TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Tarkvarateaduse instituut

Margus Salk 121047IAPB

PUUDUMISMEHHANISMI TUVASTAMINE LÜNKLIKES ANDMESTIKES JA SELLE RAKENDAMINE IMPUTEERIMISEL

Bakalaureusetöö

Juhendaja: Martin Rebane
MSc

Tallinn 2017

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Margus Salk

22.05.2017

Annotatsioon

Lõputöö eesmärgiks on jõuda meetodini, millega on võimalik tuvastada andmete puudumise mehhanismi, ning leida parim meetod puudumismehhanismide imputeerimiseks.

Puudumismehhanismide tuvastamise meetodi välja töötamiseks eemaldatakse täielikest andmestikest andmeid vastavalt puudumismehhanismi definitsioonile ning uuritakse seoseid tekkinud lünklikes andmestikes.

Parima meetod puudumismehhanismi imputeerimiseks leitakse eelnevalt tekitatud lünklike andmestike imputeerimisel ning võrdlemisel algsete andmetega.

Töö kirjutamisel valminud skriptid leiab Githubi repositooriumist https://github.com/MargusSalk/missing-data.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 34 leheküljel, 5 peatükki, 17 joonist, 13 koodinäidet.

Abstract

Detecting missing data mechanism for incomplete datasets and its applications for imputation

The aim of this thesis is to formulate a method, which could be used to detect missing data mechanism, and to decide on which imputation method is the most fitting for each missing mechanism.

Data from complete datasets are removed according to missing data mechanism definitions. Relations between attributes in these incomplete datasets are analyzed to work out a method for discovering missing data mechanisms.

Best imputation methods for different mechanisms are chosen by their performance on previously created incomplete datasets. Results are validated by comparing original and imputed data.

Scripts created as a result of this thesis are available on Github, https://github.com/MargusSalk/missing-data.

The thesis is in Estonian and contains 34 pages of text, 5 chapters, 17 figures, 13 code examples.

Lühendite ja mõistete sõnastik

MCAR	*Missing Completely at Random*, puudub täiesti juhuslikult
MAR	*Missing at Random*, puudub juhuslikult
MNAR	*Missing Not at Random*, puudub mittejuhuslikult
Imputeerimine	puuduvate väärtuste asendamine sobivate asendusväärtustega
MICE	*Multivariate Imputation by Chained Equations* – mitmekordse imputeerimise algoritm
pmm	*Predictive Mean Matching* – imputeerimismeetod, ennustatava keskmisega sobitamine

Sisukord

Koodinäidete loetelu

Jooniste loetelu

1 Sissejuhatus

Andmestike töötlemisel on üheks olulisemaks probleemiks puuduvad andmed. Automaatne lahendus puuduvate andmetega tegelemiseks kiirendaks andmetöötlusprotsesse ning aitaks ressursse kokku hoida.

1.1 Taust ja probleem

Kuna puuduvad andmed on teaduslikes andmestikes tavaline nähtus ning tihti vältimatu [1], siis andmetega töötades on oluline arvesse võtta ka puuduvaid väärtuseid. Andmete puudumise põhjused võivad tuleneda näiteks juhuslikest eksimustest andmete korjamisel, kindlat patsienti mittepuudutavast küsimusest või patsiendi otsusest jätta küsimusele vastamata [2]. Olenemata sellest, et tegu on väga levinud probleemiga, tegeletakse puuduvate andmetega üldiselt lihtsakoelistel meetoditel, mis võivad anda kallutatud lõpptulemusi [3].

1.2 Eesmärgid

Töö esimeseks eesmärgiks on välja töötada meetod, millega tuvastada puuduvate andmete mehhanismi. Meetod peaks mehhanismi otsustama puuduvate andmete esinemise ning teiste tunnuste väärtuste vahelise seoste põhjal.

Teiseks eesmärgiks on leida iga mehhanismi jaoks meetod, millega neid tõhus imputeerida (puuduvate väärtuste asendamine sobivate asendusväärtustega) oleks.

Mõlemad oleks eelduseks puuduvate andmetega tegelemise automatiseerimiseks.

1.3 Metoodika

Puuduvate andmete mehhanismi tuvastamiseks uuritakse andmestiku erinevate tunnuste puuduvate väärtuste ilmnemise seost teiste väärtustega. Erinevate mehhanismide kohta luuakse näiteandmestikud – valitakse täielik andmestik ning kustutatakse sobival viisil väärtused. Näiteandmestikud on hea põhi, mille pealt on võimalik uurimise tulemusena

valida puudumismehhanismide tuvastamise meetod. Lisaks luuakse näiteandmestike pöördprojekteerimisel meetodid, millega imputeerimine annab iga mehhanismi kohta täpseimad tulemused.

Tulemuse kontrollimiseks analüüsitakse algseid näiteandmestikke ning pöördprojekteerimisel imputeeritud andmestikke.

2 Teoreetiline ülevaade

2.1 Puuduvate andmete mehhanismid

Rubin [4] esmalt kirjeldas ning jagas erinevaid andme puudumise tüüpe nende tekkimise protsessi järgi. Hiljem tekkis neist kolm puuduvate andmetega tegelemisel üldkasutatavat mõistet: Missing Completely at Random (MCAR), Missing at Random (MAR) ja Missing Not at Random (MNAR).

2.1.1 MAR

MAR on vähem rangem puuduvate andmete mehhanism. Puuduv väärtus võib olla sõltuvuses vaadeldud muutujatest, aga mitte puuduvast väärtusest endast. Patsientide pikaajalisel ravil võib juhtuda näiteks, et patsient ei ole motiveeritud meditsiinilisse kontrolli tulema terviseprobleemide puudumise tõttu. [1] [5]

2.1.2 MCAR

MCAR on MAR erijuhtum [5] ning puuduvate andmete tugevaimaks eelduseks [1]. Puuduv väärtus ei sõltu jälgitavast parameetrist ega väärtusest endast. MCAR ei muuda andmete analüüsi tulemusi, kuna puuduvad andmed esinevad sama jaotusega kui olemasolevad andmed [1] ning viimaseid võib pidada suvaliseks alamhulgaks hüpoteetilisest terviklikust andmestikust [6]. Meditsiinilistes domeenides võib tekkida MCAR juhtum näiteks hooletusest või probleemidega andmete edastamisel.

2.1.3 MNAR

Kolmandasse puuduvate andmete mehhanismi kategooriasse kuulub MNAR. Sellel juhul sõltub väärtuse puudumine väärtusest endast ning seda peetakse väärtusliku info kaotamise tõttu mitte-ignoreeritavaks [7]. Näiteks siirdatud neeruga patsiendil ei mõõdetud kreatiniini taset siirdeelundi hülgamise tõttu.

2.2 Puudumismehhanismi tuvastamine

Puudumismehhanismi tuvastamiseks luuakse näiteandmestikud iga mehhanismi kohta, uuritakse neid, nendes esinevaid seoseid ning tehakse järeldused, mille põhjal luuakse meetod mehhanismide tuvastamiseks. Seoste tuvastamisest täpsemalt peatükis 2.4 ning näidete loomisest peatükis 3.2.

2.3 Meetodid puuduvate väärtustega tegelemiseks

Võimalused puuduvate andmetega tegelemiseks ulatuvad lünkliku objekti ignoreerimisest erinevate imputeerimismeetoditeni. Järgnevalt tuuakse välja mõningad tähtsamad meetodid.

2.3.1 Täieliku juhtumi analüüs (Complete Case Analysis)

Kõige lihtsam viis puuduvate andmetega tegelemiseks on kustutada lünklikud read. Meetodi tugevaks küljeks on selle kasutamise lihtsus [1] ja nõrkadeks külgedeks võib pidada infokadu ning potentsiaalselt kallutatud tulemusi [1] .

2.3.2 Saadaoleva juhtumi analüüs (Available Case Analysis)

Olemasolevate juhtude analüüs on täielikust vähem rangem. Kui analüüsitakse alamhulka andmeid, vaadeldakse kõiki juhtumeid täielike alamhulkadega [1]. Ei ignoreerita juhtumeid, mille puuduvad väärtused ei ole uurimisobjektiks. Peetakse väiksema statistilise võimsuse kao tõttu täieliku juhtumi analüüsist paremaks valikuks [8].

2.3.3 Ühekordne tingimusteta imputeerimine keskmisega (Single Unconditional Mean Imputation)

Leitakse olemasolevate väärtuste keskmine ja täidetakse sellega kõik lüngad [1]. Olemuselt on meetod väga lihtne, kuid sel esineb mitmeid puuduseid: imputeeritud tunnuse dispersioon väheneb [8], samas kui täpsust hinnatakse üle [8], ja antud meetodi tulemused on alati kallutatud [1].

2.3.4 Ühekordne tingimustega imputeerimine keskmisega (Single Conditional Mean Imputation)

Eelneva meetodi edasiarendus kasutab imputeerimiseks tingimuslikku keskmist. Arvutatakse erinevad keskmised sõltuvalt teise tunnuse väärtusest [1]. Sõltuvate tunnuste

valik ei ole triviaalne ning nõuab teadmisi domeenist, et olla efektiivne lahendus [1]. Lisaks väheneb ka selle meetodi kasutamisel imputeeritud tunnuse variatsioon, kuigi vähem kui tingimusteta keskmisega imputeerides [8].

2.3.5 Ennustatava keskmisega sobitamine (Predictive Mean Matching – pmm)

Sobitab puuduva väärtuse olemasoleva väärtusega, millel on lähim ennustatav keskmine [9]. Ennustuste leidmiseks kasutatakse lineaarset regressiooni [9]. Meetod jääb nõrgaks andmestikes, kus puuduval väärtusega andmetel ei ole andmestikus sarnaseid naabreid, mille pealt ennustused luua [10]. Meetod on populaarne mitmekordse imputatsiooni algoritmides, mis kasutavad seda pidevate tunnuste imputeerimisel vaikemeetodina [10].

2.3.6 Mitmekordne imputatsioon (Multiple Imputation)

Erinevalt eelnevalt mainitud meetoditele, ehitatakse antud meetodiga mitu võimalikku imputeeritud andmestikku [1]. Iga andmestikku analüüsitakse ning viimaks koondatakse tulemus üheks tervikuks [1] [8]. Kallutatud tulemusi välditakse kasutades regressioonanalüüsi [1].

2.3.7 Maksimaalse tõenäosuse meetod (Maximum Likelihood - ML)

Vastupidiselt eelnevatele ei täida maksimaalse tõenäosuse meetod lünkasid andmestikes vaid pakub olulised hinnangud regressioonimudelitele [1]. ML meetodid peaks nii MCAR kui ka MAR juhtude puhul andma kallutamata tulemusi [11] ning lisaks on ML meetodid MCAR puhul tõhusamad kui näiteks täieliku (2.3.1) ja saadaoleva (2.3.2) juhtumi analüüsi meetodid [6].

2.3.8 Sega-mustri mudel (Pattern-Mixture Model)

Antud meetodi jaoks võib puuduvate andmete mehhanism jääda tuvastamata, selle asemel kasutatakse erinevate mustrite segu puuduvate andmete kirjeldamiseks [1]. Mustrite loomine nõuab aga palju domeenispetsiifilisi teadmisi ning seetõttu jääb antud meetod töö skoobist välja.

2.3.9 Sobiva meetodi valik

Kõik eelmainitud meetodid sobivad tegelemiseks MCAR puudumismehhanismiga ning MAR jaoks sobivad ühekordne tingimustega imputeerimine keskmisega, mitmekordne imputatsioon, maksimaalse tõenäosuse meetod ja sega-mustri mudel [1]. MNAR puhul

on kõige praktilisem sega-mustri mudel [1] ja teatud modifitseerimisega mitmekordne imputeerimine [12].

Töös kasutatakse mitmekordset imputeerimise paketti ning meetod imputeerimiseks valitakse vastavalt puudumismehhanismiga sobivusele. Täpsemalt paketi valikust peatükis 4.1.

2.4 Seoste tuvastamine

Tunnustevaheliste seoste hindamiseks kasutatakse erinevat tüüpi korrelatsioonikordaja hindamise viise – lineaarne, polühoorne ning polüseriaalne. Kahe arvulise tunnuse vahelise seose arvutamiseks kasutatakse lineaarset ehk Pearsoni korrelatsioonikoefitsienti [13], arvulise ning järjestatud kategoorilise tunnuse vahelise seose jaoks polüseriaalset korrelatsioonikoefitsienti [14] ja kahe järjestatud kategoorilise tunnuse vahelise seose jaoks polühoorset korrelatsioonikoefitsienti [15].

Puuduvate andmete esinemise ning teiste tunnuste väärtuste vahelise seose leidmiseks viiakse uuritava tunnuse väärtused binaarkujule [16] – puuduv väärtus asendatakse väärtusega 1, olemasolev väärtusega 0. Järgnev meetod (Koodinäide 1) realiseerib sellise modifitseerimise R-is ning töös leiab see kasutust näiteandmestike modifitseerimisel peatükis 3.2.

```
missing_to_binary <- function(dat, attr) {
	if (class(dat[,attr]) == 'factor') {
		dat[,attr] <- as.character(dat[,attr]);
	}
	dat[,attr][!is.na(dat[,attr])] <- 0;
	dat[,attr][is.na(dat[,attr])] <- 1;
	x[,attr] <- as.factor(x[,attr]);
	return(dat);
}
modified_data <- missing_to_binary(initial_data, attribute_modified)
```

Koodinäide 1 - puuduvate andmetega tunnus binaarkujule

2.4.1 Järjestamata kategoorilised tunnused

Järjestamata kategooriliste tunnustega on seose leidmine raskendatud. Kui tunnust „Sugu“ võimalike väärtustega „M“ ja „F“ on võimalik käsitleda järjestatud suurusena väärtustega 0 ja 1, siis suurem hulk võimalikke väärtuseid nõuavad teistsugust lahendust.

Antud probleemi lahendamiseks teisendatakse kategooriliste tunnuste erinevad väärtused eraldi tõeväärtustunnusteks. Eelnevalt välja toodud näitest saaks pärast teisendamist kaks eraldi tunnust – „Sugu.M“ ja „Sugu.F“ väärtustega 0 ja 1.

Koodinäide 2 näitab kategooriliste tunnuste (v.a uuritav) muutmist tõeväärtustunnusteks.

```
# removing attribute_modified from dataset and converting other factor variables to
dummies
attr_modified_col <- select(dat, get(attribute_modified))
other_columns <- select(dat, -(get(attribute_modified)))
other_columns <- dummy.data.frame(other_columns, dummy.class="factor", sep=".",
fun=as.factor)
# add the two together
dataWithDummies <- cbind(attr_modified_col, other_columns)
```

Koodinäide 2 - kategoorilised tunnused tõeväärtustunnusteks

3 Puudumismehhanismi tuvastamine

3.1 Tehnoloogia

Arenduseks valiti R programmeerimiskeel. R on keel ja keskkond statistilistele arvutustele ning graafikale [17]. Antud keskkonnas on olemas sobivad võimalused andmete ning puuduvate väärtustega tegelemiseks. Lisaks on R-ile loodud hulgaliselt tarkvarapakette, mis töö sooritamisel kasuks tulevad.

3.2 Näidete konstrueerimine

Nagu kirjeldatud peatükis 1.3 on näiteandmestikel töös kaks eesmärki - andmestike uurimisel leida moodus, kuidas tuvastada erinevaid puudumismehhanisme, ning imputeerimismeetodite valideerimine võrreldes esialgset andmestiku imputeeritud andmestikuga.

Näidete puhul kasutatakse merikõrvade andmebaasi[1], mis sisaldab kategoorilisi ja numbrilisi tunnuseid. Andmestikus on 9 tunnust ning 4177 kirjet.

3.2.1 MCAR

MCAR mehhanismi näiteandmestiku jaoks kustutatakse valitud atribuudi väärtuseid täiesti juhuslikult kogu andmestiku ulatuses. Sel juhul ei sõltu puuduva väärtuse esinemise tõenäosus väärtusest endast ega teiste atribuutide väärtustest. Eelnimetatud on ka MCAR mehhanismi eelduseks.

Nagu järgnevalt näidatud (Koodinäide 3), kustutatakse 5 protsendil tunnuse Diameter väärtuseid. Kustutatava väärtuse indeks valitakse kogu andmestiku ulatusest.

[1] https://archive.ics.uci.edu/ml/datasets/abalone/

```
p <- 5 #percentage of missing values added

attribute_modified <- 'Diameter'

amount_removed <- as.integer(nrow(datMCAR)*p/100)

datMCAR[sample(1:nrow(datMCAR), amount_removed), attribute_modified] <- NA
```

Koodinäide 3 - MCAR puuduvate väärtuste loomine

Jooniselt (Joonis 1) on näha korrelatsioonimaatriksi üht rida, mis puudutab tunnust Diameter. Kuvatakse tunnuse puuduvate väärtuste esinemise ning teiste tunnuste vahelist korrelatsiooni. Kategooriline tunnus Sex on jagatud kolmeks tõeväärtustunnuseks, nagu on kirjeldatud peatükis 2.4.1. Puuduvate väärtuste esinemine ei oma tugevat korrelatsiooni teiste tunnustega.

```
   Diameter        Sex.F        Sex.I        Sex.M
1.00000000   0.04430123   0.02647049 -0.07051720
    Length           Ht     Whole_wt   Shucked_wt
0.03916878   0.03594862   0.03710299   0.04272286
Viscera_wt     Shell_wt        Rings
0.04162520   0.02987703 -0.03915915
```

Joonis 1 - MCAR korrelatsioonikoefitsiendid

3.2.2 MAR

MAR puhul tuleb silmas pidada, et andmete puudumise ning teiste tunnuste väärtuste vahel oleks piisavalt tugev seos. Lihtne variant on võimalik ehitada võttes sõltuvaks kategooriliste väärtustega tunnuse.

MAR näites kustutatakse 15 protsenti tunnuse Diameter väärtuseid juhul, kui tunnuse Sex väärtus on 'M' (Koodinäide 4).

```
p <- 15 #percentage of missing values added

attribute_modified <- 'Diameter'

related_attribute <- 'Sex'

related_attribute_value <- 'M'

amount_removed <- as.integer(length(which(datMAR[,related_attribute] ==
       related_attribute_value))*p/100)

datMAR[sample(which(datMAR[,related_attribute] == related_attribute_value),
       amount_removed), attribute_modified] <- NA
```

Koodinäide 4 - MAR puuduvate väärtuste loomine

Joonis 2 kujutatab MAR näite puuduvate väärtuste korrelatsioonimaatriksit võrdluses teiste tunnustega. Puuduvad andmed sisestati vaid Sex väärtuse 'M' puhul ning seetõttu esineb ka genereeritud Sex.M tunnusega väga tugev ning Sex.F ja Sex.I puhul negatiivne korrelatsioon. Teiste tunnustega tugev korrelatsioon puudub, kuid on siiski suurem, kui MCAR näites juhuslikult sisestatud puuduvate väärtustega. Põhjus võib tuleneda isaste isendite keskmiselt suuremas kaalus ning kasvus.

```
  Diameter       Sex.F       Sex.I       Sex.M
 1.0000000 -0.6104089 -0.5568154  0.9846808
    Length          Ht   Whole_wt Shucked_wt
 0.1610165  0.1387077  0.1729698  0.1665838
Viscera_wt    Shell_wt       Rings
 0.1705228  0.1609244  0.1573455
```

Joonis 2 - MAR korrelatsioonikoefitsiendid

3.2.3 MNAR

Valitakse andmestikus tunnus, mille väärtuseid kustutatakse sõltuvalt väärtustest endist.

MNAR näites kustutatakse 20 protsenti Diameter väärtuseid, mis ületavad künnise 0.5 (Koodinäide 5).

```
p <- 20 #percentage of missing values added
attribute_modified <- 'Diameter'
threshold <- 0.5 # point from which we start removing values
amount_removed <- as.integer(length(which(
       datMNAR[,attribute_modified] > threshold))*p/100)
datMNAR[sample(which(datMNAR[,attribute_modified] > threshold),
       amount_removed), attribute_modified] <- NA
```

Koodinäide 5 - MNAR puuduvate andmete loomine

Joonisel (Joonis 3) on toodud MNAR näite puuduvate väärtuste esinemise korrelatsioonikoefitsiendid. Koefitsiendid on suhteliselt suured, sest andmestikus esinevad tunnused on omavahel seotud – suurema diameetriga isenditel on ka reeglina teised mõõdud suuremad.

```
  Diameter       Sex.F       Sex.I       Sex.M
 1.0000000   0.2597816 -0.5391113   0.2034109
    Length          Ht    Whole_wt Shucked_wt
 0.5484956   0.4730804   0.6813970   0.6547707
Viscera_wt    Shell_wt       Rings
 0.6481038   0.6494325   0.3427481
```

Joonis 3 - MNAR korrelatsioonikoefitsiendid

3.2.4 Katsed alternatiivse andmestiku ning tunnustega

Lisaks merikõrvade andmestikule tehakse võrdluseks mõningad katsed veinikvaliteedi[1] andmestikuga, milles on 12 tunnust ning 1599 kirjet.

Kõigepealt tehakse veiniandmestiku katsed läbi kasutades sarnaseid andmetüüpe, mis eelnevates valitud (uuritavaks on numbriline pidev suurus). Seejärel proovitakse katseid kasutades uuritavana kategoorilist tunnust. Andmestikus on kategooriline tunnus quality, mis katsetes tõeväärtustunnusteks teisendatakse.

Esimesed katsed tehakse kasutades uuritavana tunnust alcohol. Korrelatsioonimaatriksi realt (Joonis 4) näeb, et MCAR puhul on sarnaselt algsete katsetega korrelatsioonikoefitsiendid võrdlemisi väikesed.

```
    alcohol       fixed.acidity    volatile.acidity         citric.acid      residual.sugar
 1.00000000          0.01576096         -0.02068759          0.03728318         -0.09929882
  chlorides free.sulfur.dioxide total.sulfur.dioxide             density                 pH
-0.04660387         -0.01874991          0.02518806         -0.06349724          0.00680713
  sulphates          quality.3          quality.4          quality.5          quality.6
 0.04880967          0.09979091         -0.09141148         -0.10975397         -0.02967117
  quality.7          quality.8
 0.05004633          0.13401591
```

Joonis 4 - MCAR veinikvaliteedi korrelatsioonikoefitsiendid

MAR puhul kustutati tunnuse alcohol väärtuseid, kui quality oli võrdne 6-ga. Joonis 5 näitab, et quality tunnusel 6 esineb tugev korrelatsioon uuritava tunnuse puudumisega.

```
     alcohol       fixed.acidity    volatile.acidity         citric.acid      residual.sugar
 1.000000000          0.042886908        -0.159359360          0.019081490        -0.040928203
   chlorides free.sulfur.dioxide total.sulfur.dioxide              density                 pH
-0.033638893        -0.008730020        -0.098865038          0.010048031        -0.005749261
   sulphates          quality.3          quality.4          quality.5          quality.6
 0.031163647        -0.294613394        -0.305931366         -0.324214880          0.997090845
   quality.7          quality.8
-0.331979496        -0.334320324
```

Joonis 5 - MAR veinikvaliteedi korrelatsioonikoefitsiendid

[1] https://archive.ics.uci.edu/ml/datasets/Wine+Quality

MNAR katses eemaldati alcohol väärtuseid, kui väärtus ise ületas piiri 10. Esinesid tugevamad koefitsiendid kui MCAR katsetes, kuid jäävad MAR suurematele koefitsientidele selgelt alla.

alcohol	fixed.acidity	volatile.acidity	citric.acid	residual.sugar
1.000000000	0.007034978	-0.258446938	0.091045861	0.078193395
chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH
-0.131221172	-0.088959453	-0.185425511	-0.171885228	0.054699799
sulphates	quality.3	quality.4	quality.5	quality.6
0.089209541	-0.682905768	0.064956566	-0.163938113	0.230210451
quality.7	quality.8			
0.426224963	-0.022934279			

Joonis 6 - MNAR veinikvaliteedi korrelatsioonikoefitsiendid

Alternatiivse andmestikuga sarnaseid tunnuseid uurides olid ka tulemused sarnased. Järgmisteks katseteks võeti uurimisaluseks kategooriline järjestatud tunnus quality.

MCAR ning MAR katsed andsid eelnevate katsetega analoogseid tulemusi – puuduvate andmete korrelatsioon oli MCAR puhul teiste tunnustega väike ning MAR puhul sõltuva tunnusega suhteliselt suur. MNAR puhul tekkisid mõningad erinevused. Katses (Joonis 7) kustutati quality väärtuseid, kui väärtus võrdus 6, 7 või 8-ga. Kõrgeim koefitsient on vaid veidi suurem, kui suurimad koefitsiendid MCAR katsetes – selle katse järgi on piir MCAR ja MAR eraldamiseks hägusem.

quality	fixed.acidity	volatile.acidity	citric.acid	residual.sugar
1.000000000	0.005411453	-0.087590801	0.009996783	0.149569033
chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH
0.053807365	0.022441596	-0.104808085	-0.050715627	0.024150867
sulphates	alcohol			
0.052056871	0.219273446			

Joonis 7 - MNAR veinikvaliteedi korrelatsioonikoefitsiendid, alternatiiv

MNAR katse tehakse uuesti läbi ka merikõrvade andmestiku kategoorilise tunnusega. Andmestikust kustutatakse tunnuse sugu väärtuseid, kui on tegu isaste isenditega. Katse tulemustest (Joonis 8) näeb, et puuduvate väärtuste esinemisega on teiste tunnustel nõrk korrelatsioon, mis on siiski keskmisest MCAR puhul esinevast korrelatsioonist kõrgem.

Sex	Length	Diameter	Ht	Whole_wt	Shucked_wt	Viscera_wt	Shell_wt	Rings
1.0000000	0.1635736	0.1580614	0.1596300	0.1802530	0.1813927	0.1670201	0.1662110	0.1046514

Joonis 8 - MNAR merikõrvad, alternatiiv 1

Võrdluseks tehti sama katse ka tingimusel, et kustutati isaste täiskasvanud isendite asemel ära soo väärtuseid, mis võrdusid väärtusega ’I’ (infant – imik). On näha, et antud juhul on korrelatsioon tugevam (Joonis 9).

```
      Sex     Length   Diameter         Ht   Whole_wt Shucked_wt Viscera_wt   Shell_wt      Rings
1.0000000 -0.4868071 -0.4910935 -0.4471949 -0.4780433 -0.4440626 -0.4784194 -0.4633791 -0.3752892
```

Joonis 9 - MNAR merikõrvad, alternatiiv 2

Kahe viimase katse erinevust on võimalik seletada andmestiku omapäraga. Imikueas isenditel on keskmiselt väiksemate väärtustega suuruslikud tunnused kui täiskasvanutel. Kuna isased isendid esinevad andmestikus palju laiemalt (suurem võimalike väärtuste vahemik), siis ei teki ka väga tugevat korrelatsiooni nende väärtuste eemaldamisel.

3.2.5 Järeldused puudumismehhanismi tuvastamiseks

Kirjanduse [1] põhjal võib kõige kindlama otsuse teha MCAR juhtumite kohta. Kui puuduvate väärtuste ning teiste tunnuste vahel suhteliselt tugevad seosed puuduvad, siis võib eeldada, et juhtum kuulub MCAR alla. Näiteid MCAR jaoks luues tuli välja, et terve andmestiku lõikes suvaliselt andmeid kustutades jäi korrelatsioonikoefitsientide absoluutväärtus andmete puudumise ning teiste tunnuste vahel enamasti alla 0,15.

MAR näiteandmestikes esines andmete puudumise ning sõltuva tunnuse vahel kindlalt tugevam korrelatsioonikoefitsient (suurem kui 0.9) kui teistest katsetes nähtu.

MNAR eraldamine teistest on keeruline ning üldiselt vajab see palju domeenipõhiseid teadmisi [16]. Näiteandmestikes varieerusid korrelatsiooni tugevused palju. Katsetest tuli välja, et MNAR koefitsiendid ei ulatunud piisavalt kõrgele, et seda lugeda MAR juhuks. Kui aga uuritava ning teiste tunnuste vahel märkimisväärseid seoseid polnud, siis olid MNAR juhtumid MCAR juhtumitest vaevu eraldatavad ilma välise infota. Võib öelda, et korrelatsioonidega MNAR ning MCAR eraldamiseks on tarvis selgete seoste olemasolu andmestikus.

3.2.6 Puudumismehhanismi tuvastamise realiseerimine

Puudumismehhanismi tuvastamiseks kasutatakse olemuselt lihtsat meetodit. Kontrollitakse puuduvate väärtuste esinemise ja teiste väärtuste vahelisi seoseid, leitakse kõige tugevam seos ning selle väärtuse järgi otsustatakse puudumismehhanism. Koefitsiendiga alla 0,15 liigitatakse MCAR juhtumiteks, koefitsiendiga üle 0,8 MAR juhtumiteks ning vahepealsed MNAR juhtumiteks. Koefitsientide künnised valiti katseliselt näiteandmestikega testides. Lõik realisatsioonist on toodud järgmises näites (Koodinäide 6).

```
# 1 - MCAR, 2 - MAR, 3 - MNAR
mcar_threshold <- 0.2
mar_threshold <- 0.8

missing_mechanism <- function(dat, attribute) {
  dat <- missing_to_binary(dat, attribute);
  dat <- factor_to_dummies(dat, attribute);
  cr <- hetcor(dat, std.err = F)$correlations[attribute,];
  correlations <- filter_correlations(cr, attribute);
  max <- max(correlations);
  if (max < mcar_threshold) {
    return(1);
  } else if (max < mar_threshold) {
    return(3);
  } else {
    return(2);
  }
}
```

Koodinäide 6 - puudumismehhanismi tuvastamine

3.2.7 Statistiline vs põhjuslik seos

Selles töös kasutatakse puudumismehhanismi üle otsustamiseks korrelatiivset seost puuduvate väärtuste esinemise ning teiste muutujate vahel. Sellest aga ei saa välja lugeda puudumise põhjust [25]. Näite võib tuua eelnevalt välja toodud MNAR näiteandmestiku korrelatsioonimaatriksist 3.2.3. Kuigi diameetri puuduvate väärtuste ning teiste suuruse ja kaaluga seotud tunnuste vahel esineb märgatav korrelatsioon, ei saa väita, et need kuidagi väärtuste puudumisega antud juhul seotud oleks.

3.3 Mehhanismi tuvastusmeetodi sobivus valitud lünkliku teadusandmestikuga

Andmestikus on 849 tunnust, 152 kirjet ning väga suurel hulgal puuduvaid andmeid. Eemaldatakse tunnused, milles on üle 20% väärtuseid puudu, kuna nende analüüsimine

ei anna palju infot. Alles jääb seejärel 236 tunnust. Kuna kirjeid on vähe, tähendab see seda, et on ka vähe puuduvaid ning olemasolevaid andmeid, mille pealt järeldusi teha. See viib kallutatud tulemusteni. Veel on probleemseks kohaks andmestikus esinevad rohked kategoorilised tunnused, mille kohta korrelatsioonimaatriksite genereerimine ei anna alati kõige informatiivsemaid resultaate – tulemuseks on tihti ülitugev positiivne või negatiivne korrelatsioon. Lisaks on nii suurel hulgal tunnuste jaoks korrelatsioonimaatriksite genereerimine väga ajamahukas. Arvutil, millega katseid tehti, võttis 236 tunnuse kohta maatriksi genereerimine umbes 45 minutit.

V2	V1	V5.0	V5.1	V5.2	V5.3
1.0000000	0.4949931	0.9998220	-0.9754657	-0.9772129	-0.8958083
V128.f	V128.t	V131.f	V131.t	V153.f	V153.t
0.8758643	-0.8895904	0.9200612	-0.9384913	0.9900320	-0.9929712

Joonis 10 - korrelatsioonid teadusandmestikus

Joonis 10 näitab ühe tunnuse, V2, andmete puudumise ning mõningate teiste tunnuste vahelised korrelatsioonikoefitsiendid. Nõnda äärmuslikud tulemused võivadki tekkida reaalsete juhtumite puhul väikeste valimite tõttu [26].

4 Imputeerimine

Imputeerimise eelduseks on teadmine, millise puudumismehhanismiga on tegu, kuna erinevad meetodid sobivad erinevate mehhanismidega (peatükk 2.3.9). Antud peatükis uuritakse, kuidas tulevad toime ühe laialt kasutatava paketi imputeerimismeetodid erinevate puudumismehhanismidega ning valitakse neist parim.

4.1 Võimalikud paketid imputeerimiseks

4.1.1 MICE

MICE [12] on mitmekordse imputeerimise põhjal loodud mitmekülgne pakett andmestike imputeerimiseks. Meetod kasutab iga tunnuse imputeerimiseks erinevat mudelit ning suudab imputeerida arvulisi, binaarseid, järjestamata ning järjestatud kategoorilisi andmeid. [18]. Azur jt [19] soovitavad MICE protseduure kasutada MAR puudumismehhanismi korral, aga van Buuren jt [12] kirjeldavad kasutamise võimalikkust ka MNAR puhul.

4.1.2 Amelia II

Sarnaselt MICE paketile kasutab Amelia mitmekordset imputeerimist ning eeldab, et puudumismehhanismiks on MAR [20]. Selle tugevateks külgedeks on töökindlus ning kiirus [21] [20]. Nõrkadeks külgedeks peetakse võimekust tegelema vaid pidevate muutujatega ning erinevalt MICE paketist kasutab Amelia ühise mudeliga lähenemist [20].

4.1.3 missForest

Kasutab imputeerimiseks juhusliku metsa algoritmi. Igale muutujale ehitatakse juhusliku metsa (random forest) mudel, millega ennustatakse lüngale väärtus kasutades olemasolevaid väärtuseid. Antud algoritm pakub kõrget kontrolli imputeerimisprotsessi üle ning on võimeline tegelema ka kategooriliste muutujatüüpidega. [20]

4.1.4 Hmisc

Hmisc omab kahte olulisemat imputeerimismeetodit – impute(), mis kasutab kasutaja defineeritud statistilisi meetodeid, ja aregimpute(), mis võimaldab imputeerida keskmisega kasutades aditiivset regressiooni, ennustatava keskmisega sobitamist (predictive mean matching - pmm) ja alglaadimist (bootstrapping) [20]. Hmisc võimaldab imputeerimisel ilma lisatööta kasutada binaarseid, kategoorilisi ning pidevaid tunnuseid [22].

4.1.5 mi

Kasutab mitmekordset imputeerimist ning ennustatava keskmisega sobitamist (pmm) [20]. Sarnaselt MICE paketile loob igale tunnusele eraldi mudeli imputeerimiseks [20]. mi pakett võimaldab kasutajatele põhjaliku kontrolli imputeerimisprotsessi üle, suudab töötada 11 erinevat tüüpi tunnustega ning automaatselt suudab tuvastada kaheksat eri tüüpi tunnust [23].

4.1.6 Valiku tegemine

Imputeerimiseks valiti MICE pakett selle kasutamise lihtsuse ning heade tulemuste [20] [24] tõttu. Lisaks leidub antud paketi kasutamise kohta põhjalikke artikleid ning õpetusi. MICE automatiseerib mõningad tehniliselt keerukad valikud. Näiteks valitakse automaatselt sobivad meetodid eri tüüpi muutujate imputeerimiseks [12]. Meetodid on määratud vaikimisi, kuid kõik on kasutaja poolt muudetavad [12].

4.2 Valideerimine

Imputeerimise õnnestumisest ettekujutuse saamiseks võrreldakse originaalandmestikust kustutatud andmeid imputeeritud andmetega. Selleks kasutatakse R-i paketi DMwR funktsiooni regr.eval, millega on võimalik arvutada statistilisi näitajaid regressiooni hindamiseks [24]. Täpsemalt uuritakse keskmise suhtelise absoluuthälbe suurusi. Realisatsioon R-is on toodud alljärgnevas koodinäites (Koodinäide 7). Lisaks võrreldakse imputeeritud väärtusi ning originaalandmestikust kustutatud väärtusi visuaalselt. Antud meetod annab mingisuguse ettekujutuse, kuidas imputeerimismeetodid esinevad, kuid selle järgi ei saa täie kindlusega öelda, kas üks meetod on parem kui teine. See sõltub väga palju uuritavast andmestikust. Üldiselt on vaja imputeerimisalgoritmid sobitada andmestikuga, et saada parimaid tulemusi [12].

```
actuals <- complete_dat[,attribute_modified][is.na(dat_missing[,attribute_modified])]
predicteds <- imp_dat[is.na(dat_missing[,attribute_modified]), attribute_modified]
regr.eval(actuals, predicteds)
```

Koodinäide 7 - andmete võrdlemine

4.3 MCAR

Näitamaks, kuidas võrreldakse eri meetodite esinemist, tehakse läbi MCAR andmestikuga imputeerimised vaikesätetega (ennustatava keskmisega sobitamine, peatükk 2.3.5) ning tingimusteta keskmisega. Pärast katsetamist nimetatud kahe meetodiga katsetatakse programselt läbi kõik MICE paketis olevad meetodid ning valitakse neist statistiliselt parim.

4.3.1 Üksikud katsed

MICE algoritmi kasutamine vaikesätetel (Koodinäide 8).

```
# attribute_modified <- 'Diameter'
imputed_data <- mice(datMCAR, m = 5)
imp_MCAR <- complete(imputed_data)
# check for method used
imputed_data$method[attribute_modified] # confirms "pmm" usage for imputation
```

Koodinäide 8 - MCAR imputeerimine

Korrates imputeerimisprotsessi ja regressioonvõrdlusi 10 korda ning katsete keskmiseid suhtelisi absoluuthälbeid koos analüüsides saadakse järgmised tulemused.

```
Min.      Median    Mean      Max.
0.03247   0.03842   0.03824   0.04120
```

Koodinäide 9 - MCAR pmm imputeerimise tulemused

Võrdluseks võib tuua vähem targa algoritmiga imputeerimise tulemused. Ennustatava keskmisega sobitamise asemel imputeeritakse tingimusteta keskmisega (2.3.3).

```
imputed_data <- mice(datMCAR, m = 5, defaultMethod = c("mean", "logreg", "polyreg",
"polr"))
```

Koodinäide 10 - imputeerimise vaikemeetodi vahetamine

defaultMethod parameetris vahetati meetod „pmm“ välja meetodi „mean“ vastu.

```
Min.    Median  Mean    Max.
0.2120  0.2551  0.2500  0.2821
```

Koodinäide 11 - MCAR keskmisega imputeerimise tulemused

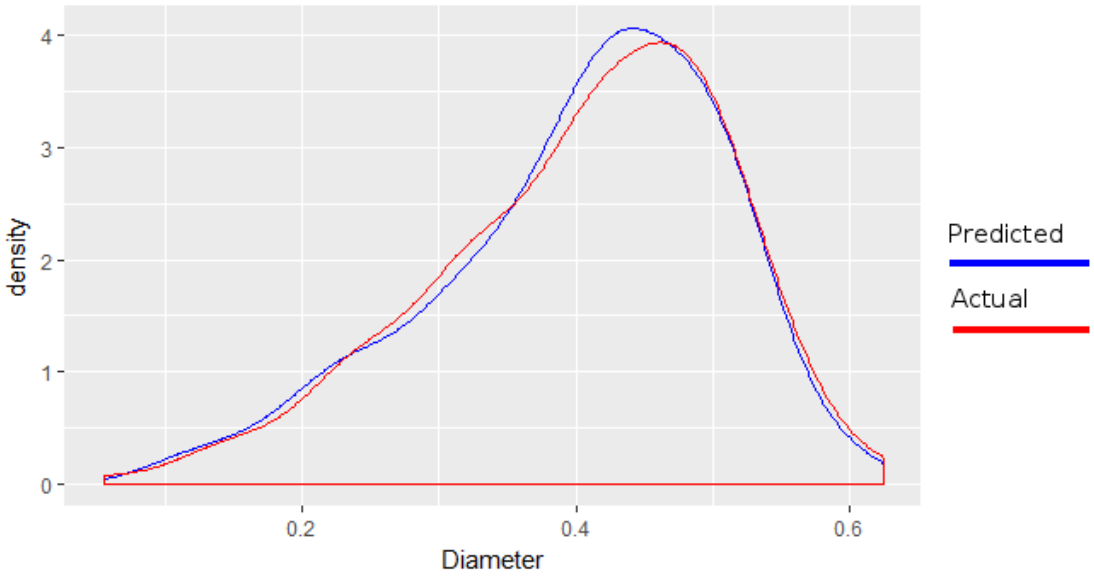
Nagu näha, siis keskmisega imputeerides on keskmine suhteline viga peaaegu 7 korda suurem.

Parema ettekujutuse saamiseks luuakse tihedusgraafikud kustutatud andmete ning imputeeritud andmete kohta. Graafikud luuakse paketi ggplot2 abil.

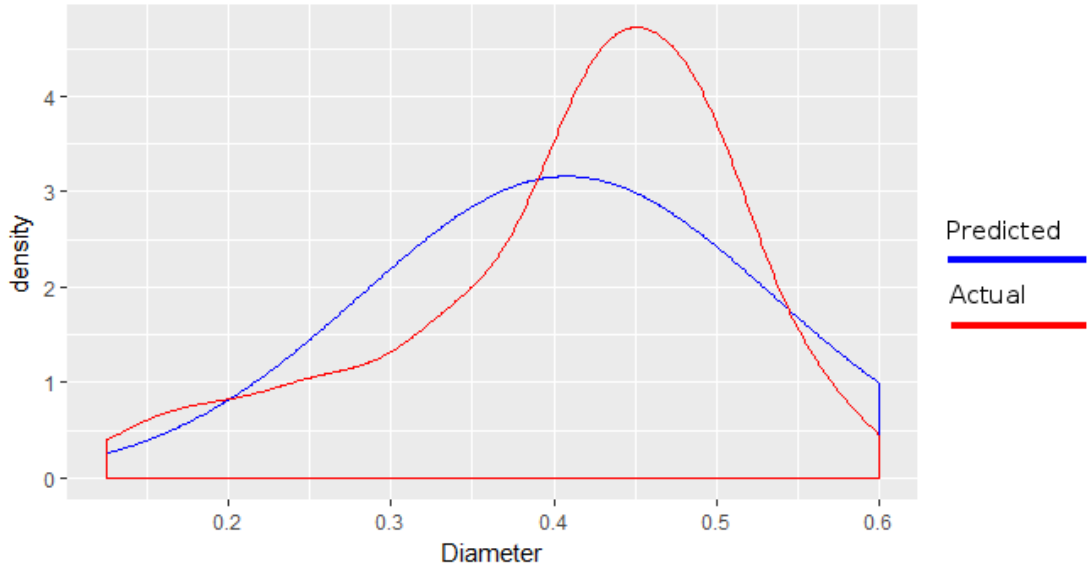
```
ggplot(predicted_diameter, aes(Diameter)) + geom_density(color='blue') +
       geom_density(data=actual_diameter, aes(Diameter), color='red')
```

Koodinäide 12 - tihedusgraafiku loomine

Graafikul (Joonis 11) on toodud ennustatava keskmisega imputeeritud andmete võrdlus kustutatud andmetega. Erinevused on suhteliselt väikesed. Võrdluseks on graafik toodud tingimusteta keskmisega imputeeritud andmete võrdlus kustutatud andmetega (Joonis 12). Graafik näitab selgelt, et keskmisega imputeerides kaotab andmestik suures osas oma eriomadusi, mida ennustatava keskmisega sobitamise algoritm suhteliselt hästi säilitada suudab.



Joonis 11 - MCAR pmm tihedusgraafik



Joonis 12 - MCAR mean tihedusgraafik

4.3.2 Parima leidmine

Puudumismehhanismi jaoks parima meetodi leidmiseks itereeritakse üle kõigi MICE teegi meetodite, mis antud andmestiku imputeerimiseks sobivad. Meetodite loetelu on toodud koodilõigus (Koodinäide 13). Meetoditega tehakse 10 testi ning võetakse nende testide keskmine mape (keskmine suhteline absoluuthälve). **Lisas 1** on toodud välja algoritm meetodite testimiseks MCAR puhul. Teiste puudumismehhanismide puhul vahetatakse välja vaid väärtuste kustutamise viis.

```
numeric_methods <- c('pmm', 'norm', 'norm.nob', 'norm.boot', 'norm.predict', 'mean',
       'quadratic', 'cart', 'rf', 'ri', 'sample', 'fastpmm')
```

Koodinäide 13 – võimalikud meetodid näiteandmestike imputeerimiseks

Jooniselt (Joonis 13) on võimalik välja lugeda, et „norm.predict“ meetod ehk lineaarregressiooni põhjal ennustatava väärtusega imputeerimine annab antud valideerimismeetodi põhjal kindlalt parimaid tulemusi. Antud meetodit kasutatakse MCAR juhtude imputeermiseks ka edaspidi.

```
      pmm      norm  norm.nob norm.boot norm.predict      mean
  0.04073   0.04452   0.04293   0.04326     0.02754   0.25490
quadratic      cart        rf        ri      sample  fastpmm
  0.07055   0.04118   0.04528   0.04687     0.33460   0.03953
```

Joonis 13 - MCAR meetodite esinemise statistika

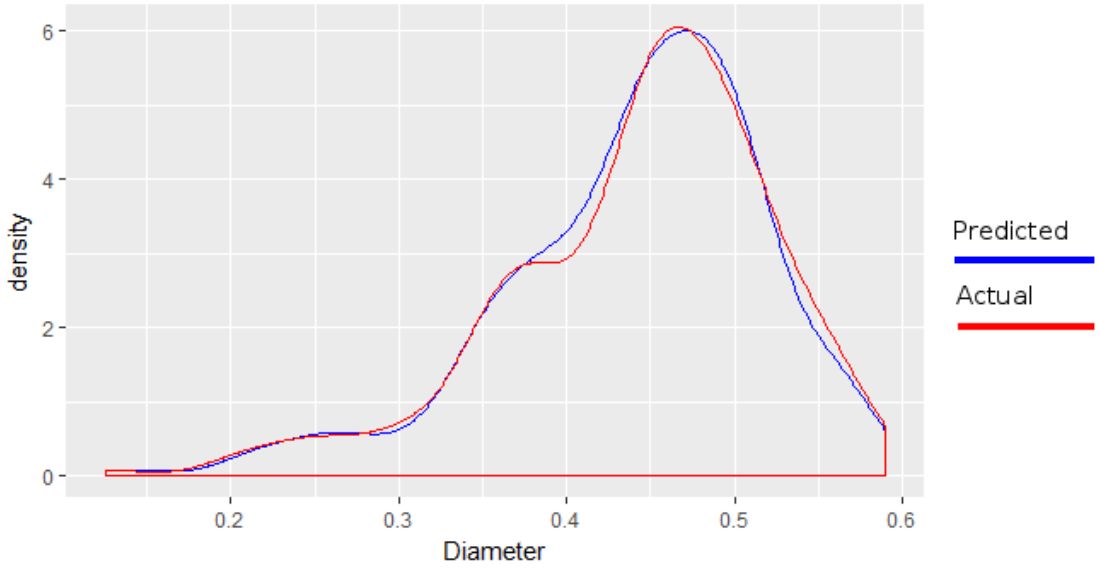
4.4 MAR

Parima meetodi leidmiseks katsetatakse imputeerimine läbi kõigi MICE teegi sobivate meetoditega.

```
      pmm      norm   norm.nob  norm.boot norm.predict      mean
  0.03765   0.04079   0.04005   0.04010   0.02645   0.19140
quadratic      cart        rf        ri    sample   fastpmm
  0.06584   0.03976   0.04321   0.04121   0.25780   0.03720
```

Joonis 14 - MAR meetodite esinemise statistika

On näha, et MAR puhul on tulemused suhteliselt sarnased MCAR katsetulemustega ning parimaks on taas „norm.predict“ meetod (Joonis 14). Järgnevalt on toodud ka ühe „norm.predict“ katse graafiline esitus (Joonis 15).



Joonis 15 - MAR norm.predict tihedusgraafik

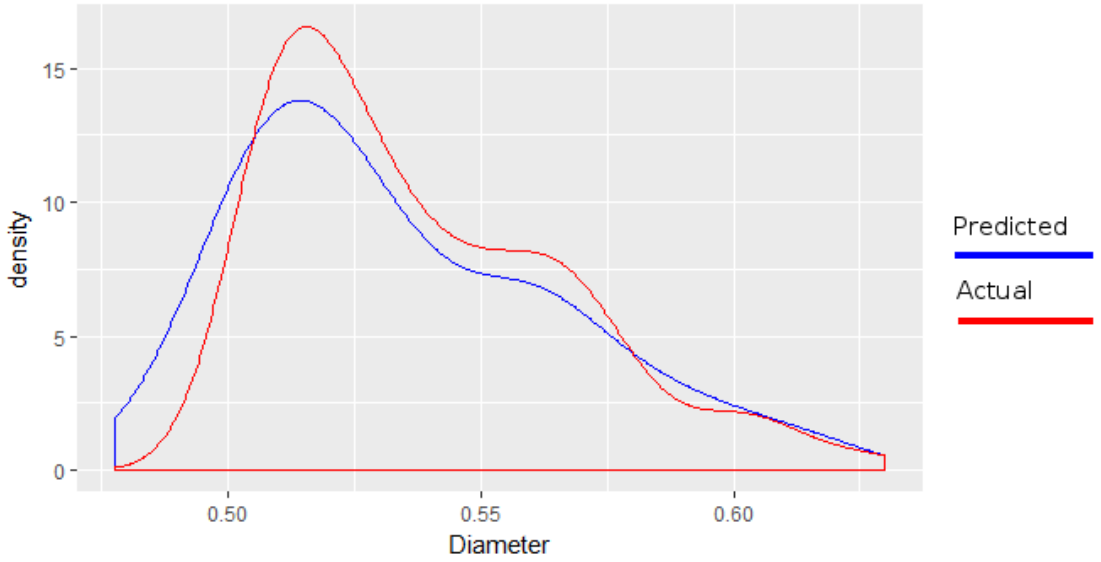
4.5 MNAR

MNAR imputeerimismeetodite katsetamiseks korratakse sama protseduuri nagu MAR puhul – kõik meetodid (Koodinäide 13) proovitaks läbi ning valitakse parim.

```
      pmm      norm   norm.nob  norm.boot norm.predict      mean
  0.03567   0.03295   0.03438   0.03451   0.02558   0.24410
quadratic      cart        rf        ri    sample   fastpmm
  0.05529   0.03679   0.04420   0.10370   0.25080   0.03420
```

Joonis 16 - MNAR meetodite esinemise statistika

Joonis 16 saab lugeda, et ka MNAR puhul on saavutatud parim tulemus „norm.predict“ meetodiga. Mõningad teised erinevused kahe eelneva puudumismehhanismiga siiski on. Kõige suurem nendest on meetodi „ri“ (random indicator) kehvem esinemine võrreldes eelnevate katsetega. Joonis 17 näitab, et parima MICE meetodi tulemused MNAR puhul jäävad tulemused teiste puudumismehhanismide imputeerimisele alla.



Joonis 17 - MNAR norm.predict tihedusgraafik

4.6 Järeldused

Antud viisil võrreldes sattus kõigi kolme puudumismehhanismi parimaks imputeerimismeetodiks MICE paketi puhul sama meetod (norm.predict). Põhjus võib tuleneda mitmest asjaolust – andmete eemaldamine on teostatud suhteliselt lihtsal viisil, andmestiku tunnused on omavahel tugevalt seotud. Lisaks ei ole valmismeetodid parimad lahendused MNAR juhtude imputeerimiseks, üldiselt vajavad need juhud algoritmide sobitamist vastavalt andmestiku omapäradele [12]. Sellest tulenevad ka suuremad erinevused eelneval graafikul (Joonis 17).

5 Kokkuvõte

Eesmärgiks oli luua meetod andmestikes puuduvate andmete mehhanismi tuvastamiseks ning valida mehhanismide imputeerimiseks sobiv meetod. Töö esimeses pooles tehti ülevaade puuduvate andmete teoreetilisest taustast ning populaarsematest meetoditest puuduvate andmetega tegelemiseks.

Mehhanismide tuvastamiseks loodi täielikest andmestikest mitmeid lünklike andmestikke vastavalt puudumismehhanismide kirjeldustele. Lünklikes andmestikes uuriti korrelatiivseid seoseid puuduvate andmete esinemise ning teiste tunnuste vahel. Tulemusena valmis eksperimentaalne meetod puudumismehhanismi tuvastamiseks. Limiteerivaks kohaks on ideaalsete näidete kasutamine uurimisel. Reaalelus kohtab harva juhtumeid, kus esineb puuduvate andmete kohta üks või teine mehhanism. Tuvastamise meetod töötas hästi väiksema arvu tunnustega andmestikes, kus tunnusete vahel esinesid piisavalt suured seosed ning tunnused olid peamiselt numbrilised.

Lisaks eelnevale leiti katseliselt parimad meetodid puudumismehhanismide imputeerimiseks MICE paketiga. Valitud valideerimismeetod on eksperimentaalne ning suhteliselt lihtne, seetõttu tuleks meetodi rakendatavust laiemalt kontrollida. Parima meetodi leidmise kaasproduktina valmis R-i kood, millega katsetatakse MICE paketi erinevaid.

Töös kirjutati valmisid mitmed R skriptid, millega viidi läbi katsed puudumismehhanismi tuvastamiseks ja imputeerimismeetodite võrdlemiseks. Lisaks sellele kirjutati skript puudumismehhanismi tuvastamise meetodi katsetamiseks ühe näiteandmestiku põhjal. Kõik skriptid on iseseisvalt käivitatavad ning väikese modifitseerimise abil kasutatavad erinevate sisseloetavate andmestikega.

Puudumismehhanismi tuvastamise ning imputeerimismeetodite testimisega on tehtud samm suurema eesmärgi poole, milleks on puuduvate andmete automaatne imputeerimine.

Kasutatud kirjandus

[1] D. Schmidt, M. Niemann ja G. L. von Trzebiatowski, „The Handling of Missing Values in Medical Domains with Respect to Pattern Mining Algorithms,“ %1 *Concurrency, Specification and Programming*, Rzeszow, 2015.

[2] J. R. Cheema, „Some General Guidelines for Choosing Missing Data Handling Methods in Educational Research,“ *Journal of Modern Applied Statistical Methods*, kd. 13, nr 2, pp. 53-75, 2014.

[3] L. O. Silva ja L. E. Zarate, „A brief review of the main approaches for treatment of missing data,“ *Intelligent Data Analysis*, kd. 18, nr 6, pp. 1177-1198, 2014.

[4] D. B. Rubin, „Inference and missing data,“ *Biometrika*, kd. 63, nr 3, pp. 581-592, 1976.

[5] J. L. Schafer ja J. W. Graham, „Missing data: Our view of the state of the art.,“ *Psychological Methods*, kd. 7, nr 2, pp. 147-177, 2002.

[6] C. K. Enders ja D. L. Bandalos, „The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models,“ *Structural Equation Modeling: A Multidisciplinary Journal*, kd. 8, nr 3, pp. 430-457, 2009.

[7] A. R. T. Donders, G. J. van der Heijden, T. Stijnen ja K. G. Moons, „Review: A gentle introduction to imputation of missing values,“ *Journal of Clinical Epidemiology*, kd. 59, nr 10, pp. 1087-1091, 2006.

[8] UCLA Statistical Consulting Group, „Multiple imputation in Stata,“ UCLA, [Võrgumaterjal]. Available: http://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/. [Kasutatud 30 04 2017].

[9] stata.com, „ Impute using predictive mean matching,“ [Võrgumaterjal]. Available: http://www.stata.com/manuals14/mimiimputepmm.pdf. [Kasutatud 04 05 2017].

[10] P. Gaffert, F. Meinfelder ja V. Bosch, „Towards an MI-proper Predictive Mean Matching,“ 25 01 2016.

[11] C. K. Enders, „A Primer on Maximum Likelihood Algorithms Available for Use With Missing Data,“ *Structural Equation Modeling*, kd. 8, nr 1, pp. 128-141, 2001.

[12] S. van Buuren ja K. Groothuis-Oudshoorn, „mice: Multivariate Imputation by Chained Equations in R,“ *Journal of Statistical Software*, kd. 45, nr 3, 2011.

[13] K. Rootalu, „Korrelatsioonikordajad,“ 2014. [Võrgumaterjal]. Available: http://samm.ut.ee/korrelatsioonikordajad. [Kasutatud 14 04 2017].

[14] U. Olsson, F. Drasgow ja N. J. Dorans, „The Polyserial Correlation Coefficient,“ *Psychometrika*, kd. 47, nr 3, pp. 337-347, 1982.

[15] F. Holgado-Tello, S. Moscoso, I. Barbero-Garcia ja E. Vila, „Polychoric versus Pearson correlations in Exploratory and Confirmatory Factor Analysis with ordinal variables,“ *Quality and Quantity*, kd. 44, nr 1, pp. 153-166, 2010.

[16] K. Grace-Martin, „How to Diagnose the Missing Data Mechanism,“ [Võrgumaterjal]. Available: http://www.theanalysisfactor.com/missing-data-mechanism/. [Kasutatud 14 04 2017].

[17] „What is R?,“ [Võrgumaterjal]. Available: https://www.r-project.org/about.html. [Kasutatud 25 03 2017].

[18] „mice: Multivariate Imputation by Chained Equations,“ [Võrgumaterjal]. Available: https://CRAN.R-project.org/package=mice. [Kasutatud 15 04 2017].

[19] M. J. Azur, E. A. Stuart, C. Frangakis ja P. J. Leaf, „Multiple Imputation by Chained Equations: What is it and how does it work?,“ *International Journal of Methods in Psychiatric Research,* kd. 20, nr 1, pp. 40-49, 2011.

[20] Analytics Vidhya Content team, „Tutorial on 5 Powerful R Packages used for imputing missing values,“ Analytics Vidhya, 04 03 2016. [Võrgumaterjal]. Available: https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/. [Kasutatud 17 04 2017].

[21] „Amelia II: A Program for Missing Data,“ [Võrgumaterjal]. Available: http://gking.harvard.edu/amelia. [Kasutatud 17 04 2017].

[22] N. J. Horton ja K. P. Kleinman, „Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models,“ *The American Statistician,* kd. 61, nr 1, pp. 79-90, 2007.

[23] Y.-S. Su, A. Gelman, J. Hill ja M. Yajima, „Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box,“ *Journal of Statistical Software,* kd. 45, nr 2, 2011.

[24] S. Prabhakaran, „Missing Value Treatment,“ 25 04 2016. [Võrgumaterjal]. Available: https://www.r-bloggers.com/missing-value-treatment/. [Kasutatud 25 04 2017].

[25] „Correlation does not imply causation,“ [Võrgumaterjal]. Available: https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation. [Kasutatud 25 04 2017].

[26] U. Science, „Real world results,“ Understanding Science, [Võrgumaterjal]. Available: http://undsci.berkeley.edu/article/real_world_results. [Kasutatud 20 05 2017].

Lisa 1 – MICE imputeerimismeetodite testimine

```
for (method in numeric_methods) {
  mape_array <- c()
  for (iteration in 1:10) {
    modified_datMCAR <- datMCAR
    # delete values
    modified_datMCAR[sample(1:nrow(modified_datMCAR), amount_removed),
attribute_modified] <- NA
    #impute values
    imputed_data <- mice(modified_datMCAR, m = 5, defaultMethod = c(method, "logreg",
"polyreg", "polr"))
    imp_MCAR <- complete(imputed_data)
    # comparison
    actuals <-
       original[,attribute_modified][is.na(modified_datMCAR[,attribute_modified])]
    predicteds <-
       imp_MCAR[is.na(modified_datMCAR[,attribute_modified]), attribute_modified]
    mape <- regr.eval(actuals, predicteds)["mape"]
    # add mape to array
    names(mape) <- NULL
    mape_array <- c(mape_array, mape)
  }
  # get mean from summary of mean absolute percentage errors
  mean <- summary(mape_array)['Mean']
  names(mean) <- NULL
  result_array <- c(result_array, mean)
}
# results
names(result_array) <- numeric_methods
```