Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Hanna Britt Parman

Ettevõtte kompleksse kalendriaasta aruande (EKOMAR) puuduvate müügitulu väärtuste imputeerimine majandusaasta aruannetele tuginedes

Matemaatilise statistika eriala
Bakalaureusetöö (9 EAP)

Juhendajad: Kristi Lehto, MSc.
Mare Vähi, MSc.

Tartu 2020

ETTEVÕTTE KOMPLEKSSE KALENDRIAASTA ARUANDE (EKOMAR) PUUDUVATE MÜÜGITULU VÄÄRTUSTE IMPUTEERIMINE MAJANDUSAASTA ARUANNETELE TUGINEDES

Bakalaureusetöö

Hanna Britt Parman

Lühikokkuvõte

Töö eesmärk on leida eeskiri ettevõtte kompleksse kalendriaasta aruande (EKOMA-Ri) müügitulu puuduvate väärtuste imputeerimiseks. Siiani on imputeerimiseks kasutatud keskväärtust ja enne kihtidepõhise keskväärtuse leidmist on erindeid eemaldatud eksperthinnangu alusel, selle asemel proovitakse nüüd leida kindlat eeskirja mingi protsendi suuremate (ja väiksemate) vaatluste eemaldamiseks. Saadud tulemuse täpsust hinnatakse, kasutades majandusaasta aruande andmestikku, kus on olemas umbes poolte imputeeritud väärtustega ettevõtete tegelik müügitulu. Üldkogumi kihtideks jaotamisel katsetatakse kahte alternatiivset meetodit: senine kihistamise meetod ning uus meetod töötajate arvu ja Eesti Majanduse Tegevusalade Klassifikaatori (EMTAKi) kolmekohalise koodi kombinatsioonina. Imputeerimisel kasutatakse kihi 10% või 20% kõige suurematest väärtuste eemaldamist kui ka kihi 10% või 20% kõige suuremate ja väiksemate väärtuste eemaldamist enne keskväärtuse arvutamist.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: imputeerimine, trimmitud keskväärtus, andmeanalüüs.

IMPUTATION OF MISSING TURNOVER VALUES IN THE COMPREHENSIVE ANNUAL QUESTIONNAIRE FOR ENTERPRISES (EKOMAR) BASED ON ANNUAL ACCOUNTS

Bachelor thesis

Hanna Britt Parman

Abstract

The objective of this Bachelor's thesis is to determine a rule for imputing turnover values in the comprehensive annual questionnaire for enterprises (EKOMAR). Experts' judgement has been used to exclude outliers from the mean imputation thus far, but in this thesis various alternatives of excluding a fixed percent of the largest (and smallest) values are proposed. The outcome is evaluated by comparing the imputed values with the values from the annual reports, where approx. half of the imputed values have a corresponding true value. For the stratification, two alternative methods are used: the current stratification method and a new alternative based on the number of employed people in the enterprise and the triple digit code of the Estonian Classification of Economic Activities (EMTAK). Imputation is done by removing 10% or 20% of the largest values from the stratum or by removing 10% or 20% of the largest and smallest values from the stratum before mean calculation and imputation.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: imputation, trimmed mean, data processing

Sisukord

Sissejuhatus

Reaalses elus kasutavates andmekogumites on tavaline, et andmestikud on lünklikud ehk leidub puuduvaid väärtusi. Andmete analüüsimiseks on aga enamasti vaja täielikku andmekogumit. Andmestiku lünklikkusega tegelemisel on mitmeid erinevaid võimalusi, klassikalisemad meetodid on kõigi puuduvate väärtustega objektide välja jätmine või keskväärtusega imputeerimine [1].

Kuigi need meetodid on laialt kasutust leidvad, siis ei tähenda see seda, et need oleksid parimad viisid väärtuste imputeerimiseks. Puuduvate väärtustega objektide eemaldamisega on võimalik andmestikust olulist informatsiooni välja jätta. Nii objektide eemaldamisega kui ka keskväärtusega (mediaani, moodiga) imputeerimisel võib hiljem saadud andmestiku analüüsimisel saada nihkega hinnangud keskväärtusele ja dispersioonile. [2]

Lünklike andmestike töötlemine on andmeanalüüsis läbiv probleem, millega valesti ümber käimisel võib hiljem hinnanguid arvutades nihe tekkida. Seetõttu on leitud ka võimsamaid lahendusi andmete imputeerimiseks, mis õigesti rakendamisel aitavad taastada andmestiku tegeliku kuju ja selle abil hinnangute täpsust tõsta. Sellised meetodid on näiteks mitmene imputeerimine (MI), mitmese imputeerimise algoritm (MICE). [3]

Oluline on ka see, et lünklike andmestike analüüs ei hõlma ainult andmete imputeerimist, vaid algab esmalt puuduvate väärtuste diagnostikaga. Diagnostika on vajalik, kuna puuduvaid väärtusi saab klassifitseerida nii täiesti juhuslikult puuduvateks väärtusteks (MCAR), juhuslikult puuduvateks väärtusteks (MAR) või mittejuhuslikult puuduvateks väärtusteks (NMAR). Pärast andmestiku analüüsimist on võimalik leida sobivaim meetod lünkliku andmestiku töötlemiseks. [3, 4]

Antud bakalaureusetöö keskendub Statistikaameti ettevõtte kompleksse majandusaasta aruande (EKOMAR) andmestiku puuduvate müügitulu väärtuste jaoks parima imputeerimismeetodi leidmisele. EKOMARi andmete kokkupaneku ajal imputeeritakse nende ettevõtete väärtused, kes ei ole sellel hetkel majandusaasta aruannet esitatud(väikese töötajate arvuga kihtides). Osa ettevõtteid, kelle andmed EKOMARist puuduvad, on oma majandusaasta aruande hiljem siiski esitanud. Seega on nende ettevõtete puhul võimalik imputeeritud väärtusi ja nende jaotust tegelike väärtustega võrrelda. Hetkel imputeeritakse EKOMARi andmestikus kihi keskmisega, kuna andmestikus on palju näitajaid ja see võimaldab kõiki imputeerida ühtselt. Müügitulu imputeerimiseks võib leida lisainfor-

matsiooni, kuid kuna kõik näitajad ei ole seotud müügituluga, on siiamaani eelistatud kihikeskmisega imputeerimist. EKOMARi koostamisel kasutatakse lisainfona registrite ja infosüsteemide keskuse äriregistrist saadud ettevõtete majandusaasta aruannete andmeid ning maksu- ja tolliametilt saadud käibedeklaratsiooni, tulu- ja sotsiaalmaksu, kohustusliku kogumispenisoni makse ja töötuskindlustusmakse deklaratsiooni ning füüsilise isiku tuludeklaratsiooni[5].

1 Puuduvate väärtuste liigitus

Puuduvate väärtuste analüüsimisel on üks olulisemaid probleeme selle selgeks tegemine, kas puudumine on juhuslik, või on selle mustri taga sõltuvus kas tunnuse enda või mõne teise tunnuse väärtusest.

Statistik Donald B. Rubin liigitas aastal 1976 oma teoses "Inference and Missing Data" puuduvad väärtused kolme kategooriasse. [6]

Kategooriate kirjeldamiseks kasutame tähistuseks andmemaatriksit $\mathbf{X}$, mis on täielike ja puudulike andmete järgi kaheks jagatud:

$$\mathbf{X} = (\mathbf{X}_{OBS} + \mathbf{X}_{MIS}),$$

kus $\mathbf{X}_{OBS}$ tähistab olemasolevaid ja $\mathbf{X}_{MIS}$ puuduvaid väärtusi.

Olgu puudumisindikaatorite maatriks $\mathbf{M} = \{m_{ij}\}$, kus $m_{ij} = 1$, kui andmete maatriksis on väärtus olemas ja $m_{ij} = 0$, kui väärtus puudub.

Täiesti juhuslik puudumine (Missing Completely at Random e MCAR) kirjeldab juhtu, kui

$$P(\mathbf{M}|\mathbf{X}) = P(\mathbf{M}).$$

Kuna selline puudumine on andmetest endast sõltumatu, siis on see kolmest kõige lihtsam juht. Ainuke probleem võrreldes täielike andmetega on informatsiooni kadu väärtuste puudumise tõttu. Siiski tuleb mainida, et kuigi sellise puudumisega on imputeerimisel lihtsaim ümber käia, siis on see reaalsete andmestike peal tihti ebarealistlik.[7]

Juhuslik puudumine (Missing at Random e MAR) kirjeldab juhtu, kui

$$P(\mathbf{M}|\mathbf{X}) = P(\mathbf{M}, \mathbf{X}_{OBS}).$$

Selle juhu puhul on seega väärtuse puudumine sõltuv mingi teise tunnuse väärtusest ning see kategooria on laiem kui täiesti juhuslik puudumine ning ka realistlikum. Seetõttu ka paljud tänapäevased andmete asendamise meetodid eeldavad, et andmed puuduvad juhuslikult.[7]

Mittejuhuslik puudumine (Not Missing at Random e NMAR) kirjeldab juhtu, kui

$$P(\mathbf{M}|\mathbf{X}) = P(\mathbf{X}_{OBS}, \mathbf{X}_{MIS}).$$

Seega väärtused puuduvad mittejuhuslikult, kui puuduvate väärtuste puudumise tõenäosus ei ole kõigi väärtuste puhul võrdne, kuid selle põhjus on teadmata. See juht on kõige keerulisem ja selle lahendamiseks tuleks enne asendamise üritamist siiski välja selgitada tegurid, mis võiks puudumiste mustrit seletada.[7]

2 Keskmisega imputeerimine

Käsiraamatu "Handbook on Methodology of Modern Business Statistics"[6] imputeerimisest rääkivas peatükis on välja toodud kuus erinevat alapeatükki erinevate imputeerimisviisidega. Neli peatükki on üldisemate imputeerimismeetodite kohta: deduktiivne, mudelipõhine ja doonoripõhine imputeerimine ning Little' ja Su meetod. Ülejäänud kaks peatükki on kordusmõõtmiste imputeerimise ning redigeerimispiirangutega väärtuste imputeerimise kohta. Hetkel EKOMARi andmestikus imputeerimiseks kasutatav keskmisega imputeerimise meetod kuulub mudelipõhise imputeerimise alapeatüki alla.

2.1 Meetodi kirjeldus

Keskmisega imputeerimise kirjutamise meetodi kirjeldus on samuti koostatud käsiraamatu "Handbook on Methodology of Modern Business Statistics"[6] põhjal. Käsiraamat on koostatud Eurostati poolt rahastatud programmi Memobust raames.

Keskmisega imputeerimisel asendatakse iga puuduv väärtus kõigi olemasolevate vaatluste keskmisega. Seega tunnuse y i-ndat väärtust, kui see on puuduv, imputeeritakse vastavalt:

$$\tilde{y}_i = \overline{y}_{OBS} = \frac{\sum\limits_{k \in OBS} y_k}{n_{OBS}},$$

kus OBS on olemasolevate vaatluste alamhulk.

Keskmisega on võimalik imputerida ka väiksemates kihtides, mis on imputeeritava tunnuse mõttes võimalikult homogeenne. Seega tunnuse y i-ndat väärtust, mis kuulub kihti h imputeeritakse vastavalt:

$$\tilde{y}_{hi} = \overline{y}_{h;OBS} = \frac{\sum\limits_{k \in h \cap OBS} y_{hk}}{n_{h;OBS}},$$

kus n_{OBS} on väärtust omavate vaatluste arv kihis h.

Ettevõtlusuuringute kontekstis soovitatakse käsiraamatus defineerida kihid majandustegevuse ja töötajate arvu järgi, kuna nende tunnuste järgi grupeeritult uuringu tulemit tavaliselt ka väljastatakse. Nende tunnuste järgi defineeritud kihinumbri järgi kihistamist ja kihikeskmisega imputeerimist kasutatakse EKOMARi andmestikus ka praegu.

2.2 Eelised ja puudused

Keskmisega imputeerimise suureks eeliseks on meetodi lihtsus, kuid selle juures on ka olulisi puudusi. Üks tähtsamatest on see, et keskmisega imputeerimine toob endaga kaasa ka osade dispersioonil põhinevate parameetrite nihke, kuna dispersioon väheneb. Keskmisega imputeerimine ei arvesta ka objektide suhteid, mis omakorda vähendab objektide vahelist korrelatsiooni. Keskväärtuse parameeter jääb imputeerimisel aga paika, mis on selle meetodi üks eeliseid. [6]

Kuigi ka kihtides imputeerimine toob endaga kaasa dispersiooni vähenemise, siis hästi jaotatud kihtide puhul on võimalik saada võrreldes tavalise keskmise imputeerimisega siiski parem tulemus. Juhul kui tunnuse dispersioon kihtide vahel on palju suurem kui kihtide sees, võimaldab see dispersiooni ja selle põhjal arvutatud teisi statistikuid täpsemalt hinnata. [6]

2.3 Trimmitud keskmine

Keskväärtus ja dispersioon on kasulikud statistikud juhul, kui andmed on normaaljaotusega. Teistsuguste ja eriti mittesümmeetriliste jaotuste puhul tasub kaaluda ka alternatiivseid statistikuid. Üks võimalik asendus tavalisele keskmisele on trimmitud keskmine, mille puhul eemaldatakse enne keskmise arvutamist mingi protsent suuremaid ja väiksemaid vaatlusi.[8]

3 Andmestike kirjeldus

Kasutatud algandmestikud on ettevõtte kompleksse majandusaasta aruande (EKOMAR) [5] andmestik aastatel 2013-2017 ja majandusaasta aruande (MAA) kasumiaruande 2009-2018 aastate andmestik. EKOMARis vaadeldakse riigi ja kohalike omavalitsuste ettevõtteid ning vähemalt 20 hõivatuga eraettevõtteid kõikselt, ülejäänud ettevõtetest tehakse stratifitseeritud lihtne juhuslik valik ettevõtte tegevusala ja hõivatute arvu järgi.

Analüüsis kasutatakse EKOMARi andmestikus järgmiseid tunnuseid:

- YEAR - aasta,
- EMTAK - ettevõtte tegevusala EMTAK 2008 järgi[9],
- SA_ID - ettevõtte anonümiseeritud ID-kood,
- tarvank - ettevõtte hõivatud töötajaid,
- kihinr - kihinumber,
- C_10 - müügitulu,
- TIN_IMPU - andmete päritolu.

Tunnuse EMTAK abil kodeeritakse EKOMARi andmestikku veel neli tunnust (Lisa 1):

- EMTAK_grupp_1 - tähtkood,
- EMTAK_grupp_2 - kahekohaline numberkood,
- EMTAK_grupp_3 - kolmekohaline numberkood,
- tootaja_grupp - EMTAK koodi järgi grupeeritud hõivatute arv ettevõttes (EMTAK $<$ 40000 puhul tasemetega 1-9, 10-19 ja 20 või rohkem töötajat ning EMTAK $>$ 41000 puhul tasemetega 1, 2-9, 10-19 ja 20 või rohkem töötajat)

MAA andmestikust kasutatakse järgmiseid tunnuseid:

- PERIOD_NM - aasta
- SA_ID - ettevõtte anonümiseeritud ID-kood

- Ka_50_1 - müügitulu

Andmete päritolu tunnusel TIN_IMPU on kolm võimalikku väärtust: tühik, kui andmed on pärit EKOMARi küsitlusest, MAA, kui andmed pärinevad MAAst, ning $KESK$, kui andmed on imputeeritud. Andmed on imputeeritud keskväärtusega kihtides, mis on määratud tunnuse $KIHINR$ järgi. Tunnus $KIHINR$ on moodustatud ettevõtte tegevusala, hõivatute ja omandivormi (riigi ja kohaliku omavalitsuse ettevõtted ning eraettevõtted) järgi. Enne keskväärtuse arvutamist on vahel kihtidest eemaldatud mõned erandid, kuid seda tehes ei ole järgitud otsest eeskirja vaid eemaldamine on tehtud statistiku eksperthinnangu alusel.

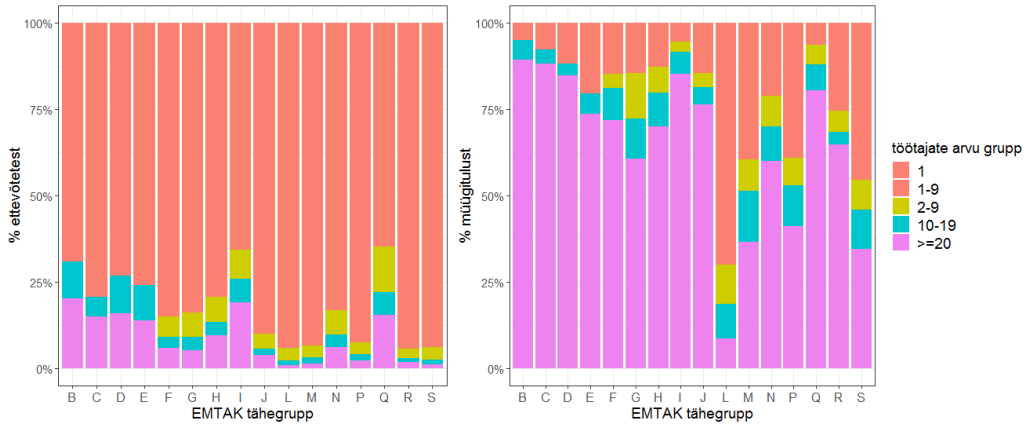
Üle 20 hõivatuga ettevõtete andmed saadakse EKOMARi küsitlusest ja mittevastanute andmed asendatakse administratiivandmete, eelmise perioodi või olemasolevate lühiajastatistika andmetega. Ülejäänud ettevõtete EKOMARi küsitlusse kaasamise metoodikat on aastatel 2013, 2014 ja 2015 muudetud. Aastal 2013 lõpetati alla kümne hõivatuga tööstusettevõtete kaasamine EKOMARi küsitlusse, aastal 2014 lõpetati ka ühe hõivatuga ehitus- ja jaekaubandusettevõtete kaasamine küsitlusse ning aastast 2015 ei kaasata ka ühe hõivatuga teisi kaubandus- ja teenindusettevõtteid. Nende andmeid hinnatakse, kasutades äriregistrile esitatud majandusaasta aruandeid.

4 Imputeerimine

4.1 EKOMARi andmestiku analüüs

Alates aastast 2015 kehtiva metoodika järgi on puuduvaid väärtuseid vaja imputeerida tööstustegevusaladel 1-9 töötajaga ettevõtete seas ja ehitus-, kaubandus- ja teenindustegevusaladel ühe töötajaga ettevõtete seas, kuna neid ettevõtteid küsitlusse ei kaasata ning kasutatakse MAA andmeid, mida kõik ettevõtted aga õigeaegselt ei esita. Töötajate arvude kihtides 2-9 ja 10-19 valikuuringu kihtides imputeerimist ei kasutata, andmete laiendamiseks arvutatakse igale vastanud ettevõttele kaal. Need ettevõtted, kellel on 20 või rohkem töötajat, peavad esitama oma andmed EKOMARi küsitluses. Seetõttu analüüsitakse EKOMARi andmestiku puhul edaspidi vaid aastate 2015-2017 andmeid, kus tunnuse *tootaja_grupp* väärtus on kas 1 või $1-9$.

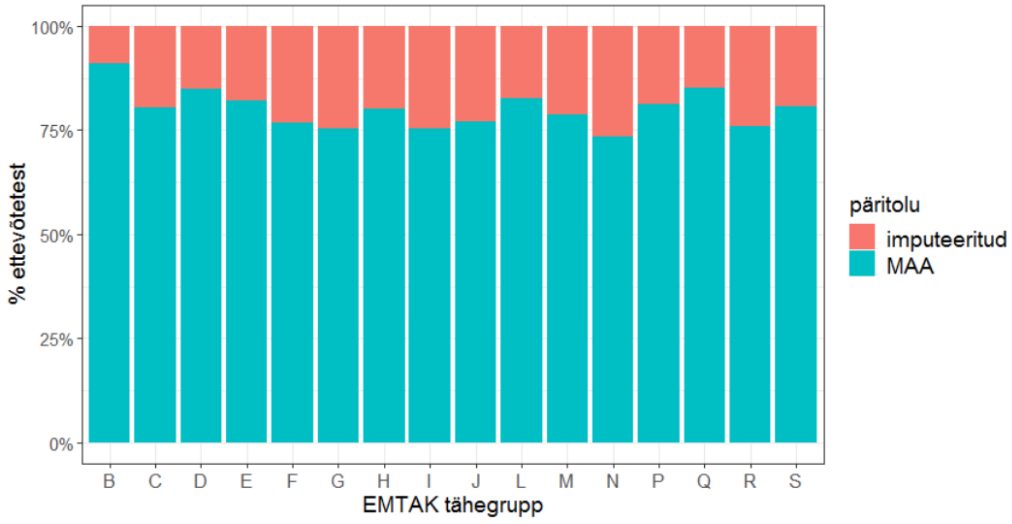
Jooniselt 1 on näha, et selliseid ettevõtteid on EKOMARi andmestikus igas tähtgrupis kõige rohkem. Nende ettevõtete müügitulu osakaal on aga enamikes tähegruppides väiksem kui 25%.



Joonis 1. Ettevõtete arvu ja müügitulu osakaal töötajate arvu grupi järgi aastate 2015-2017 andmete põhjal

Esmalt vaadatakse joonise 2 abil puuduvate väärtuste osakaalu igas EMTAKi tähtkoodi järgi defineeritud grupis. Enim ehk umbes 27% vaatlustest on imputeeritud tähtkoodi N puhul ehk haldus- ja abitegevuste sektoris. Vähim on imputeerimiste osakaal aga sektoris B ehk mäetööstuse sektoris, kus on imputeeritud umbes 9% vaatlustest. Sektoris B on ka

kõigist sektoritest töötajate grupis 1-9 kõige vähem ettevõtteid.



Joonis 2. 1 ja 1-9 töötajaga ettevõtete imputeeritud vaatluste osakaal müügitulust EKOMARi andmestikus 2015-2017 aastate andmete põhjal

Kui ettevõte on esitanud MAA pärast EKOMARi esitamise tähtaega, siis on võimalik võrrelda selle ettevõtte EKOMARi imputeeritud andmeid tegelike MAA andmetega. Samas on ettevõtteid, kes ei ole MAAd esitanud ka hiljem ehk nende imputeeritud väärtuste tegelik väärtus on siiani teadmata. Võimalik, et ettevõte on tegevuse lõpetanud, kuigi aruande aastal veel tegutses, või on andmed mõnel muul põhjusel esitamata. Edaspidi nimetatakse EKOMARi müügitulu, mis on saadud MAA põhjal, õigeaegselt esitatud MAA müügituluks. Seda müügitulu, mis on EKOMARi andmestikus imputeeritud, kuid mille kohta on MAA andmestikus väärtus olemas, nimetatakse edaspidi hiljem esitatud MAA müügituluks.

EKOMARi ja MAA andmestikke saab ühendada aasta ja ettevõtte ID-koodi järgi ning nende sellisel ühendamisel on näha, et EKOMARi andmestikus leidub kolm vaatlust, mille puhul on tunnuse TIN_IMPU väärtus $KESK$ ehk märgitud, et nende vaatluste väärtused on kihikeskmisega imputeeritud, kuid vaatluse müügitulu on võrdne MAA andmestiku mittenullilise müügituluga. Seetõttu asendatakse tunnuse TIN_IMPU väärtuseks nende vaatluste puhul MAA.

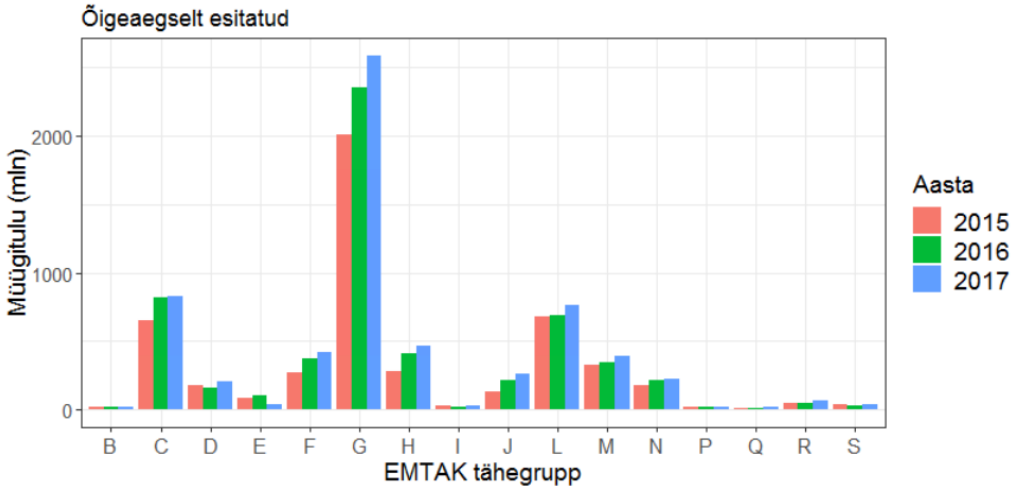
Imputeeritud firmade andmeid EMTAKi tähtkoodi järgi vaadates tuleb välja, et sama aasta MAA müügitulu olemasolu on iga tähtkoodi puhul sarnane ehk umbes pooltel juhtudest on sama aasta MAA müügitulu hilinenud väärtus olemas (täpsed tähtkoodi järgi moodustatud

Tabel 1. EKOMARi imputeeritud üksuste arv hilisema MAA olemasolu järgi EMTAKi tähekoodi gruppides töötajate arvuga 1 ja 1-9 gruppides, 2015-2017

hilisem MAA	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**
puudu	12	1630	29	62	1740	3719	829	361	1183	1047	2602	1234	196	91	511	564
olemas	15	1674	50	54	1587	3475	712	335	1435	1414	3637	1122	271	98	719	606

gruppide protsendid on vahemikus 46,6%-63,3%).

Aastad 2015-2017 on olnud majanduse stabiilse kasvu aastad, seda kinnitab ka joonis 3, on näha, et müügitulu on enamikes sektorites kasvanud.



Joonis 3. MAA õigeagselt esitanud ettevõtete müügitulu summa aastate lõikes

Hilisemalt esitatud MAA kasvutrend ei ole aga kõigis tegevusalades proportsionaalne õigeaegselt esitatuga. Osade sektorite puhul on hiljem esitanute hulka mõnel aastal sattunud üksikud suure väärtusega müügitulu vaatlused, mis suurendavad oluliselt kihi müügitulu summat ja seetõttu ka keskväärtust.

Kuna keskväärtust ei imputeerita andmestikus tähtkoodi lõikes, vaid täpsemates EMTAKi kolmekohaliste koodi gruppides, siis vaadatakse keskväärtust esmalt kolmekohalise EMTAKi koodi järgi.

Vaadeldavate andmete puhul leidub kolmekohalise EMTAK koodi ja aasta kombinatsiooni järgi 639 erinevat kihti, nendest 513 puhul on võimalik võrrelda hiljem esitatud MAA kesk-

mist müügitulu ja õigeaegselt esitatud MAA keskmist müügitulu (s.t et kihis on vähemalt üks imputeeritud vaatlus, mille kohta on hiljem esitatud MAA müügitulu ning vähemalt üks õigeaegselt esitatud MAA vaatlus).

Tabel 2. Õigeaegselt ja hiljem esitatud MAA keskmise müügitulu võrdlus EMTAK kolmekohalise koodi ja aasta järgi moodustatud kihtides

aasta	**õigeaegse MAA keskmine müügitulu suurem**	**n**	**õigeaegse MAA keskmine müügitulu suurem/n**
2015	115	172	0.67
2016	119	167	0.71
2017	96	174	0.55

Niimoodi kihistades on igal aastal õigeaegselt esitatud MAA keskmine müügitulu enamikes kihtides suurem kui hiljem esitatud MAA keskmine müügitulu.

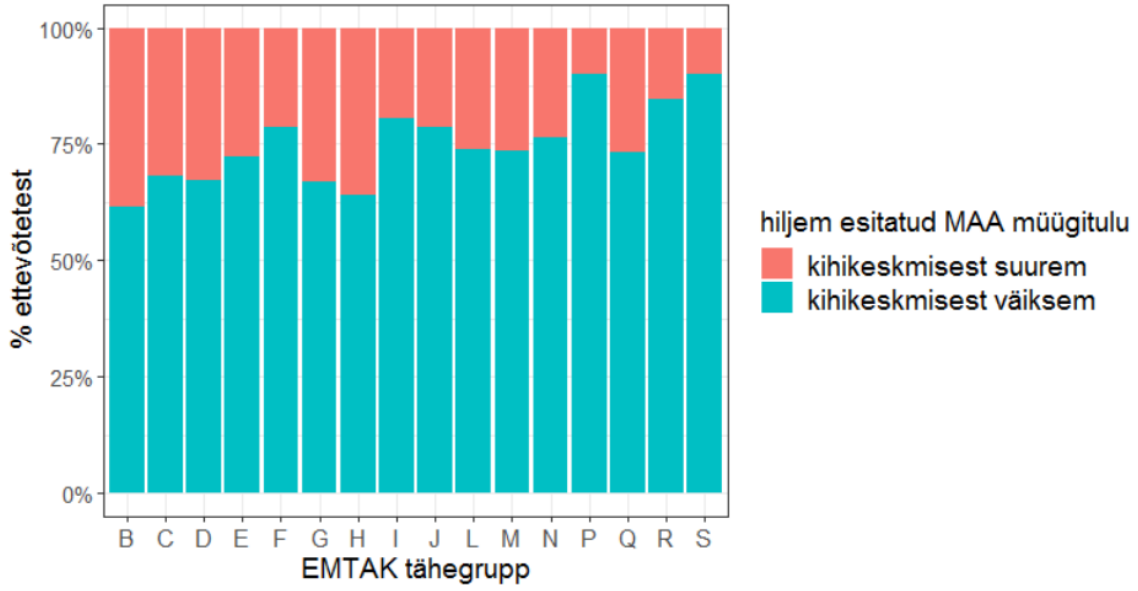
Igas kihis leitakse ka kihinumbri järgi defineeritud kihtides õigeaegselt esitatud MAA müügitulu põhjal kihikeskmine ja võrreldakse MAA hiljem esitanud ettevõtete müügituluga. Joonisel 4 on näha, et kõigis tähegruppides on enamikul juhtudel hilinenud müügitulu kihikeskmisest väiksem.

Joonisel 5 vaadatakse eraldi ka sektorit G ehk hulgi- ja jaekaubanduse ning mootorsõidukite ja mootorrataste remondi sektorit, mille müügitulu ja töötajate arv on kõige suurem. Ka siin on näha, et igas kahekohalise EMTAKi järgi määratud grupis on enamik ettevõtteid need, mille hiljem esitatud MAA müügitulu on õigeaegselt esitatud MAA müügitulu kihikeskmisest väiksem.

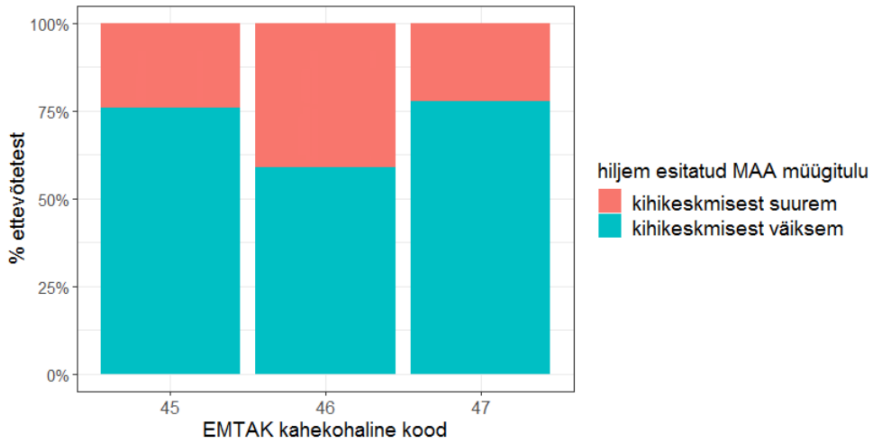
Seega jooniste 4 ja 5 põhjal saab teha järelduse, et hiljem MAA esitanud ettevõtete müügitulu on õigeaegselt MAA esitanud ettevõtete keskmisest müügitulust väiksem. See tähendab, et keskväärtusega imputeerides hinnatakse paljude ettevõtete müügitulu üle.

Analüüsitakse ka väiksemate müügituluga ettevõtete jaotust. Andmestikust eraldatakse kõik puuduvate väärtustega ettevõtted, kelle MAA põhjal esitatud müügitulu ei ole üle miljoni euro.

Ka jooniselt 6 saab kinnitust sellele, et kõige üldisemas plaanis imputeeritakse äärmiselt

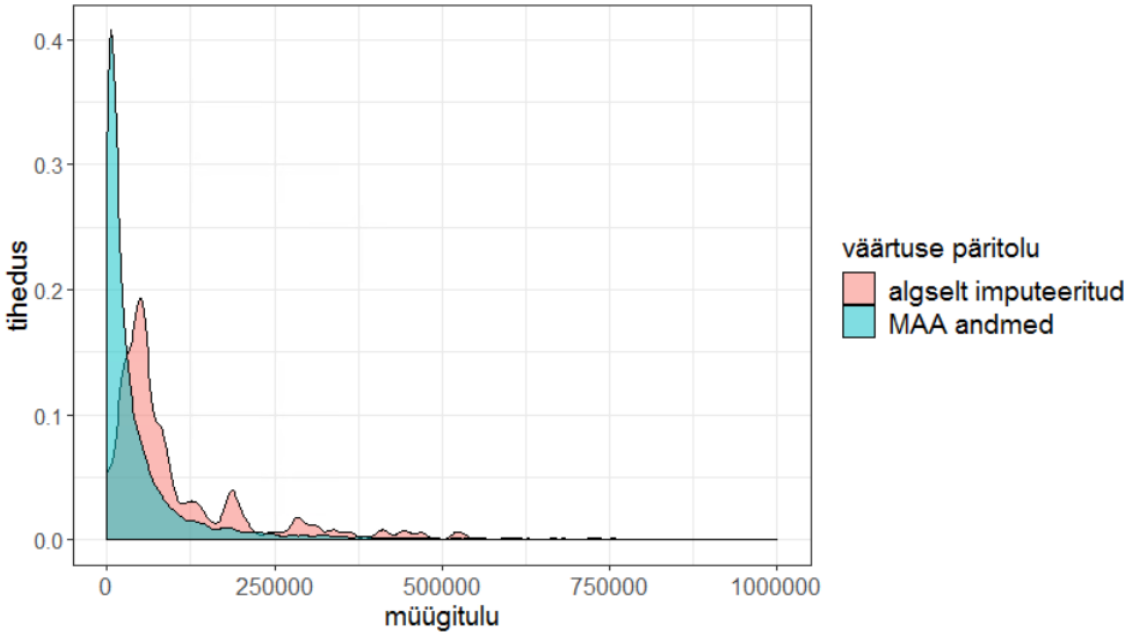


Joonis 4. MAA hiljem esitanud ettevõtete müügitulu võrdlus õigeaegselt esitatud MAA kihikeskmisega EMTAK tähegruppides



Joonis 5. MAA hiljem esitanud ettevõtete müügitulu võrdlus õigeaegselt esitatud MAA kihikeskmisega EMTAK tähegrupis G kahekohalise EMTAK tähekoodi lõikes

väikse müügituluga ettevõtetele liiga suur müügitulu. Seetõttu võib imputeerimisel trimmitud keskväärtuse kasutamine keskväärtuse asemel anda tõelisele jaotusele sarnasema jaotuse.



Joonis 6. MAA andmete ja EKOMARist pärit müügitulu jaotus (müügitulu kuni väärtuses miljon eurot)

4.2 Imputeerimine

Analüüsi põhjal on selge, et üksikud ekstreemsed vaatlused mõjutavad kihikeskmist oluliselt, seetõttu tasuks need keskmise arvutamisel välja jätta.

Kuna siiamaani on EKOMARi andmestikust enne keskväärtuse arvutamist kihtides äärmuslikke vaatlusi eemaldatud manuaalselt ilma kindla eeskirjata, siis proovitakse nüüd leida parim fikseeritud viis väärtuste eemaldamiseks. Nendes kihtides, kus on üle kümne olemasoleva vaatluse, kasutatakse müügitulu imputeerimisel trimmitud keskmist. Seega igast sellisest kihist eemaldatakse olemasolevate väärtuste seast mingi protsent äärmuslikest väärtustest ja imputeeritakse siis olemasolevate väärtuste põhjal arvutatud uue kihikeskmisega. Kihtides, kus on kümme või vähem olemasolevat vaatlust, imputeeritakse tavalise kihikeskmisega.

Parima lahenduse leidmiseks viiakse läbi 10% ja 20% eemaldamine nii kõige suurematest väärtustest kui ka kõige suurematest ja väiksematest väärtustest. Sellised imputeerimised tehakse läbi nii algses andmestikus tunnuse $KIHINR$ järgi defineeritud kihtides kui ka tunnuste $EMTAK\_grupp\_3$ ja $tootaja\_grupp$ abil defineeritud kihtides. Uus kihista-

mise meetod jaotab ettevõtted väiksemasse arvu eri kihtidesse kui algne. Uue tunnuse *impu_kiht* järgi jagunevad need ettevõtted, mille *tootajate_grupp* väärtus on kas 1 või $1-9$, 315 erinevasse kihti. Algse kihtide jaotuse järgi jaotatakse kõigi aastate peale kokku samad ettevõtted 1177 erinevasse kihti. Seega on uue jaotuse järgi erinevaid kihte peaaegu kolm korda vähem.

Uute kihtide puhul leidub aastate 2015-2017 kohta kokku 11 kihti, kus on ainult puuduvad väärtused, ehk neis kihtides pole võimalik midagi imputeerida. Ka tunnuse $KIHINR$ järgi on hetkel selliseid kihte, kus imputeerimine on võimatu, selle lahendamiseks on imputeeritud mõne teise lähedase kihi keskväärtusega.

Seega saadakse kokku kaheksa alternatiivset viisi puuduvate väärtuste imputeerimiseks (Lisa 2).

4.2.1 Müügitulu kogusumma

Alternatiivseid imputeeritud väärtusi saab esmalt võrrelda müügitulu kogusummade abil. Kogusummasid vaadatakse aastate ja EMTAK tähtkoodi lõikes.

Võrdlusesse võetakse need väärtused, mis on andmestikus imputeeritud, kuid mille kohta on esitatud hilinenud MAA. Võrreldakse algse EKOMARi andmestiku imputeeritud väärtuste müügitulu kogusummat ja alternatiivsete imputeerimismeetodite abil saadud müügitulu summa erinevust tegelikust müügitulu summast, mis on saadud MAA andmestikust. Müügitulu kogusumma andmed on antud tabelis 3.

Sektori B ehk mäetööstuse puhul töötas kõige paremini 20% suurimate ja väiksemate vaatluste trimmimine uutest moodustatud kihtidest. Aastate 2015. ja 2017. puhul on kogusumma palju lähedasem tegelikule summale kui praegune kihikeskmine. Aastal 2016 on see aga umbes poole võrra väiksem tegelikust väärtusest ja praegune keskmine on palju lähedasem.

Sektori C ehk töötleva tööstuse puhul on praegune imputeerimisviis kokkuvõttes kõige täpsem, vaid 2016. aastal on 10% ülevalt ja alt trimmimine uutes kihtides saavutanud tegelikule kogusummale lähema tulemuse.

Sektori D ehk elektrienergia, gaasi, auru ja konditsioneeritud õhuga varustamise puhul on 10% ülevalt ja alt trimmimine uute kihtide järgi saavutanud tegelikule kogusummale

Tabel 3. Hiljem esitatud MAA müügitulu kogusumma võrdlus algse imputeeritud väärtustega ning uute alternatiivsete trimmitud keskmise meetoditega vanades ja uutes kihtides

aasta	EMTAK tähegrupp	MAA hilinenud müügitulu	EKOMARi andmestikus imput.	10% trim. mõlemalt poolt v. kihtides	20% trim. mõlemalt poolt v. kihtides	10% trim. ülevalt v. kihtides	20% trim. alt v. kihtides	10% trim. mõlemalt poolt u. kihtides	20% trim. mõlemalt poolt u. kihtides	10% trim. ülevalt u. kihtides	20% trim. alt u. kihtides
2015	B	342004	547027	1103932	1049657	1053957	962306	986306	887851	886284	728697
	C	71614770	67170404	92844258	81290222	85709781	69504483	97054596	81294408	88149457	66400211
	D	8089365	4043759	3635418	3505751	3478393	3211256	5281216	3382290	4619607	2602812
	E	1279861	2307712	4022064	3634824	3630034	2976046	4658035	3997467	4199076	3270724
	F	36135086	36447344	42930818	36538383	39451898	29688565	39677225	32257824	36065291	25072405
	G	305341735	202159465	249284074	192843137	229260271	162215685	155057108	113068613	138044331	86674768
	H	42330868	28493022	69941833	64743427	66920916	59739398	28870558	23778137	26163883	19388820
	I	5817890	3310793	5406890	5087614	5014325	4372162	3972473	3613381	3542105	2841280
	J	21511588	19717998	22656651	20163659	20452550	16178356	19988286	17639182	17788368	13630471
	L	48537362	67696344	52153344	38648017	47801912	29905455	44696480	32918441	41315651	25280714
	M	61791133	44626874	50606231	44606096	45496007	35303677	45342173	40049816	40626198	31252439
	N	38020820	27357667	28357390	23721086	25446291	18897836	23473840	18713434	20917043	14563691
	P	2017054	2139616	2593868	2273803	2330874	1810737	2527009	2209124	2265450	1748456
	Q	805352	851530	1262548	1162773	1172427	1004443	926540	857084	840041	700574
	R	10472605	7799221	7212475	6398574	6451460	5064485	6822637	5982369	6104995	4708722
	S	21573531	4194841	4559942	3820383	4095973	3076950	3705487	3532226	3332918	2844310
2016	B	1133522	1023148	1185972	1094227	1084523	914695	1406127	1315470	1307433	1141143
	C	70226578	88800964	115343037	102108800	106788985	87808327	120398442	101889339	109457900	83792142
	D	22835221	22240971	46545090	45996097	46225274	45533433	13710828	7147620	12104259	5478453
	E	926576	2973600	3687776	3274383	3334034	2717533	4053714	3309396	3647287	2639589
	F	48629258	42933139	49695093	41581140	45725552	33740795	53377065	40768591	49080195	31741619
	G	532311323	229869037	230841686	191218914	209047090	152748087	223292929	157109541	198466107	119977438
	H	44054992	38392982	41475822	33380567	37195902	26362778	37605104	28761501	33577706	22452821
	I	3373195	4058779	4542696	4215695	4069455	3383569	4647772	4261740	4156094	3385471
	J	27869084	24341636	27595494	24220240	24714604	19230317	26158120	22285858	23346793	17297229
	L	60489027	71677817	56856593	43295356	52140356	33087434	52744234	38502851	46940723	29454625
	M	70819902	59613478	60507354	53505603	54272688	41911517	55546992	48344138	49772285	37670211
	N	34674776	26962396	29759267	25785133	26707218	20557263	29357258	22722263	26108706	17530073
	P	28836767	2781810	2686231	2398393	2409481	1910356	2722571	2395277	2444708	1908665
	Q	1787365	1246524	2316214	2166859	2161662	1885369	1613575	1461198	1450580	1167464
	R	9479869	10249483	10251549	9110918	9202247	7307498	8995020	7832939	8070481	6225022
	S	14639636	3833767	4800109	4605480	4317112	3746288	4585299	4400326	4129580	3566974
2017	B	1074704	1905949	1850209	1581203	1683112	1277658	1725691	1443392	1546295	1147629
	C	78507963	73850431	141062047	121997394	129697624	103390337	144088428	118356997	130124501	95802188
	D	8217510	28552604	34443303	32500818	33281452	31002809	19501333	8800973	17330686	6668684
	E	6077622	3937326	8107345	7440149	7310726	6098807	10173332	8942748	9082345	7086892
	F	63060949	54132702	69621950	57984308	63880645	46473273	67448939	55127140	61488521	42889131
	G	804696325	249305167	492513538	397955671	453520838	334961739	312824137	211660228	278151086	161287364
	H	143137547	48880749	79134728	63640583	70662595	49962572	57194101	45286739	51287283	35590773
	I	3740789	3478358	6658090	6032447	6000758	4834687	6501150	5818852	5805354	4565941
	J	63618849	31111533	46372203	38199907	41308850	29685874	39145022	33315173	34893980	25702765
	L	77220372	84253346	74522218	54886944	68360331	41966564	64377036	48401603	59457984	37144752
	M	487426476	55159719	82058785	71906986	73520598	56345593	75140473	64425767	67318326	50252701
	N	38959698	29491339	46752799	40521676	41826967	32342936	38722219	30343679	34437624	23407143
	P	3582518	2651224	4011702	3620554	3596220	2882080	3923743	3521240	3523351	2805536
	Q	1926120	1055940	2263885	2011819	2064653	1657209	1876462	1682279	1679131	1321790
	R	13072695	8693305	17091805	15339126	15281109	12047122	12753330	10833330	11415674	8523360
	S	8104107	3860828	7744377	6960966	6925919	5543062	6136947	5848580	5526874	4728416

lähema tulemuse aastatel 2015 ja 2017. Ainult ülevalt 10% trimmimine uutes kihtides on küll 2017. aasta tegelikele väärtustele veel lähemal, kuid 2015. aastal on see praeguste

imputeeritud väärtuste summast kaugemal.

Sektori E ehk veevarustuse ning kanalisatsiooni, jäätme- ja saastekäitluse sektori puhul on 2015. aastal kõige lähedasem tegelikule kogusummale 20% ülevalt ja alt algsetes kihtides trimmimisel saadud summa. Aastal 2016 on praegune imputeerimisviisi kogusumma kordades tegelikust suurem, seega on kõik alternatiivsed viisid saanud tegelikule lähedasema tulemuse ja kõige lähedasema mõlemad kihistamisviisid koos 20% ülevalt trimmimisega. Aastal 2017 on algne imputeerimisviis andnud täpseima kogusumma.

Sektori Q ehk tervishoiu ja sotsiaalhoolekande valdkonna puhul oli 10% eemaldamine suurimatest ja väiksematest väärtustest täpsem kui algne imputeerimisviis aastal 2015. Teistel aastatel ükski alternatiivne viis algsest paremat tulemust ei andnud.

Ülejäänud sektorite F, G, H, I, J, L, M, N, P, R ja S puhul ei saavutanud ükski alternatiivne meetod ühelgi aastal täpsemat kogusummat kui algne imputeerimismeetod. Need sektorid saab omakorda jaotada kolmeks.

Esimeste sektorite puhul oli igal aastal EKOMARis algne imputeeritud väärtuste kogusumma tegelikust müügitulu summast väiksem. Seetõttu oli iga trimmitud keskväärtusega asendatud väärtus omakorda veel väiksem ja seega tegelikust väärtusest veel kaugem. Sellised sektorid olid G, H, J, M, N ja S.

Teisel juhul oli igal aastal algne imputeeritud väärtus küll suurem kui tegelik, kuid uued alternatiivsed meetodid hindasid tegelikku väärtust omakorda nii palju väiksemaks, et algne imputeeritud väärtus oli kokkuvõttes ikkagi täpsem. Selline sektor on L.

Kolmandate sektorite puhul oli tulemus olenevalt aastast erinev, ehk segu esimeste ja teiste sektorite omadustest. Sellised sektorid olid F, I, P ja R.

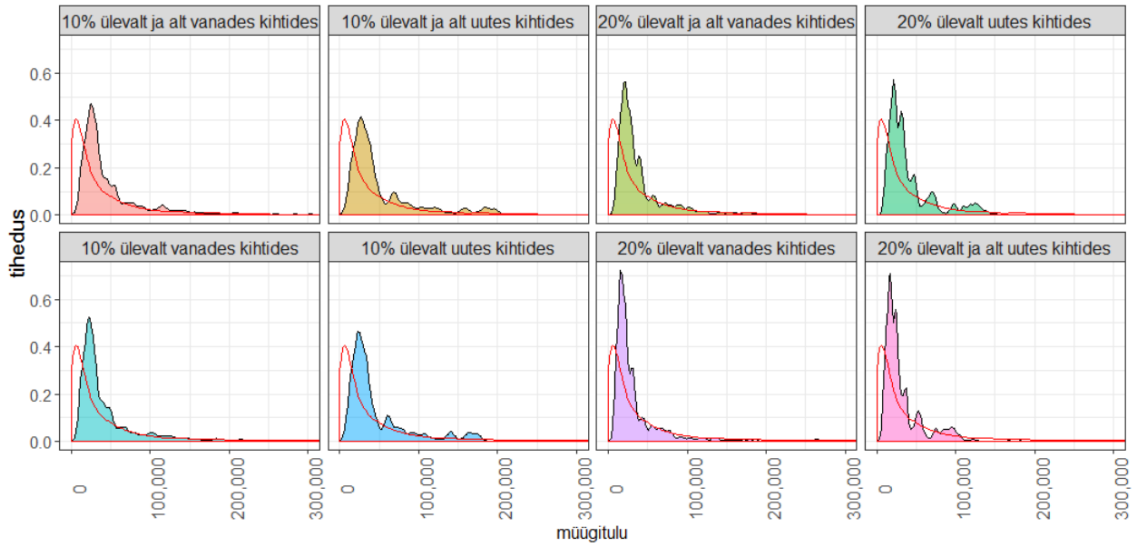
Kõigist sektoritest, kus paremat kogusumma hinnangut ei saavutatud, võib välja tuua suurima ettevõtete arvu ning müügituluga sektori G ehk hulgi- ja jaekaubanduse ning mootorsõidukite ja mootorrataste remondi valdkonna.

4.2.2 Müügitulu jaotus

Vaadatakse ka alla miljoni eurose müügituluga ettevõtete müügitulu:

Andmestiku eelneva analüüsi põhjal sai selgeks, et nende imputeeritud ettevõtete müügitulu põhjal, kellel on olemas MAA müügitulu väärtus, on väike müügitulu rohkematel

ettevõtetel, kui seda hetkel imputeeritakse. Seetõttu võiks (eriti suurte väärtustega) erindite eemaldamine imputeeritud väärtuste jaotust muuta tegelikkusega sarnasemaks. Seega võrreldakse nüüd nende puuduvate väärtustega ettevõtete puhul alternatiivsete imputeerimismeetoditega saadud väärtuste jaotust võrreldes tegeliku jaotusega MAA andmete põhjal.



Joonis 7. MAA andmete (punane kontuurjoon) ja kõigi alternatiivsete imputeerimisviiside andmete jaotuse võrdlus (müügitulu kuni väärtuses miljon eurot) aastatel 2015-2017

Võrreldes kõiki alternatiivseid imputeerimismeetodeid joonisel 7 tuleb välja, et 20% eemaldamine kas ainult suurematest või suurematest ja väiksematest kihi väärtustest toob kaasa selle, et imputeeritakse liiga väikseid väärtuseid. Ainult 10% vastavate väärtuste eemaldamise tulemusel on saadud sarnasem jaotus tegelikule. Kõige lähedasem on olnud olnud uute kihtide jaotuse järgi 10% eemaldamine, kõige sarnasem tegelikule jaotusele on olnud 10% suurimate ja väiksemate väärtuste eemaldamine.

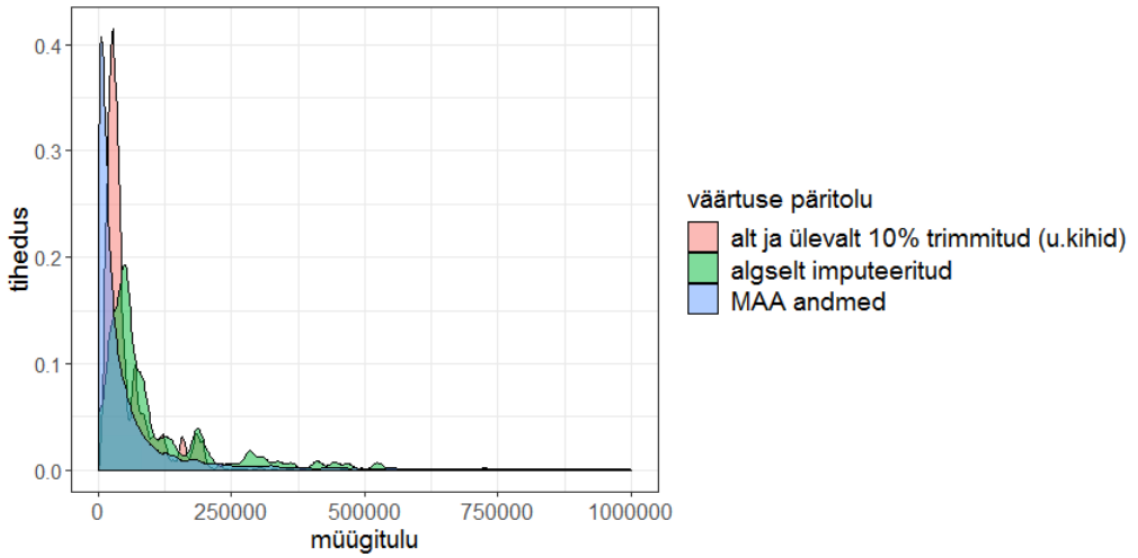
Kui nüüd joonise 8 abil võrrelda kõige lähedasema jaotusega alternatiivset viisi ehk uutes kihtides 10% eemaldamist suurimatest ja väiksematest väärtustest ja algset imputeerimisviisi MAA andmetega, siis on selge, et alternatiivse viisi jaotus on oluliselt sarnasem tegelikkusele.

Kuna sektor G on suurima müügitulu ja töötajate arvuga, siis vaadatakse selle jaotust ka

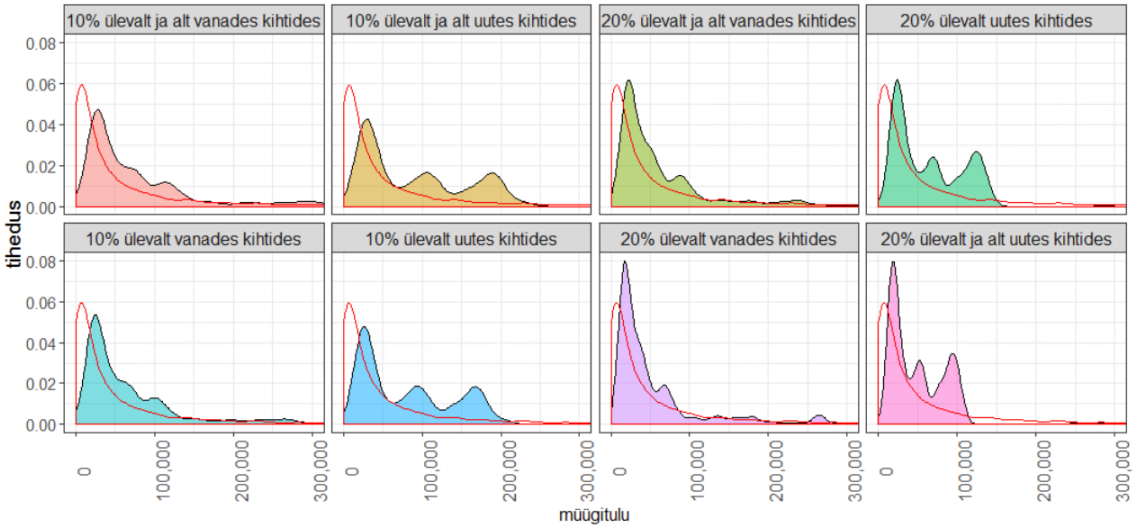
eraldi.

Kuna joonisel 9 on välja toodud ainult ühe sektori müügitulu, siis on võrreldes kõigi andmete joonisega 7 näha erinevust, et imputeeritud keskväärtuste künkad joonistuvad palju selgemalt välja. See tuleb rohkem välja eriti uute kihtide jaotuste puhul, mis põhineb antud juhul vaid kolmekohalisel EMTAK koodil. Seega on jaotus vanade kihtide puhul tegelikkusega sarnasem, 20% suurimate ja väikseimate väärtuste eemaldamine kihtidest tundub sektori G puhul andvat kõige lähedasema jaotuse tegelikule.

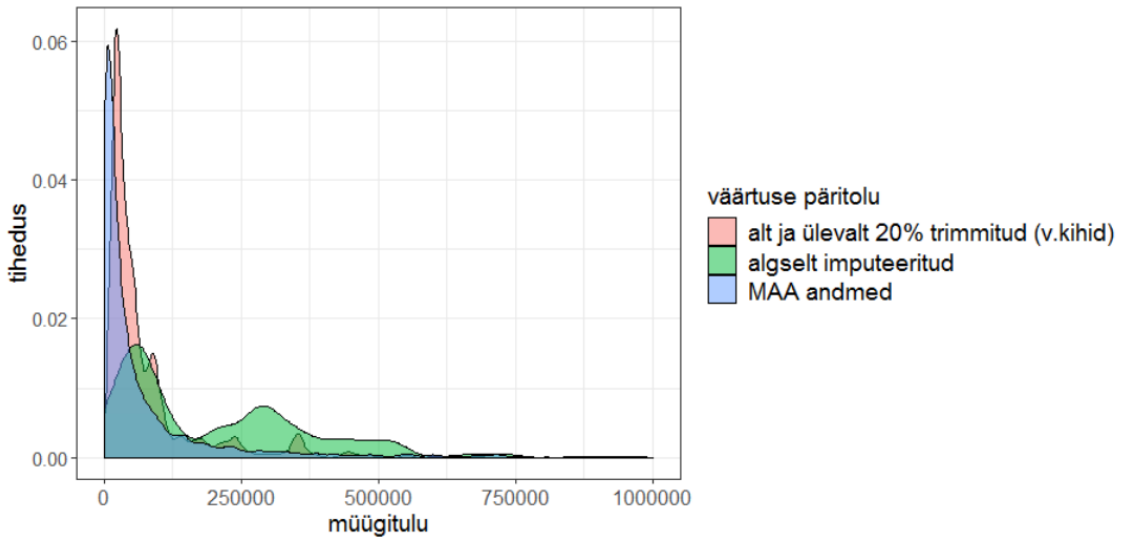
Sektori G puhul on võrreldes algse imputeerimisviisiga saavutatud parima alternatiivse viisiga samuti oluliselt lähedasem tulemus tegelikule jaotusele, mida on näha jooniselt 10.



Joonis 8. Algse imputeerimisviisi, 10% alt ja ülevalt trimmitud alternatiivse imputeerimisviisi ja MAA andmete jaotuse võrdlus(müügitulu kuni väärtuses miljon eurot) aastatel 2015-2017



Joonis 9. Sektori G MAA andmete (punane kontuurjoon) ja kõigi alternatiivsete imputeerimisviiside andmete jaotuse võrdlus (müügitulu kuni väärtuses miljon eurot) aastatel 2015-2017



Joonis 10. Sektori G algse imputeerimisviisi, 20% alt ja ülevalt trimmitud alternatiivse imputeerimisviisi ja MAA andmete jaotuse võrdlus(müügitulu kuni väärtuses miljon eurot) aastatel 2015-2017

Kokkuvõte

Töös imputeeriti nii algsetes kui uutes defineeritud kihtides erinevate kihtidepõhiste trimmitud keskmisega. EMTAK tähegrupi tasemel võrreldi saadud väärtuste kogusummat nii algsete imputeeritud väärtuste kui hilinenud MAA andmete põhjal saadud kogusummaga. Võrreldi ka alla miljonilise müügituluga ettevõtete jaotust.

Andmestiku analüüsi puhul tuli välja tõsiasi, et kui hakata andmeid vaatama väiksemates gruppides ehk juba näiteks EMTAK tegevusala esimese taseme järgi, siis mõned erandlikud vaatlused võivad üldpilti oluliselt muuta.

Seetõttu oli näha, et kui müügitulu oli hiljem esitanud isegi EMTAK tähegruppi mõttes äärmiselt suure müügituluga ettevõte, siis olid imputeeritud kogusummad tegelikust oluliselt väiksemad. Kogusummade puhul ei tulnud välja sellist alternatiivset meetodit, mis oleks olnud igas tähegrupis algsest viisist kogusumma imputeerimisel parem. Ka ei saanud otseselt üldises plaanis eelistada ühte kihtide jaotust teisele, kuna mõnes kihis oli kõige lähedasem uutes kihtide järgi imputeeritud väärtuste summa ja mõnes teises jälle vanade kihtide abil. Paljudes gruppides ei õnnestunud aga ühegi meetodiga kogusummat paremini imputeerida. Selle põhjuseks oli enamasti see, et üksikud erandlikud suured summad olid kogusumma suureks muutnud ja seda on imputeerimisel võimatu arvestada. Tähegruppidest, kus alternatiivsed meetodid paremaid tulemusi ei andnud, võib välja tuua suurima müügitulu ja ettevõtete arvuga sektori G ehk hulgi- ja jaekaubanduse ning mootorsõidukite ja mootorrataste remondi sektori.

Algselt oli keskväärtusega imputeerimise puhul näha, et väikese müügituluga ettevõtetele imputeeritakse suurem müügitulu, kui neil tegelikult on. Seetõttu saadi ka trimmimisel alla miljonilise müügituluga ettevõtete jaotus sarnasem, kui algse meetodiga. Suurima ettevõtete arvu ning müügitulu summaga sektori G puhul tuli välja, et imputeerimine uutes suuremates kihtides tõi kaasa nö küngaste ilmumise graafikule. Vanades kihtides imputeerimine andis ühtlasema ja tegelikule sarnasema jaotuse.

Seega ühesugune eeskiri ei tööta kõigis tegevusalades ühtviisi hästi. Keskväärtuse trimmimine näitas parimat tulemust kihtides B, C, D, E ja Q, kus vähemalt ühel aastal oli see täpsem müügitulu kogusumma hinnang kui algne imputeerimisviis. Keskväärtusega (sh trimmitud keskväärtusega) imputeerimise eelis on lihtsus ja see, et seda saab kasutada kõigil andmestiku näitajatel hoolimata sellest, kas need on müügituluga seotud või mitte.

Keskväärtusega imputeerimine ei tööta hästi, kui MAA on esitamata suurema müügituluga ettevõtetel.

EKOMARi puhul on kindlasti ka väga palju imputeerimisega seotud uurimise tulevikupotentsiaali. Esiteks on antud töös uuritud vaid müügitulu imputeerimist ehk teada ei ole see, kuidas trimmitud keskväärtuse kasutamine müügitulu järgi teiste tunnuste jaotust mõjutab. Teiseks on töö sisendiks järgmisele analüüsi sammule, mis hõlmab Maksu- ja Tolliameti andmete kaasamist paremaks imputeerimiseks, et MAA hiljem esitanud ettevõtte müügitulu suurust veel paremini hinnata.

Kasutatud kirjandus

[1] Scheffer, J. (2002). "Dealing with Missing Data". *Research Letters in the Information and Mathematical Sciences*, 3, lk. 153–160.

[2] Zhang, Z. (2016). "Missing data imputation: focusing on single imputation". *Annals of Translational Medicine*, 4.

[3] Lang, K. M. ja Little, T. D. (2018). "Principled Missing Data Treatments". *Prevention Science: The Official Journal of the Society for Prevention Research*, 19.3, lk. 284–294.

[4] Rubin, D. B. (1976). "Inference and Missing Data". *Biometrika*, 63.3, lk. 581–592.

[5] *ESMS metaandmed: Ettevõtete majandusnäitajad (aasta)* (2020). `https://www.stat.ee/esms-metaandmed?code=20300` (vaadatud 10.05.2020).

[6] *Memobust Handbook on Methodology of Modern Business Statistics* (2014). `https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en` (vaadatud 07.05.2020).

[7] Buuren, S. van (märts 2012). *Flexible Imputation of Missing Data*. CRC Press.

[8] Tukey, J. W. ja McLaughlin, D. H. (1961). "The variance of means of symmetrically trimmed samples from normal populations, and its estimation from such trimmed samples. (Trimming/winsorization I)".

[9] *Eesti Majanduse Tegevusalade Klassifikaator 2008* (2008). `https://www.rik.ee/et/e-ariregister/emtak-tegevusalad` (vaadatud 07.05.2020).

Lisad

Lisa 1. EKOMARi andmestiku uute tunnuste kodeerimine ja töös kasutatava alamandmestiku määramine (rakendustarkvara R)

```
\library{dplyr}
\library{tidyr}

# algse andmestiku nimi on EKOMAR_2013_2017_anon
#esimene EMTAK kiht(tahekiht)

EMTAK_grupp_1 <- cut(x=EKOMAR_2013_2017_anon$EMTAK, breaks=c(0,
    50000, 100000, 350000,360000,410000,450000,490000,550000,
580000,640000,680000,690000,770000,840000,850000,860000,900000,
    940000,970000,990000,Inf), include.lowest=T,right = FALSE,
labels=c("A""B""C""D""E""F""G""H""I""J""K""L""M""N""O""P""Q""R""S
    ""T""U"))
EKOMAR_EMTAK_gruppides <- cbind(EKOMAR_2013_2017_anon,EMTAK_grupp
    _1)

# eraldatakse 2015-2017 andmed
EKOMAR_2015_2017_anon <- EKOMAR_EMTAK_gruppides %>% filter(YEAR
    >=2015)

# valede andmetega andmeread parandatakse
EKOMAR_2015_2017_anon = EKOMAR_2015_2017_anon %>% mutate(TIN_IMPU
    =ifelse(YEAR==2017 & SA_ID %in% c("10322778","10396592","
    10493893"),"MAA",TIN_IMPU))

# teine ja kolmas EMTAK kiht(2-kohaline numbrikood)
EKOMAR_2015_2017_anon<- EKOMAR_2015_2017_anon %>%
```

```
mutate(
EMTAK_grupp_2=as.factor(ifelse(nchar(as.character(EMTAK))==5,
    paste(0,substr(EMTAK,1,1),sep = ""),substr(EMTAK,1,2))),
EMTAK_grupp_3=as.factor(ifelse(nchar(as.character(EMTAK))==5,
    paste(0,substr(EMTAK,1,2),sep = ""),substr(EMTAK,1,3))))


# tunnus tootaja_grupp
EKOMAR_2015_2017_anon=EKOMAR_2015_2017_anon %>%
mutate(tootaja_grupp=ifelse(EMTAK_grupp %in% c("B","C","D","E"),
ifelse(TARVANK<=9,"1-9",ifelse(TARVANK<=19,"10-19",">=20")),
ifelse(EMTAK_grupp=="F" & as.numeric(as.character(EMTAK_grupp_3))
    <=410, ifelse(TARVANK==1,"1",ifelse(TARVANK==2,"2",ifelse(
    TARVANK<=10,"3-10","11-20"))), ifelse(TARvANK==1,"1",ifelse(
    TARVANK<=9,"2-9",ifelse(TARVANK<=19,"10-19",">=20"))))))
# loplik andmestik, millel imputeerimist proovitakse

imputeerimise_loplik = EKOMAR_2015_2017_anon %>% filter(tootaja_
    grupp %in% c("1","1-9"))
```

Lisa 2. Imputeerimisel kasutatud funktsioonid ja saadud uued andmestikud (rakendustarkvara R)

```
# imputeerimisfunktsioon tunnuse KIHINR jargi:

imputeerimisfunkts_kihinr <- function(protsent,alt_ka){
# argumentideks eemaldamise protsent ja see, kas trimmitakse vaid
    ulevalt voi ka alt ("ei"/"jah")
#kihinr jargi leian kihid, kus on olemas andmeridu rohkem kui 10

kihinr_analyysl <- imputeerimise_loplik %>%
filter(TIN_IMPU!="KESK") %>%
```

```
group_by(KIHINR,YEAR) %>%
summarise(n=n(),
protsentiil_all=quantile(C_10,1-protsent),
protsentiil_yleval=quantile(C_10,protsent)) %>% mutate(rohkeml0=
    ifelse(n>10,"jah","ei"))

#leian kihikeskmise, ja seal kus rohkem 10, eemaldan suured (ja
    ka vaikesed)
kihinr_analyys2 <- imputeerimise_loplik %>%
full_join(kihinr_analyysl,by=c("KIHINR","YEAR")) %>%
filter(TIN_IMPU!="KESK",!(alt_ka=="jah" & rohkeml0=="jah" & C_10<
    protsentiil_a11),!(rohkeml0=="jah" & C_10>protsentiil_yleval))
    %>% group_by(KIHINR,YEAR) %>%
summarise(n=n(),
kihikesk=mean(C_10,na.rm = T))
# lopuks tagastakse imputeerimise_loplik andmestik koos tunnusega
    uusimpu
koos_kihinr <- kihinr_analyys2 %>% select(KIHINR,kihikesk,YEAR)
   %>% right_join(imputeerimise_loplik,by=c("KIHINR","YEAR")) %>%
mutate(uusimpu=ifelse(TIN_IMPU=="KESK",kihikesk,NA))
return(koos_kihinr)
}

# imputeerimisfunktsioon tunnuste tootaja_grupp ja EMTAK_grupp_3
    abil:

imputeerimisfunkts_omakihid <- function(protsent,alt_ka){ #leian
    kihid, kus on olemas andmeridu rohkem kui 10 EMTAK_grupp_3_
    analyysl <- imputeerimise_loplik %>%
filter(TIN_IMPU!="KESK") %>%
group_by(EMTAK_grupp_3,YEAR,tootaja_grupp) %>% summarise(n=n(),
protsentiil_all=quantile(C_10,1-protsent), protsentiil_yleval=
```

```
    quantile(C_10,protsent)) %>% mutate(rohkeml0=ifelse(n>10,"jah"
    ,"ei"))
#leian kihikeskmise, ja seal kus rohkem 10, eemaldan suured(ja ka
     vaikesed)
EMTAK_grupp_3_analyys2 <- imputeerimise_loplik %>% full_join(
    EMTAK_grupp_3_analyysl,by=c("EMTAK_grupp_3","YEAR","tootaja_
    grupp")) %>%
filter(TIN_IMPU!="KESK",!(alt_ka=="jah" & rohkeml0=="jah" & C_10<
    protsentiil_a11),!(rohkeml0=="jah" & C_10>protsentiil_yleval))
     %>% group_by(EMTAK_grupp,EMTAK_grupp_3,YEAR,tootaja_grupp)
    %>%
summarise(n=n(),
kihikesk=mean(C_10,na.rm = T))

# lopuks tagastakse imputeerimise_loplik andmestik koos tunnusega
     uusimpu
koos_EMTAK_grupp_3 <- EMTAK_grupp_3_analyys2 %>%,
select(EMTAK_grupp_3,kihikesk,YEAR,tootaja_grupp) %>%
right_join(imputeerimise_loplik,by=c("EMTAK_grupp_3","YEAR","
    tootaja_grupp")) %>%
mutate(uusimpu=ifelse(TIN_IMPU=="KESK",kihikesk,NA))
return(koos_EMTAK_grupp_3)
}


# imputeerimisfunktsioone kasutades saadud erinevad imputeeritud
     vaartustega andmestikud
imp_kihinr_10_ay=imputeerimisfunkts_kihinr(0.9,"jah") imp_kihinr_
    20_ay=imputeerimisfunkts_kihinr(0.8,"jah") imp_kihinr_10_y=
    imputeerimisfunkts_kihinr(0.9,"ei") imp_kihinr_20_y=
    imputeerimisfunkts_kihinr(0.8,"ei") imp_omakiht_10_ay=
    imputeerimisfunkts_omakihid(0.9,"jah") imp_omakiht_20_ay=
```

```
imputeerimisfunkts_omakihid(0.8,"jah") imp_omakiht_10_y=
imputeerimisfunkts_omakihid(0.9,"ei") imp_omakiht_20_y=
imputeerimisfunkts_omakihid(0.8,"ei")
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Hanna Britt Parman,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Ettevõtte kompleksse kalendriaasta aruande (EKOMAR) puuduvate müügitulu väärtuste imputeerimine majandusaasta aruannetele tuginedes", mille juhendajad on Kristi Lehto ja Mare Vähi, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, 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.

Hanna Britt Parman, 19.05.2020