TARTU ÜLIKOOL

Sotsiaalteaduste valdkond

Ühiskonnateaduste instituut

Ühiskonna ja infoprotsesside analüüsi õppekava

Henry Lass

Andmekvaliteedi hindamise protsessi automatiseerimine ja selle olulisus organisatsioonile

Magistritöö

Juhendaja(d): Toomas Saarsen,PhD

Terje Trasberg, PhD

Tartu 2025

SISUKORD

SISSEJUHATUS

Uuringud (Batini & Scannapieco, 2016; Rahm & Do, 2000; Otto, 2011) on üha enam keskendunud andmete kvaliteedi tagamise ja kontrollimise meetodite automatiseerimise võimalustele ja väljakutsetele organisatsioonides. Andmekvaliteedi mõistmine on organisatsioonide tõhususe ja konkurentsivõime edendamiseks oluline, sest kvaliteetsed ja usaldusväärsed andmed võimaldavad teha informeeritud juhtimisotsuseid, vähendavad vigadest tulenevaid riske ja aidates optimeerida ressursikasutust (Etikala, 2025). Eesti kontekstis on usaldusväärsed ja ajakohased andmed olulised näiteks Statistikaametile, mille olemus on seista kvaliteetsete andmete eest kogu riigis ning pakkuda järjepidevalt olulist, usaldusväärset ja tähendusega informatsiooni (Statistikaameti kodulehekülg, 2025).

Andmete kvaliteedi probleem on laialdaselt täheldatud nii avalikus sektoris kui ka erasektoris, kus organisatsioonid koguvad ja töötlevad järjest suuremaid andmemahte ning seisavad silmitsi vajadusega tagada andmete õigsus, terviklikkus, järjepidevus ja ajakohasus. Mitmed autorid (Rahm & Do, 2000; Otto, 2011; Etikala, 2025) on väitnud, et automatiseeritud andmekontrollid aitavad oluliselt vähendada tööjõukulusid ning tõstavad protsesside kiirust ja kvaliteeti.

Kuigi varasemad uuringud on käsitlenud automatiseeritud andmekvaliteedi kontrollimeetodeid (Batini & Scannapieco, 2016; Otto, 2011), ei ole need piisavalt keskendunud andmestike struktuuri ja organisatsioonispetsiifiliste eripärade mõjuga meetodite valikule ja rakendatavusele. Viidatud uuringute tulemused tõstatavad uusi kriitilisi küsimusi, näiteks millised organisatsioonilised ja andmete struktuuriga seotud tegurid määravad ära meetodite sobivuse ja kuidas arendada paindlikke lahendusi, mida saab kohandada erinevaid vajadusi või eesmärke silmas pidades. Lisaks, kuigi Rahm & Do (2000) toetavad universaalse lahenduse otsimist, näitavad hilisemad tõendid (Otto, 2011), et pigem on vaja paindlikku ja kontekstitundlikku lähenemist. Selleks, et keskenduda organisatsioonispetsiifiliste eripäradest tulenevate meetodite valikule ning rakendatavusele, on töös võetud näiteks Statistikaamet, kus erinevaid automatiseeritud andmekontrolli meetodeid sõltuvalt organisatsiooni eripäradele katsetada.

Magistritöö eesmärk on katsetada, milliseid automatiseeritud andmekvaliteedi kontrolli- ja puhastusmeetodeid sobib Statistikaameti kontekstis rakendada erineva struktuuriga andmestike puhul ning töötada välja andmekontrolli raamistik ja praktiline lahendus, mis võimaldab meetodeid paindlikult kohandada vastavalt organisatsiooni ja andmekomplekti eripäradele.

Uurimiseesmärgi täitmiseks on töös püstitatud kaks empiirilist uurimisküsimust:

1. Millised on kasutegurid, mida automatiseeritud andmekontrollid pakuvad Statistikaameti näitel?
2. Milliseid organisatsiooni eripärasid tuleb arvestada andmete kontrollimise meetodite rakendamisel?

Magistritöö koosneb viiest peamisest peatükist: teoreetiline raamistik, probleemiseade ja metoodika, praktilise osa kirjeldus ning viimaks tulemused ja arutelu. Esmalt antakse ülevaate olemasolevast kirjandusest andmekvaliteedi ja automatiseeritud kontrollimeetodite ning andmekontrolli protsessi teemal. Seejärel kirjeldatakse töös rakendatavat metoodikat, sealhulgas andmekvaliteedi kontrollimeetodeid, mida reaalselt katsetati kvaliteediraportite koostamiseks ja viimaks tutvustatakse selles peatükis analüüsimeetodeid. Empiirilises osas esitatakse kvalitatiivse ja kvantitatiivse analüüsi tulemused, tuvastatakse valitud kontrollimeetodite tugevused ja piirangud ning esitatakse välja töötatud raamistik. Arutelu ning tulemuste peatükis vastatakse püstitatud uurimisküsimustele. Töö annab panuse nii teaduslikult, pakkudes metoodilisi soovitusi andmekvaliteedi kontrolliks, kui ka praktiliselt, aidates organisatsioonidel kavandada ja ellu viia andmekvaliteedi automatiseerimist ja juhtimist.

1. KIRJANDUSE ÜLEVAADE

Kirjanduse ülevaates avatakse andmetöötluse ja -kvaliteedi temaatikat. Selgitatakse, miks on andmed, täpsemalt andmete kvaliteet oluline. Antakse ülevaate erinevatest probleemidest, millega võivad andmekasutajad kokku puutuda ning millised on varem uuritud ja rakendatud meetodid, mis on end tõestanud andmete kvaliteedi hindamisel.

1.1 Andmete kvaliteedi hindamise eesmärk

Andmepõhises maailmas on andmekvaliteedi hindamine saanud vältimatuks eelduseks usaldusväärsete ja tõhusate otsuste tegemisel (Charles, 2024). Kasvavad andmemahud ja andmete kogumise mitmekesisus on muutnud kvaliteedi tagamise järjest keerulisemaks, mistõttu organisatsioonid, kes soovivad teha kvaliteetseid andmepõhiseid otsuseid, peavad pöörama tähelepanu kasutatavate andmete terviklikkusele ja järjepidevusele (Charles, 2024; Batini & Scannapieno, 2016). Mida rohkem kasutatakse erinevatest allikatest pärit andmeid, seda raskem on säilitada nende ühtset struktuuri ja kvaliteeti (Batini & Scannapieno, 2016). Samas pakub allikate mitmekesisus ka võimalust andmete võrdlemiseks ja seeläbi kvaliteedi parendamiseks, valides analüüsiks parimad võimalikud andmekomplektid (Batini & Scannapieno, 2016).

Andmekvaliteedi hindamine hõlmab küsimusi andmete täielikkuse, õigsuse ja kasutuskõlblikkuse kohta (Sügis, Tampuu, Aljanaki, Fišel & Kull, 2025). Kvaliteedi hindamiseks tuginetakse sageli erinevatele aspektidele, nagu täpsus, korrektsus, täielikkus, asjakohasus, minimaalsus, selgus, ligipääsetavus, järjepidevus, kasulikkus ja usaldusväärsus (Batini & Scannapieno, 2016). Need dimensioonid võimaldavad andmeid süsteemselt hinnata, võttes arvesse nii tehnilisi, kui ka sisulisi kriteeriume. Hindamise oluline osa on konteksti mõistmine: näiteks võib meditsiiniekspert tuvastada andmetes puudusi, mida tehniline kontroll ei märka (Sügis jt, 2025).

Magistritöö eesmärgi valguses on andmekvaliteedi hindamise mõistmine esmatähtis, kuna erineva struktuuriga andmestike puhul võivad olulised olla erinevad kvaliteedidimensioonid. Selleks, et välja töötada paindlik andmekontrolli raamistik, on vaja mõista, millised kvaliteedinõuded on

erinevates olukordades prioriteetsed ning kuidas neid hinnata (Charles, 2024; Batini & Scannapieno, 2016).

1.2 Andmete kontrollimise meetodid

Andmekvaliteedi hindamiseks kasutatakse erinevaid kontrollimeetodeid, mis jagunevad laias laastus süntaktilisteks ja semantilisteks kontrollideks (Charles, 2024; Batini & Scannapieco, 2016). Süntaktiline kontroll keskendub sellele, kas andmed vastavad etteantud struktuurilistele reeglitele, näiteks kas kuupäevaväljad sisaldavad korrektseid kuupäevi või kas numbrilised väljad on tõepoolest numbrilised (Batini & Scannapieco, 2016). Praktikas on süntaktilist kontrolli lihtsam läbi viia, sest sageli piisab teadaolevate lubatud väärtuste komplektiga võrdlemisest (Batini & Scannapieco, 2016).

Semantiline kontroll seevastu hindab väärtuste sisulist õigsust, näiteks kas inimese sünniaeg on kooskõlas muude andmetega tema kohta (Batini & Scannapieco, 2016). Semantilise kontrolliga proovitakse teha selgeks, kas andmestikus esinevad väärtused on kooskõlas sellele vastava päriselu olukorraga (Ahiagble & Stein, 2022). Semantiline kontroll nõuab sügavamat mõistmist ja konkreetsete väärtuste teadmist, mistõttu on sellised kontrollid töömahukamad ja keerulisemad (Batini & Scannapieco, 2016).

Kuna käesoleva magistritöö eesmärgiks on katsetada, millised on sobivad automatiseeritud kontrolli- ja puhastusmeetodid erineva struktuuriga andmestike jaoks Statistikaametis, peab välja selgitama, millised kontrollitüübid on automaatseks rakendamiseks sobivamad ning kuidas nende tulemusi kontekstipõhiselt tõlgendada (Charles, 2024).

Automaatse kontrolliga kaasnevad ka mitmed puudused. Kontrollisüsteemide ülesseadmine on üldjuhul väga kulukas ja keerukas (Charles, 2024). Semantiliste automaatsete kontrollide ülesseadmine nõuab üldjuhul täiendavat ressurssi töötajalt või eksperdilt, kellel on vastava valdkonna sügavad teadmised (Ahiagble & Stein, 2022). Valdkonna teadmiste nõude tõttu on tõenäoliselt selliseid automatiseerimise projekte keerulisem sisse osta. Lisaks sellele, et kontrollide ülesseadmine on aeganõudev ning keeruline võib näiteks liiga rangelt seadistatud kontroll genereerida palju veateateid, mis nõuavad inimeste tähelepanu (Rahm & Do, 2000). Samuti ei pruugi loodud kontrollreeglid olla piisavalt paindlikud, et arvestada andmete konteksti või eripäradega (Charles, 2024). Muutuvate andmestruktuuride ja ärivajaduste tõttu tuleb automaatseid kontrolle pidevalt kohandada, et need säilitaksid oma tõhususe (Charles, 2024).

Just nende aspektide mõistmine on oluline, et magistritöö raames katsetatavad andmekontrolli meetodid ja uus kontrolliprotsess võimaldaks automaatsete meetodite paindlikku ja vajaduspõhist kasutust (Charles, 2024; Batini & Scannapieco, 2016).

1.2.1 Andmekvaliteedi kontrollimise automatiseerimine

Käsitsi andmete kontrollimine on suurte andmemahtude puhul väga ressursi- ja ajakulukas ning veaohtlik, on fookus üha enam liikunud automaatsetele kontrollimeetoditele (Charles, 2024; Epperson jt, 2023). Automaatse andmekontrolli puhul kasutatakse tarkvaralisi tööriistu, mis aitavad valideerida andmeid etteantud reeglite ja mustrite põhjal ilma, et inimene peaks protsessi sekkuma (Charles, 2024). Lisaks toetavad paljud automaatsed andmekvaliteedi tööriistad ülevaadet ja monitooringut kontrolli kvaliteedi osas, mis võimaldavad reeglite uuendamist vastavalt vajadusele, näiteks uute andmete saabumisel, mis on varasemast erinevad ning olemasolevad kontrollid ei ole enam tõhusad (Zhou jt., 2024). Siit selgub, et kuigi liigutakse üha enam automaatsemate kontrollide poole, on tarvis, et loodud kontrollid oleksid siiski paindlikud ja võimaldaksid vajadusel kontrollimeetodeid täpsustada või täiendada (Ahiagble & Stein, 2022).

Automaatse andmekontrolli peamised eelised on seotud efektiivsuse, skaleeritavuse, järjepidevuse ja kulutõhususega (Charles, 2024). Suurtes andmemahtudes võimaldab automaatne kontroll läbi viia sadu või tuhandeid kontrolle kiiresti ja ilma tööjõukulu lineaarse kasvuta (Charles, 2024). Sama kinnitavad ka Epperson jt. (2024), et automaatne andmete profileerimine genereerib enamike kasutajate endi loodud ülevaadetest automaatselt, kiirendades analüüsiprotsessi ning vähendades vajadust käsitsi koodi kirjutamiseks. Samuti välistab automaatne kontroll subjektiivsed vead, mida inimene võib teha, rakendades rangelt ja järjepidevalt samu reegleid igal korral (Charles, 2024).

Automatiseeritud andmekontrollide hulka kuulub mitmesuguseid meetodeid, millest olulisemad on:

1. Probleemsete väärtuste tuvastamine

- **Puuduvate väärtuste analüüs**, mis on sisuliselt süsteemne kontroll, mille eesmärk on selgitada välja, kas puuduvate andmete jaotus järgib mingit loogikat või viitavad puuduvad väärtused vigadele või puudustele andmetes (Sügis jt, 2024).
- **Unikaalsuse kontroll**, mille eesmärk on tuvastada andmetes duplikaadid ja kindlustada, et iga reaalne üksus oleks andmetes esindatud vaid ühe kirjena (Batini & Scannapieco, 2016; Statistikaamet, 2020).
- **Tunnuste jaotuse kontroll**, ehk automaatne võrdlus eeldatud ja tegeliku jaotuse vahel, näiteks normaaljaotusega võrdlemine (Sügis jt, 2024).
- **Volatiilsuse kontroll**, mis on sarnane eelnevale, tunnuste jaotuse kontrollile, kuid selle eesmärk on avastada ootamatuid kõikumisi andmetes (Ahiagble & Stein, 2022).
- **DBSCAN-klasterdamine** moodustab sarnaste andmepunktidega klastrid, eesmärk leida anomaaliad või erindeid (Ahiagble & Stein, 2022).

2. Struktuuri ja formaadi kontrollimine

- **Formaadikontroll**, kus hinnatakse, kas andmeväljade väärtused vastavad etteantud tehnilistele nõuetele (Charles, 2024; Statistikaamet, 2020).
- **Andmestiku struktuuri vastavuse kontroll**imisel hinnatakse andmestiku struktuuri vastavalt etteantud reeglitele ja ootustele, see kontroll aitab tagada andmekoosseisu korrektsuse (Batini & Scannapieco, 2016).

3. Sidususe kontroll

- **Koodide valideerimine** võimaldab teha kindlaks, kas sisestatud väärtused vastavad etteantud koodiloendile (Charles, 2024).
- **Ärireeglite kontroll**, millega tuvastatakse ja kontrollitakse andmetes esinevad seosed ning luuakse andmekvaliteedi valideerimiseks reeglid ja kontekst (Ahiagble & Stein, 2022).
- **Väärtusvahemike kontroll**, mis kontrollib arvuliste tunnuste puhul, kas arvulised või tekstilised väärtused jäävad lubatud piiridesse (Statistikaamet, 2020).

- **Kategooriate täielikkuse kontroll**, mille eesmärk on teha kindlaks, kas kategoriaalsete tunnuste väärtused vastavad lubatud väärtuste loetelule (nt sugu: ainult „mees“ või „naine“) (Sügis jt, 2024).
- **Andmete järjepidevuse kontroll ajas**, mille puhul jälgitakse, kas väärtuste muutused eri ajahetkedel on loogilised ja ootuspärased (Statistikaamet, 2020; Batini & Scannapieco, 2016).
- **Kohustuslike väljade täielikkuse kontroll**, mille kaudu hinnatakse puuduvate väärtuste olemasolu ja mõju (Statistikaamet, 2020). Sarnast semantilist kontrolli on rakendanud ka Ahiagble ja Stein (2022), et kontrollida andmestiku kontekstis oluliste tunnuste täielikkust.

Mainitud meetodite rakendamisel tuleb arvestada, et erinevatest allikatest pärit andmestikud on ülesehituselt, kui ka sisult väga erinevad. Seetõttu on selle töö kontekstis oluline pöörata tähelepanu just neile kontrollimeetoditele, mis on paindlikud erinevatele andmestikele rakendamiseks. Mõned esile toodud automaatsed kontrollimeetodid on liiga jäigad, näiteks kohustuslike väljade täielikkuse kontroll eeldaks seda, et iga andmestiku puhul on vaja seadistada, milliste väljade puhul konkreetset kontrolli teostama peab. Selle asemel on mõistlik teostada kõigi tunnuste puhul puuduvate väärtuste analüüs, mis tagab, et kõigi kohustuslike väljade kontroll on tehtud, ilma, et peaks iga andmestiku puhul kontrolli eraldi seadistama. Paljude meetodite puhul on kontrollimisel esile toodud konteksti vajadus, näiteks väärtusvahemike kontrolli puhul on tarvis aru saada, mis on lubatud piirid, millesse peavad kontrollitavad väärtused jääma. Samas on sellist süntaktilist kontrolli lihtsam läbi viia, sest piisab teadaolevate lubatud väärtuste komplektiga võrdlemisest (Batini & Scannapieco, 2016).

Eelneva kokkuvõtteks on kirjanduses esile toodud kaks peamist andmete kontrollimise tüüpi - semantilised ja süntaktilised kontrollid (Charles, 2024; Batini & Scannapieco, 2016). Sageli tuginetakse andmete kvaliteedi hindamiseks omakorda erinevatele dimensioonidele, nagu täpsus, korrektsus, täielikkus, asjakohasus, minimaalsus, selgus, ligipääsetavus, järjepidevus kasulikkus ning usaldusväärsus, mis jagunevad semantilisteks ja süntaktilisteks kontrollideks (Batini & Scannapeio, 2016).

Süntaktilised kontrollid seatakse üles tagamaks andmestiku struktuuri ja formaadi (kuupäevaliste tunnuste vormistus, lubatud väärtuste kontroll) ning selliste kontrollide automatiseerimine on lihtsamini teostatav, sest üldjuhul on teada reeglid või mustrid, millele andmestik vastama peab.

Semantilised kontrollid aga hindavad andmete sisulist õigsust, näiteks väärtuste kontekstile vastavalt loogiline jaotus või kuupäeva tunnuse kooskõla muude tunnustega andmestikus (näiteks kui isik on väga noor ning talle on andmestikus märgitud ebaloogiliselt kõrge haridustase või on sünnikuupäev hiljem kui abiellumise kuupäev) (Batini & Scannapieno, 2016). Semantilised kontrollid on kontekstispetsiifilisemad ning nõuavad ärireeglite või muu sarnase seadistamist, sest eelnevalt toodud näide noore inimese kõrgest haridusest ei ole tehniliselt tuvastatav probleem - isik võib olla sündinud sel ajal, mis on andmetes märgitud ning samuti on võimalik, et isikul on andmetes märgitud haridustase, kuid sisuliselt ei ole kirje õige. Selliste kontrollimeetodite rakendamine nõuab seega valdkonnast tulenevaid reegleid ning üldjuhul ka inimese sekkumist või tõlgendamist.

Järgnevalt on toodud Tabel 1, kus on koondatud kokku, millise kvaliteedidimensiooni puhul on tegu semantilise või süntaktilise kontrolliga, milline on automatiseerimise rakendatavus ning millist automatiseeritud andmekontrollimeetodit konkreetse dimensiooni kontrollimiseks rakendada. Automatiseeritavuse hinnang on antud põhinedes sellele, kuivõrd on tarvis lisaks kontrollimeetodi rakendamisele siiski inimese sekkumist kvaliteedidimensiooni vastavuse kinnitamiseks.

Tabel 1. Kvaliteedidimensioonid ning neile vastavate automatiseeritud kontrollimeetodite ülevaade.

Kvaliteedidimensioon	Kas süntaktiline või semantiline?	Kontrollimeetod	Automatiseeritavus
Täpsus	Süntaktiline	• Formaadikontroll	Kõrge
	Semantiline	• Tunnuste jaotuse kontroll • DBSCAN-klasterdamine	Keskmine
Korrektsus	Semantiline	• Koodide valideerimine • Väärtusvahemike kontroll • Ärireeglite kontroll	Keskmine
Täielikkus	Süntaktiline	• Puuduvate väärtuste analüüs	Kõrge
	Semantiline	• Kohustuslike väljade täielikkuse kontroll	Keskmine
Asjakohasus	Semantiline	• Tunnuste jaotuse kontroll (eeldab eelmääratud kvooti, millele peab vastama)	Madal
Minimaalsus	Süntaktiline	• Unikaalsuse kontroll	Kõrge
Selgus	Semantiline	• Kategooriate täielikkuse kontroll	Keskmine
Ligipääsetavus	Süntaktiline	• Formaadikontroll	Kõrge
	Semantiline	• Andmestiku struktuuri kontrollimine (vastavalt reeglitele või ootustele)	Kõrge
Järjepidevus	Semantiline	• Andmete järjepidevuse kontroll ajas • Volatiilsuse kontroll	Keskmine
Kasulikkus	Semantiline	• Tunnuste jaotuse kontroll (vastavalt eeldustele)	Madal
Usaldusväärsus	Süntaktiline	• Unikaalsuse kontroll	Keskmine
	Semantiline	• Koodide valideerimine	Keskmine

Allikad: Batini & Scannapieno, 2016; Statistikaamet, 2020; Ahiagble & Stein, 2022; Charles, 2024; Sügis jt,, 2025

Meetodite valikul on tarvis alustada sellest, mis aspektid on automatiseeritavad ning mis on prioriteetne konkreetses organisatsioonilises või uurimuse kontekstis. Kõrge automatiseeritavusega kontrollimeetodid on formaadikontroll, unikaalsuse kontroll, puuduvate väärtuste analüüs ning andmestiku struktuuri kontrollimine. Kolme esimese meetodi puhul on tegu

süntaktilise kontrolliga ning viimane, andmestiku struktuuri kontrollimine (just võttes arvesse ootuseid või reegleid, mis andmestikule seatakse) liigitub semantilise kontrolli hulka, arvestades, et struktuuri kontrollimise puhul on tarvis teada, millise struktuuriga andmestik olema peab ning selleks on vaja, et oleks olemas dokumentatsioon andmestiku kohta. Kuna on oluline arvestada ka andmestiku kontrollimise eesmärkidega, näiteks andmestiku kasulikkuse või asjakohasusega, mille puhul on automatiseeritavus madal, on paratamatu, et tuleb ka nende kvaliteedidimensioonide kontrollimisega tegeleda ning seega kasutada selliseid kontrollimeetodeid, mille puhul täielikku automatiseerimist ei saa realiseerida.

1.3 Andmete kontrollimise protsess

Andmekontrolli protsess on süsteemne ja mitmeetapiline tegevus, mille eesmärk on tagada andmete kvaliteet kogu andmetega töötamise elutsükli vältel (Anaconda Foundation, 2020; Epperson, Gorantla, Moritz & Perer, 2023). Tõhus andmekontroll ei tähenda üksnes konkreetsete vigade leidmist, vaid hõlmab ka põhjalikku andmete tundmaõppimist, probleemkohtade kaardistamist ja andmekvaliteedi parandamist enne, kui andmeid kasutatakse analüüsiks või mudelite treenimiseks.

Andmekontrolli protsessi võib üldiselt kirjeldada järgmiste etappidena:

1.3.1 Andmete profileerimine

Esimene samm andmekontrolli protsessis on andmete profileerimine, mille käigus analüüsitakse andmestikke, et mõista nende struktuuri, väärtuste jaotust, tüpoloogiat ja võimalikke anomaaliaid (Epperson jt, 2023; Anaconda Foundation, 2020). Profileerimine hõlmab näiteks: puuduvate väärtuste tuvastamist, erindeid ehk tavapärasest väga palju erinevaid väärtuseid (Rootalu, 2014) ning loogiliste vastuolude leidmist, väärtuste jaotuse ja sageduste hindamist ning andmetüüpide ja formaadi vastavuse kontrollimist. Profileerimise eesmärk on luua esmane ülevaade andmete kvaliteedist ja suunata tähelepanu võimalikele probleemidele, mis vajavad edasist töötlemist (Epperson jt, 2023).

1.3.2 Kontrollireeglite määramine

Järgmise sammuna määratletakse vastavalt andmekirjeldustele ning vajalikud andmekontrolli reeglid ja seatakse paika automaatsed või poolautomaatsed kontrollimeetodid (Charles, 2024). Siin sammus otsustatakse, milliseid konkreetseid kontrollimeetodeid (unikaalsuse kontroll, formaadikontroll, väärtusvahemike kontroll jne) rakendatakse vastavalt andmestiku struktuurile ja organisatsiooni vajadustele (Batini & Scannapieco, 2016; Statistikaamet, 2020).

Kontrollireeglite määramisel on oluline silmas pidada, et erinevate andmetüüpide ja eesmärkide puhul võivad vajalikud kontrollid olla erinevad, ehk ei saa lähtuda vaid ühest kontrolliprotseduurist. Näiteks struktureeritud andmete puhul on võimalik rakendada ulatuslikku formaadikontrolli, kuid poolstruktureeritud andmete puhul tuleb rohkem toetuda loogilistele mustritele ja semantilisele valideerimisele (Charles, 2024).

1.3.3 Kontrolli läbiviimine

Kui eelnevad sammud on läbitud, viiakse need andmestiku peal läbi. Automaatsete tööriistade abil kontrollitakse, kas andmed vastavad etteantud reeglitele, tuvastades puudused ja vastuolud (Charles, 2024; Kandel, Parikh, Paepcke, Hellerstein & Heer, 2012). Automaatsete tööriistade kasutamine võimaldab genereerida kiiresti kokkuvõtteid, näiteks tunnuste väärtuste järjepidevuse ülevaadet, jaotushistogramme kvantitatiivsete veergude jaoks või sagedustabeleid kategooriliste tunnuste puhul (Kandel jt, 2012; Epperson jt, 2023; Siddiqui, Kim, Lee, Karahalios & Parameswaran, 2016). Selle etapi tulemusel saadakse nimekiri andmekvaliteedist, kust ilmnevad võimalikud probleemid, mis vajavad kas automaatset parandamist või edasist analüüsi.

1.3.4 Probleemihaldus

Kontrolli läbiviimise tulemusel tuvastatud probleemide põhjal viiakse läbi edasised analüüsid ning vajadusel parandused, näiteks duplikaatides eemaldamine, puuduvate väärtuste täitmine või parandamine, formaadi normaliseerimine ja anomaaliate käsitlemine (Batini & Scannapieco, 2016; Statistikaamet, 2020). Probleemihalduse etapis on seega oluline hoida selge ülevaade tehtud muudatustest ja vajadusel dokumenteerida kõik olulisemad andmetes tehtud muudatused.

1.3.5 Kontrolli kordamine

Pärast probleemihalduse etappi viiakse kontrollid uuesti läbi, et veenduda, kas probleemid andmetes on lahendatud ja kas andmete kvaliteet vastab soovitud tasemele (Charles, 2024). Andmete kvaliteedi jälgimine ei lõpe ühekordse kontrolli ja puhastamisega: eriti suurte või pidevalt uuenevate andmestike puhul peab kontrolliprotsess olema pidev ja korduv (Heer, 2019; Perkel, 2018). Automatiseeritud andmekvaliteedi kontrollimise süsteemid võimaldavad luua dünaamilisi lahendusi, mis tuvastavad kvaliteediprobleeme andmestike muutumisel.

2. PROBLEEMISEADE

Organisatsioonides toodetakse ja kasutatakse tänapäeval järjest rohkem andmeid, mida kasutatakse organisatsiooni efektiivsuse, tulemuslikkuse või kasumi tõstmise eesmärgil, seega on andmete usaldusväärsus, kvaliteet ja ajakohasus järjest olulisemad (Batini & Scannapieco, 2016). Eestis on üheks organisatsioon, kus kasutatakse väga palju andmeid, Statistikaamet - kogutakse, töödeldakse ja toodetakse suurel hulgal andmeid. Statistikaameti eesmärk on seista selle eest, et Eesti andmed oleksid usaldusväärsed, kvaliteetsemaid ja ajakohased (Statistikaameti koduleheküIg, 2025). Selleks, et suurel hulgal ajakohaseid ja kvaliteetseid andmeid koguda, töödelda ja kasutada, on oluline leida viise automatiseerimiseks ja anda manuaalsem osa tööst masinale. Üks võimalik automatiseeritav protsess töös andmetega on andmete kvaliteedi kontrollimine (Charles, 2024). Kuigi automatiseeritud andmekvaliteedi kontrollimine on mitme uuringu kohaselt (Rahm & Do, 2000; Otto, 2011; Etikala, 2025) toonud kaasa protsessi kiirenemise ja tööjõukulude vähenemise, tõdevad ka Rahm ja Do (2000) enda töös, et ühetaolise, universaalse lahenduse väljatöötamine on keeruline, sest kontrolle on tihti tarvis rakendada eripäraste andmekomplektide ning organisatsioonide kontekstis. Seega, organisatsioonilised ning andmete sisu- või kontekstipõhised eripärad mõjutavad seda, milliseid meetodeid saab rakendada andmekvaliteedi kontrollimiseks (Otto, 2011). Statistikaameti puhul on üheks selliseks eripäraks asjaolu, et andmed, mida kasutatakse on pärit paljudest erinevatest välistest andmekogudest ning enamike andmetes esinevate probleemide korral peab Statistikaamet pöörduma andmed esitanud andmekogu poole, seetõttu ei ole võimalik kasutada andmekvaliteedi kontrollimiseks meetodeid, mille eesmärk on andmeid automaatselt puhastada või muuta. Meetodid, mida saab Statistikaametis kasutada, peavad võimaldama kasutajal saama selge ülevaate esinevatest probleemidest, et neid probleeme hõlpsasti mõista ja andmekoguga kommunikeerida. Kui ei saa rakendada selliseid meetodeid, mis andmeid puhastavad või muudavad, on tarvis teada, mis on kasutegur, mida automatiseeritud andmekvaliteedi kontrollide rakendamine Statistikaametile pakub. Siiski, et mõista, kuidas kujundada laialt kohaldatav raamistik, mis sisaldab erinevate meetodite kombineerimist vastavalt organisatsiooni või andmekomplekti iseärasustele ilma, et peaks iga juhtumi korral uut lahendust välja töötama, on vaja uurida, milliseid meetodeid saab

erinevat tüüpi andmete puhul rakendada ning kuidas on neid meetodeid arvesse võttes võimalik luua kontrolliraamistik.

Töö eesmärk on katsetada andmekvaliteedi kontrolli- ja puhastusmeetodeid, mida sobib Statistikaametis rakendada eri tüüpi andmete puhul, et tagada andmete kvaliteet ning kujundada kontrolliprotsess, mis toetaks meetodite kohandamist vastavalt organisatsiooni ja andmetest tulenevatele vajadustele. Selleks, et eesmärki täita, on töös kasutatud Statistikaametile laekuvaid erineva struktuuriga registriandmeid, millele on rakendatud valitud andmekvaliteedi kontrollimise meetodeid.

Töös uuritakse ja vastatakse järgnevatele uurimisküsimustele:

1. Millised on kasutegurid, mida automatiseeritud andmekontrollid pakuvad Statistikaameti näitel?
2. Milliseid organisatsiooni eripärasid tuleb arvestada andmete kontrollimise meetodite rakendamisel?

3. MEETOD

Metoodika on automatiseeritud andmekvaliteedi kontrollimeetoditele ja -protsessile keskendunud, mille käigus iga kontrollitava andmestiku kohta genereeritakse kvaliteediraport. Automatiseeritud andmekontrolli meetodid on kirjanduse ülevaates esitatud lähenemisviisidel põhinevad, tagades seeläbi teoreetiliselt ja praktiliselt usaldusväärse aluse kasutatavate analüüsimeetodite jaoks. Seejuures osutusid valituks automatiseeritud kontrollid, mis võimaldavad ressurssi kokku hoida ning tagada standardiseeritud protsessi. Kontrollimise protseduuri ja tehnilise lahenduse töötas välja töö autor käesoleva töö raames selleks, et rakendada akadeemilisi teadmisi praktilises kontekstis ühes asutuses. Automatiseeritud andmekontrolliga analüüsitakse iga andmestiku põhjal mitmeid kvaliteedinäitajaid (näiteks täpsus, terviklikkus, asjakohasus ja järjepidevus) ning koostatakse nende alusel kvaliteediraport. Tulemuseks saadud raportid annavad detailse ülevaate andmestiku tugevustest ja puudustest ning võimaldavad süsteemset hindamist vastavalt kehtestatud kvaliteedinäitajatele nagu puuduvate väärtuste arv andmestikus ja kirjete unikaalsus ning andmete terviklikkus.

Andmed, mille peal katsetatakse erinevaid kontrolli- ja puhastusmeetodeid, on Statistikaametis olemasolevad andmed, mis on erineva struktuuriga ning sisuga. Kättesaadavad andmed võimaldavad rakendada erinevaid kontrollimeetodeid ning näha, kuidas need töötavad ning millist kasu neist on. Esimesele uurimisküsimusele vastamiseks analüüsitakse varasematel aastatel asutusse esitatud andmetes esinenud vigu, et võtta nende kontrollimine aluseks uue kontrolliprotsessi kujundamisel. Uue kontrolliprotsessi kujundamiseks analüüsitakse lisaks andmekvaliteediga seotud varasemaid uurimusi ning katsetatakse varem leitud meetodeid reaalsetel Statistikaametisse laekuvatel andmetel. Dokumenteeritakse ja võrreldakse meetodite sobivust, tulemuslikkust ja kulutõhusust erinevate andmetüüpide korral. Lisaks on kasulikkuse hindamiseks kättesaadavad andmed kontrollile kulunud aja ja leitud vigade kohta läbi viidud katsetuse uue automatiseeritud andmekontrolli protsessiga ning võrdluseks on olemas ka aasta varasemad andmed ajakulu kohta, kui tehti veel andmekontrolle enne uue lahenduse katsetamist. Ajaandmete analüüsiks koondati andmed ühte andmestikku ning arvutati erinev kirjeldav statistika, näiteks aritmeetiline keskmine, mis on üldlevinud statistiline näitaja, ning mediaan, mis

ei ole niivõrd mõjutatud mõne andmestiku väga lühikesest või pikast kontrollimise ajast (Rootalu, 2014). Kasulikkuse ning piirangute kohta küsiti tagasiside ka ekspertidelt, kes seda lahendust proovida said. Tagasiside andmine oli vabatahtlik ning kirjalik. Tagasiside puhul esitati küsimused, mis oli ekspertide arvates uue lahenduse puhul kasutegurid ja millest tuntakse uue lahenduse puhul puudust. Kirjalik tagasiside saadi kahelt eksperdilt, nende vastuste põhjal on teostatud horisontaalne sisuanalüüs, mis võimaldab kirjalike tagasisidede seast koondada kokku kõik konkreetse teema kohta käiva informatsiooni, seega on võimalik saadud tagasiside põhjal koondada uue automatiseeritud lahenduse kasutegurid ja puudujäägid (Kalmus, Masso, Linno, 2015).

Teisele uurimisküsimusele vastamiseks on sisend samuti peamiselt Statistikaametist - andmestikud, mida on tarvis kontrollida, olemasolev praktika ja protsess, varasemad levinumad kvaliteediprobleemid andmetes ning käesoleva töö käigus automatiseeritud protsessi osa. Analüüsiti, millised tegurid mõjutavad erinevate andmekontrolli meetodite rakendatavust ja kasutusele võtmist organisatsioonis. Samuti, mil määral mõjutab meetodite valikut organisatsiooni laekuvate või organisatsioonis kasutatavate andmete maht, nende laekumise sagedus, seatud kvaliteedikriteeriumid ning asutuse olemusest tulenevad piirangud. Hinnatakse uue lahenduse tulemuslikkuse ja kvaliteetsuse aspekte ehk kujundatud raamistik on piisavalt intuitiivne, et seda kasutama hakates on võimalik kiiresti saada kasu kontrolliprotsessi ressursi kokkuhoiu või kvaliteetse tulemuse näol ning, kas lahendus võimaldab paindlikkust meetodite valikul vastavalt konkreetsele andmekontekstile.

Organisatsioon, kus uurimistööga seotud andmekvaliteedi kontrollimise automatiseerimist katsetati ja uuriti on Statistikaamet, sest töö autor on ise selle organisatsiooni andmeanalüütik ning on ise töös käsitletava probleemiga vahetult kokku puutunud ning soov ning vajadus uue lahenduse järele oli aktuaalne ning ajakohane. Samuti võimaldas autori teadmine organisatsiooni toimimise kohta luua katseks lahendus, mis pakuks reaalses olukorras head alternatiivi juba olemasolevale praktikale. Uus lahendus ning tööprotsess eeldas põhjalikku taustauuringut erinevatest meetoditest ning varasematest uuringutest ja teisest küljest praktiliselt tööriista loomist ning selle abil erinevate meetodite katsetamist.

Loodud andmekontrolli lahenduse tulemuseks on andmeanalüüsi tarkvaras loodud põhjalik programm, millega saab genereerida soovitud andmestiku kohta kvaliteediraportid. Lisaks koguti kasutajatelt, kes uut lahendust enda töös rakendada said, tagasiside lahenduse tugevustest ja nõrkustest. Need on aluseks nii kvantitatiivsele, kui ka kvalitatiivsele analüüsile. Kvantitatiivses analüüsiosas keskendutakse raportites esinevate mõõdikute hindamisele, mille peal valitud

andmekvaliteedi kontrollimise meetodit rakendati ning mis tulemusi see meetod andis. Samuti on lahenduse efektiivsuse hindamiseks olemas andmed aja kohta, mis kulus uue lahendusega andmete kontrollimisele ning võrdluseks on olemas andmed aja kohta, mis kulus aasta varem vana protsessiga. Töö kvalitatiivne analüüsiosa tugineb valdkonna ekspertide tagasisidele, tõlgendustele ja hinnangutele uuele automatiseeritud protsessile ning nende sidumine kirjanduslike allikatega. Ekspertide arvamused aitavad tuvastada, kas automatiseeritud meetodid on suutnud korrektselt ja asjakohaselt kajastada andmestike kvaliteedis veendumiseks vajalikke aspekte ning kas nende meetodite tulemused vastavad meetodi eesmärgile.

Allikate otsimiseks ning ideede kogumiseks kasutati generatiivset tehisintellekti ChatGPT 4o ning ChatGPT 4.5. Peamiselt kasutati tehisintellekti mõne artikli või teema kohta täiendavate materjalide otsimiseks ning ka enda leitud allikatest tööga haakuvate, relevantsete teemade tuvastamiseks.

4. PRAKTILINE OSA

4.1 Statistikaameti olemasolev praktika

Statistikaametis kasutatakse paljudest erinevatest registritest pärit andmeid, et toota riiklikku statistikat. Rahvastikustatistika puhul kasutatakse ainuüksi Eestis alaliselt elavate inimeste nimekirja kokkupanekuks 18 erinevas registrist pärit andmestikke, mida on kokku üle 30 (Statistikaamet, 2022). Selleks, et paljusid erinevaid andmestikke kombineerivaid meetodeid saaks Statistikaametis rakendada, on kriitiline, et kõik kasutatavad andmestikud oleksid kvaliteetsed.

Statistikaametisse, kitsamalt rahvastikustatistika valdkonda, laekuvad igal aastal sarnaste struktuuridega andmekomplektid. Tehnilise eelkontrolli raames rakendatakse esmajärjekorras süntaktilisi kontrolle, mille eesmärk on veenduda, et andmed vastavad etteantud struktuurilistele reeglitele. Süntaktiline kontroll hõlmab näiteks tunnuste olemasolu kontrolli, kuupäevade formaadi kontrolli, väljade täituvuse ja lubatud tähemärkide kasutamise kontrolli. See etapp aitab kiiresti sõeluda välja ilmseid tehnilisi vigu, enne kui andmestikud liiguvad edasi sisuliseks kontrolliks. Nii Batini ja Scannapieco (2016) kui ka Charles (2024) on toonitanud süntaktilise kontrolli olulisust kui esmast ja vältimatut sammu andmekvaliteedi tagamise protsessis, mis pärinevad samadest andmekogudest. Näiteks saadab Haridus- ja Teadusministeerium igal aastal andmestiku, mis sisaldab eelmisel aastal mõne haridustaseme omandanud isikute andmeid koos konteksti andvate lisatunnustega. Kõigi laekuvate andmekomplektide puhul on andmekogudega varasemalt kokku lepitud, millises vormingus ja sisuga andmed laekuvad, sealhulgas näiteks haridustasemete kodeeringud ja kuupäevavahemikud (01.01–31.12). Lisaks isikupõhistele andmetele laekuvad ka hoonete ja muude aadressiobjektide andmestikud.

Kuna laekuvad andmed on üldjuhul sarnased eelnevate perioodide andmetega ja struktuurikokkulepped on juba varasemalt tehtud, võimaldab see rakendada võrreldavust varasema perioodiga. Selline lähenemine toetab andmekvaliteedi hindamist aja jooksul, võimaldades tuvastada süsteemseid muutusi või kõrvalekaldeid, mille olulisust on rõhutanud ka varasem kirjandus (Batini & Scannapieco, 2016; Sügis jt, 2025).

Andmete kontrollimine valdkonna analüütikute poolt toimub alles pärast seda, kui andmestik vastab kõigile seatud tehnilistele nõuetele andmete andmebaasi laadimisel ning kui andmed on jõudnud Statistikaameti andmebaasi. Statistikaametis on olnud ka juhtumeid, kus andmeesitaja on andmestiku esitamisel imputeerinud esitatavasse andmestikku vaikeväärtuseid puuduvate väärtuste asemele. Näiteks isiku sünniaja tunnus on Statistikaameti poolt määratud kohustuslikuks tunnuseks, mis peab olema kõigil andmestikus olevatel isikutel täidetud, kuid andmeesitaja enda andmebaasis sellist piirangut seatud ei ole ning sinna on sattunud puuduva sünniaja infoga isikud. Seega on andmeesitaja teinud puuduvate väärtuste kontrolli läbimiseks näiliselt andmed korda, et need vastaksid Statistikaameti seatud kriteeriumile, kuid sisuliselt ei ole sellised väärtused kasutatavad. Seega, kuigi tehnilised süntaktilised kontrollid on esmatähtsad hindamaks andmete kvaliteeti, rõhutavad Sügis jt (2025) vajadust kontekstipõhise lähenemise järele, sest semantilisi puuduseid andmetes võivad märgata vaid valdkonna eksperdid. Statistikaametis tegelevad sisulise andmete kontrolliga analüütikud, kes põhinevad kontrollimisel oma erialastele ja kogemuslikele teadmistele. Kui andmed on Statistikaameti andmebaasis, jõuavad andmestikud analüütikute töölauale, kes kasutavad andmeid vastavalt oma töövaldkonnale või varasemale kogemusele konkreetse andmestikuga. Kuna Statistikaametis on varem kasutatud erinevaid tarkvarasid, nagu SPSS ja SAS, viis iga analüütik andmekontrolli läbi oma äranägemise ja eelistatud tööriista abil. Selle tulemusel on mõned andmestikud aastate jooksul läbinud põhjalikumaid kontrolle kui teised, sõltuvalt analüütiku põhjalikkusest ja töökorraldusest.

Pärast Statistikaameti otsust viia andmetöötlus RStudio keskkonda avanes analüütikutel võimalus standardiseerida andmekontrollide metoodikat ning rakendada samu kontrolle erinevate andmestike puhul. Siiski jäi kontrollimeetodite rakendamine iga analüütiku individuaalseks otsuseks. Mõned analüütikud viisid läbi väga detailseid kontrolle iga tunnuse tasandil, mis oli ajamahukas, samas kui teised keskendusid vaid tuntud ja olulisematele tunnustele, jättes muu tähelepanuta. Sarnast olukorda, kus puudub ühtne ja süsteemne andmekvaliteedi kontrolli lähenemine, on kirjeldatud ka varasemas kirjanduses kui ühe peamise riskitegurina, mis võib viia andmekvaliteedi probleemide varjatud püsimiseni (Epperson jt, 2023).

Positiivseks küljeks RStudio kasutuselevõtuga oli, et kord välja töötatud kontrollise skriptid võimaldasid järgmisel aastal uue perioodi andmeid kontrollida sama loogika ja protsessiga. See tagas järjepidevuse andmekontrollis konkreetse andmestiku piires. Probleemiks aga jäi, et sarnased kontrollid loodi küll konkreetsetele andmestikele, kuid ei laiendatud neid teistele sarnase struktuuriga andmestikele. Killustatuse vältimise olulisust on toonitanud ka Batini ja Scannapieco

(2016), rõhutades vajadust standardiseeritud ja koordineeritud andmekvaliteedi kontrolliprotsesside järele.

Andmekvaliteedi kontrollimise protsessi killustatus võib viia olukorrani, kus ühe andmestiku puhul on väga põhjalikult kontrollitud näiteks kodakondsuse tunnuse järjepidevust ajas, samas kui teise andmestiku puhul sellist kontrolli ei tehta. Oht seisneb selles, et kui statistikaprojektidesse kaasatakse uusi andmeallikaid, võib kasutusele sattuda andmestikke, mille kvaliteeti pole piisavalt põhjalikult kontrollitud, ning sellega kaasnevad riskid ebatäpsete või eksitavate statistiliste järelduste tegemiseks. Kuigi sellist juhtumit pole seni realiseerunud, on varasemad uuringud (Epperson jt, 2023) näidanud, et selliste probleemide ilmnemisel võivad olla negatiivsed tagajärjed ka edaspidistele protsessidele ja otsustele, mis nende andmetega tehakse. Olemas oli seega vajadus kontrollireeglite määramiseks andmete kontrollimise protsessis ehk teha otsus, milliseid konkreetseid kontrollimeetodeid on tarvis rakendada vastavalt andmestiku struktuurile ja organisatsiooni vajadustele (Batini & Scannapieco, 2016; Statistikaamet, 2020).

Kui andmetes avastati vigu, hinnati vigade ulatust ja mõju statistikatootmisele. Põhitunnuste vigade korral oli nõutav parandamine; abitunnuste vigade puhul piirduti probleemide dokumenteerimise ja kasutajate teavitamisega. Põhitunnusteks nimetatakse neid tunnuseid, mille kohta konkreetsest andmeallikast andmeid vajatakse, seega on põhitunnus olenevalt andmestikust erinev. Kui andmestik on peamiseks allikaks näiteks isiku emakeele määramiseks, nimetatakse emakeele tunnust põhitunnuseks, see peab olema kindlasti korrektne. Kui samas andmestikus on lisaks emakeelele ka isiku ütluspõhine perekonnaseisu tunnus, mille jaoks on Statistikaametis ka usaldusväärsemaid allikaid, nimetatakse selles kontekstis perekonnaseisu tunnust abitunnuseks. Kui samale objektile eksisteeris mitu allikat, eelistati parandada peamised allikad. Alternatiivsete andmete puudumisel ja sobivate tingimuste korral kasutati ka imputeerimist ehk täideti puuduvad andmed näiteks eeldatud või keskmiste väärtustega.

Kui andmekvaliteedis tuvastatakse probleem(e), on probleemihaldus Statistikaameti kontekstis mõnevõrra erinev organisatsioonidest, kus hallatakse ise andmeallikaid. Probleemihaldus toimub üldiselt kontrollide tulemusel tuvastatud probleemide edasise analüüsi ning vajadusel paranduste tegemisega, näiteks puuduvate väärtuste täitmisega või formaadi normaliseerimisega (Batini & Scannapieco, 2016; Statistikaamet, 2020). Statistikaamet saab enda andmed erinevatelt majavälistelt registritelt, seega registrisse kantavate andmete ja nende töötlemise üle ametil kontrolli ei ole. Kui andmetes esinev probleem on selline, mis nõuab andmete puhastamist või parandamist oluliste puuduste või vigade osas, peab amet paluma registripidajal, kes andmed esitas, puudused likvideerida ning andmed uuesti esitada. See aspekt rõhutab andmete

usaldusväärsuse ja päritolu hindamise tähtsust, mida kirjanduses on samuti esile tõstetud (Batini & Scannapieno, 2016). Andmed, mida Statistikaametis kasutatakse on pseudonümiseeritud. Sel põhjusel on vigade parandamine mõnevõrra keerulisem - andmekogule ei saa konkreetsete kirjete parandamiseks viidata, andmete kontrollimisega on vaja aru saada, millest viga tuleneb ning viga kirjeldama, et andmekogu saaks selle süsteemselt parandada. Pseudonümiseerimise eesmärk on muuta isikuandmed mitteindentifitseeritavaks (Andmekaitse Inspektsioon, 2024). Lisaks, kui tegu on sisulise veaga, on võimalus, et andmekogule ongi andmeesitaja (isik, ettevõte, muu asutus) esitanud andmed sellisel kujul ning andmekogul ei olegi võimalik andmeid parandada. Lihtsamate tehniliste probleemide puhul andmetes, nagu tühjad kirjed või duplikaatsed kirjed, ei ole alati tarvis andmekogult andmeid uuesti pärida, sellised probleemid on võimalik Statistikaametis ära lahendada, kuid ka sellest teavitatakse andmekogu, et nad saaksid ka enda andmebaasis olevates andmetes probleemi ära lahendada.

Olemasoleva praktika puhul on selge, et andmete kontrollimise protsess on mõnes etapis üsna killustatud, näiteks kontrollireeglite määramine on iga analüütiku ja iga andmestiku puhul üsna omanäoline ning seega ei ole võimalik üldiselt öelda, et kõik kasutusse minevad andmestikud on kontrollireeglitele vastavad.

4.2 Mida otsustati katsetada

Andmekvaliteedi kontrollimise praktiline osa on Statistikaametis tehtud esmaste andmekontrolli läbiviimise automatiseerimine, automaatsete väljundite loomine ning töötajate tagasiside kogumine ja analüüsimine ühtse automatiseeritud lahenduse kasutamisele, mille rakendamine oli osa tööülesandest andmete kvaliteedi kontrollimisel. Vajadus standardiseeritud ja automatiseeritud esmase andmekontrolli järele tulenes laialdasest andmete sissevoolust erinevatest majavälistest registritest, kus kiire ja kvaliteetne kontroll on kriitilise tähtsusega, eriti rahvastikustatistika koostamisel. Automaatne andmekvaliteedi kontroll pakub muuhulgas efektiivsuse, järjepidevuse ja kulutõhususe kasvu (Charles, 2024). Samuti otsustati automatiseerida just kontrolli läbiviimise osa protsessist, sest automaatsete tööriistade kasutamine kontrolli läbiviimisel võimaldab genereerida kiiresti erinevaid kokkuvõtteid, näiteks tunnuste väärtuste järjepidevuse ülevaadet, jaotushistogramme või sagedustabeleid (Kandel jt, 2012; Epperson jt, 2023; Siddiqui, Kim, Lee, Karahalios & Parameswaran, 2016).

Kuna andmekomplektid on pärit usaldusväärsetest allikatest ei eeldata Statistikaametis, et andmestikud on vigased, pigem vastupidi, eeldus on, et andmed on korras, kuid siiski on oluline

eeldust kontrollida. Konkreetselt rahvastikustatistikas kasutatavad andmekomplektid saabuvad iga aasta alguses ning nende kvaliteet peab vastama kõrgetele nõuetele. Lisaks kvaliteedi tagamisele oli oluline probleemide varajane avastamine, mis võimaldaks neid analüüsida, lahendada ning vajadusel vastutavate osapooltega kooskõlastada. See vajadus toetab uurimisküsimust, mis puudutab andmekvaliteedi kontrollimeetodite kohandamist vastavalt organisatsiooni ja andmekomplekti eripäradele. Eeltoodud vajadused on samuti kooskõlas Charles (2024) ja Batini & Scannapieco (2016) rõhutatud andmete järjepidevuse ja terviklikkuse säilitamise olulisusega.

4.3 Kuidas katsetus läbi viidi

4.3.1 Tööriista loomine

Kuna rahvastikustatistikas kasutatavad andmed laekuvad registritelt igal aastal perioodil jaanuarist märtsini, alustati uue lahenduse arendamist 2024. aasta septembris, et vältida ajalist kattumist andmete kontrollimise kõrghooajaga. Lahendus rakendati esmakordselt 2025. aasta alguses laekunud andmete kontrollimiseks. 2025. Aasta oli andmekvaliteedi kontrollimise automatiseerimise katsetamiseks sobiv aasta ka seetõttu, et võrreldes eelmise aastaga ei tellitud täiesti uusi andmestikke, seega oli kõigi andmestike andmekvaliteedi kontrolli puhul arvestada, et kontrollitavast andmestikust on olemas versioon ka varasemate perioodide kohta.

Valitud tehniline platvorm lahenduse väljatöötamiseks oli R Markdown, kuna see võimaldab hõlpsasti luua automatiseeritud raporteid kasutades RStudio keskkonda. Peamine eesmärk uue lahenduse puhul on esmajoones andmete profileerimine. Profileerimisel luuakse esmane ülevaade andmete kvaliteedist ja suunata tähelepanu võimalikele probleemidele, mis vajavad edasist töötlemist (Epperson jt, 2023). Kuigi olemas on ka teisi sarnase funktsionaalsusega rakendusi, osutus valituks just R Markdown, sest suur osa Statistikaameti andmetöötlusest ja -analüüsist teostatakse RStudio keskkonnas ning see on töötajatele tuttav. Valik tuttava keskkonna kasuks tehti eesmärgiga, et töötajad tunneksid end tööriista kasutamisel mugavalt ning saaksid lahendust rakendada ilma vajaduseta õppida tundma täiesti uut tarkvarakeskkonda.

Asutuses viidi vahetult enne andmekontrolli tööriista arendamist läbi R Markdowni koolitus, mis toetas lahenduse kiiret ja sujuvat juurutamist. Koolitusel tutvustati, kuidas R Markdowni on asutuses juba varem kasutatud ning millised võimalused on kättesaadavad versioonide ja lisapakettide kaudu. See andis tööriista arendamiseks vajalikud teadmised ja tööriistakasti, mille abil sai andmekontrolli lahenduse välja töötada. Selline valik ja ettevalmistus on samuti kooskõlas

kirjanduses käsitletud vajadusega kasutada tööriistu, mis vähendavad tehniliste vigade esinemist ja võimaldavad kiiret andmete mõistmist (Kandel jt, 2012; Heer, 2019). Sellest tulenevalt on ootus tööriistale, et selle kasutusele võtmine ei suurenda kogu kontrollimise protsessile kuluvat aega, sest isegi kui töötajatel kulub tööriista rakendamise õppimisele rohkem aega kui neil kuluks juba harjumuspäraseks saanud kontrollide läbiviimisele, annab uus lahendus kiire ülevaate kõigist tehnilist laadi probleemidest, mis andmetes esineda võivad.

4.3.2 Tööriista funktsionaalsus ja rakendamine

Tööriista sisendiks loodi dialoogiaken (Joonis 1), kus kasutajatel on tarvis sisestada enda andmebaasi kasutajanimi, seejärel aasta, millal kontrollitavad andmed laekusid ning viimaks valida andmebaasi skeem ning konkreetne andmestik, mida kontrollida soovitakse. Menüüpõhine valik tõstis kasutajamugavust ja vähendas vigade ohtu. Tööriista käivitamisel hoiti sisendite küsimine minimaalsena, et raporti saaks genereerida ka näiteks uus töötaja, kellele on antud andmebaasis ligipääs kontrollimist vajavale andmestikule, mille kohta tal varasemad teadmised puuduvad. Seetõttu ei palutud täpsustada, milline on andmestikus olevate objektide unikaalsust määrav tunnus või muu spetsiifiline sisuline teadmine andmestiku kohta.

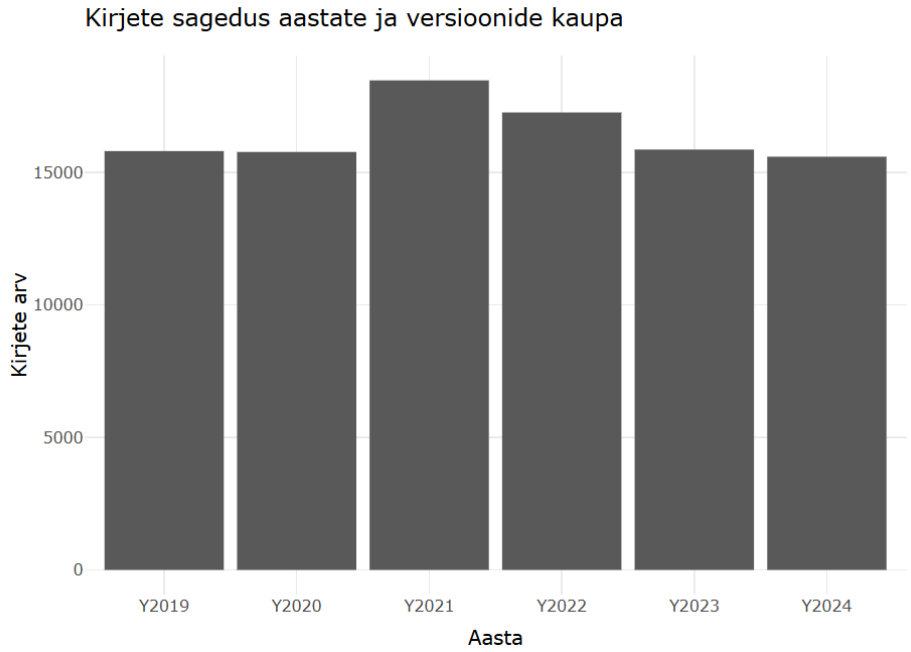
Joonis 1. Andmekontrolli tööriista dialoogiaken (täidetud näidisandmetega).

Järgmise sammuna määrati andmete kontrollimiseks reeglid ning valiti kontrollimeetodid. Kontrollireeglite määramisel on oluline vastavalt andmekirjeldustele panna paika vajalikud andmekontrolli reeglid ning panna paika automaatsed või poolautomaatsed kontrollimeetodid, mida rakendada (Charles, 2024).

Raporti struktuur pakkus esmalt andmestiku kirjete arvu ja duplikaatide ülevaadet, järgnesid tunnuste jagamine kategooriatesse ja nende spetsiifiline kontroll. Kuna sageduste võrdlemine tabeli kujul vajab rohkem süvenemist ning inimesel on kerge midagi märkamata jätta, on enamike sagedustega seotud meetodite puhul valitud visuaalne lahendus, et tööriista kasutajal oleks kohe konkreetne pilt ees olukorrast andmestikus. Siiski on mõne andmetüübi võrdluseks raportis ka tabeli kujul väljundid, sest unikaalseid väärtuseid võib olla palju ning nende sagedus ei ole selle andmetüübi puhul peamine kontrollitav aspekt.

Duplikaatsete väärtuste kontrolliks on raportis üks rida teksti - “Kas andmestikus on unikaalsed kirjed?” - ning selle järel on kas jaatus või eitus, vastavalt kontrolli tulemusele. Tegu on sisuliselt väga lihtsa kontrolliga, mis kontrollib iga rea unikaalsust, kuid aitab kontrollida minimaalsuse ja usaldusväärsuse kvaliteedidimensiooni. Kirjete unikaalsuse kontroll on väga oluline ning kõrgelt automatiseeritav samm andmekvaliteedi kontrollimiseks, mis võimaldab tuvastada andmetes esinevad duplikaadid (Batini & Scannapieco, 2016).

Erinevate perioodide kohta käivate andmete puhul on oluliseks kvaliteedidimensiooniks andmete täpsus ja järjepidevus. Kirjete arvu analüüsimiseks rakendati jaotuste analüüsi, mida esitati tulpdiagrammidena, kus y-teljel esitati andmestiku aasta ja x-teljel kirjete arv (Joonis 2). See lähenemine võimaldab visuaalselt hinnata, kas kirjete arv on perioodist perioodi püsinud järjepidev või esinenud olulisi muutusi. Valitud meetod tugineb Sügis jt (2024) ja Ahiagble ja Stein (2022) kirjeldatud tunnuste jaotuse ja volatiilsuse kontrollile, mille eesmärk on tuvastada andmetes ootamatuid kõikumisi. Näiteks, kui analüüsi käigus märgati, et kirjete arv oli pigem kasvutrendis, viitas see andmestiku loogilisele arengule ja toetas järjepidevuse eeldust. Selline perioodidevaheline võrdlus on samuti kooskõlas kirjanduses (Epperson jt, 2023) rõhutatud vajadusega hinnata andmete ajas muutumist ja tuvastada võimalikke struktuurseid nihkeid andmestikes.



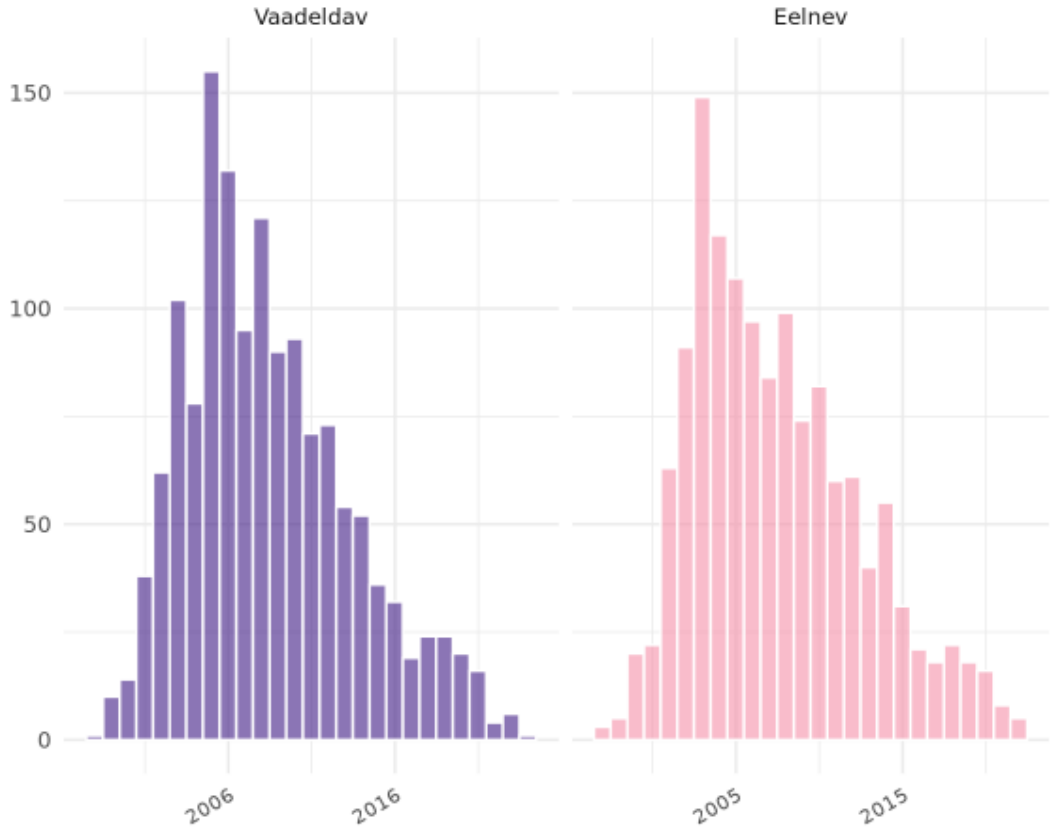
Joonis 2. Andmestiku kirjete arv erinevate vaatlusperioodide kaupa.

Kategoriaalsete tunnuste puhul on olulised kvaliteedidimensioonid selgus ja järjepidevus. Sellest tulenevalt analüüsiti väärtuseid ja nende järjepidevust eelnevate perioodidega (Joonis 3). Kontrollimeetod tugineb Sügis jt (2024) kirjeldatud kategooriate täielikkuse kontrollile, mille eesmärgiks on veenduda kõigi kategoriaalsete tunnuste väärtuste vastamist lubatud väärtuste komplektile. Kuna aga kõigi tunnuste puhul ei ole antud ette konkreetset koodiloendit, loodi raportisse tabel, kus on ridades kõik kategoriaalsed tunnused ning järgnevates veergudes kõik uued unikaalsed väärtused, mis vastavas tunnuses olid. Viimases tabeli veerus on kuvatud, kas on mõni väärtus, mis eksisteeris varem, kuid on nüüd puudu ning ka teistpidi, kas on mõni täiesti uus väärtus kategoriaalses tunnuses, mida varem esinenud ei ole. Selline lähenemine võimaldab läbi viia kategoriaalsete tunnuste väärtuste kontrollimeetodit ilma, et oleks vaja spetsiaalselt iga kategoriaalse tunnuse jaoks olemasolevat koodiloendit.

Muutuja	Lisandunud	Eemaldatud
PERENIMI	Puudub	Puudub
EESNIMI	Puudub	Puudub
IK	Puudub	Puudub
EMAKEEL	ukraina / vene, baski, kirgiisi, bosnia	suahiili, kikuju, somaali, horvaadi, sindhi, pandžabi, moldova, Eesti keel teise keelena
SUGU	Puudub	Puudub
KODAKONDSUS	Tadžikistan, Etioopia, Jeemen, Tšehhi, Alžeeria, Montenegro, Laos, Kõrgõzstan, Jordaania, Uganda, Guatemala, Lõuna-Korea	Horvaatia, Singapur, Põhja-Makedoonia, Nicaragua, Island

Joonis 3. Andmestikus olevate kategoriaalsete tunnuste analüüsimiseks loodud lahendus.

Kuupäevatunnuste puhul rakendati visuaalset analüüsimeetodit, täpsemalt histogrammi, mille eesmärk oli hinnata, kas kirjete jaotus vastava kuupäeva tunnuse põhjal järgib loogilist mustrit (Joonis 4). Loogiline muster võib iga andmestiku ja tunnuse puhul olla erinev, näiteks hariduse omandamise kuupäevade puhul on ootuspärane, et suur osa kuupäevi langeb juunikuu lõppu, mil Eestis tasemehariduse õppeaasta üldjuhul lõpeb. Selleks, et vähendada kontekstispetsiifiliste teadmiste vajalikkust, toodi kuupäevade jaotuse analüüsis iga tunnuse puhul kõrvale ka eelmisel perioodil sama tunnuse jaotus. Analüüsitava ja eelneva perioodi võrdlemine andis kiire ülevaate, kas kirjete jaotus oli järjepidev. Perioodi võrdlemine on kooskõlas Batini ja Scannapieco (2016) esile toodud järjepidevuse kontrolli põhimõtetega, andes võimaluse saada kiire ülevaade andmestiku võimalike kvaliteediprobleemide kohta erinevate perioodide kohta käivate andmete puhul. Samuti tugineti Sügis jt (2024) kirjeldatud tunnuste jaotuse kontrollile. Lisaks graafilisele väljendusele on raportis ka kirjas kui paljude kirjete puhul on konkreetses kuupäeva tunnuses puuduvaid väärtuseid. Kuupäevatunnuste kontrollimisel on kontrollitavad kvaliteedidimensioonid asjakohasus ning täielikkus (Batini & Scannapieco, 2016).



Joonis 4. Andmestikus olevate kuupäeva tunnuste analüüsimiseks loodud visuaalne lahendus.

Arvuliste tunnuste kontroll käsitles kontrollitava andmestiku puhul iga arvulise tunnuse puhul puuduvate väärtuste analüüsi ning maksimaalsete ja minimaalsete väärtuste võrdlust eelneva perioodi sama andmestiku arvuliste tunnustega nagu Joonisel 5. Puuduvate väärtuste analüüs aitab tuvastada, kas tunnuses esinevad puuduvad väärtused viitavad tõenäoliselt mingile veale, mis on süsteemne või juhuslik (Sügis jt, 2024). Maksimaalsete ja minimaalsete väärtuste võrdlemine toetub väärtusvahemike kontrollimeetodile, millega kontrollitakse tunnuste väärtuste paigutumist oodatud piiridesse (Batini & Scannapieco, 2016). Selleks, et väärtusvahemikud oleksid kontekstis, on, nagu ka eelnevate rakendatud meetodite puhul, lisatud võrdlus eelneva perioodiga. Kvaliteedidimensioonid, mida kontrolliti olid korrektsus ning täielikkus (Batini & Scannapieco, 2016).

	Eelnev			Vaadeldav		
Muutuja	**Puuduvad väärtused**	**Minimaalne**	**Maksimaalne**	**Puuduvad väärtused**	**Minimaalne**	**Maksimaalne**
OPPEASUTUS_ID	0	1	2083347	0	1	2123708
JURIIDILINE_AADRESS_ASUKOHT	2	120	9835	2	120	9835

Joonis 5. Andmetes olevate arvuliste tunnuste analüüsiks loodud risttabel.

Viimaks arendati eraldi kontrollimoodul aadressiobjekti võtmete kontrollimiseks, kasutades Aadressiandmete Süsteemi (ADS). Eraldi oli vaja arendada aadressiga seotud tunnuste kontroll, kuna ADS ühtlustab kõiki aadresse sisaldavaid andmeid. Igas andmestikus, kus on kirjega seotud mõni asukoht, näiteks asutuse asukoha puhul, esines sellele vastava aadressiobjekti võti, mille kaudu sai ühendada üldisest ADS andmestikust konkreetseid aadressiobjekte. Aadressiobjektide liikideks on näiteks eluruumid, maaüksused ja maakonnad. Kõige parema kvaliteedi aadressiobjekti võtmete puhul oli võtme vasteks hoone täpsusega aadressiobjekt. Seega on siin olulisel kohal Batini ja Scannapieco (2016) poolt esile toodud täpsuse dimensioon.

Aadressiobjekti võtmete kontrolli jaoks loodi automaatkontrolli etapp, mis tuvastas, kas andmetes esineb aadressivõtit sisaldavaid tunnuseid ning kui esines, analüüsiti esmalt sisuliste aadressivõtmete osakaalu ning viidi läbi võrdlus eelmise aastaga, et luua kontekst, mis kvaliteet käesoleval andmestikul on. Samuti hinnati vasteta aadressivõtmete osakaalu, tuvastamaks võimalikud kvaliteediprobleemid, mis võivad mõjutada andmestiku kasutatavust ja usaldusväärsust (Joonis 6). Kontrollimeetod tugineb Batini ja Scannapieco (2016) ja Statistikaameti (2020) kirjeldatud formaadikontrolli põhimõttele, et hinnata väärtuste vastavust tehnilistele nõuetele.

Aadressivõtmete kontroll

Kontrollitavates andmetes on SA_ADR_VOTI, kontroll käivitub.

Sisulisi võtmeid: 1941 (99.9%),
eelmise aasta andmetel: 1975 (99.9%)
Aadressivõtmetes UNK (215760) väärtuseid: 2 (0.1%),
eelmise aasta andmetel: 2 (0.1%).

Tühjasid aadressivõtmeid oli vaadeldavas perioodis: 0. Eelmises perioodis: 0

Vaadeldaval perioodil ADS_OID vähemalt hoone täpsusega: 1927 (99.18%).
Eelneval perioodil ADS_OID vähemalt hoone täpsusega: 1964 (99.35%).

Aadressiobjekti liigi kood	Aadressiobjekti liik	Vaadeldav	Vaadeldava perioodi osakaal, %	Eelnev	Eelneva perioodi osakaal, %
ER	Eluruum	243	12.51	238	12.04
EE	Elu ja/või ühiskondlik hoone	350	18.01	341	17.25
MR	Mitteeluruum	14	0.72	15	0.76
ME	Mitteelukondlik hoone	1320	67.94	1370	69.30
CU	Maaüksus	14	0.72	11	0.56
LP	Liikluspind	0	0.00	0	0.00
VK	Väikekoht	0	0.00	0	0.00
LO	Linnaosa	0	0.00	0	0.00
AY	Asustusüksus	0	0.00	0	0.00
OV	Omavalitsus	0	0.00	0	0.00
MK	Maakond	0	0.00	0	0.00
UNK	Teadmata	2	0.10	2	0.10

SA_ADR_YLDISTATUD väärtus > 0 näitab, et väljund on leitud sisendi (teksti)osa(de) väljajätmise ehk üldistamise teel
See tähendab, et mida väiksem on selle muutuja väärtus, seda vähem on tehtud lisatöötlust ja seda kvaliteetsemad on algandmed.
Võrdleme vaadeldavat ja eelnevat perioodi

SA_ADR_YLDISTATUD	arv, Vaadeldav	osakaal, Vaadeldav	arv, Eelnev	osakaal, Eelnev
0	1928	99.33	1965	99.49
1	13	0.67	10	0.51

0 ADS_OID-i kaotasid kehtivuse enne 2024-12-31. 0 neist 2025. a jooksul. 1 ADS_OID-i hakkasid kehtima hiljem kui 2024-12-31.

Joonis 6. Andmete olevate aadressivõtmete kontrollimiseks loodud analüüsimeetod.

4.4 Kuhu katsetusega jõuti

Automatiseeritud lahendus läks 2025. aasta alguses käiku ning tulemused näitasid, et automatiseeritud lahendus võimaldab rakendada erineva struktuuriga andmestike puhul ühtseid ja vastavalt vajadusele kohandatavaid kontrollimeetodeid, võttes arvesse organisatsiooni ja andmekogude eripära. Kontrollimeetodeid rakendati vastavalt andmestikus olevale tunnusetüübile, seega oli iga tunnusetüübi jaoks olemas kontrollimiseks sobiv meetod. Loodud lahendus toetab kirjanduses esile toodud vajadusi standardiseeritud ja kohandatava andmekvaliteedi kontrolli protsessi järele, tõstes oluliselt andmekontrolli tõhusust ja järjepidevust Statistikaameti töös (Batini & Scannapieno, 2016; Charles, 2024; Epperson jt, 2023). Andmete visuaalne analüüs ja profileerimine järgib kirjanduses esile toodud soovitusi andmete jaotuse analüüsimiseks ja kirjeldava statistika kasutamiseks varajases kontrollifaasis (Seltman, 2018; Epperson jt, 2023). Kasutegur Statistikaametile selle katsetusega on standardiseeritud ja kohandatav andmekvaliteedi kontrolli protsess ning tööriist, mille tulemusel on andmete kvaliteedikontroll järjepidev, ühtlaselt dokumenteeritud.

Katsetuse tulemuseks olevaid raporteid analüüsides on võimalik tuvastada peamised probleemid, mis andmestikes esinevad. Seeläbi saab näiteks ebasobivate väärtuste esinemisel või kirjete arvu perioodiliti põhjendamatu muutuse korral indikatsiooni, et andmestik ei ole kvaliteetne ega kasutatav ning on tarvis kontakteeruda andmeesitajaga.

4.5 Ekspertide tagasiside

Ekspertidel paluti jätta tagasiside vabas vormis, et saada erinevaid vaatenurki loodud andmekontrolli automatiseerimise lahendusele. Lahenduse positiivse poole pealt tõid eksperdid esile meetodite tulemuste esitamine visuaalselt, näiteks kui analüütik teadis mingi andmestiku puhul, et andmestikus peab kirjete arv igal perioodil olema kõrgem kui eelnev, andis visuaalne kirjete arvu graafik kiiresti aimu, kas see vastab tõele. Samuti toodi esile, et kuupäevade jaotumise võrdlus eelmise perioodiga andis võimaluse aastasiseseks jälgimiseks, mida varem käsitsi kontrolle tehes pea kunagi ei vaadatud. Kategoriaalsete muutujate kontrollimeetod võimaldas ekspertide sõnul kerge vaevaga tuvastada, kas mõnes tunnuses, mille aluseks on koodiloend, on uusi elemente, millega peab ka väljundi tegemisel arvestama. Samuti oldi rahul võimalusega kontrolliprotsessi korrata kui esimesel kontrollimisel avastati vead ning andmed parandati või paluti andmekogul andmed uuesti esitada. See näitab, et andmekontrolli protsessis paranes probleemide esinemisel uuesti laekunud andmestike kontrollide kordamine veendumaks, kas

probleemid andmetes on lahendatud ja kas andmete kvaliteet vastab soovitud tasemele (Charles, 2024). Lisaks saab sel viisil kaht kvaliteediraporti võrreldes näha, kas esimesel korral esinenud vead on kõrvaldatud ning muud aspektid on andmestiku puhul jäänud samaks. Toodi esile erinevate automaatikaga seotud anomaaliate tuvastamist, millele varem käsitsi kontrolle tehes suurt tähelepanu ei pööratud. Anomaaliate tõttu tööriist tõrkus, sest kontrolli eeldus on, et andmete andmebaasi laadimine on tehtud standardsel kujul ning kindlatele tingimustele vastavalt.

Andmekontrolli lahenduse kasutajad pidasid olulisteks puuduseks liiga jäigalt ette seatud tingimusi. Kuna lahendus loodi igal aastal laekuvate andmestike kontrollimiseks, ei ole konkreetsest tööriistast kasu näiteks iga kuu või iga päev uuenevate andmete kvaliteedi kontrollimisel. Veel mainiti, et kuupäevasid sisaldavate tunnuste kontrollimeetod on küll hea visuaalseks ülevaateks, kuid kui kuupäevade vahemik on väga pikk, ei ole graafik eriti informatiivne, sest x-teljele kuvatavad kuupäeva väärtused ei ole eristatavad või on liiga suure intervalliga. Samuti oli loodud lahenduses duplikaatide kontrollimise etapp, kuid kasutajad oleksid soovinud, et duplikaatide kontroll oleks samuti konteksti jaoks võrdluses eelnevatel perioodidel laekunud andmetega, et mõista, kas duplikaatide sisaldumine andmetes on esmakordne või on duplikaadid ka varasemalt laekunud andmetes. Veel oli oluline puudus ekspertide sõnul erinevat tüüpi tunnuste puhul puuduvate väärtuste mitte kuvamine kvaliteediraportis. Kasutajad tõid esile veel erinevaid tehnilist laadi tõrkeid ja probleeme, mis takistasid loodud tööriista iseseisvalt kasumist ning nõudsid tööriista autori sekkumist, et see töökorda seada.

Kasutajad tundsid veel puudust klassifikaatorite kontrollimisest neis tunnustes, kus on väärtused piiratud kindlaks määratud väärtuste loendiga. Sellised kontrollid olid nad juba ise varem loonud ning seetõttu oli neil tarvis siiski lisaks uuele tööriistale kasutada ka varem enda loodud lahendusi, et teha andmetele täielik kvaliteedi analüüs. Oluline puudus, mida kasutajad veel märkisid, uue lahenduse osas oli seotud selle keerulise ülesehitusega ning kui seal oleks vaja kasvõi mõni väike muudatus teha, kuluks neil endil selleks palju aega ning ilmselt peaks sellega tegelema tööriista autor.

Kasutajate juba harjumuspärane kontrollimine varasematel aastatel tõsteti samuti esile. Lisaks eelnevas lõigus kirjeldatud klassifikaatori kontrollile toodi esile seda, et kasutajad on juba mitu aastat teinud RStudios ise kontrolle. Neist kontrollidest on kujunenud neile endile olulised aspektid või veaohtlikud kohad, mida nad soovivad kindlasti enda kontrollitavate andmestike puhul kontrollida. Loodud ühtne lahendus sellist paindlikkust kasutajatele ei pakkunud.

Kasutajate tagasiside oli ka kogu kontrollimise perioodi jooksul vahetu ning enamik tehnilist laadi probleemid said lahenduse juba andmete kontrollimise perioodil. Näiteks puuduvate väärtuste kuvamise funktsioon sai lisatud kuupäeva tunnustele. Metoodilist laadi probleemid, näiteks konkreetset tüüpi tunnuste puhul rakendatava kontrollimeetodi valiku puhul jäid kogu lahenduse kasutamise perioodi vältel siiski käiku samad meetodid, et nende meetodite rakendamise efektiivsust oleks võimalik põhjalikult hinnata. Samuti ei olnud töö raames võimalik kasutada töötajate ressurssi selleks, et katsetada ühe andmestiku kontrolli käigus mitut tööriista erinevate kontrollimeetodite komplektiga, sest töötajate eesmärk oli andmed võimalikult efektiivselt ning hästi kontrollitud saada. Sellest tulenevalt töötati välja vaid üks tööriist valitud meetoditega, mida katsetati terve kontrolliperioodi vältel.

4.6 Uurimus

Peamised näitajad, mida saab automatiseeritud kontrollimeetodite rakendamisel ja varasema praktika vahel dokumentatsiooni või andmete pealt võrrelda, on seotud ajaandmete-, vigade arvuning vigade sisuga. Andmed mainitud võrdluste ja analüüsi jaoks on kättesaadavad üks aasta enne uute meetodite rakendamist ning esimese aasta kohta, mil uusi meetodeid rakendati. Kokku analüüsiti ligi 40 erineval hetkel laekunud andmestikku. Neist 36 olid sellised, mille puhul oli võimalik kõrvutada kahe aasta andmeid.

4.6.1 Ajaandmete võrdlus

Tabelis 2 on kahe aasta ajaandmete analüüsi tulemuste kokkuvõte, kus on arvutatud kõigi võrreldavate andmestike pealt erinevad näitajad, mis võimaldavad saada ülevaate ajalisest muutusest kahe aasta vaates, 2024. aastal toimus kontrollimine veel iga töötaja enda nägemuse järgi ning 2025. aastal rakendati esmakordselt uut tööriista, mis käesoleva tööga seoses välja töötati. Teises ja kolmandas veerus on vastavalt 2024. ja 2025. aasta ajaandmed, viimases veerus on kahe aasta ajaline muutus.

Tabel 2. Andmekvaliteedi kontrollimisele kulunud aja andmete kokkuvõte.

Näitaja	Kontrollimisele kulunud aeg 2024 (h)	Kontrollimisele kulunud aeg 2025 (h)	Muutus (h)
Summa	206.5	163.9	-42.6
Keskmine	5.9	4.7	-1.2
Mediaan	4.0	3.0	-0.9
Miinimum	1.0	0.5	-12.7
Maksimum	23.0	25.0	19.0

Esimene näitaja Tabelis 2 on kogu aeg, mis kulus kõigi võrreldavate andmestike kontrollidele. 2024. Aastal kulus andmestike kontrollimisele kokku 206,5 tundi, samade andmestike kontrollimiseks 2025. aastal uue kontrollimise lahenduse abil kulus kokku 163,9 tundi. Muutus kahe aasta vaates on 42,6 tundi. 2025. aastal kulus andmestike kontrollimisele seega ligikaudu viiendiku võrra vähem aega. Kogu aja kokkuhoidu uue tööriista kasutuselevõtuga selgitada ei saa, sest ka andmestike kvaliteet on aasta-aastalt erinev, mõnel aastal on üks andmestik vigane ning nõuab palju aega, et vigasid mõista ning andmekoguga kommunikeerida, teisel aastal võib sama andmestik olla kohe esimesel laekumisel korrektne ja kasutatav. Teisalt ei ole ka reaalne, et andmete kontrollimisele kulunud aeg langeks mitmekordselt, sest lisaks kontrollietappidele, mida

on võimalik automatiseerida, jääb siiski kontrollimise protsessi sisse ka eksperdi sisuline andmeanalüüs.

Teine näitaja, mida analüüsiti, oli keskmine aeg, mis ühe andmestiku kontrollimiseks kulus (Tabel 2). Keskmine aeg, mis kulub ühe andmestiku kontrollile langes 5,9 tunnilt 2024. aastal 4,7 tunnile 2025. aastal. Seega keskmiselt langes ühe andmestiku kontrollimisele kulunud aeg 1,2 tunni võrra, ehk samuti ligi viiendiku võrra. Siiski, nagu Tabelis 2 minimaalset ja maksimaalset aega vaadates on näha, varieerub ühe andmestiku kontrollimise aeg 2024. aastal 1-23 tunni vahel ning 2025. aastal 0,5-25 vahel. Kuna varieeruvus ühe andmestiku kontrollimisele kuluva aja puhul on niivõrd suur, kaasati analüüsi ka mediaan, mida ekstreemsed väärtused ei mõjuta. 2024. aastal oli ühele andmestikule kuluva aja mediaaniks 4 tundi, 2025. aastal 3 tundi. Muutuse mediaan oli 0,9 tundi, mis on samuti viiendiku võrra madalam.

Tabelis 2 jääb silma, et minimaalne aeg, mis kulus ühe andmestiku kontrollimiseks vähenes poole tunni võrra. Samas tõusis maksimaalne ühe andmestiku kontrollimiseks kuluv aeg kahe tunni võrra. Samuti, nagu Tabelist 2 näha, on maksimaalne tõus ühe andmestiku kontrolli puhul 19 tundi, see viitab asjaolule, et ühel eelneval aastal kontrollitud andmestikus probleeme ei esinenud ning tööriista katsetamise aastal avastati andmestikus probleeme, mille tõttu pikenes oluliselt andmestiku kontrollimisele kuluv aeg. Kõige suurem langus kontrollimise ajas on 12,7 tundi, sarnaselt maksimaalsele tõusule, ei ole siin tegu katsetuse suure efektiivsusega vaid sel aastal oli andmeesitajalt saadud andmestik hea kvaliteediga ning ei olnud tarvis andmeid mitu korda küsida ja kontrollida.

4.6.2 Vead andmestikes

Andmestikud, mida kontrolliti on Statistikaametile laekunud juba ligikaudu 10 aastat, seega on enamik uue andmestikuga seotud vigu juba minevikus parandatud ning põhilised süntaktilistest vigadest tulenevad probleemid on uurimuse ajaks juba kõrvaldatud. Aasta enne automatiseeritud kontrollimeetodite rakendamist, 2024. aastal oli tarvis andmeesitajalt küsida andmed uuesti kuuel korral. 2025. aastal küsiti andmed uuesti kolmel korral, kaks neist andmestikest oli tarvis küsida uuesti nii 2024. kui ka 2025. aastal. Peamised probleemid, mis esinesid, mille tõttu oli vaja andmeid uuesti küsida olid seotud duplikaatsete kirjetega andmetes, seda esines mõlemal vaatlusaastal ning see on üldiselt lihtne probleem, mida kontrollida ning mille puhul andmete uuesti laekumisel ei ole tarvis põhjalikku kontrolli uuesti teostada. See näitab aga, et ka uus lahendus püüab selle probleemi edukalt kinni ning seda kontrolli ei ole vaja töötajal enam ise teostada. Veel esines ajalise kattuvuse ebakõla, peamiselt sisaldasid andmed kirjeid, mis olid väljaspool kokkulepitud ajaraami, näiteks 2024. aasta kirjete hulgas oli ka 2025. aasta alguse kirjeid, seda oli võimalik hõlpsasti uue andmekontrolli tööriistaga tuvastada. Üks oluline probleem andmetes, mille automatiseeritud tööriist hästi kinni püüdis, oli seotud kategoriaalse tunnuse ebasobivate väärtustega. Nagu peatükis 3.3.2 oleval Joonisel 3 kuvatud, oli väga hõlpsasti näha, millised väärtused olid ühes tunnuses eelmisel perioodil laekunud andmestikus ning millised väärtused on seal kontrollitavas andmestikus. Kui kategoriaalne tunnus peab vastama kindlale koodiloendile, annab tööriistas olev kontrollimeetod selge indikatsiooni, et tunnuses on koodiloendisse mittekuuluvaid väärtuseid.

5. TULEMUSED JA ARUTELU

Uurimistöö eesmärgiks oli katsetada, milliseid automatiseeritud andmekvaliteedi kontrolli- ja puhastusmeetodeid saab Statistikaametis rakendada erineva struktuuriga andmete puhul ning kuidas kujundada raamistik ja praktiline lahendus, mis võimaldab kontrolli kohandamist vastavalt organisatsiooni ja andmekogude eripärale. Teoreetilises osas esile toodud erinevad kontrollimeetodid ning empiiriline katse organisatsiooni näitel tõestasid, millised automatiseeritud kontrollimeetodeid olenevalt vajadusest kasutada saab ning teisalt, kuidas on võimalik luua standardiseeritud, kuid siiski paindlik ja erinevate andmestike jaoks rakendatav andmekontrolli lahendus, mis annab võimaluse koostada standardiseeritud kontrollitulemus ja dokumentatsioon erineva struktuuri- ja sisuga andmestike käsitlemiseks.

5.1 Automatiseeritud andmekontrolli kasutegurid

Esimesele uurimusküsimusele "**Millised on kasutegurid, mida automatiseeritud andmekontrollid pakuvad Statistikaameti näitel?**" vastuseks on mitu kasutegurit, mis töö praktilisest osast selgus. Praktilises osas teostatud tehniline kontroll süntaktiliste kontrollide kaudu vastab kirjanduses (Batini & Scannapieco, 2016; Charles, 2024) esile toodud põhimõtetele, mille kohaselt tuleb esmalt välistada struktuursed vead enne sisulise andmekontrolli algust. Kui varasemalt oli igal analüütikul tarvis tehniline kontroll endale ise välja töötada, mis võis viia selleni, et kõik kontrollisid nii, nagu oskasid või tahtsid, siis nüüd on kontrollid standardiseeritud ning on teada, et konkreetsed kontrollid on iga andmestiku peal teostatud. Samuti toetab perioodiliselt laekuvate andmete võrdlemine varasemate aastate andmestikega kirjanduses kirjeldatud praktikat ajalisest määratlusest sõltuvate andmekvaliteedi muutuste tuvastamiseks (Sügis jt, 2025). Seega on esimeseks kasuteguriks standardiseeritud kontrolliprotseduur, mis annab kindlust, et iga kontrollitud andmestik on läbinud samad kontrollid.

Teiseks kasuteguriks on ajaandmete analüüsist selgunud ajalise ressursi vähenemine kahe vaatlusaasta võrdluses ligi 20% võrra nii kogu kontrollile kulunud ajast, kui ka keskmiselt ühe andmestiku kontrollimiseks kulunud ajast. See on eriti märkimisväärne, sest Statistikaametis võeti

kasutusele uus lahendus esimest korda 2025. aastal, töötajad said tööriista proovida esmakordselt alles uute andmete laekumisel ning kontrollile kulunud aeg sisaldab kogu kontrollimise protsessi, seega ka uue tööriista õppimist ja rakendamist. Lisaks on kvaliteetsed andmed ka edasiste etappide puhul, näiteks statistika tootmisel, aluseks optimaalseks ressursikasutuseks (Etikala, 2025). Seega on teiseks andmekvaliteedi kontrollimise automatiseerimise kasuteguriks ressursi kokkuhoid.

Kolmandaks selgub andmestikes esinevate vigade ülevaatest, et automatiseeritud andmekontrolli tööriista kasuteguriks on suutlikkus tuvastada probleeme sama tõhusalt nagu seda tegi varem inimene, näiteks duplikaatsete kirjete kontrolli puhul, siis seda aspekti ei ole vaja enam manuaalselt käsitleda. Ning kuna näitena toodud viga esines mõlemal vaatlusaastal, on tegu ühe olulise kontrolliga, mida on tarvis kindlasti iga andmestiku puhul rakendada.

Töö kasuteguriks saab lugeda veel praktilise vajaduse ja teoreetiliste suundumuste ühteliitmist. Rakendatud lahendus (R Markdowni põhine raportite loomine) oli kooskõlas soovitustega kirjandusest kasutada tehnilisi vähese õppevajadusega tööriistu (Kandel jt, 2012; Heer, 2019). Kasutajatelt saadud tagasiside näitas muuhulgas, et tuttava tarkvarakeskkonna kasutamine vähendas vastuseisu muutustele ja parandas kasutuselevõtu kiirust.

Automatiseeritud andmekontrollide rakendamisel Statistikaametis esines ka nõrkuseid ja piiranguid. Sisulise kontrolli osas ilmnes probleemina, et kuigi välja töötatud lahendus tagas enamike andmestike piires järjepidevuse, ei olnud loodud lahendused kõikidesse andmestikesse üldistatavad. See kinnitab varasemate uuringute (Epperson jt, 2023) väiteid, et ilma koordineeritud lähenemiseta võib organisatsioonis tekkida "vaikne" andmekvaliteedi langus. Näiteks oli tööriista sisse kirjutatud palju eelduseid andmestiku kohta, põhjuseks see, et tööriista kasutaja saaks minimaalse sisendiga andmekvaliteedi raporti koostada. Luues lahendus, mis nõuab detailset kasutaja sisendit, on andmekvaliteedi kontrollimine küll mõnevõrra detailsem, kuid kindlasti ressursikulukam.

Veel on nõrkuseks töös loodud automatiseeritud andmekontrolli lahenduse ja uue protsessi puhul juurutamisele üsna vähese tähelepanu pööramine ning sellest tulenevalt ilmselt ka kõik töötajad, kes oleksid saanud seda lahendust enda töös kasutada ning sellest kasu saada, seda ei teinud. Selleks oleks võinud pikemalt teha eelnevat teavitustööd ning käesolevas töös uuritud aspekte täpsemalt lahti selgitada. Näiteks andmekvaliteedi standardiseeritud kontrollide kasulikkust ning ressursivõitu, mis selline lahendus pikemas plaanis pakkuda saab.

Peamiseks piiranguks töös loodud lahenduse puhul laiemas kontekstis on asjaolu, et välja töötatud kontrolliraamistik rakendub praegusel kujul konkreetses organisatsioonilises ja andmestike

kontekstis ning ei pruugi ilma kohandamiseta olla otse rakendatav teistes asutustes või teistlaadi andmestike puhul, sest tööriista kirjutati sisse palju Statistikaameti spetsiifilisi eeldusi andmete kohta. Lisaks näitas tulemuste analüüs, et kuigi automatiseerimine vähendas töömahukaid korduvaid kontrolle, jäi vajadus inimkontrolli järele, eriti kontekstispetsiifiliste anomaaliate tuvastamisel.

Kasutajatelt saadud tagasiside, kes said lahendust proovida, oli positiivne just nende kontrollimeetodite osas, mille tulemused esitati raportis visuaalselt ning mis võimaldasid neil andmestiku sisu teades kerge vaevaga hinnata, kas andmestiku kvaliteet on ajas püsiv või on seal mingeid eripärasid, millele tuleb tähelepanu pöörata. Sellest tulenevalt võiks kaaluda edaspidi ka teiste kontrollimeetodite tulemuste visuaalse esitlemise kasutamist. Kategoriaalsete ning arvuliste tunnuste analüüsiks oleks ilmselt ka olnud võimalik tulemuste visuaalse esitlemise lahendus leida, et kasutaja saaks ka nende tunnuste puhul kiire ja sisulise ülevaate tunnuste kvaliteedist. Siiski on visuaalsel tulemuste esitlemisel oluline, et selle aluseks olev kontrollimeetod toimiks õigesti, tulemused oleksid korrektsed ning loetavad. Kasutajatelt saadud tagasisidest tuli veel välja, et kui graafiku loomine on täiesti automaatne, võib tulemuste pealt automaatselt genereeritud graafik olla loetamatu, näiteks kui tunnuses olevaid unikaalseid väärtuseid on väga palju või mõne üksiku väärtuse esinemine andmestikus on teistest oluliselt erinev.

5.2 Organisatsiooni eripärad andmete kontrollimise meetodite rakendamisel

Teise uurimisküsimuse "**Milliseid organisatsiooni eripärasid tuleb arvestada andmete kontrollimise meetodite rakendamisel?**" puhul analüüsiti nii Statistikaameti eripära asutusena, kui ka protsesse ja andmetega seotud iseärasusi.

Esimene eripära Statistikaameti puhul on andmestikud, mida kasutatakse. Andmestikud on pärit välistest andmekogudest, seega ei olnud võimalik võtta kasutusele andmekvaliteedi kontrollimiseks meetodeid, mis andmeid muudavad või puhastavad. Kui midagi on tarvis Statistikaametile laekunud andmetes muuta või korrigeerida, on tarvis kontakteeruda andmekoguga ning paluda andmekontrolli käigus selgunud probleem lahendada ning andmed uuesti Statistikaametile edastada. Eelneva tõttu on Statistikaametis kasutatavate kontrollimeetodite valik mõnevõrra piiratum kui neis asutustes, kes haldavad andmestikke ise ning saavad seeläbi paindlikumalt ning otsesemalt sekkuda andmestiku sisu parandamisse.

Teine eripära, mis takistab kasutamast meetodeid, mis konkreetsete probleemsete kirjete salvestavad ning näiteks andmeesitajale edastavad, on asjaolu, et andmekogudel on isikustatud andmestikud ning kui need andmestikud Statistikaametisse laekuvad, pseudonümiseeritakse kõik andmed andmekaitse eesmärgil. See tähendab, et igale isikule luuakse pseudonüüm, mille tagajärjel Statistikaameti analüütikutel puudub võimalus kirjet konkreetse isikuga seostada. Statistikaameti töötajatel, kes andmeid sisuliselt kontrollivad, ei ole seega võimalik välise andmeesitajaga suhtluses viidata konkreetsetele isikutele või kirjetele, millega probleem esineb, sest andmeesitajal ei ole asutuses kasutatavaid pseudonüüme. Statistikaameti töötaja peab andmekontrolli käigus leitud probleemi korral sisuliselt aru saama probleemi olemusest ning seda andmekogule selgitama, et neil oleks võimalik probleem keskselt lahendada.

Eripära, millega tuleb samuti arvestada on seotud organisatsioonis kasutatavate andmete olemusega. Kuigi kirjandusest ning varasematest uuringutest tuli välja veel analüüsimeetodeid, näiteks ärireeglite kontroll, DBSCAN-klasterdamine või kohustuslike väljade kontroll, siis nende meetodite kasutamine ei läinud käiku. Samuti ei läinud täies mahus käiku töös käsitletud organisatsioonis kasutatavate andmestike varieeruva struktuuri ning omapära tõttu näiteks unikaalsuse kontrollimeetod, seda küll rakendati üldisel tasandil, mis kontrollis, et andmestikus oleks iga kirje unikaalne, kuid loodud lahendus ei suutnud tuvastada igas kontrollitavas andmestikus unikaalset objekti, mille unikaalsust andmestikus kontrollida. Samuti ei rakendatud kohustuslike väljade kontrolli, sest kohustuslikud väljad ei ole keskselt määratud ning see kontroll eeldaks oluliselt suuremaid eelteadmisi ning tööriistale antavat sisendit kontrollitava andmestiku kohta. Selleks, et unikaalseid objekte või kohustuslike väljade kontrolli läbi viia, oleks vaja tööriista kasutajalt oluliselt rohkem sisendit kontrollitava andmestiku kohta, sest nii unikaalseid objekte kirjeldavad tunnuseid kui ka kohustuslikud tunnused on andmestike puhul erinevad. Tööriista puhul oli aga lihtne kasutamisele võtmine oluline, sest kui tööriist nõuab palju sisendit, näiteks oleks tarvis määrata iga kord tööriista kasutades kontrollitava andmestiku puhul täpne tunnus või tunnuste kombinatsioon, mis moodustab ühe unikaalse objekti ning oleks tarvis märkida kõik kohustuslikud väljad andmetes, mis ei või olla täitmata, oleks protsess ajamahukam ning eeldaks kasutajalt, et oleks laekunud andmestikuga põhjalikult tutvutud.

Viimaks on organisatsioonis tarvis vähemalt üldisemal tasandil mõista, mis andmetega on tegu ning selle põhjal seada andmete kontrollimise eesmärgid. Statistikaameti puhul kaasatud andmekontrolli automatiseerimise katsetusse andmestikud, mis laekuvad kord aastas ning kirjeldavad mõnd sündmuste, isikute või objektide nimekirja. Statistikaametile on oluline, et laekunud andmed oleksid täielikud, täpsed, asjakohased, korrektsed ning järjepidevad, seega on tarvis läbi viia andmekvaliteedi kontrollid kirjete arvu, unikaalsuse, järjepidevuse, kirjeldavate tunnuste jaotuse ning sisu kohta.

KOKKUVÕTE

Magistritöö eesmärk oli katsetada, andmekvaliteedi kontrolli- ja puhastusmeetodeid, mida sobib Statistikaametis rakendada eri tüüpi andmete puhul, et tagada andmete kvaliteet ning kujundada kontrolliprotsess, mis toetaks meetodite kohandamist vastavalt organisatsiooni ja andmetest tulenevatele vajadustele. Uurimisprobleem seisneb asjaolus, et aina kasvavate andmemahtude ja organisatsioonispetsiifiliste eripärade tõttu ei ole inimesel enam ise võimalik kõiki probleeme andmetes tuvastada, samuti ei ole ühetaolised andmekvaliteedi kontrollimise tööriistad alati usaldusväärsed ning on vaja kohandatavat lähenemisviisi.

Metoodiliselt keskendus töö automatiseeritud kontrollimeetodite rakendamisele Statistikaametis läbi viidavate andmekvaliteedi kontrollidele, mille käigus iga andmestiku kohta genereeriti raport erinevate näitajate (terviklikkus, ajakohasus, järjepidevus) alusel. Automatiseeritud kontrollimeetodite rakendamisel hinnati kasutegureid, mida lahendus pakub. Praktikas testiti kirjanduse põhjal valitud meetodeid reaalsetel Statistikaameti andmetel: analüüsiti vigade tüüpe, dokumenteeriti meetodite tulemuslikkust ning kaardistati mõjutegurid (andmete maht, laekumise sagedus, organisatsiooni piirangud).

Praktilise osa tulemustest selgus, et kasutegurid, mida automatiseeritud andmekvaliteedi kontrollimeetodite rakendamine Statistikaameti näitel pakuvad, on esiteks standardiseeritud andmekvaliteedi kontrollimise protsess ning dokumentatsioon. Teiseks kasuteguriks on uue tööriista rakendamisel andmete kvaliteedi kontrollimisele kuluva aja vähenemine viiendiku võrra. Kolmandaks kasuteguriks on käsitöö vähendamine kontrollide puhul, mida on võimalik lihtsasti automatiseerida ning seeläbi protsessi ühtluse tagamine. Veel on kasuteguriks teoreetiliste suundumuste ja praktilise vajaduse ühteliitmine.

Uurimuse käigus selgus, et uue lahenduse puhul on ka hulk nõrkuseid ja piiranguid, millega peab arvestama. Esiteks on uue lahenduse nõrkuseks asjaolu, et katsetuse käigus loodi lahendus, mis töötaks võimalikult paljude andmestikega, seetõttu aga oli spetsiifilisemate ja omapärasemate andmestike puhul tarvis teha mõningaid kohandamisi andmestiku struktuurile vastavalt, mis oli uue lahenduse puhul üsna tülikas. Kasutajate tagasiside tõi esile visuaalsete meetodite esitlemise

ja erinevate analüüsikomponentide väärtuse, aga ka vajaduse paindlikuma seadistuse järele. Teise nõrkusena selgus katsetuse käigus, et vähese juurutamise ja töötajate kaasamise tõttu varasemates etappides ei kasutanud kõik töötajad, kellele uus tööriist katsetamiseks pakuti, seda tööriista kasutama ning seetõttu oli kogu potentsiaalne kasu katsetusest ilmselt väiksem.

Piiranguks läbiviidud katse puhul on ühele konkreetsele organisatsioonile ning selles kasutatavatele andmestikele keskendumine ning seetõttu ei pruugi täpselt sama lahendus sobida teistes organisatsioonides kasutamiseks. Siiski on käesolevas töös läbiviidud katsetus heaks alguspunktis organisatsioonidele, kes kaaluvad enda andmekvaliteedi kontrollimise protsessi automatiseerimist.

Eripärad, millega tuleb organisatsioonil andmekvaliteedi kontrollimeetodite valikul arvestada on esiteks seotud andmestike päritoluga. Kui andmestikud on organisatsiooni enda omad või enda hallata, on võimalik rakendada andmekvaliteedi kontrollimisel rohkem meetodeid, mis andmeid ka parandavad või muudavad, mis muudab kogu protsessi sujuvamaks ja kindlasti ka kiiremaks. Statistikaameti puhul selliseid meetodeid rakendada ei saanud, sest andmestikud on väliste andmekogude valduses ning probleemide ilmnemisel oli tarvis, et andmekogu parandaks ise vead ning esitaks Statistikaametile parandatud andmestikud uuesti. Teiseks, eripära, millega tuleb arvestada, on organisatsioonis või organisatsioonile rakenduvad andmekaitse nõuded. Statistikaameti puhul pseudonümiseeritakse kõik isikustatud andmestikud, seega on konkreetsete kvaliteediprobleemidega kirjete tuvastamise kontrollimeetodite asemel tarvis kasutada kontrollimeetodeid, mis võimaldavad tuvastada probleemi põhjuse või sisu. Veel on oluline mõista organisatsioonis kasutatavate andmete olemust, kuigi teoreetiliselt on väga palju erinevaid kontrollimeetodeid, on oluline valida meetodid, mida on võimalik kontrollitavate andmestike peal rakendada. Viimaks on igas organisatsioonis kasutusel veidi erinevad andmestikud ning eesmärgid, milleks neid andmeid kasutada, ka need aspektid mõjutavad seda, milliseid andmekvaliteedi kontrollimeetodeid andmestikel on võimalik rakendada.

Töö edasiarendusena oleks huvitav viia katsetus läbi laiemas kontekstis, esmajoones näiteks Statistikaametis ka teistes osakondades, kellele laekuvad samuti erinevatest andmekogudest andmestikud, mille kvaliteeti on tarvis kontrollida. Samuti võiks laiem kontekst olla ka mingi hulk riigiasutusi, kus oleks võimalik katsetus läbi viia. Teisest küljest oleks huvitav kindlasti viia sarnane uurimus läbi asutuses, mis haldab andmestikke ise ning kus ei oleks kontrollimeetodite valikul piiravaks teguriks see, et andmestike sisu ei või muuta. Käesolev magistritöö annab siiski sisendit andmekvaliteedi kontrollimise automatiseerimise kasuteguritest organisatsiooni näitel ning ka meetodite valiku jaoks mõtteainet teistele asutustele.

SUMMARY

Automation of the data quality assessment process and its importance for the organization

The aim of this master's thesis was to test automated data quality assessment and cleansing methods suitable for implementation at Statistics Estonia's diverse datasets. Its primary objectives were to identify which assessment and cleaning techniques are suitable across different data types, ensure sustained data quality, and to design a flexible assessment process that adapts to both organizational requirements and dataset-specific characteristics. The research problem arises from rapidly growing data volumes and organization-specific features, which render purely manual checks infeasible and one-size-fits-all tools unreliable, highlighting the need for a customizable and adaptable approach.

The research implemented a suite of automated assessment routines, syntactic and semantic checks, completeness and consistency assessments, within an R Markdown framework. For each incoming dataset, the system generated standardized reports covering completeness, timeliness, and consistency metrics. Then those methods were applied to real data: cataloguing error types, documenting each technique's effectiveness, and mapping influencing factors such as data volume, submission frequency, and organizational constraints.

The results showed that automating the data quality assessment workflow delivered several tangible benefits for Statistics Estonia. First, it produced a standardized, well-documented process through which every incoming dataset was profiled, validated against predefined syntactic and semantic rules, and reported on uniformly, eliminating the previous ad hoc variation between individual analysts' scripts. Second, the new R Markdown–based tool reduced the total time spent on data checks by 20 percent, as repetitive tasks like format validation, missing-value analysis, range checks, duplicate detection and basic consistency tests were executed programmatically rather than by hand. Third, by removing routine manual work, thereby ensuring the uniformity of the process. Finally, this initiative successfully bridged established theoretical frameworks with real-world practical needs.

During the research, the new solution revealed important limitations and contextual considerations. Although the solution was architected to accommodate a wide range of dataset structures, specialized or unusually structured files still required manual configuration and script adjustments, which was quite cumbersome according to the experts. User feedback highlighted the value of presenting visual methods, as well as the need for more flexible settings. Due to the low level of employee involvement and implementation in the early stages, the solution suffered from uneven staff engagement: some analysts continued to rely on legacy procedures rather than adopting the new tool, thereby the total benefit from the new tool was probably lower.

Key considerations for selecting data quality assessment methods include the origin and governance of the datasets. When an organization owns or directly manages its data, it can deploy a broader array of techniques, including those that actively correct or transform records, making the overall process both smoother and faster. In contrast, Statistics Estonia must rely on externally managed registers; any detected errors require the data provider to remediate and resubmit the corrected files rather than allowing in-place modification. A second critical factor is the data protection framework under which the organization operates. At Statistics Estonia, all personal data are pseudonymized, precluding assessment methods that pinpoint individual records for correction; instead, the focus shifts to methods capable of diagnosing the root cause or nature of quality issues without re-identifying subjects. Finally, it is essential to understand the specific characteristics and intended uses of the data in question. Although a wide variety of assessment techniques exist in theory, only those compatible with the organization's actual datasets and use cases should be chosen. Moreover, because each organization—and even each department—may work with distinct data structures and analytical goals, method selection must be tailored to the combination of dataset design and downstream requirements.

In further studies extending the test to other Statistics Estonia divisions and to peer government bodies, especially those managing their own data, would test the approach's generalizability. Such studies could explore more advanced semantic checks and investigate deeper analyst integration to enhance uptake. This thesis lays a foundational framework and highlights critical factors for organizations aiming to automate their data quality control processes.

KASUTATUD KIRJANDUS

Ahiagble, A. P., & Stein, H. (2022). Approaches for automated data quality analysis: Syntactic and semantic assessment. In GI-Jahrestagung (pp. 1023–1036). Gesellschaft für Informatik. https://dl.gi.de/items/31e34a8f-e71d-44f4-bcaf-5eca9b554335

Anaconda Foundation. (2020). *The state of data science 2020: Moving from hype toward maturity*. https://www.anaconda.com/state-of-data-science-2020

Andmekaitse Inspektsioon. (2024). Mõisted. Kasutatud 25.05.2025. https://www.aki.ee/isikuandmed/kkk/moisted

Batini, C., & Scannapieco, M. (2016). Data and information quality: Dimensions, principles, and techniques. Springer.

Charles, E. (2024). Data Validation Techniques for Ensuring Data Quality.

Epperson, W., Gorantla, V., Moritz, D., & Perer, A. (2023). Dead or Alive: Continuous Data Profiling for Interactive Data Science. IEEE Transactions on Visualization and Computer Graphics. PP. 1-11. 10.1109/TVCG.2023.3327367.

Etikala, S. (2025). *Enterprise Data Quality Management: A Technical Deep Dive into Automated Governance*. International Journal on Science and Technology, 16(2), Article 3050.

Heer, J. (2019). Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences, 116*(6), 1844–1850. https://doi.org/10.1073/pnas.1807184115

Kalmus, V., Masso, A., & Linno, M. (2015). Kvalitatiivne sisuanalüüs. Sotsiaalse analüüsi meetodite ja metodoloogia õpibaas. Retrieved May 26, 2025, from https://samm.ut.ee/kvalitatiivne-sisuanalyys/

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the*

International Working Conference on Advanced Visual Interfaces (AVI '12) (pp. 547–554). Association for Computing Machinery. https://doi.org/10.1145/2254556.2254659

Otto, B. (2011). Organizing data governance: Findings from the telecommunications industry and consequences for large service providers. Communications of the Association for Information Systems, 29(3).

Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. *Nature News.* https://doi.org/10.1038/d41586-018-07196-1

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3–13.

Rootalu, K. (2014). Kirjeldav statistika. Sotsiaalse analüüsi meetodite ja metodoloogia õpibaas. Retrieved May 26, 2025, from https://samm.ut.ee/kirjeldav-statistika/

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* Association for Computing Machinery. https://doi.org/10.1145/3411764.3445518

Seltman, H. (2018). *Experimental design and analysis.* Carnegie Mellon University. http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

Siddiqui, T., Kim, A., Lee, J., Karahalios, K., & Parameswaran, A. (2016). Effortless data exploration with zenvisage: An expressive and interactive visual analytics system. *Proceedings of the VLDB Endowment, 10*(4). https://doi.org/10.14778/3025111.3025126

Statistikaamet. (2020). *Andmekvaliteedi juhis 2020.* https://www.stat.ee/sites/default/files/2022-03/Andmekvaliteedi%20juhis_2020.pdf

Statistikaamet. (2022). 2021. aasta registripõhise loenduse metoodika kirjeldus. Kasutatud 20.05.2025. https://www.stat.ee/sites/default/files/2022-06/Registripõhise%20loenduse%20metoodika%20raport.pdf

Statistikaameti koduleheküIg. (2025). Kasutatud 20.05.2025, https://stat.ee/et/statistikaamet/meist

Sügis, E., Tampuu, A., Aljanaki, A., Fišel, M., & Kull, M. (2025). *Praktiline andmeteadus: Kõrgkooliõpik* (uuendatud jaanuar 2025). Tartu Ülikooli arvutiteaduse instituut.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Henry Lass,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Andmekvaliteedi hindamise protsessi automatiseerimine ja selle olulisus organisatsioonile", mille juhendajad on Toomas Saarsen ja Terje Trasberg, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;
2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;
3. olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;
4. kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Henry Lass

28.05.2025