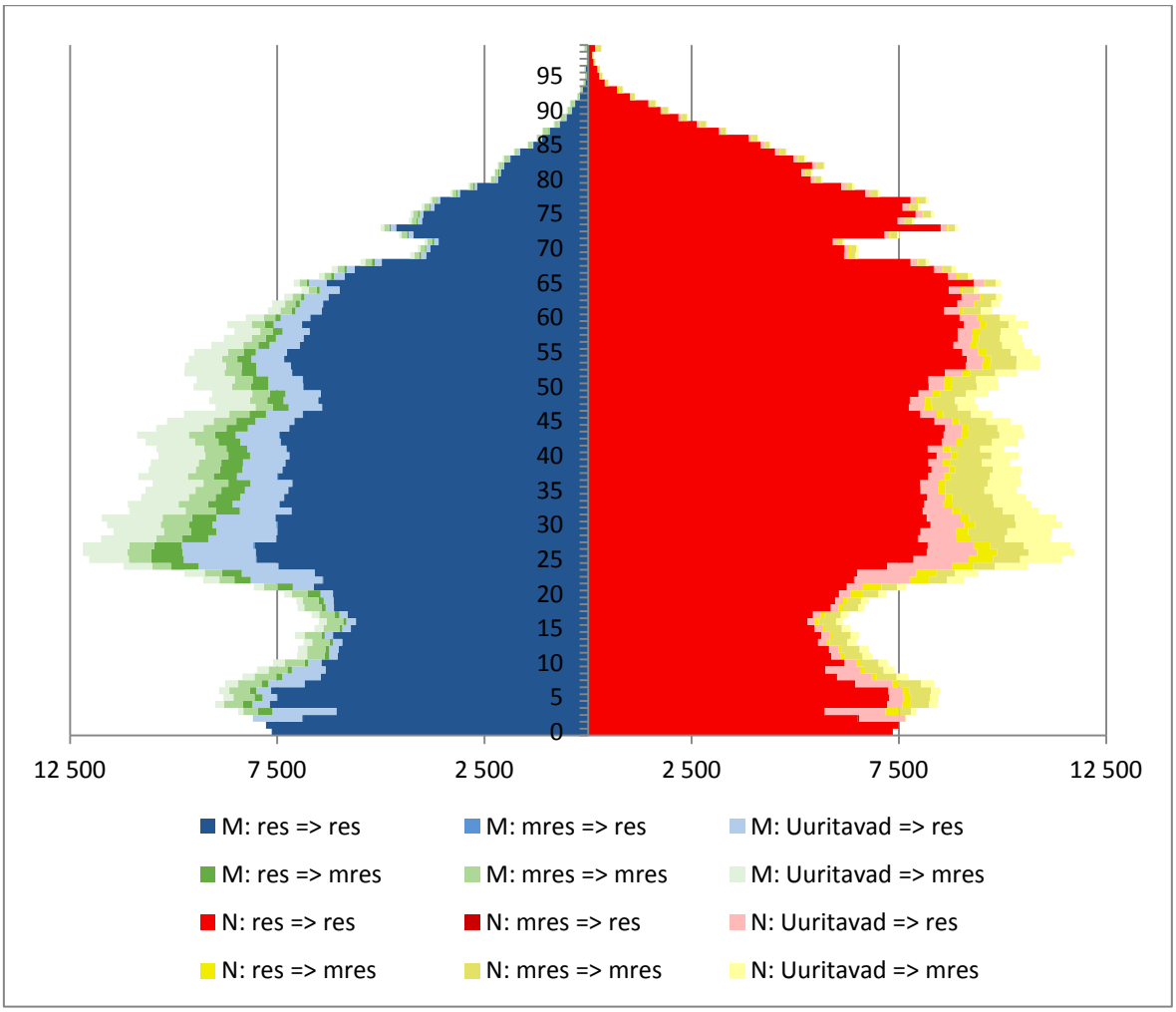
Edaspidi on joonistel ja tabelites kasutatud lühendeid res, mis tähendab residente; mres, mis tähendab mitteresidente ja uuritavad, mis tähendab isikuid, kellele oli vaja residentsus prognoosida. Tähistus „res => mres“ tähistab isikuid, kes olid antud töö alapunktis 1.2 määratud residentideks, kuid mudelite põhjal on järeldatud, et registri andmetele tuginedes käituvad need isikud kui mitteresidendid.

Rahvastikupüramiidis (joonis 4) on eraldi värvidega välja toodud, kas isik prognoositi residendiks või mitte. Naised, kes prognoositi Eestisse elama, on tähistatud punaste toonidega. Naised, kes peaksid prognooside kohaselt Eestist olema lahkunud, on toodud kollastes toonides. Vastavalt meeste puhul on kasutatud siniseid ja rohelisi toone. Peaaegu olematud on rühmad, kus mitteresidendid on prognoositud residentideks nii meeste kui naiste puhul.

Kõige rohkem uurimise all olevatest isikutest prognoositi residentideks 23–62-aastaseid mehi. Selles vanuserühmas oli ka kõige rohkem neid, keda ei suudetud määrata kindlateks residentideks ja mitteresidentideks ehk uuritavate rühm oli kõige suurem. Kõige vähem prognoositi residente aga vähemalt 63-aastaste hulgas, sest selles vanuserühmas oli varasemalt eeldatud peaaegu kõik kindlateks residentideks.

Kõige rohkem kindlaid residente on prognoositud mitteresidentideks 23–30-aastaste meeste seas. Probleemidele antud vanuserühmas viitasid ka headusenäitajad. Antud rühma

jaoks tuleb tulevikus lisaks otsida veel residentsust näitavaid infoallikaid. Lisainfot võivad kindlasti anda töö alguse jaoks puudu jäänud registrid, mis sisaldavad infot isikut tõendavate dokumentide vahetamise kohta Eestis ning vangis ja kriminaalhoolduses olevate isikute kohta. Kui isik on Eestis vangis, siis on ta kindlasti Eesti resident, kuid ta ei pruugi olla esindatud teistes Eesti registrites.



Joonis 4. Rahvastikupüramiid algsete ja prognoositud residentsuse jaotusega

Järgnevalt on võrreldud RRis registreeritud elukohta algse residentsuse määratluse ja prognoositud tulemuste järgi. Mudeleid rakendati kõigile RRis olevatele isikutele, mitte ainult uuritavatele. Tabelist 7 on näha, et mudelite põhjal on rohkem kui 90 000 inimesel RRis tõenäoliselt vale elukoht ja nad ei ela enam Eestis. Samas tuleb meeles pidada, et siin hulgas võib olla inimesi, kes tegelikult siiski elavad Eestis, aga neid ei esine piisavalt teistes registrites, et seda kindlaks teha.

Tabel 7. Prognoositud mitteresidentide elukoha jaotus

	RRi elukoht			Kokku
	Eesti	Puudu	Välismaa	
Uuritavad =>mres	61 824	438	16 829	79 091
	32,7%	0,2%	8,9%	41,9%
mres => mres	0	2 931	71 733	74 664
	0,0%	1,6%	38,0%	39,5%
res => mres	28 657	1 132	5 357	35 146
	15,2%	0,6%	2,8%	18,6%
Kokku	90 481	4 501	93 919	188 901
	47,9%	2,4%	49,7%	100,0%

Mudelite järgi on Eestis elavateks määratud üle 12 000 isiku, kelle registreeritud elukoht on puudu või on välismaa (tabel 8). Siin hulgas võib olla vähesel hulgal ka inimesi, kes on oma elukoha RRis korrektselt registreerinud, kuid lahkunud alles 2014. aastal ja aasta esimeses pooles erinevates registrites esinenud. Samuti võib sellesse gruppi kuuluda inimesi, kes käivad jätkuvalt aegajalt Eestis ja on kajastatud siinsetes registrites, kuigi välismaa elukoht RRis on korrektne. Kui määrata antud isikud Eesti residentideks, siis on vaja neile määrata ka Eestis täpsed elukohad, mis omakorda on aluseks perekondade ja leibkondade määramiseks registripõhises loenduses.

Tabel 8. Prognoositud residentide elukoha jaotus

	RRi elukoht			Kokku
	Eesti	Puudu	Välismaa	
Uuritavad => res	87 250	1 369	4 633	93 252
	6,8%	0,1%	0,4%	7,3%
mres => res	0	313	998	1 311
	0,0%	0,0%	0,1%	0,1%
res => res	1 174 409	1 004	3 982	1 179 395
	92,2%	0,1%	0,3%	92,6%
Kokku	1 261 659	2 686	9 613	1 273 958
	99,0%	0,2%	0,8%	100,0%

3.2. Võrdlus teiste statistiliste meetodite abil leitud lahendustega

Kursuses Andmetöötlusmeetodid analüüsis samu andmeid SA turvalistel töökohtadel Kristi Tüli logistilise ja lineaarse regressiooniga, Maia Arge diskriminantanalüüsiga ning Kaidi Jõgi ja Hindrek Teder klasteranalüüsiga. Antud magistritöös on toodud võrdlus kahe esimese tudengi tööga. Klasteranalüüsis kasutati analüüsimiseks valimeid, sest *SAS Enterprise Guide* ei suutnud töödelda nii suuremahulisi andmeid ja seetõttu võrdlust tehtud ei ole.

Toodud aine raames kasutasid tudengid teistsuguseid eeldusi kui antud magistritöö regressioonanalüüsi osas, mis on kirjeldatud alapeatükkides 1.2 ja 1.5. Tudengid otsustasid kindlateks residentideks valida 2 rühma: tabeli 1 järgi rühmad numbritega 12 ja 13 (antud töös kasutati kolme rühma 12, 13 ja 14). Kindlateks mitteresidentideks valiti samad rühmad 1 ja 2. Lisaks hindasid tudengid teistsuguseid vanusrühmi (autori regressioonanalüüsis oli kokku 10 erinevat vanus ja/või soorühma, vt 1.5):

1. 0 – 6 aastased
2. 7 – 18 aastased
3. 19 – 39 aastased mehed
4. 19 – 39 aastased naised
5. 40 – 62 aastased mehed
6. 40 – 62 aastased naised
7. vähemalt 63 aastased

Edaspidi on tabelites tähistatud antud magistritöö autori tulemused kui logistiline regressioon 1 ja Kristi Tüli saadud tulemused kui logistiline regressioon 2.

Kõige rohkem RRi väljavõttes olevaid isikuid ehk 88,4% määras residentideks diskriminantanalüüs (tabel 9). Kõige vähem ehk 86,4% logistiline regressioon 2. Antud töö autori poolt tehtud analüüs määras residentideks 87,1% RRi isikutest.

Tabel 9. Erinevate meetodite tulemused

	res		mres	
Logistiline regressioon 1	1 273 958	87,1%	188 901	12,9%
Logistiline regressioon 2	1 260 816	86,2%	202 043	13,8%
Lineaarne regressioon	1 263 454	86,4%	199 405	13,6%
Diskriminantanalüüs	1 292 838	88,4%	170 021	11,6%

Kõikide meetoditega osutusid residentideks 1 250 247 inimest ehk 85,5% kõigist RRi väljavõttes olnud isikutest (tabel 10). Mitteresidentideks on mudelid määranud 158 491 inimest ehk 10,8% vaadeldud kogumist. Kõige suurem erinevus ülejäänud lahendustest on diskriminantanalüüsi korral, kus 21 924 inimest määratakse residentideks, aga kõik regressioonmudelid määravad need isikud mitteresidentideks. Kokku on üle 54 000 isiku, kelle vähemalt üks meetod määrab Eesti residendiks.

Tabel 10. Erinevate meetodite tulemuste võrdlus

Logistiline regressioon 1	Logistiline regressioon 2	Lineaarne regressioon				Kokku
		mres		res		
		Diskriminantanalüüs		Diskriminantanalüüs		
		mres	res	mres	res	
mres	mres	158 491	21 924	4 268	3 448	188 131
		10,8%	1,5%	0,3%	0,2%	12,9%
	res	25	330	18	397	770
		0,0%	0,0%	0,0%	0,0%	0,1%
res	mres	3 265	7 218	385	3 044	13 912
		0,2%	0,5%	0,0%	0,2%	1,0%
	res	2 022	6 130	1 547	1 250 347	1 260 046
		0,1%	0,4%	0,1%	85,5%	86,1%
Kokku		163 803	35 602	6 218	1 257 236	1 462 859
		11,2%	2,4%	0,4%	85,9%	100,0%

Kasutatud meetoditest kõige sarnasemalt prognoosivad kaks logistilist regressiooni, erinevad tulemused on saadud üle 14 000 isikul ehk 1% uuritavast kogumist (tabel 11). Tabelist 10 on näha, et antud töö autori tehtud mudelid prognoosivad rohkem inimesi Eesti residentideks võrreldes Kristi Tüli poolt loodud mudelitega. Kõige suurem erinevus on lineaarse regressiooni ja diskriminantanalüüsi puhul, kus erinevalt on prognoositud üle 41 000 isiku ehk peaaegu 3% uuritavast kogumist.

Tabel 11. Erinevate meetodite võrdlus

Meetodid	Isikud	
Logistiline regressioon 1 vs logistiline regressioon 2	14 682	1,0%
Logistiline regressioon 2 vs lineaarne regressioon	19 652	1,3%
Logistiline regressioon 1 vs lineaarne regressioon	26 766	1,8%
Logistiline regressioon 1 vs diskriminantanalüüs	33 318	2,3%
Logistiline regressioon 2 vs diskriminantanalüüs	39 246	2,7%
Lineaarne regressioon vs diskriminantanalüüs	41 820	2,9%

3.3. Võrdlus SA avaldatud rahvaarvuga

SA rahvaarv koostatakse alates eelmisest loendusest isikupõhiselt. 2012. aasta rahvaarvule lisati 30 760 inimest, kes moodustasid REL2011 alakaetuse [12]. Edaspidi on iga aasta lisatud isikupõhisele loendile registreeritud rahvastikusündmuste põhjal sünnid, surmad ja elukohavahetuste (rände) andmed. Toodud metoodika põhjal elab 2015. aasta alguses Eestis 1 313 271 inimest.

01.01.2015 seisuga SA avaldatud rahvaarvus on 3392 inimest, keda ei ole sama seisuga RRis. Erinevus on tingitud rahvaarvu leidmise metoodikast ning mõningasest RRi alakaetusest. Edasisest analüüsist on need isikud välja jäetud.

Kõige rohkem hindab residentideks diskriminantanalüüs - 97,1% hinnatud residentidest on SA rahvaarvus esindatud, aga mitteresidentideks hindab ainult 86,2% kõigist SA rahvaarvu mittekuuluvatest isikutest (tabel 13). Antud juhul käituvad sarnaselt logistiline regressioon 2 ja lineaarne regressioon. Põhjusteks, et kaks logistilist regressiooni ei ole SA rahvaarvuga sarnased, on erinevate vanusrühmade kasutamine ja erinevus kindlate residentide õpperühmas.

Tabel 12. SA rahvaarvu ja meetodite abil leitud tulemuste tundlikkus ja sensitiivsus

	Tundlikkus	Sensitiivsus
Logistiline regressioon 1	95,5%	85,3%
Logistiline regressioon 2	95,1%	89,9%
Lineaarne regressioon	95,1%	88,1%
Diskriminantanalüüs	97,1%	86,2%

Kuna ükski mudel ei anna väga head lähedust SA rahvaarvuga, siis järgnevalt on vaadatud kõiki eri meetoditel saadud tulemusi koos ehk on moodustatud tunnus, mis näitab, mitmes meetodis hinnati isik residendiks.

RRi väljavõttes olevatest isikutest 89,5% on esindatud SA rahvaarvus, mis kinnitab fakti, et RRis on ülekaetus (tabel 13). RRi väljavõttes olevatest isikutest 84,6% on esindatud nii SA rahvaarvus, kui ka kõik statistilised meetodid on määranud nad residentideks. Samas leidub üle 13 000 isiku, kes on esindatud RRis ja kõik 4 statistilist meetodit määravad isikud residentideks, kuid SA rahvaarvus neid ei ole. Seevastu leidub üle 32 000 isiku, kes on SA rahvaarvus, kuid keda registri andmete põhjal prognoositud mudelitega ei määrata residentideks.

Tabel 13. Statistiliste meetodite summa ja SA rahvaarvu võrdlus

Summa*	SA rahvaarv				Kokku	
	mres		res			
0	125 869	8,6%	32 622	2,2%	158 491	10,8%
1	5 714	0,4%	23 768	1,6%	29 482	2,0%
2	5 944	0,4%	7 477	0,5%	13 421	0,9%
3	2 364	0,2%	8 754	0,6%	11 118	0,8%
4	13 089	0,9%	1 237 258	84,6%	1 250 347	85,5%
Kokku	152 980	10,5%	1 309 879	89,5%	1 462 859	100,0%

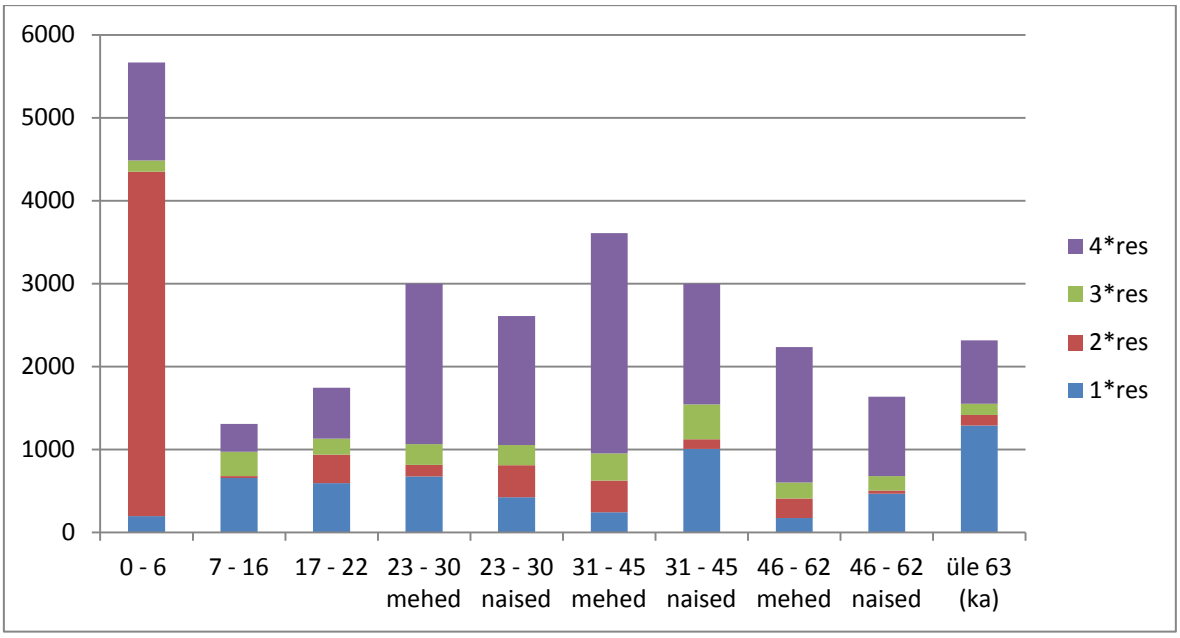
*Summa tähistab väärtust, mitme statistilise meetodiga isik residendiks määrati.

Järgnevalt on vaadatud tabelis 12 toodud gruppide jaotust töö autori poolt kasutatud vanus- ja/või soogruppide järgi. Välja on jäetud grupid, kus kõik neli meetodit hindasid isiku residendiks või mitteresidendiks ja sama tulemus on ka SA rahvaarvus.

Joonisel 5 on toodud nende üle 27 000 isiku jaotus, keda SA rahvaarvus ei olnud aga vähemalt üks statistiline meetod prognoosis residendiks. Vanuse- ja/või soorühmade lõikes eristub teistest märgatavalt 0–6-aastaste rühm, kes on vähemalt kahe meetodiga prognoositud Eesti alalisteks elanikeks. Vaadates andmeid täpsemalt selgub, et suuremas osas (75%) on tegemist pärast eelmise loenduse momenti (31.12.2011) sündinud lastega, kellel RRis elukoht puudub või on välismaa. Võib arvata, et väikelaste vanemad on lapsed registreerinud õigele aadressile, kui on tegemist registreeringuga välismaale. Seega need lapsed tegelikult Eestis ei ela. Vaadates täpsemalt, millised statistilised meetodid need lapsed residentideks hindavad, siis üle 99% eksivad logistiline regressioon 1 ja diskriminantanalüüs. Alla 5% eksivad logistiline regressioon 2 ja lineaarne regressioon ehk antud juhul prognoosivad õigesti residentidest mitteresidentideks.

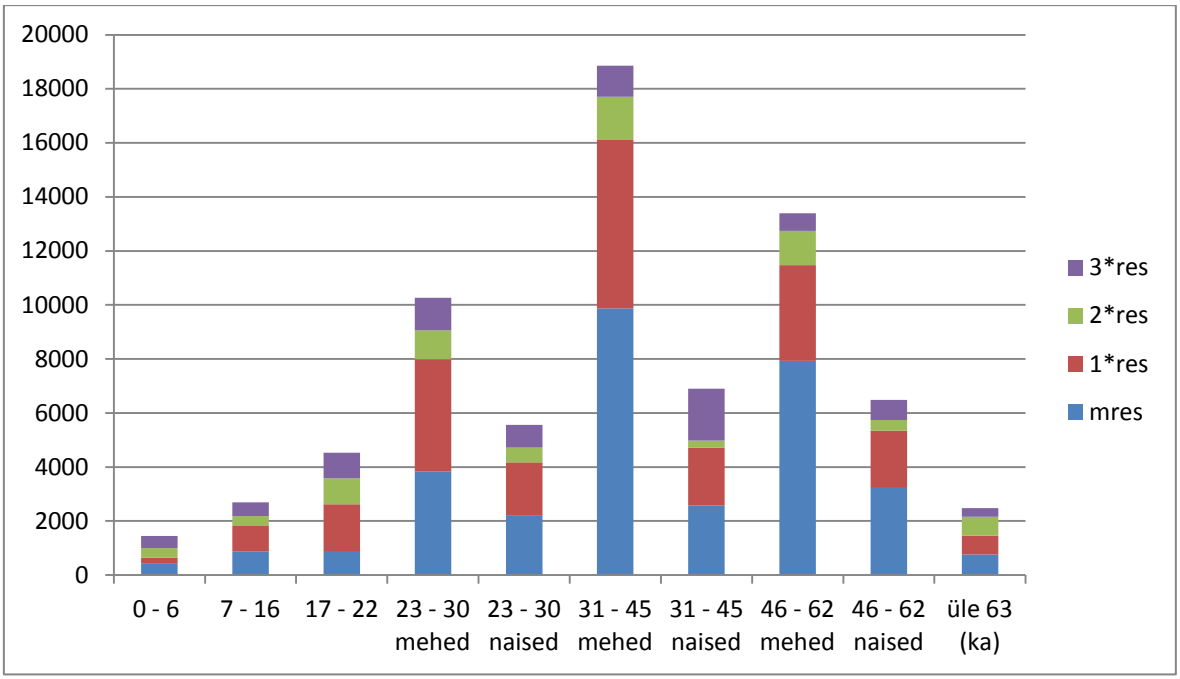
Kui on saabunud antud tööst välja jäänud registrite andmed ja andmetele rakendatakse uuesti statistilise meetodeid, siis tuleks muuta alapunktis 1.2 tehtud kontrollgruppide valikul tehtud eeldusi: alla 3- aastased lapsed, kelle RRis registreeritud elukoht on puudu või välismaal, lugeda Eestist lahkunuteks.

Ülejäänud vanus- ja/või soogruppides ei paista erilisena käituvaid rühmi. Kõige rohkem hindavad kõik neli statistilist meetodit residentideks 31–45-aastaseid mehi, keda SA rahvaarvus ei kajastu. Tulevikus on plaanis täiendada senist SA rahvaarvu metoodikat registripõhise loenduse jaoks väljatöötatud metoodikaga.



Joonis 5. SA rahvaarvu mitteresidentide jaotus prognoositud tulemuste summa ning vanus- ja/või soorühmade järgi

Joonisel 6 on nende üle 72 000 isiku jaotus, kes kuulusid SA rahvaarvu, kuid vähem kui neli statistilist meetodit prognoosisid residendiks. Kui eelmisel joonisel oli näha, et neli meetodit hindasid residentideks 2500 meest vanuses 31–45 aastat, keda SA rahvaarvus ei olnud, siis joonisel 6 on näha, et SA rahvaarvu kuulub peaaegu 10 000 samas vanuses meest, keda ükski meetod ei prognoosinud residendiks. Vaadates täpsemalt, mitmesse registrisse need mehed kuuluvad, siis 44% neist kuuluvad ainult RRi, 34% kuulub lisaks RRile veel ühte registrisse, 17% kuulub lisaks RRile veel kahte registrisse ja 5% kuulub koos RRiga viide registrisse. Samasugust käitumist on näha ka 23–30- ja 46–62-aastaste meeste puhul. Toodud tulemust võib parandada puudu jäänud registri andmete lisamisega mudelitesse.



Joonis 6. SA rahvaarvu residentide jaotus prognoositud tulemuste summa ning vanus- ja/või soorühmade järgi

3.4. Järeldused ja ettepanekud

Kõige paremini sai registrite andmete põhjal prognoosida 7–16-aastaste residentsust Eestis. Kõige keerulisem aga oli 23–62-aastaste meeste prognoosimisega. Antud tulemusi võiks parandada isikut tõendavate dokumentide andmekogu, vangide ja kriminaalhooldusaluste registri ning kohustusliku kogumispensioni registri ja RRis toimunud sündmuse (abiellumine, elukoha vahetus Eesti siseselt) andmete lülitamine analüüsi. Lisaks võiks otsida veel registrite andmeid, näiteks valimas käinute nimekiri, jahimeeste/relvalubade register jne.

Neljast erinevast lähenemisest residentsuse prognoosimisel andsid kõige lähedasemaid tulemusi kaks logistilist regressiooni erinevate eeldustega. SA rahvaarvuga hindas nendest kahest sarnasemalt antud magistritöö autori poolt läbi viidud logistiline regressioonanalüüs.

Kõige lähedasema tulemuse SA avaldatud rahvaarvuga andis diskriminantanalüüs, mis hindas õigesti 97% residentidest ja 86% mitteresidentidest. Kuna võrdlus SA rahvaarvuga vanus- ja/või soorühmades tõi välja, et antud meetod hindas alla 3-aastaste väikelaste hulga suuremaks tänu tehtud eeldustele, siis tuleks muuta residentide ja mitteresidentide kontrollgruppe.

Lisaks võib kontrollgruppide määratlemisel kasutada ka teiste registrite andmeid. Näiteks on kindlalt Eestis alaliselt elavad isikud need, kes on riigi poolt ööpäevaringselt erihoolekandel või vangis.

Pärast paranduste tegemist tuleb läbi viia uus analüüs lisandunud andmete, parandatud kontrollgruppidega ning mõlemas vanus- ja soorühmade jaotuses.

KOKKUVÕTE

Järgmine rahva ja eluruumide loendus, mis kuulub 2020/2021 loendusvooru, planeeritakse Eestis läbi viia registripõhiselt. Üks olulisemaid teemasid loenduste puhul on üldkogumite määratlemine, et kõik isikud ja eluruumid saaksid igas riigis ühekordselt loendatud. Käesoleva magistritöö eemärk oli määratleda 2014. aasta lõpu seisuga registrite andmete põhjal loenduse isikute üldkogum. Saadud tulemusi on planeeritud kasutada 2015. aasta lõpus toimuval esimesel prooviloendusel.

Töös kasutati 11 Eesti registri andmeid. Põhjalikumalt toodi ülevaade logistilise regressioonanalüüsi tulemustest inimeste residentideks määramise kohta. Kõige paremini eristusid registri andmete põhjal 7–16-aastaste noorte residendiks olemine võrreldes Eestist lahkunutega. Kõige keerulisemaks osutus 23–62-aastaste meeste residentsuse prognoosimine.

Lisaks magistritöö autorile analüüsisid samu koondandmeid, mis antud töö autor kokku pani, TÜ magistrandid aine Andmetöötlusmeetodid raames. Erinevate analüüside võrdlus tõi välja, et kõige sarnasemalt käitusid logistilise regressioonanalüüsid erinevate vanus- ja soorühmade ning kontrollgruppide korral. Kõige rohkem erinesid teistest meetoditest diskriminantanalüüsiga saadud tulemused.

SA avaldatud rahvaarvuga võrreldes andis kõige sarnasemaid tulemusi diskriminantanalüüs. Võrdlus tõi välja, et alla 3-aastaste puhul tuleks muuta kontrollrühmi, sest hetkel olid määratud residentideks ka need väikelapsed, kelle RRis registreeritud elukoht puudus või oli välismaal, kuigi võib arvata, et suuremas osas on neil elukoht õigesti registreeritud. Kõige suuremad erinevused SA rahvaarvuga ilmnesid 23–62-aastaste meeste hulgas. Antud probleemi tuleks proovida lahendada antud tööst välja jäänud registri andmete ning RRi sündmuste andmete lisamisega.

Antud töö tulemus on oluline samm registripõhise loenduse isikute üldkogumi määramisel. Plaanis on jätkata töös toodud ettepanekutega, et isikute üldkogumit täpsustada.

Kasutatud kirjandus

[1] **Tiit, Ene-Margit** (2014), 2011. aasta rahva ja eluruumide loendus. Metoodika

http://www.stat.ee/77706 (viimati vaadatud 30.05.2015)

[2] Euroopa Nõukogu ja Parlamendi Määrus EÜ nr 763/2008 (2008)

http://eur-lex.europa.eu/legal-content/ET/TXT/PDF/?uri=CELEX:32008R0763&rid=5

(viimati vaadatud 20.04.2015)

[3] Euroopa Komisjoni Määrus EÜ nr 1201/2009 (2009)

http://eur-lex.europa.eu/legal-content/ET/TXT/PDF/?uri=CELEX:32009R1201&rid=9

(viimati vaadatud 20.04.2015)

[4] **Puur, Allan; Sakkeus, Luule; Aben, Siim** (2013), REGREL metoodika väljatöötamise projekti lõpparuanne

http://www.stat.ee/dokumendid/76831 (viimati vaadatud 26.05.2015)

[5] Rahvastikuregistri seadus,

https://www.riigiteataja.ee/akt/122112013002?leiaKehtiv

[6] **Tiit, Ene-Margit; Meres, Koit; Vähi, Mare** (2012), Rahvaloenduse üldkogumi hindamine. Eesti Statistika Kvartalikiri 3/2012

http://www.stat.ee/57669 (viimati vaadatud 30.05.2015)

[7] **Tiit, Ene-Margit; Vähi, Mare** (2012), Rahvaloendajate tegevus küsitluse järel. Eesti Statistika Kvartalikiri 2/2012

http://www.stat.ee/57667 (viimati vaadatud 30.05.2015)

[8] **Allison, Paul D.** (1999), *Logistic Regression Using the SAS System; Theory and Application. Cary, NC: SAS Institute Inc.*

[9] **Kaart, Tanel** (2012), Binaarsete tunnuste analüüsimeetodid, Õpiobjekt

http://ph.emu.ee/~ktanel/bin_tunnuste_analyys/bin_tunnuste_analyys.pdf

(viimati vaadatud 26.05.2015)

[10] **Allison, Paul D.** (2014), Measures *of Fit for Logistic Regression, SAS Global Forum* 2014,

http://statisticalhorizons.com/wp-content/uploads/GOFForLogisticRegression-Paper.pdf

(viimati vaadatud 26.05.2015)

[11] *Usage Note 39109: Measures and tests of the discriminatory power of a binary logistic model*, http://support.sas.com/kb/39/109.html (viimati vaadatud 13.05.2015)

[12] SA rahvaarvu metoodika,

http://pub.stat.ee/px-web.2001/Database/Rahvastik/01Rahvastikunaitajad_ja_koosseis/04Rahvaarv_ja_rahvastiku_koosseis/RV_021.htm (viimati vaadatud 29.05.2015)

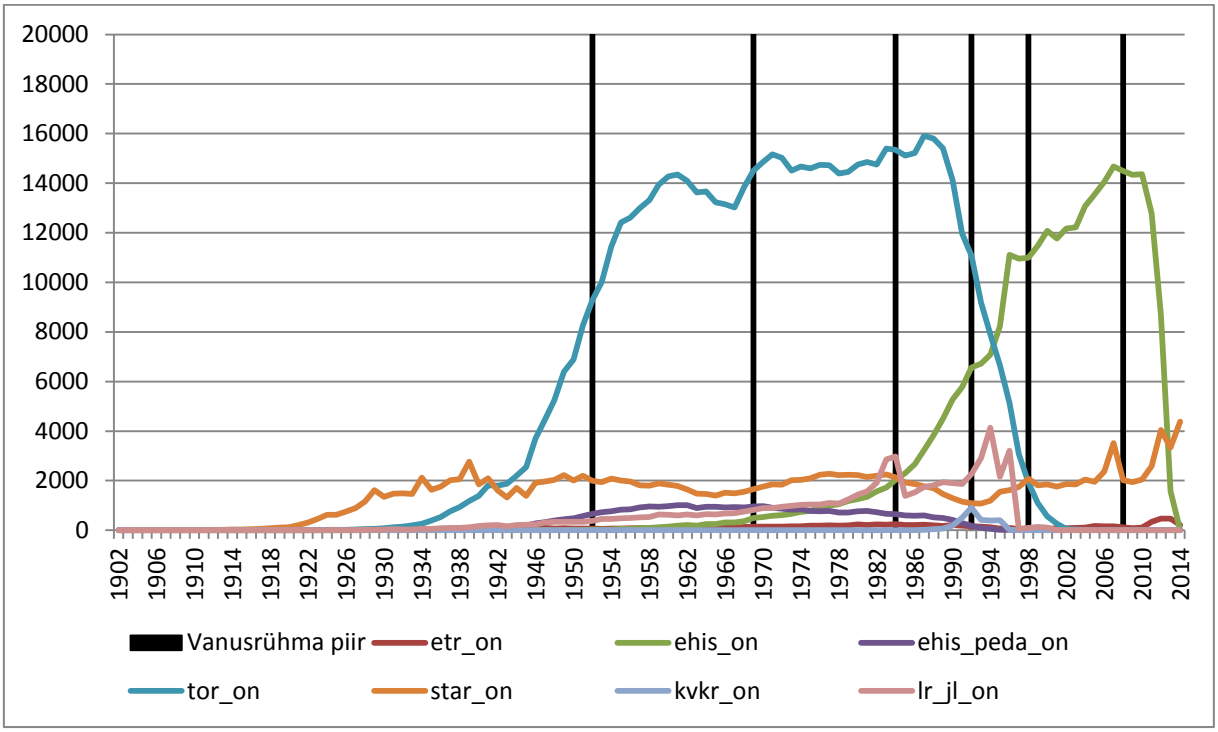
LISA 1. Kasutatavate tunnuste loend ja kirjeldus

Nr.	Tunnuse nimi	Allikas	Sisu
1	IK_ID	RR	Anonüümitud isikukood
2	SUGU	RR	RRi isikukoodist tuletatud sugu
3	SYNNIAEG	RR	RRi isikukoodist tuletatud sünniaeg
4	VANUS	Arvutatud	Vanus seisuga 01.01.2015
5	elukoht	RR	= {„Eesti“; „Välismaa“; „Puudu“}
6	etr_on	ETR	0 = ei olnud registris
			1 = isik oli ETRi väljavõttes (loa/õiguse lõppaeg oli puudu või hiljem kui 01.01.15 (ka))
7	ehis_on	EHIS	0 = ei olnud registris
			1 = õpib EHISe andmetel 01.01.2015
8	ehis_peda_on	EHIS	0 = ei olnud registris
			1 = EHISe andmetel on pedagoog 31.12.2014 seisuga
9	tor_on	TÖR	0 = ei olnud registris
			1 = isik oli TÖRis perioodil 01.01-01.12.2014
10	star_on	STAR	0 = ei olnud registris
			1 = isik on STARis
11	kvkr_on	KVKR	0 = ei olnud registris
			1 = isik on ajateenistuses või asendusteenistuses 2014 aastal
12	lr_jl_on	liiklusregister	0 = ei olnud registris
			1 = isik on vahetanud 2014 aasta jooksul vähemalt korra Eestis juhiluba
13	lr_om_on	liiklusregister	0 = ei olnud registris
			1 = isik on müünud või ostnud sõiduki või on olnud liisingu sõiduki kasutaja või on olnud kasutaja sõidukile, mille müüs/ostis juriidiline isik

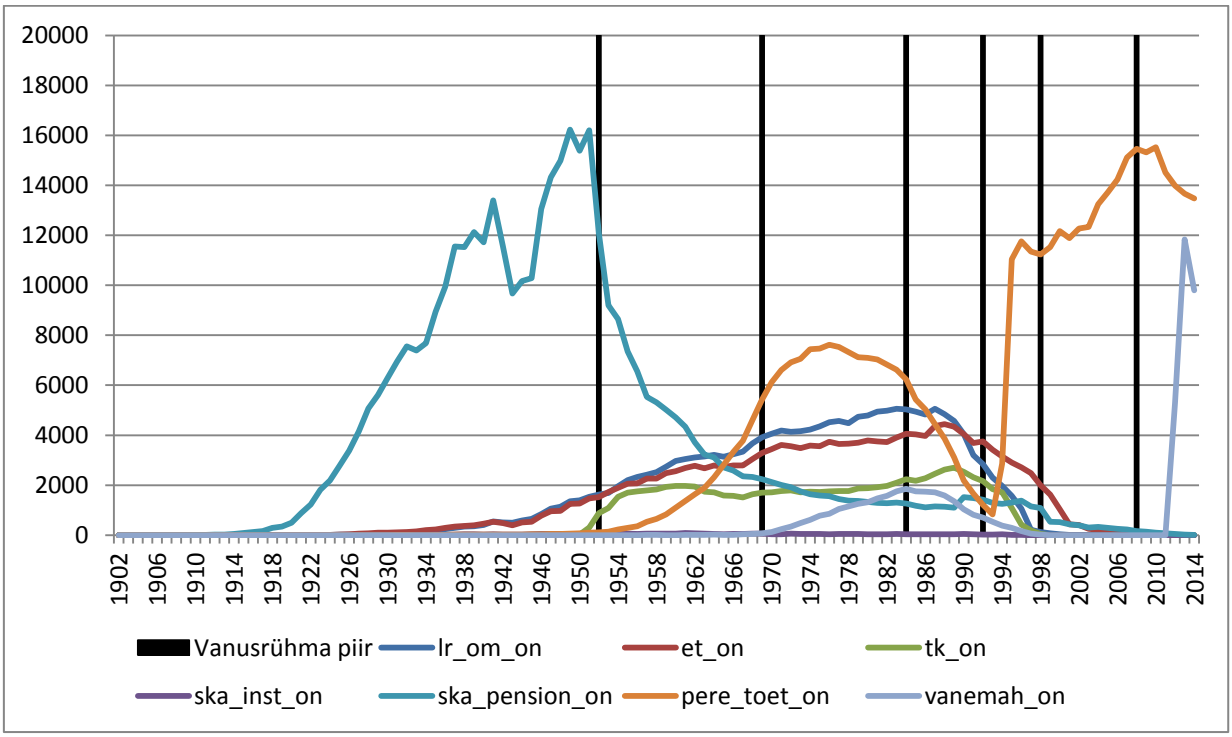
Nr.	Tunnuse nimi	Allikas	Sisu
14	et_on	e-toimik	0 = ei olnud registris
			1 = isik on osalenud kohtuistungil või ülekuulamisel; isik on toimepannud kuriteo/väärteo ja seostatud sündmusega; isik on võtnud vastu dokumendi isiklikult, mille puhul kättesaamise info on nõutav (va elektroonilisel teel saadetud teated)
15	tk_on	EMPIS	0 = ei olnud registris
			1 = isik on olnud töötu või tööotsija 2014 aastal
16	ska_inst_on	PKR	0 = ei olnud registris
			1 = isik on riigi poolt ööpäevaringsel erihoolekandel
17	sots_toet_on	PKR	0= ei olnud registris
			1 = isikule makstakse sotsiaaltoetust riigi poolt või on toetuse saaja eestkostja
18	ska_pens_on	PKR	0 = ei olnud registris
			1 = isikule makstakse riiklikku pensioni (va need, kellele makstakse välisriigi kontole või elavad välismaal)
19	pere_toet_on	PKR	0 = ei olnud registris
			1 = isikule makstakse peretoetust või on laps, kelle eest seda makstakse
20	vanemah_on	PKR	0 = ei olnud registris
			1 = isikule makstakse vanemahüvitist või on laps, kelle eest seda makstakse
21	hambaravi_on	KIRST	0 = ei olnud registris
			1 = isikule on kompenseeritud hambaravi või proteese vahemikus 01.01.2014-31.12.2014
22	digiretsept_on	KIRST	0 = ei olnud registris
			1 = isik on 2014 aasta jooksul välja ostnud digiretsepti (väljaostja isik, mitte kellele retsept kirjutati)

Nr.	Tunnuse nimi	Allikas	Sisu
23	raviarve_on	KIRST	0 = ei olnud registris
			1 = isikul on alustatud 2014 aastal raviarve
24	lapsvabastus_on	KIRST	0 = ei olnud registris
			1 = isik on vabastatud töölt sünnitamise või lapsendamise tõttu (omab sünnitus-või lapsendamislehte)
25	toovoimetus_on	KIRST	0 = ei olnud registris
			1 = isikul on olnud 2014 aastal vähemalt korra töövõimetusleht (haigushüvitis, hooldushüvitis)
26	kindlustus_on	KIRST	0 = ei olnud registris
			1 = isikul on Haigekassa kindlustatus 2014 aastal vähemalt korra (va need, kellel olid ainult järgnevad kindlustused: Isik kuni 19-aastaseks saamiseni, Välismaa üliõpilane, Eesti pensionär teises EL liikmesriigis ja EL liikmesriigis elav pereliige)
27	ryhm_et	arvutatud	1 – 14, vt tabel 1.
28	reg	arvutatud	Tunnuste nr 6 – 26 summa.

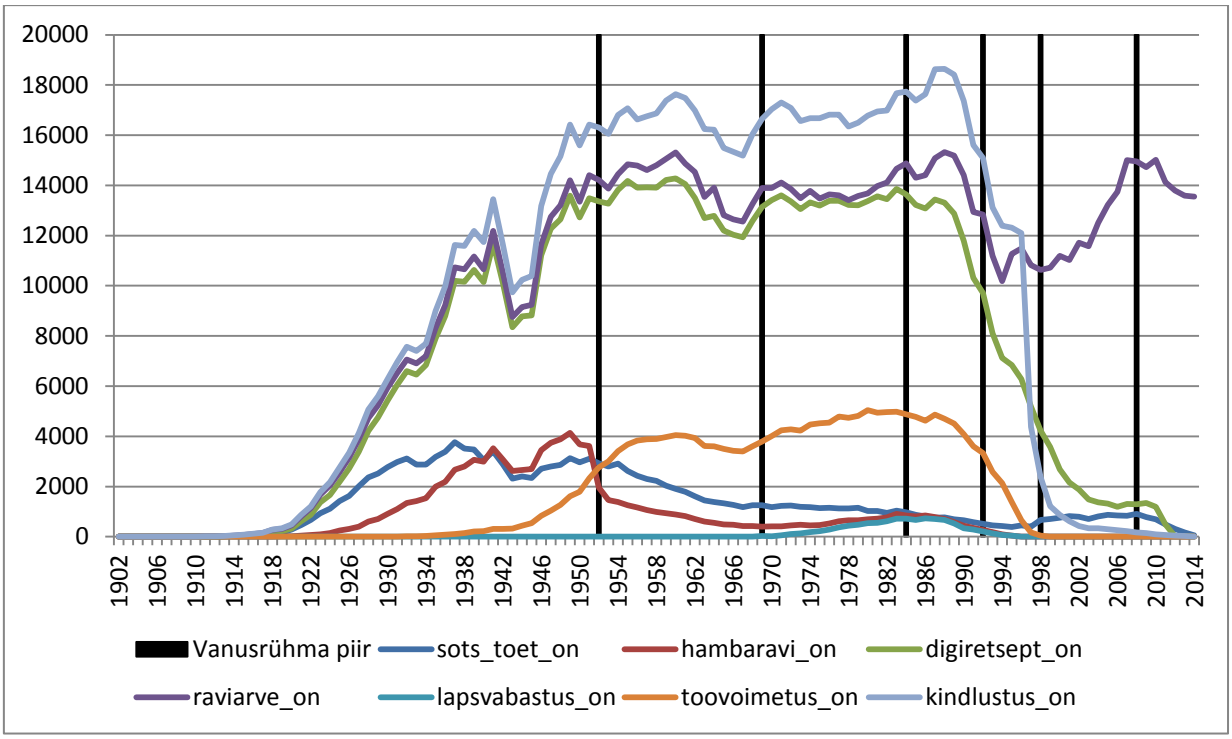
LISA 2. Joonised registrite aktiivsusest sünniaastati



Joonis 1. Inimeste registrites esinemine (1)



Joonis 2. Inimeste registrites esinemine (2)



Joonis 3. Inimeste registrites esinemine (3)

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Ethel Maasing

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Eesti alaliste elanike määratlemine registripõhises loenduses,

mille juhendaja on Mare Vähi,

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, 01.06.2015