RAHVALOENDUSE ÜLDKOGUMI HINDAMINE

Ene-Margit Tiit,
Tartu Ülikool, Statistikaamet

Koit Meres **Mare Vähi**
Statistikaamet **Tartu Ülikool**

Mis on statistikas üldkogum, mis valim? Missugune on rahvaloenduse üldkogum? Kui täpne on rahvaloendusel saadud rahvaarv ning missuguseid statistilisi mudeleid saab selle täpsustamiseks kasutada? Artikkel annab metoodilise ülevaate üldkogumi määramisest, alakaetuse parandamise ning rahvaarvu korrigeerimise võimalustest.

Üldkogumi mõiste

Üldkogum on statistikas mõiste, mis tähendab kõigi uuritavate objektide hulka. Valikuuringute puhul uuritakse küll osa üldkogumist, valimit, kuid alati on uuringu eesmärgiks saada teavet üldkogumi kohta. Eesmärk saavutatakse sobivalt defineeritud (disainitud) valimi ja selle disainile vastavate üldistusreeglite abil. Tänapäeval on valikuuringud laialt levinud ja neist saadakse väga oluline osa teadmistest ühiskonna ja majanduse kohta. Kuid valikuuringute kõrval on tähtis koht ka kõiksetel uuringutel. Nende puhul uuritakse kõiki üldkogumisse kuuluvaid objekte, st tegemist on olukorraga, kus valim ja üldkogum ühtivad. Kõige tuntum ja olulisem kõikne uuring on rahvaloendus, mille eesmärgiks on saada otsest ja vahetut infot riigi (piirkonna) kõigi elanike kohta.

Rahvaloenduse üldkogum

Rahvaloenduse üldkogumiks on vastava riigi või piirkonna kogu rahvastik. Kuna rahvastik pidevalt muutub – inimesi sünnib ja sureb (näiteks Eestis on päevas keskmiselt 40 sünni- ja surmajuhtu), aga ka rändab nii riiki sisse kui ka välja –, siis on üldkogumi üheseks mõistmiseks tarvis kasutada väga täpset määratlust. Rahvaloenduse üldkogum määratakse loendusmomendi (loenduse kriitilise momendi) seisuga. Eestis oli viimase rahvaloenduse (REL2011) kriitiline moment 31. detsembril kell 00.00. Kuigi tegemist oli rahva ja eluruumide loendusega, käsitleme käesolevas kirjutises ainult üht osa sellest – rahvaloendust – ja rääkides üldkogumist, mõistame selle all isikute üldkogumit ning jätame kõrvale eluruumid ja leibkonnad. Rahvastiku üldkogumisse kuuluvad kõik need Eesti riigi elanikud, kes loendusmomendil olid elavate kirjas. Loendusmomendist hiljem sündinud või sellest varem surnud üldkogumisse ei kuulu.

Võrreldes ajakriteeriumiga on üldkogumi määratlemisel märksa keerukam paiksuse kriteerium. Rahvaloenduse üldkogumit on ajalooliselt käsitletud kahel viisil. Ühel juhul loendatakse nn kohalolevat või faktilist rahvastikku. Selle kindlakstegemiseks selgitatakse välja kõik isikud, kes viibivad loendusmomendil vastava riigi pinnal, sealhulgas ka kõik lühiajalised külalised hotellides, rongides ja laevadel (mis on selle riigi territooriumil). Kohaloleva rahvastiku loendamine eeldab võrdlemisi lühikest loendusaega (üks-kaks, erandjuhul kuni kümme päeva). Teine, nüüdisajal enam kasutatav variant on lugeda alalist rahvastikku. Alalise rahvastiku määratlemine on keerukam, kuid võimaldab loendada märksa pikema aja jooksul ja on ka sisukam edasiste rakenduste mõttes.

Alaline rahvastik

Alalise rahvastiku hulka kuuluvad kõik isikud, kes elavad loendushetkel alaliselt vastavas riigis. Riigis alaliselt elavateks loetakse inimesed, kes on selles riigis elanud vähemalt ühe aasta (12 kuu) jooksul või kes on selles riigis elanud küll vähem aega, aga kavatsevad vähemalt

12 kuud elada. Siinjuures ei ole oluline, kas isik elab riigis seaduslikult või illegaalselt. Küll aga on rahvusvaheliselt kokku lepitud mõned erijuhud käsitlemaks isikuid, kes elavad või tegutsevad mitmes riigis (nn hargmaised isikud). Kui isikul on perekond, kes elab ühes riigis, ja isik ise töötab teises riigis, kuid veedab suurema osa oma töövabast ajast koos oma perekonnaga, siis loetakse ta selle riigi alaliseks elanikuks, kus elab tema perekond. See määratlus kehtib ka siis, kui isik on teises riigis tööl käinud rohkem kui aasta jooksul. Kui aga inimene õpib ülikoolis või keskkooli järel kutsekoolis ja õpingud kestavad vähemalt aasta, siis loetakse ta õpingutepaiga (riigi) alaliseks elanikuks sõltumata sellest, kui sageli ta külastab teises riigis viibivaid omakseid (vanemaid). Eraldi reeglid on kehtestatud ka diplomaatidele, välisesinduste töötajatele ja sõjalistes missioonides osalejatele, kes üldjuhul loetakse koduriigi alalisteks elanikeks.

Alalise rahvastiku määratlemise juures on kõige keerukam selgitada, kas riigist vähem kui aasta tagasi lahkunud inimene kavatseb välismaale jääda vähemalt 12 kuuks või mitte. Sisseränanute puhul on see lihtsam, sest neilt (või nende leibkonnaliikmetelt) on seda loenduse käigus võimalik küsida. Väljaränanutega pole aga üldjuhul võimalik kontakti saada. Kuigi suurel osal sellistest inimestest on kodumaal sugulasi, ei tarvitse nemad lahkunute pikemaajalisi plaane teada ja nende kaudu saadav teave pole alati ei täielik ega ka täiesti vastav.

Miks eelistatakse tänapäeval uurida alalist rahvastikku? Põhjuseks on inimeste nüüdisajal väga suureks kasvanud liikuvus, mistõttu kohalolev rahvastik võib lühikese aja jooksul üsna suurel määral varieeruda. Turismipiirkondades võib elanike arv hooajati mitmekordistuda, mõne suure ristluslaeva saabumine väikelinna võib selle elanikkonda märgatavalt suurendada, ülikoolilinnad tühjenevad õppevaheaegadel jne. Kuna rahvaloenduse üldkogum saab tavaliselt aluseks edasisele rahvastikustatistikale (muidu poleks see kallis uuring majanduslikult õigustatud), eeldab see, et üldkogumi suurus peab olema võimalikult püsiv. Seetõttu sobib alaline elanikkond tänapäeval rahvastiku arvestuse aluseks märksa paremini kui faktiline. Varasematel ajalooperioodidel, kui rahvastik oli võrdlemisi paikne, erines alaline elanikkond faktilisest üsna vähe ning nende erisus ei mõjutanud rahvastiku arvestust kuigivõrd.

Ometi pole ka alalise rahvastiku kasutamine rahvastikustatistikas tänapäevalgi vaba probleemidest ja küsitavustest, kusjuures segaduste allikaks on nimelt paiksuse kriteerium. Erinevalt loenduste üldisest tavast küsitakse rahvaloendusel objektiivse info kõrval (isik on teatava aja jooksul mingis riigis viibinud) ka subjektiivset infot (isik kavatseb riiki teatavaks ajaks jääda). Kavatsusi puudutav väide on enam-vähem tõsiseltvõetav siis, kui küsimusele vastab kõnealune isik ise, kui aga vastab mõni teine isik (nt leibkonnaliige, mis on loenduse puhul üldiselt vastuvõetav), ei tarvitse vastus olla tõene. Kahjuks on aga (ajutiselt) oma alalisest elukohast eemal viibivate isikute puhul paratamatu, et nende eest vastab keegi teine. Selle tõttu on vahetegemine ajutiselt (alla 12 kuu) eemal viibijate ja lahkunute (st enam kui 12-kuulist või jäädavat eemalviibimist kavandanute) vahel ütluste põhjal üsnagi problemaatiline.

Enamiku Eestis varem toimunud loenduste puhul on määratud niihästi alaline kui ka faktiline rahvastik. Rahvastikusündmuste aluseks võeti varasematel loendustell faktiline rahvastik, kuid ajapikku see muutus. 2000. aastal esitati valdav osa väljunditest alalise rahvastiku kohta. Alalise ja faktilise rahvastiku erinevus oli siis kõigi aegade suurim: alaline rahvastik ületas faktilist enam kui 13 000 inimese võrra (ligi 1% rahvastikust).

Rahvaloendusel saadud rahvaarv ei ole täpne

Rahvaloendus peaks põhimõtteliselt andma rahvastiku kohta objektiivse ja kõigist välisteguritest sõltumatu pildi, sealhulgas täpse alalise rahvaarvu. See on nii aga üksnes juhul, kui õnnestub loendada kõik selles riigis alaliselt elavad inimesed (püsielanikud), loendatute hulka ei sattu ühtegi liigset isikut ja kedagi ei loendata korduvalt. Kahjuks see üldjuhul ei õnnestu.

Üks kõige olulisemaid täpsuse näitajaid on loenduse kaetus, mis iseloomustab loendatud isikute arvu L ja loendamisele kuuluvate isikute arvu – üldkogumi – N vahekorda. Kaetust iseloomustab suhe L / K ja vastavalt sellele, kas see suhe on väiksem kui 1 või suurem kui 1, on tegemist ala- või ülekaetusega. Ülekaetust on võimalik olulisel määral vältida, kui rahvastik on isikukoodiga tuvastatav, kuid tänapäeva loenduse suurimaks probleemiks on alakaetus, mis näitab, kui suur

osa loendamisele kuuluvast rahvastikust (üldkogumist) jäi tegelikult loendamata. Alakaetuse määraks on suhe (N – L) / N, mida tavaliselt väljendatakse protsentides.

Rahvaloenduste korraldajad kogu maailmas on üksmeelel selles, et rahvaloendustel muutub inimeste kättesaamine aina keerulisemaks. Põhjusi on mitmesuguseid, aga kõige olulisemad on neist kaks: esiteks inimeste suur liikuvus, mitmes elukohas ja isegi mitmes riigis elamine ja töötamine ning ühtlasi perekondade ja leibkondade vormide mitmekesistumine; teiseks inimeste suurenenud privaatsusetaotlus, soovimatus oma andmeid teistele (loendajale, riigivõimule) teatada. Hirm, et loendusandmeid võidakse kasutada isiku huvide vastu, pole teavitustööst ja turvameetmetest hoolimata tänapäevalgi kadunud.

Eesti rahvaloenduse tulemuste täpsus

Rahvaloenduse täpsust saab hinnata mitmeti. Kui kahe loenduse vahel ei ole toimunud drastilisi rahvastikusündmusi, sobib kõrvutada eelmise loenduse põhjal tehtud jooksva statistika tulemusi loendustulemustega. Sisuliselt mõõdetakse sellega kahe järjestikuse loenduse andmete kooskõla. Nii oli võimalik hinnata 1934. aasta loenduse, samuti nõukogude ajal toimunud loenduste (1970, 1979 ja 1989) täpsust. Selgus, et loenduse täpsus oli hea 1934. aastal (alakaetus u 1500 inimest, seegi suures osas seletatav vastsündinute registreerimise viivitustega), samuti 1979. aastal (erinevus alla 1000 inimese).

Teine võimalus loenduse täpsust hinnata on kasutada järelloendust. Selle puhul selgitatakse valimi põhjal tehtud järelloendusel välja isikud, kes jäid põhiloendusel loendamata, ja laiendatakse nende arv valikuteooria eeskirjade kohaselt kõigile loendamisele kuuluvatele isikutele. Seda meetodit kasutati 2000. aasta rahvaloendusel. Selgus, et loendus oli alakaetud ehk et osa loendamisele kuuluvatest isikutest oli jäänud loendamata. Alakaetuse hinnanguks oli 1,2%, kusjuures hinnangu autori sõnutsi oli see alakaetuse alampiir. Seega oli selge, et loenduse tulemusena saadud üldkogumi arvukus erineb tegelikust rahvaarvust vähemalt 15 000 inimese võrra. Seda teadmist rahvastikustatistika näitajate täpsustamiseks siiski ei kasutatud.

Kolmas võimalus rahvaloenduse täpsust hinnata on kasutada lisainfot, näiteks registreid. Võrreldes loendatud isikute arvu mõne kogu rahvastikku esindava registri aktiivsete kirjete arvuga, on põhimõtteliselt võimalik saada hinnang loenduse kaetusele ja osalt ka muudele kvaliteedinäitajatele. On aga selge, et registreid saab loenduse kvaliteedi hindamiseks kasutada üksnes siis, kui registrite endi kvaliteet on küllalt usaldusväärne ja kõrge.

Alakaetuse parandamise võimalused ja meetodid

Sõltuvalt lisateabe olemasolust on loenduse alakaetuse parandamiseks mitu võimalust, kuid senises rahvaloenduste praktikas pole neid kuigi sageli kasutatud. Üks võimalus on kasutada kaalusid (sarnaselt valikuuringutega). Näiteks kui on selge, et teatavas asulas on 3% elanikest jäänud loendamata, siis omistatakse igale asula elanikule kaal 1,03 ja parandatakse sellega asula elanike üldarv, kusjuures elanike soo-vanusjaotus, aga ka teiste loendusel mõõdetud tunnuste jaotus jääb täpselt selliseks, nagu see loendusel kindlaks tehti. Tulemus on (enam-vähem) õige siis, kui loendamata jäämine on täiesti juhuslik ehk ei sõltu elanike soost, vanusest ja teistest tunnustest – näiteks kui ühe loendaja andmed on puudu jäänud ja selle loendaja piirkond ei erine millegi poolest asula üldpildist. Enamasti on loendamata ja loendatud isikud siiski pisut erinevad: loendamata kipuvad jääma pigem nooremad ja liikuvamad inimesed. Seepärast pole niisuguse metoodika rakendadamine alati otstarbekas. Ka pole seda mõistlik rakendada väikeste asulate korral.

Lisavõimalusi rahvaloenduse alakaetuse täpsustamiseks pakuvad riiklikud registrid. Väga hästi toimiva ja täieliku registrite süsteemiga riikides on juba loobutud tavapäraste loenduste korraldamisest. Selle asemel tehakse loendusetaolisi kokkuvõtteid registrite põhjal. Esimestena asusid sellele teele Põhjamaad – Soome, Rootsi ja Taani. Kuigi tänapäeval on sellesse loetelusse lisandunud veel riike, ei ületa ka 2011. aasta loendusvoorus registripõhiselt loendust korraldanud riikide arv kümmet. Küll aga on võimalik registreid kasutada loendusandmete

parandamiseks ja täiendamiseks. Selle tegevuse idee on väga lihtne. Kui oletada, et iga vastava riigi elanik on kantud (ühte või mitmesse) registrisse ja surres või riigist lahkudes kustutatakse ta sealt, siis pakub niisugune register suurepärase võimaluse loendusandmeid täiendada. Sellise registri abil saab küll täpsustada üldkogumi loetelu (ja sellesse kuuluvate isikute arvu), kuid üheski registris ei ole andmeid kõigi loendusküsimuste kohta. Seetõttu tuleb osale loendusankeedi küsimustele otsida vastused teistest allikatest. Kõige informatiivsem on paljudes riikides olemas olev rahvastikuregister, samuti mitmesugused sotsiaal- ja arstiabi registrid. Kuigi registreid on tänapäeval paljudes riikides, pole sageli analüüsitud ja hinnatud nende kvaliteeti, katvust ja koostoime võimalusi. Probleem võib tekkida ka andmekaitsega: kuigi Euroopas on erinevate andmekogude isikuandmete seostamine statistika eesmärgil erandina lubatud, võivad üksikutes riikides seadused ka rangemad olla.

Rahvaloenduse korraldajate dilemma

Tänapäevaste rahvaloenduste korraldajad seisavad dilemma ees, kas lugeda rahvaarvuks loendatud isikute arv või seda parandada, eriti juhul, kui on selge, et tegemist on arvestatava alakaetusega. Varasematel loendustel üldiselt seda probleemi ei olnud. Esiteks, juhuslikel põhjustel tekkinud üle- ja alakaetus kompenseerusid vastastikku ja teiseks polnud registrite näol olemas alternatiivseid infoallikaid. Rahvastiku väiksema liikuvuse ja võimalik, et ka suurema seaduskuulekuse tõttu olid eksimused siis tõenäoliselt väiksemad, kvaliteedinormid aga leebemad.

Niihästi loendusandmete vahetu kasutamine kui ka nende korrigeerimine põhjustavad probleeme.

- Loendusandmete vahetu kasutamise korral on peamiseks probleemiks see, et rahvastikuandmed on teadvalt ekslikud. Alakaetus 1–2% võib tähendada arvestatavaid nihkeid teiste tunnuste jaotustes, näiteks võib mõni soo-vanuserühm olla 5–10% võrra tegelikust väiksem või mõne piirkonna elanikkond loenduse andmetel tegelikust märksa napim (kui alakaetust on osaliseltki põhjustanud loendajate tegevus). Ekslikud rahvastikuandmed põhjustavad aga ka oluliste rahvastikunäitajate – sündimus- ja suremus- ning isegi majandusnäitajate (SKP isiku kohta) – ekslikkust.
- Loendusandmete parandamisega seostub terve hulk probleeme. Esiteks puudub selleks standardne ja rahvusvaheliselt soovitatav metoodika, mis tuleb igas riigis vastavalt olemasolevatele ressurssidele (teabeallikatele) välja töötada. Teiseks peab see metoodika olema küllalt läbipaistev ja arusaadav, et vältida kahtlustusi hinnangute poliitilise kallutatuse suhtes. Kolmandaks on tarvis täpsustada mitte üksnes rahvaarvu, vaid alternatiivsetest andmeallikatest tuleb leida ka lisatavatele isikutele loendusel küsitud oluliste tunnuste väärtused.

Kõigi nende probleemide tõttu on loendusandmete parandamist rahvaarvu täpsustamise eesmärgil seni üsna vähe praktiseeritud. Võib siiski oletada, et 2011. aasta loendustulemuste korral tehakse seda senisest rohkem. Juba on loenduse põhjal leitud rahvaarvu registrite põhjal parandanud Läti statistikaamet.

Eesti 2011. aasta rahvaloenduse alakaetus ja rahvaarvu korrigeerimise küsimus

Eesti 2011. aasta rahvaloenduse tulemused, mis 31. mail 2012 ametlikult avaldati, on ilmselt alakaetud. Seda kinnitasid hulgaliselt Statistikaametisse laekunud signaalid, samuti mitmed meediakajastused. Peale tavapäraste loendamata jäämise asjaolude (ajutine kodunt eemal-viibimine, soovimatus loendajaga suhelda ja oma andmeid avaldada, loendajate eksimused ja tegematajätmised) ilmnes 2011. aasta loendusel veel üks põhjus, mille tõttu osa inimesi jäi loendamata. Loenduse esimesel etapil toimunud e-loendus oli väga edukas ja internetis loendas end ligemale 66% elanikest. Kõik need inimesed märkisid ise endi elukoha. Kuigi oli palutud märkida tegelik, mitte registreeritud elukoht, leidus neid, kes märkisid enda registreeritud elukoha

(või ka mõne muu elukoha), kus nad aga ise ei elanud. Kui selles elukohas tegelikult elas leibkond, kes end internetis ei loendanud, võiski ta loendamata jääda, sest loendajad ei külastanud eluruume, mis e-loendusel olid korrektselt loendatud. Niisugusel viisil loendamata jäänud inimeste kohta laekus pärast loenduse lõppu rohkesti teateid.

Loendusmeeskonnal tuleb nüüd langetada otsus, kas lugeda rahvastiku arvuks loendatud isikute arv, mis on teadaolevalt tegelikust rahvaarvust väiksem, või püüda seda parandada. Standardset ettekirjutust rahvusvahelistelt organisatsioonidelt selle kohta ei ole. Kui mõne riigi loendusmeeskond otsustab oma loendustulemusi parandada, siis on see vastuvõetav, kuid on ka võimalik esitada korrigeerimata rahvaarvud hoolimata sellest, et loenduse alakaetus on teada.

Kuna Eesti 2011. aasta rahvaloenduse registreeritud metoodikas oli ette nähtud registrite kasutamine loenduse eri etappidel, on ka loendustulemuste täpsustamine registrite abil põhimõtteliselt seaduslik. Loendustulemuste parandamise küsimust arutati REL-i teadusnõukogus ja kuigi lõplikku otsust 25. juunil 2012 toimunud nõukogu koosolekul vastu ei võetud, kaldusid nõukogu liikmete arvamused pigem hinnangute parandamise poole. Nõukogu liikmed rõhutasid niihästi koosolekul kui ka sellele eelnenud ja järgnenud mõttevahetuses ettevaatlikkuse vajadust otsuste langetamisel, eelistatavalt võimalikult väikest parandust (pigem jätta isikud, kelle staatus on mõneti kahtlane, püsielanike hulka lugemata), otsustamismetoodika läbipaistvuse ja veenvuse tähtsust ning vajadust seda põhjalikult meediale selgitada.

Küsimus otsustati Statistikaametis 29. augustil 2012 toimunud koosolekul. Rahvaloenduse andmeid ei muudeta, kuid 2012. aasta detsembris avaldatakse alakaetuse andmed kogu riigi kohta, samuti vanuserühmade ja kohalike omavalitsuste kaupa, mis võimaldab asjahuvilistel arvutada kõigis neis alajaotustes välja hinnangulise nihketa (tegeliku) rahvaarvu.

Niisugune rahvaarvu parandamine on Eesti loenduste praktikas esmakordne. Kuigi juba eelmise, 2000. aasta loenduse järelloendusel selgus alakaetus, rahvaarvu ei parandatud. Selleks polnud võimalustki. Esiteks puudus aktsepteeritav metoodika, teiseks puudusid alternatiivsed andmeallikad usaldusväärsete ja kontrollitud-auditeeritud registrite näol. Ka ei lubanud tollased andmekaitse seadused eri registrite andmestike ühistöötlust (linkimist), selle toimingu jaoks vajalik krüptimismetoodika ei olnud veel rakendusteni jõudnud.

Selle tulemusena on eelmisest loendusest möödunud ligi 12 aasta jooksul Eestis teada olev rahvaarv olnud tegelikust mõnevõrra väiksem. Kui arvestada, et loenduse alahinnang oli 1,2–1,5%, siis on alust arvata, et perioodi alguses (2000. aastate alguspoolel) elas Eestis ligikaudu 20 000 inimest rohkem, kui on kirjas Statistikaameti veebilehel. Aastatega on eelmisest rahvaloendusest tulenev tegeliku ja ametliku rahvaarvu erinevus vähenenud ja võimalik, et pöördunud koguni vastassuunaliseks. Selle peamiseks põhjuseks on teine, vastupidise toimega rahvastikuprotsess – välisränne –, mille saldo on vaatlusperioodil negatiivne ja mis on osaliselt registreerimata. Selgub, et 2011. aasta loenduse hinnangute parandamise juures oleks paratamatu ka 2000. aasta loenduse tulemustele hinnangu andmine ja võimalik, et ka vahepealsete rahvastikuarvude teatav täpsustamine.

Täpsustatud rahvaarvu avaldamise kasuks räägib eeskätt vägagi loomulik taotlus saada rahvastikust õige, võimalikult tegelikkusele vastav pilt, mis moodustaks parima võimaliku aluse riigi ja kohaliku elu korraldamiseks. Samuti on väga oluline soov lähendada üksteisele seni Eestis kasutatud kolme rahvaarvu – Statistikaameti rännet arvestamata, Statistikaameti rännet arvestavat ja rahvastikuregistri Eesti elanike arvu. Kõigi kolme rahvaarvu erinevus ulatub üle 20 000, mis on 1–2% rahvaarvust. Erinevuste aluseks ongi ülalmainitud asjaolud: ühelt poolt 2000. aasta rahvaloenduse alakaetus, teiselt poolt rände arvestamine või arvestamata jätmine. Siinjuures väärib märkimist, et eri vanuste puhul on allikate rahvaarv erinev. Näiteks ületab rahvastikuregistri Eesti elanike arv (mis on kokkuvõttes suurim) Statistikaameti rännet mittearvestavat rahvaarvu eelkooliealiste laste puhul, kuid on Statistikaameti rahvaarvust väiksem nooremas koolieas laste puhul. Ka 25–30-aastaste noorte arvukus on rahvastikuregistri andmestikus väiksem kui Statistikaameti omas. See näitab, et ükski praegu kasutatav andmestik ei ole veatu, seda enam, et neist rahvaarvudest ühegi puhul ei ole arvestatud registreerimata rännet. Kindlasti ei vasta ükski olemasolev andmestik tegelikule rahvastiku olukorrale soo-

vanusjaotuse ega paiknemise mõttes, kuigi rahvastiku üldarv võib mõne puhul olla võrdlemisi lähedane Eesti tegelikule rahvaarvule 2011. aasta rahvaloenduse loendusmomendil.

Eesti võimalused 2011. aasta rahvaloenduse käigus määratud rahvaarvu parandada

2011. aasta rahvaloenduse eel tehti Eestis tõsist tööd registreid analüüsides ja täiustades. Võrreldes teiste riikidega on Eesti registritel niihästi plusse kui ka miinuseid.

- Põhilised Eesti isikuandmeid sisaldavad registrid on identifitseeritud isikukoodi abil, seega on need omavahel seostatavad.
- Eestis on välja töötatud aadressistandard (ADS), mis võimaldab ühtse skeemi alusel kirjeldada kõikide eluruumide, aga ka muude oluliste paikade (nt töökohad) aadresse.
- Eestis on toimiv rahvastikuregister, millesse kantakse jooksvalt kõik rahvastikusündmused (sünnid, surmad, registreeritud elukohavahetused).
- Eestis on kõiki õppureid, õpetajaid ja haridusdokumente sisaldav hariduse infosüsteem (EHIS), suurest hulgast alamregistritest koosnev tervisekindlustuse (Haigekassa) register, maksukohustuslaste andmeid sisaldav Maksu- ja Tolliameti register, mitmesuguste toetuste ja pensionide andmeid sisaldav sotsiaalkindlustuse register ja rida teisi registreid (vt ka „Rahvaloendajate tegevus küsitluse järel". Eesti Statistika Kvartalikiri nr 2, 2012).

Nende positiivsete tahkude kõrval tuleb aga tähelepanu pöörata ka varjukülgedele ja puudujääkidele.

- Kõik Eesti registrid on võrdlemisi noored: enamik neist on asutatud käesoleval sajandil, mistõttu nende kasutamise ja koos analüüsimise, seega ka vigade avastamise kogemus on väike.
- Mõnede registrite andmete ajakohastamine jätab soovida. Näiteks tervisekindlustatute registris võivad teatud vanuses inimesed olla ka siis, kui nad on Eestist lahkunud.
- Suurimaks puuduseks Eesti põhilises registris – rahvastikuregistris – on erisus 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 sisse-kirjutuse 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.

Erinevad metoodikad loenduse alakaetuse vea parandamiseks

On selge, et rahvaloendajad ei saa puuduvaid inimesi n-ö välja mõelda. Rahvastiku üldkogumile saab lisada üksnes niisuguseid isikuid, kes on Eesti registrites ja kelle kohta on alust langetada otsus, et nad olid loendusmomendil Eesti püsielanikud.

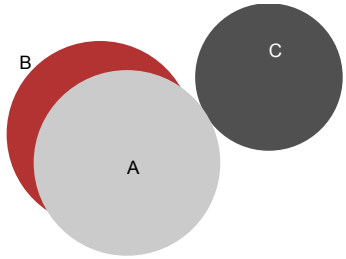
A. Kõige loomulikum on võtta analüüsimisele need isikud, kes on rahvastikuregistris kirjas Eesti elanikena, kuid kelle kohta pole loendustulemusi.

B. Teine võimalus on analüüsida lisaks eelnimetatutele ka neid Eesti isikukoodiga inimesi, kes on küll kantud rahvastikuregistrisse, kuid kes registri andmetel ei ela Eestis (elavad välismaal või on elukohariik märkimata).

C. Peale rahvastikuregistris märgitud isikute on võimalik analüüsida ka selliseid isikuid, kellel on Eesti isikukood ja kes on kirjas mõnes teises Eesti registris.

Joonis 1. Isikud, kelle puhul on võimalik rakendada otsustusreeglit selgitamaks, kas nad võiksid olla loendamata jäänud Eesti püsielanikud

Figure 1. Application of the judgement rule to determine whether a person could be a permanent resident of Estonia who was not enumerated



Lisateavet kõigi nende isikute kohta on võimalik saada kõigist (ülejäänud) riiklikest registritest. 2011. aasta rahvaloenduse andmete täiendamiseks on otstarbekas kasutada nende isikute 2011. aasta jooksul registrites jäädvustatud tegevusi. Kindlasti ei ole kõik registrid püstitatud eesmärgi mõttes samaväärsed. Osa registreid nõuavad inimese enda aktiivsust ja dokumentaalselt tõestatud Eestis elamist, mõnede puhul aga pole see ilmtingimata tarvilik. Registriandmete erinev usaldusväärsus võib tuleneda ka analüüsitava inimese vanusest. Näiteks omavad lapsed tervisekindlustust automaatselt, kuid tööealised mitte: tervisekindlustuse tagamiseks peab inimene töötama või õppima jne. Mõnevõrra rohkem annab teavet arsti külastus Eesti ravikindlustuse alusel – seda esineb välismaal viibivate inimeste puhul harvem, kuigi ka see pole võimatu. Suhteliselt kindel tunnus inimese Eestis elamise kohta on see, kui ta õpib Eesti haridusasutuses päevaõppes. Samuti võib võrdlemisi usaldusväärseks Eestis elamise tunnuseks lugeda kohaliku omavalitsuse toetuse määramist.

Registrite kasutamiseks on põhimõtteliselt kaks sisuliselt erinevat võimalust: eksperthinnangud ja statistilised mudelid.

Eksperthinnangud rahvastiku üldkogumi täpsustamiseks

Pärast registrite põhjalikku (sisulist ja vormilist) analüüsi on võimalik määrata registrite hulgas usaldusväärsemad ja selgitada nende omavahelised seosed, mille tulemusena info ühest registrist teise vahetult üle kantakse, samuti registrid, millesse võib üsna lihtsalt sattuda teave isikute kohta, kes tegelikult ei kuulu Eesti püsielanike hulka. Vältides registritevahelistest seostest tulenevaid ülevõimendusi, on võimalik koostada eksperthinnangud, otsustamaks isikute kuuluvuse üle Eesti püsielanike hulka 2011. aasta rahvaloenduse kriitilisel hetkel. Niisuguseid hindamiseeskirju on võimalik koostada põhimõtteliselt erinevate isikukogumite jaoks (vt joonis 1).

Jämeda hinnanguna on hulka A kuuluva isiku püsielanikuks tunnistamise aluseks 2011. aastal vähemalt kahes, hulka B või C vähemalt kolmes küllalt usaldusväärses registris aktiivselt esinemine.

Eksperthinnangule tugineva meetodi eeliseks on selle lihtne mõistetavus: arusaamiseks pole tarvis mingeid statistikateadmisi. Selle metoodika kõige tõsisemaks puuduseks on aga subjektiivsus. Ühe või teise registri andmete usaldusväärsust või sõltumatust on üksnes „pehmete" meetoditega võrdlemisi raske tõestada. Pole võimalik kinnitada ka seda, et leitud eeskiri on optimaalne, st põhjustab väikseimaid hinnanguvigu.

Statistilisel mudelil põhinevad hinnangud üldkogumi täpsustamiseks

Võimalik on ka teine tee, mille puhul eksperdi subjektiivsus ei mõjuta kuigivõrd tulemust. Selle aluseks on sobivaima eristava eeskirja statistiline määramine. Kirjeldame järgnevas diskriminant-analüüsi meetodit. Selle meetodi rakendamisel koostatakse otsustamiseks vajalik algoritm nn õppeandmestiku põhjal. Pärast loenduse toimumist on võimalik defineerida kaks selgelt määratletud isikute rühma. Üks on „püsielanikud" (P), need on rahvastikuregistri andmetel 1.01.2012 seisuga Eesti elanikud, kes on püsielanikena loendatud ja loendamisel ise vastanud. Teine on „lahkunud" (L), need on lahkununa loendatud (kas ise välismaal vastanud või neid on märkinud lähisugulased) isikud, kes rahvastikuregistri andmetel 1.01.2012 Eestis ei elanud. Nimelt nende „õpperühmade" abil moodustataksegi sobivaim eristav eeskiri, mis võib tugineda kas lineaarsele või logistilisele mudelile. Alljärgnevas kirjeldatakse lineaarse mudeli moodustamist. Mudeli argumentideks (kirjeldavateks tunnusteks) kasutatakse teavet aktiivsuse kohta registrites 2011. aastal, nagu seda tehakse ka eksperthinnangu korral. Siinjuures võib potentsiaalsete argumentide loetelusse lülitada lisaks registrisse kuulumise tunnustele ka mitmesuguseid registrite põhjal moodustatud koondtunnuseid ja indekseid, mis arvestavad näiteks vastavas registris või selle alamregistrites esinemise kordsust või ajastust. Oluline on aga see, et argumentide valik mudelisse ja neile omistatavate kaalude määramine toimub automaatselt. Algoritm toimib nii, et esimesena valitakse mudelis tunnus, mis püsielanikke ja lahkunuid kõige tugevamini eristab, järgmisel sammul lisatakse tunnus nii, et tekib kõige tugevamini rühmi eristav tunnustepaar jne – nii kaua, kuni tunnuste lisamine enam mudelit oluliselt ei paranda. Kuivõrd registrites esinemise aktiivsus sõltub oluliselt isiku vanusest ja osalt ka soost, on mõistlik koostada eeskiri üksikute soo-vanuserühmade jaoks eraldi. Selleks jaotatakse kõik isikud vanusrühmadesse, arvestades tööealiste puhul ka sugu. Rühmade piiride määratlemisel võetakse arvesse tegelikku erinevates registrites esinemise sagedust. Osutus otstarbekaks moodustada kokku üheksa rühma, lisatud on ka nende tinglikud nimetused:

1. Lapsed (vanus 0–6 aastat);
2. Õppurid (vanus 7–19 aastat);
3. Noorukid (mehed, vanus 20–29 aastat);
4. Neiud (naised, vanus 20–29 aastat);
5. Keskeas mehed (vanus 30–39 aastat);
6. Keskeas naised (vanus 30–39 aastat);
7. Vanemad mehed (vanus 40–59 aastat);
8. Vanemad naised (vanus 40–59 aastat);
9. Eakad (vanus vähemalt 60 aastat).

Vajalik tunnuste arv oli eri rühmade puhul 4–7 (vt tabel 1).

Tabel 1. Automaatselt eristavasse lineaarsesse funktsiooni määratud tunnuste kordajad
Table 1. Parameter coefficients assigned automatically to the discriminating linear function

Tunnus *Parameter*	Soo-vanuserühm *Age-sex group*								
	Laps *Child*	Õppur *Student*	Nooruk *Young man*	Neiu *Young woman*	Keskeas Mees *Middle-aged man*	Keskeas naine *Middle-aged woman*	Vanem Mees *Older man*	Vanem Naine *Older woman*	Eakas *Elderly person*
Vabaliige *Free member*	1,735	1,505	1,826	1,555	1,784	1,459	1,872	1,687	1,836
HK1	0,075	0,210	0,047	0,204	0,024	0,142	0,021	0,906	0,118
HK2	0,176		0,096	0,187	0,147	0,293	0,087	0,185	0,0412
HK3		0,028							
EH1	0,008	0,790	0,010	0,013					
EH2		0,083	0,020	0,030	0,009	0,014	0,006		
MTA			0,012	0,010	0,027	0,024	0,010	0,011	
Sotst1	0,157	0,120				0,040		0,009	
Sotst2		0,004							0,001
STAR			0,013		0,016		0,010	0,005	0,001
Mntam			0,009	0,008	0,011	0,006	0,003		

Haigekassa andmete põhjal moodustati kolm tunnust: HK1 on binaarne tunnus, mis näitab haigekassa registrisse kuulumist 2011. aastal, HK2 iseloomustab erinevatesse alamregistritesse kuulumiste arvu ja HK3 sisaldab üksnes usaldusväärseimat infot isiku kindlustatuse kohta. Hariduse kohta on samuti kaks indeksit, neist EH1 on binaarne (ei/jah), EH2 iseloomustab ka kordsust (näiteks, isik, kes ühtaegu õpetab ja õpib, saab kõrgema hinnangu). MTA iseloomustab sissetuleku saamist Eesti ettevõttest, Sotst1 tähistab peretoetuse (see võib iseloomustada nii last kui ka lapsevanemat), Sotst2 sotsiaaltoetuse, STAR kohaliku omavalitsuse määratud toetuse või hüvitise saamist. Mntam on indeks, mille väärtuse määrab isiku sissekanne liikluskindlustuse registrisse. Potentsiaalsete argumentide loetelu, millest mudeli kordajaid otsiti, oli tegelikult märksa pikem – sellesse kuulusid pensionid, vanemahüvitis, töövõimetus- ja puudetoetused jne –, kuid need tunnused ei lisanud rühmade eristamisel mudelisse valitutele täiendavat infot.

Tabelis 1 on mõnel juhul mudelis mitu sama registri põhjal moodustatud tunnust. Kuna need ei ole erimärgilised, pole sel juhul tegemist nn multikollineaarsusega (mis vähendab mudeli täpsust ja muudab selle tõlgendamise raskeks), vaid asjaoluga, et vastavasse registrisse kuulumise mõju ei ole lineaarne. Väärib tähelepanu, et kõigi vanuserühmade puhul osutub haigekassa registritesse kuulumine suurima eristava väärtusega tunnuseks. Väga olulise kaaluga on ka hariduse infosüsteemi põhjal moodustatud indeksid, kuigi see register ei ole keskealiste ja eakate osas kattev. Sotsiaaltoetustest on suurima eristava väärtusega peretoetus, ootuspäraselt toimib kõigi tööealiste puhul olulise eristajana ka kuulumine maksumaksjate registrisse (MTA). Hoolimata kahtlustest (võimalust sooritada eksamid ja kindlustada auto Eestis kasutavat sageli ka välismaalased) lisab liikluskindlustuse registris vaatlusaastal aktiivselt osalemine infot inimese püsielanike sekka kuulumise kohta, olles küll enamikus mudelites kõige väiksema mõjujõuga. Rea eeldatavalt oluliste registrite väljajäämist eristavate tunnuste hulgast seletab tunnustevaheline statistiline seotus – näiteks pensioni- või vanemahüvitise registrisse kuulumisest järeldub isiku kuulumine haigekassa registrisse, seega ei lisa vastav register uut teavet.

Automaatselt mudelisse valitud tunnustest moodustub prognoosiv funktsioon, mida võib (iga soo-vanuserühma korral) ette kujutada sirgena kahe punkti vahel (vt joonis 2). Need punktid on „Keskmine lahkunu“ ja „Keskmine püsielanik“, joonisel kujutatud ringidena. Iga õpperühma kuuluva isiku jaoks arvutatakse prognoosiva funktsiooni väärtus ehk punkt sellel lõigul. Joonisel markeerivad neid punkte pisikesed ellipsid. Püsielanike puhul on prognoosiva funktsiooni väärtus

(prognoos) lähemal „Keskmisele püsielanikule“, lahkunute prognoosid on aga üldiselt lähemal „Keskmisele lahkunule“.

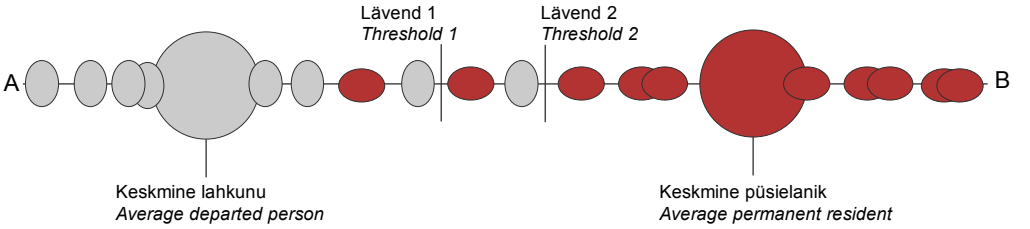
Niiviisi saab prognoosi arvutada mitte üksnes õpperühma kuuluvatele isikutele, vaid ka ülejäänutele (kasutades selleks tabelis 1 antud tunnuste kordajaid) ja vastavalt sellele, kus nende prognoos lõigul paikneb, loetakse isik kas lahkunuks või püsielanikuks. Mudeli konstruktsioonist järeldub, et püsielanike hulka saab lugeda ainult neid isikuid, kes esinevad mudelis märgitud registrites (vähemalt olulisemates), lahkunute hulka aga need, kes neis registrites ei esine või esinevad väheolulistes.

Lävendi määramine

Peale diskrimineeriva funktsiooni on tarvis määrata ka lävendväärtus, mille järgi toimub otsustamine (vt joonis 2).

Joonis 2. Kaks võimalikku lävendit. Prognoosiv funktsioon (lõik AB), lahkunute ja püsielanike keskmised (ringid) ja üksikutele lahkunutele ning püsielanikele vastavad punktid prognoosival lõigul. Lävend jagab lõigu AB kaheks osaks, lävendist vasakul paiknevad isikud loetakse lahkunuteks, lävendist paremal paiknevad – püsielanikeks.

Figure 2. Two potential thresholds. Predictive function (line AB), average values for the departed and permanent residents (circles), and points corresponding to individual departed persons and permanent residents on the prediction line. The threshold divides the line AB in two parts. Persons to the left of the threshold are counted as departed, while those to the right are counted as permanent residents.



Kui optimaalsele mudelile vastav prognoosiv lõik on määratud, tuleb paika panna ka lävend. Lävendi määramisel on ülesande püstitajal teatav valikuvabadus, ent siiski ei saa seda teha subjektiivse valikuga. Lävendi määramisel tuleb arvesse võtta tõsiasja, et statistilise otsustuse juures on paratamatu eksimine. Lävendi valikust sõltub aga otsustusvigade tõenäosus. Siit järeldub, et lävendi määramisel on otstarbekas lähtuda otsustusvigade tõenäosusest.

Selle ülesande lahendamisel on kaks võimalikku viga.

- Esimest liiki viga tehakse siis, kui isik loetakse püsielanikuks, aga ta on tegelikult Eestist lahkunud (elab püsivalt välismaal).
- Teist liiki viga tehakse siis, kui isik, kes tegelikult on Eesti elanik, loetakse lahkunuks (välismaal elavaks).

Otsustusreeglit saab teha kahel viisil. Esimese lähenemise puhul loetakse mõlemad vead samaväärseks ja püütakse leida lävend, mille korral mõlema vea tõenäosus on võrdne ja võimalikult väike. Selle meetodi eeliseks on, et saadav rahvaarvu hinnang on maksimaalselt täpne.

Teine võimalus põhineb ettevaatlikkusel rahvaarvu hindamisel. Selle puhul hoitakse esimest liiki vea tõenäosus võimalikult väiksena, mille tagajärjel paratamatult suureneb teist liiki vea tõenäosus. Sel juhul kokkuvõttes alahinnatakse rahvaarvu. Näiteks võib seada eesmärgiks, et

isiku ekslikult püsielanike hulka lugemise tõenäosus ei tohi olla suurem kui 0,05. Niisugune lubatava vea valik on statistikas väga tavaline. See tähendab, et kui näiteks hindamisele kuulub 10 000 isikut, siis mudeli põhjal otsustades loetakse neist ekslikult püsielanikeks 500 (kuid pole teada, missugused need on). Kui samal ajal teist liiki vea tõenäosus on näiteks 0,09, määrab otsustuseeskiri 900 püsielanikku lahkunuks ja rahvaarvu hinnangusse tekib eksitus 400 inimese võrra.

Võib küsida, kas otsustuseeskirja ei saa moodustada nii, et vigu üldse ei tekiks? Kahjuks ei ole see statistiliste otsustuste puhul üldiselt võimalik. Põhjuseks on asjaolu, et kõigi tunnuste väärtused on paratamatult juhuslikud. Näiteks pole midagi võimalik teha selle vastu, et osa püsielanikke, kes on rahvastikuregistris Eesti elanikud ja on ennast ka loendanud (sealjuures ise loendajale vastanud või Eestis loendusankeedid täitnud), ei ole 2011. aastal üheski registris ühegi aktiivse tegevusega kirjas. Seega joonisel 2 kujutatud lävend 2 ei ole enamasti praktiliselt saavutatav.

Otsustusvigade arvutamine

Otsustusvigade arvutamiseks on põhimõtteliselt kaks võimalust: kasutada objektide paiknemise hindamisel teoreetilist jaotust (selleks on tavaliselt normaaljaotus) või arvutada vead õpperühmade põhjal empiiriliselt.

Esitame tabelis 2 lineaarse mudeli jaoks lävendi väärtused tingimusel, et teoreetiliselt oleksid vigade tõenäosused võrdsed. Lisaks teoreetilistele vigade tõenäosustele on arvutatud ka selliselt konstrueeritud lävendi puhuks empiirilised vigade tõenäosused ja ekslikult määratud isikute tõenäosused õpperühmas.

Empiiriline esimest liiki viga tekib siis, kui õpperühma osasse L (lahkunud) kuuluv isik määratakse otsustusreegli alusel rühma P (püsielanik). Sellise sündmuse esinemissagedus on kasutatava õppeandmestiku korral 11%, mis on võrdlemisi halb tulemus. Kuna aga rühma L osatähtsus õppeandmestikus on võrdlemisi väike, siis kogu õpperühma mõjutab see viga siiski võrdlemisi vähe ja niisuguseid ekslikult rühma P määratud isikuid on kogu õppeandmestikust vaid 0,14%. Empiiriline teist liiki viga tekib siis, kui rühma P kuuluv isik (püsielanik) määratakse otsustusreegli alusel rühma L (lahkunud). Sellise sündmuse suhteline sagedus on alla 4% ja kogu õppeandmestikus on niisugusel viisil ekslikult lahkunuks tunnistatute osatähtsus 3,8%. Esitatud arvutustest järeldub, et vigade empiirilised tõenäosused erinevad keskmiselt teoreetilistest tõenäosustest 2–3 korda, kusjuures oodatust suurem on nimelt esimest liiki vea tõenäosus. Samas on selge, et rühmade L ja P arvukuse erinevuse tõttu ei ole ekslikult määratud P-isikute osatähtsus õppeandmestikus suur. Kuna õppeandmestik moodustab reaalsest loendusandmestikust võrdlemisi suure osa, siis kehtivad need hinnangud ligilähedaselt ka reaalse loendusandmestiku korral.

Tabelist 2 on näha, et kõigi soo-vanuserühmade puhul on lävend kahe rühma keskväärtuste vahel, (näha ka joonisel 2), paiknedes enamasti lähemal rühma P keskmisele kui rühma L keskmisele. Rühma P kuuluvatest punktidest ühe osa sattumine lävendist allapoole (joonisel 2 vasakule) tuleneb juba mainitud asjaolust, et kõik Eesti riigis elavad ning loendatud isikud ei ole vaatlusaastal endast registritesse jälgi jätnud ning seda viga on registrite põhjal praktiliselt võimatu vähendada. Küll aga on võimalik vähendada lävendi nihutamisega esimest liiki viga ehk viga, mis tekib rühma L kuuluva isiku määramisel rühma P.

Tabel 2. Võrdsete teoreetiliste vigade tingimusel määratud lävendid ja empiirilised vead
Table 2. Thresholds and empirical errors determined on the condition of equal theoretical errors

	Soo-vanuserühm *Age-sex group*									
	Laps *Child*	Õppur *Student*	Nooruk *Young man*	Neiu *Young woman*	Keskeas Mees *Middle-aged man*	Keskeas Naine *Middle-aged woman*	Vanem Mees *Older man*	Vanem Naine *Older woman*	Eakas *Elderly person*	Keskmine *Mean*
L keskmine *D mean*	1,804	1,654	1,849	1,610	1,811	1,502	1,890	1,720	1,899	
P keskmine *P mean*	1,997	1,995	1,983	1,981	1,980	1,983	1,989	1,989	1,999	
Lävend *Threshold*	1,953	1,908	1,916	1,832	1,898	1,782	1,943	1,885	1,987	
Teoreetiline vea tõenäosus (ühine) *Theoretical probability of error (combined)*	0,043	0,016	0,081	0,034	0,076	0,011	0,083	0,024	0,126	0,063
Empiiriline 1. liiki vea tõenäosus *Empirical probability of type 1 error*	0,101	0,050	0,154	0,079	0,147	0,073	0,0166	0,079	0,380	0,109
Empiiriline 2. liiki vea tõenäosus *Empirical probability of type 2 error*	0,028	0,028	0,096	0,079	0,075	0,027	0,067	0,035	0,005	0,039
Ekslikult määratud P osatähtsus õppeandmestikus, % *Share of erroneously assigned P in training data, %*	0,04	0,04	0,31	0,25	0,36	0,25	0,21	0,12	0,05	0,14
Ekslikult määratud L osatähtsus õppeandmestikus, % *Share of erroneously assigned D in training data, %*	2,81	2,82	9,42	7,62	7,28	2,59	6,64	3,43	4,71	3,81

Lävendi määramine vastavalt etteantud vea tõenäosusele

Järgnevalt vaatleme võimalust piirata otsustamisel esimest liiki vea tõenäosust. Vahetult on see võimalik teoreetilise vea puhul. Eeldame järgnevas arutelus, et teoreetilise otsustusvea tõenäosus püsielaniku määramisel ei tohi ületada väärtust 0,02. Seda piiri nimetatakse olulisuse nivooks. Püstitatud tingimusele vastava otsustusreegli saamiseks tuleb määrata kõigis vanuserühmades uus lävend. Uue lävendi puhul kahes soo-vanuserühmas (õppurid ja keskealised naised) esimest liiki otsustusvea tõenäosus suureneb. Alati pole etteantud olulisuse nivoole vastava otsustusreegli, st lävendi valik siiski võimalik. See on nii siis, kui rühmad eristuvad väga halvasti, sest rühmakeskmised asuvad lähestikku ja objektid paiknevad läbisegi.

Tabel 3. Lävendi arvutamine juhul, kui esimest liiki vea teoreetiline tõenäosus on piiratud arvuga 0,02.

Table 3. Threshold calculation if the theoretical probability of type 1 error is limited to 0.02.

	Soo-vanuserühm *Age-sex group*									
	Laps *Child*	Õppur *Student*	Nooruk *Young man*	Neiu *Young woman*	Keskeas Mees *Middle-aged man*	Keskeas Naine *Middle-aged woman*	Vanem Mees *Older man*	Vanem Naine *Older woman*	Eakas *Elderly person*	Keskmine *Mean*
Lävend 2 *Threshold 2*	1,983	1,899	1,948	1,860	1,936	1,752	1,969	1,892	1,997	
Teoreetiline 1. liiki vea tõenäosus *Theoretical probability of type 1 error*	0,020	0,020	0,020	0,020	0,020	0,020	0,020	0,020	0,100	0,039
Teoreetiline 2. liiki vea tõenäosus *Theoretical probability of type 2 error*	0,292	0,009	0,232	0,068	0,221	0,004	0,277	0,033	0,499	0,220
Empiiriline 1. liiki vea tõenäosus *Empirical probability of type 1 error*	0,057	0,050	0,083	0,079	0,106	0,103	0,101	0,075	0,090	0,087
Empiiriline 2. liiki vea tõenäosus *Empirical probability of type 2 error*	0,064	0,027	0,168	0,079	0,090	0,022	0,101	0,038	0,146	0,086
Ekslikult määratud P osatähtsus õppe-andmestikus, % *Share of erroneously assigned P in training data, %*	0,02	0,04	0,17	0,24	0,25	0,34	0,12	0,11	0,01	0,11
Ekslikult määratud L osatähtsus õppe-andmestikus, % *Share of erroneously assigned D in training data, %*	6,33	2,65	16,43	7,69	8,79	2,16	9,95	3,74	14,62	8,50

Kui võrrelda lävendiga 1, paikneb uus lävend (lävend 2) enamikul juhtudel lähemal rühma P keskpunktile. Selle tagajärjel väheneb esimest liiki vea tõenäosus, kuid teist liiki vea tõenäosus kasvab märgatavalt. Erandiks olid rühmad, kus juba esimese lävendi puhul oli esimest liiki vea tõenäosus väiksem kui 0,02: uue lävendi rakendamisel nendes esimest liiki vea tõenäosus suurenes ja teist liiki vea tõenäosus vähenes. Kõige keerukam on olukord aga viimase soo-vanuserühmaga (eakad), kus rühmad halvasti eristuvad (rühmakeskmiste vahe on vaid 0,1).

Selle rühma puhul polnud võimalik määrata lävendit kahe keskväärtuse vahel nii, et esimest liiki vea tõenäosus oleks 0,02, ning selle asemel on kasutatud olulisuse nivood (maksimaalset lubatavat esimest liiki vea tõenäosust) 0,1. Ka sel juhul on teist liiki vea tõenäosus ligi 0,5.

Kokkuvõttes selgub, et uue otsustuseeskirja korral erinevad empiirilised vead teoreetilistest rohkem kui esimese eeskirja puhul, kuid on saavutatud hea tasakaal esimest ja teist liiki vigade empiiriliste tõenäosuste vahel. Ekslikult P rühma määratud isikute osatähtsus kogu õppeandmestikus on vaid kümnendik protsenti, mis on kokkuvõttes hea näitaja.

Logistilised otsustusmudelid

Koostati ka kaks logistilist mudelit. Neist esimese puhul lähtuti ligikaudu võrdsetest teoreetilistest vigadest, teise puhul piirati esimest liiki vea tõenäosus väärtusega 0,02 (kui võimalik). Esimese logistilise mudeli puhul erines programmi poolt automaatselt valitud tunnuste hulk mõnevõrra lineaarse mudeli omast (ei kasutatud tunnust HK, üksikutesse rühmadesse lisandusid täiendavad registrid, nagu vanemahüvitis, pension, puudetoetus, töövõimetus). Teise logistilise mudeli puhul oli mõnevõrra vähendatud õpperühma koostist, säilitades üldised põhimõtted. Logistilisi mudeleid iseloomustavad järgmised veahinnangud (vt tabel 4).

Tabel 4. Logistiliste mudelite veahinnangud
Table 4. Error estimates of the logistic models

	Soo-vanuserühm *Age-sex group*									
	Laps *Child*	Õppur *Student*	Nooruk *Young man*	Neiu *Young woman*	Keskeas Mees *Middle-aged man*	Keskeas Naine *Middle-aged woman*	Vanem Mees *Older man*	Vanem Naine *Older woman*	Eakas *Elderly person*	Keskmine *Mean*
1. mudel *Model 1*										
Teoreetiline 1. liiki vea tõenäosus *Theoretical probability of type 1 error*	0,11	0,02	0,02	0,02	0,02	0,02	0,02	0,02	0,02	0,2
Teoreetiline 2. liiki vea tõenäosus *Theoretical probability of type 2 error*	0,39	0,01	0,47	0,18	0,35	0,01	0,44	0,09		0,5
Empiiriline 1. liiki vea tõenäosus *Empirical probability of type 1 error*	0,15	0,13	0,08	0,10	0,09	0,10	0,07	0,12	0,13	
Empiiriline 2. liiki vea tõenäosus *Empirical probability of type 2 error*	0,01	0,005	0,13	0,04	0,11	0,02	0,13	0,03	0,29	

Tabel 4. Logistiliste mudelite veahinnangud Järg **–** *Cont.*
Table 4. Error estimates of the logistic models

	Soo-vanuserühm *Age-sex group*								
	Laps *Child*	Õppur *Student*	Nooruk *Young man*	Neiu *Young woman*	Keskeas Mees *Middle-aged man*	Keskeas Naine *Middle-aged woman*	Vanem Mees *Older man*	Vanem Naine *Older woman*	Eakas *Elderly person*	Keskmine *Mean*
2. mudel *Model 2*										
Teoreetiline 1. liiki vea tõenäosus *Theoretical probability of type 1 error*	0,03	0,02	0,02	0,02	0,02	0,02	0,02	0,02	0,28
Teoreetiline 2. liiki vea tõenäosus *Theoretical probability of type 2 error*	0,26	0,0001	0,54	0,16	0,45	0,008	0,47	0,15	0,25
Empiiriline 1. liiki vea tõenäosus *Empirical probability of type 1 error*	0,07	0,10	0,05	0,10	0,07	0,12	0,06	0,13	0,44
Empiiriline 2. liiki vea tõenäosus *Empirical probability of type 2 error*	0,02	0,004	0,11	0,03	0,10	0,01	0,11	0,02	0,004

Logistilised mudelid otsustasid residentide hulka paigutada mõnevõrra rohkem potentsiaalseid isikuid (vt tabel 5, kus võrdluse aluseks (100%) on lineaarne mudel 1).

Tabel 5. Mudelite põhjal tehtud otsustuste võrdlus
Table 5. Comparison of judgements made by models

Mudel	Lineaarne 1 *Linear 1*	Lineaarne 2 *Linear 2*	Logistiline 1 *Logistic 1*	Logistiline 2 *Logistic 2*	*Model*
Residente, %	100	84	116	103	*Residents, %*

Kõige „kriitilisemat“ mudelit (Lineaarne 2) prooviti ka kõigi tegelikult loendatud püsielanike andmestikul. Mudel otsustas, et nende hulgas on mitteresidente ligi 10%, mis näitab, et see mudel annab ilmselt nihkega hinnangu. Hästi on omavahel kooskõlas mudelid Lineaarne 1 ja Logistiline 2, kus olulist nihet ei ilmnenud.

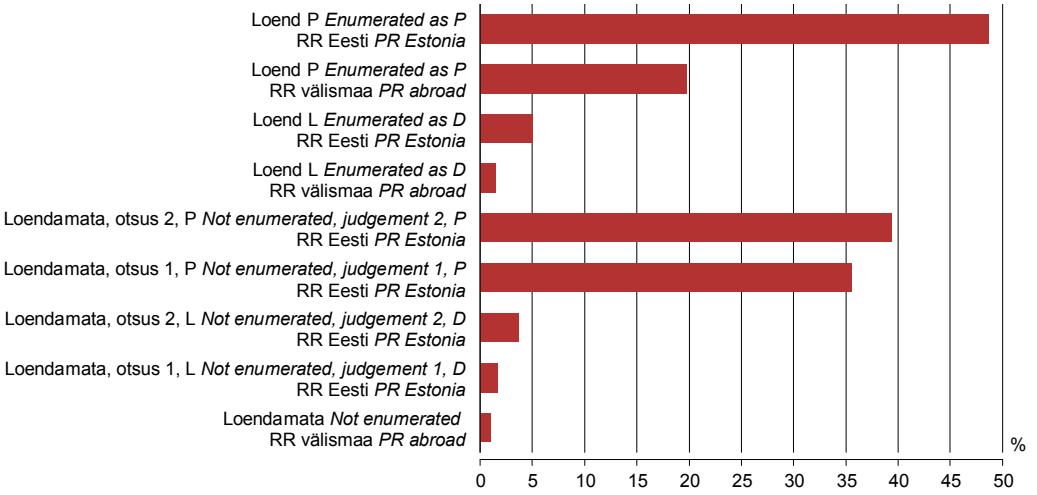
Otsustustulemuste kontrollimine

Järgmine samm korrektses otsustusprotsessis on otsustustulemuste kontrollimine. Praegusel juhul on selleks võimalus kasutada registriandmeid loendusmomendi järel, s.o 2012. aasta alguses. On selge, et kõik püsielanikudki ei ole sellel ajavahemikul registrites aktiivselt osalenud, kuid ootuspäraselt peaks erinevate isikurühmade osalusmäärade erinevus andma täiendavat

teavet isikute Eestis kohal- või eemalviibimise kohta. Kasutada saame siin kaht koondtunnust, neist üks on aktiivsus 2012, mis sisaldab mitme registri andmeid, teine on ravil käik ajavahemikus novembrist 2011 kuni märtsini 2012. Kumbagi neist tunnustest ei ole kasutatud otsustusfunktsioonide koostamisel.

Jooniselt 3 on näha, et loendatud Eesti püsielanikest, kes elavad rahvastikuregistri andmetel Eestis, on registrite andmetel aktiivsed olnud ligi pooled (48%). Peaaegu poole väiksem on registriaktiivsus neil, kes on küll püsielanikena loendatud, kuid kes rahvastikuregistri andmetel Eestis alaliselt ei ela. See on väga väike isikuterühm (ca 0,5%), kuhu kuulub näiteks välistudengeid, ja selle madalat aktiivsust seletavad mitteelaniku seisundiga seotud barjäärid. Lahkunutena loendatud isikud, kes rahvastikuregistri andmetel elavad Eestis, esinevad registrites umbes kümme korda vähem kui püsielanikena loendatud, mis on igati mõistetav. Veelgi harvemini sattusid 2012. aastal registritesse need lahkunutena loendatud, kelle elukoht rahvastikuregistri andmetel ei ole Eestis.

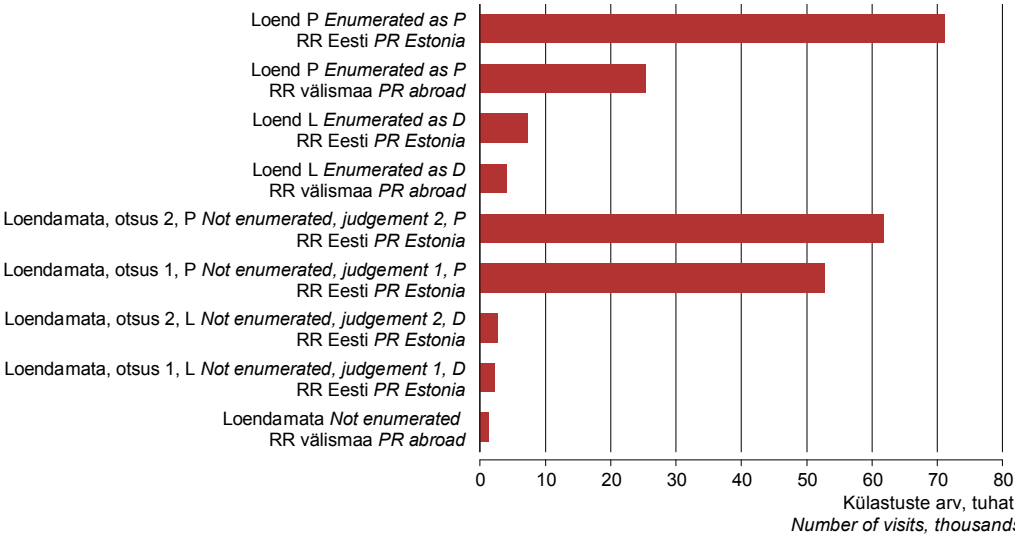
Joonis 3. Isikurühmade suhteline aktiivsus Eesti registrites, 2012. aasta algus
Figure 3. Relative activity of different groups in Estonian registers, the beginning of 2012



Loendamata isikuid kajastab viis alumist tulpa. Kõrgeim, enam kui 80% võrreldes loendatud püsielanikega, on rangema otsustusreegli (lävend 2) alusel püsielanikeks arvatud loendamata isikute registriaktiivsus. Ootuspäraselt on sellest pisut madalam (90%) leebema otsustusreegli (lävend 1) põhjal püsielanike hulka arvatud isikute registriaktiivsus. See, et tegemist on mõnevõrra väiksema registriaktiivsusega võrreldes loendatud isikutega, on sisuliselt mõistetav: loendamata isikud ongi loomuldasa vähem aktiivsed kui loendatud inimesed. Hulka L ehk lahkunute hulka arvatud isikute registriaktiivsus on mõlema otsustusreegli puhul lähedane lahkunutena loendatud rahvastikuregistris olevatele Eesti elanikele. Otsustusreegleid ei ole rakendatud isikutele, kes rahvastikuregistri andmetel ei ole Eesti elanikud. Kuigi 1% ka nende hulgast on esinenud 2012. aastal registrites, võib seda lugeda pigem juhuslikuks, järelikult pole otstarbekas nende hulgast püsielanikke otsida (vt ka joonis 1).

Teine, eelmise tunnusega küll osaliselt seotud, kuid võrdlemisi ilmekas tunnus näitab isikute arstikülastusi (Eesti raviasutustes Eesti kindlustusega) perioodil novembrist 2011 kuni märtsini 2012 (vt joonis 4).

Joonis 4. Aktiivsete haigekassa teenuste esinemissagedus, november 2011 – märts 2012
Figure 4. Frequency of active health insurance services, November 2011 – March 2012



Joonisel 4 esitatakse aktiivsuse andmed samade isikuterühmade kohta nagu joonisel 3 ja ka pilt on üsna sarnane. Veelgi veenvam on rangema otsustusreegliga (lävend 2) püsielanike hulka arvatud isikute käitumise sarnasus loendatud püsielanikega ja kahe isikuterühma – lahkunute ning püsielanike – erinemine. Jooniselt ilmneb, et otsustusreeglite põhjal määratletud rühmad eristuvad registrikäitumise poolest 2012. aastal mõnevõrra selgemini kui loendusandmete põhjal määratletud isikud.

Kahe erineva otsustusreegli (lävend 1 ja lävend 2) erinevus põhjustas rakendamisel tegelikule loendamata rahvastikuregistris olevale Eesti elanike hulgale umbes 1/6 hulka P määratud isikute paigutamist hulka L. Koguandmestiku seisukohast on siiski tegemist vähem kui 0,5%-ga elanikkonnast.

Järeldus

Kasutades Eesti registreid, eriti rahvastikuregistrit, ning usaldusväärset matemaatilise statistika metoodikat, on võimalik Eesti rahvastiku üldkogumi hinnangut oluliselt parandada, lisades 2011. aasta rahva ja eluruumide loenduse käigus loendatud isikutele isikud, kes jäid loendamata, kuid kes suure tõenäosusega elavad Eestis (loenduse alakaetus).

Metoodiliselt on otstarbekas:

- kasutada õpperühmadele tuginevat otsustusreeglit (diskriminantanalüüsi), mille korral valitakse kõikvõimalikest registritunnustest programmiliselt välja sobivaimad;
- moodustada eraldi otsustusreeglid sobivalt valitud soo-vanuserühmades;
- kasutada esimest liiki viga piiravat otsustusreeglit tagamaks tulemuse suurem usaldatavus n-ö tundlike järelduste suhtes;
- rakendada otsustusreeglit ainult rahvastikuregistris olevatele Eesti elanikele, kes ei ole loendatud (erandina ka nendele, kelle puhul loendustulemused on vasturääkivad);
- esitada valitud ja rakendatav otsustusreegel avalikult (trüki- ja sotsiaalmeedias) ja kommenteerida seda üksikasjalikult;
- tulemusi esitades anda ka veahinnangud.