TURUN YLIOPISTO

Datan imputointi eikä amputointi

Menetelmiä ja strategioita puuttuvan datan käsittelyyn

Kandidaatin tutkielma

Turun yliopisto

Tietotekniikan laitos

Tietojenkäsittelytieteen tutkinto-ohjelma

Laatija:

Verneri Kallio

Ohjaajat:

Professori Tapio Pahikkala

11 2024

Kandidaatin tutkielma
Tietotekniikan laitos
Turun yliopisto

Oppiaine: Tietojenkäsittelytiede
Tutkinto-ohjelma: Tietojenkäsittelytieteen tutkinto-ohjelma
Tekijä: Verneri Kallio
Otsikko: Datan imputointi eikä amputointi
Sivumäärä: 28 sivua, 1 liitesivu
Päivämäärä: 11 2024

Abstrakti

Data-analytiikka on ala, joka kasvaa jatkuvasti, sillä dataa kerätään enemmän ja monipuolisemmin erilaisten laitteiden avulla. Samalla datan puuttuvien tai virheellisten arvojen oikeanlainen käsittely muuttuu tärkeämmäksi, sillä yhä enemmän päätöksiä ja tutkimuksia tehdään sen pohjalta. Tutkielmani pyrkii luomaan yleiskäsityksen siitä, millaisia menetelmiä ja strategioita käytetään puuttuvan datan hallitsemiseen ja hyödyntämiseen. Tutkielmassa keskitytään rakentamaan pohjustus, jotta voi ymmärtää puuttuvan datananalyysin teorian käsitteellisellä tasolla, joten syvempi matemaattinen näkökulma rajataan pois monien aiheiden kohdalla. Datamekanismit kuten MCAR, MAR ja MNAR ovat oleellisia puuttuvan datan oikeanlaiseen käsittelyyn, sillä niiden avulla voimme paremmin ymmärtää millaisia tekniikoita kannattaa soveltaa erilaisten vaillinaisten datasettien kohdalla. Tämän ymmärryksen parantamiseksi on olemassa monia muita strategioita kuten puuttuvuuden kuvioiden huomioiminen sekä selkeiden tavoitteiden määrittäminen imputoinneille. On kehitetty monenlaisia imputointi- ja poistomenetelmiä vuosikymmenien aikana ja vaikka monet niistä ovat vanhentuneet tai harvoin optimaalisin valinta niin niiden opettelu on edelleen hyödyksi, sillä ne auttavat käsittämään miten ja miksi edistyneemmät menetelmät toimivat. Nämä tutkielmassa esitetyt kehittyneemmät menetelmät ovat moni-imputointi ja MICE-algoritmi. Tutkielmassa keskitytään jatkuvaan ja numeeriseen taulukkomuodossa olevaan keinotekoiseen dataan, mutta useimpia esitettyjä tekniikoita ja strategioita voidaan soveltaa monen erilaisen datatyypin kohdalla.

Asiasanat: imputointi, moni-imputointi, data-analytiikka, datamekanismi, MCAR, MAR, MNAR, MICE

Sisällysluettelo

1 Johdanto

Puuttuva data on edelleen suuri ongelma monissa kvantitatiivisissa tutkimuksissa eri aloilla. Käsittelemätön tai huonolla tavalla käsitelty puuttuva data voi merkittävästi vääristää myöhempiä analyysejä ja tuloksia. On siis ensiarvoisen tärkeätä panostaa datan laatuun jo aikaisessa vaiheessa tutkimusta. Eräs tekniikka tämän laadun varmistamiseksi on nimeltä imputointi, joka tarkoittaa arvon tai arvojen sijoittamista puuttuvan datan tilalle siten, että ne simuloivat todellisia arvoja mahdollisimman hyvin. [7]

On keksitty monenlaisia menetelmiä puuttuvan datan hallitsemiseen. Modernimmat menetelmät kuten moni-imputointi (MI) ja suurin uskottavuus (ML) ovat yleensä paljon parempia vaihtoehtoja perinteisiin menetelmiin verrattuna kuten poistaminen ja keskiarvon imputointi. Tämä johtuu muun muassa siitä, että kehittyneemmät menetelmät eivät vaadi yhtä tiukkoja oletuksia puuttuvasta datasta ja täten virhealttius vähenee, joista perinteiset tekniikat usein kärsivät. [7]

Kirjoitelmassa keskitytään kolmeen tapaan käsitellä puuttuvaa dataa. Ensimmäinen on täydellisten havaintorivien analyysi (CCA) ja se sisältää pelkästään täydellistä dataa, sillä kaikki puuttuvaa dataa sisältävät rivit on poistettu. Tämä johtaa usein harhaan, sillä poistetuilla riveillä on usein dataa, jota olisi voinut hyödyntää. Toinen tekniikka on yksittäisimputointi, jossa puuttuvien arvojen tilalle laitetaan yksittäinen ennustus perustuen muihin arvoihin kuten esimerkiksi taulukon sarakkeen keskiarvo. Tällä tekniikalla säilytetään vaillinaiset havainnot tai datarivit, mutta se on hyvin epätarkka menetelmä. Kolmas menetelmä on moni-imputointi, jossa jokaista uupuvaa arvoa kohden tehdään monta ennustusta. [12]

Tutkielman tarkoitus on auttaa käsitteellisesti ja myös teknisesti ymmärtämään moderni ja laajassa käytössä oleva menetelmä moni-imputointi ja antaa yleiskuva vanhemmista ja yksinkertaisemmista tekniikoista sekä miksi nämä ovat yleensä selkeästi huonompi valinta. Käydään myös läpi teoreettiset perustukset puuttuvan datan analysoinnissa, kuten datamekanismi ja eräs suosittu moni-imputointia soveltava algoritmi nimeltä MICE. Lisäksi otetaan selvää erilaisista strategioista, kuten herkkyysanalyysi, joka kannattaa huomioida parhaan puuttuvuuden menetelmän valinnassa.

Syy sille miksi tutkielman aihealue kiinnostaa minua ja miksi tutkin sitä on se, että puuttuva data on suuri ongelma data-analytiikassa niin tutkimuksissa kuin työelämässä, mutta tästä huolimatta sille ei usein annetta tarpeeksi huomiota ja ongelmat suurenevat entisestään. Alan tieteelliset julkaisut ja kirjat ovat usein aika teoreettisia ja muualta löytyvät lähteet ovat taas melko suppeita, joten olen yrittänyt tällä tutkimuksella löytää näiden väliltä sopivan välimaaston.

Tutkielman tutkimus tehtiin kirjallisuusanalyysinä ja lähteet haettiin UTU Volterin kautta. Lähteet löydettiin hakusanoilla kuten practical guide | handbook to missing data, MCAR, MNAR, MAR, missing data imputation, imputation, missing pattern, multiple imputation ja MICE. Lähteiden karsinta suoritettiin alussa otsikon ja abstraktin avulla. Seuraavaksi tein hyödyllisimmistä lähteistä muistiinpanoja ja aloin aikaisessa vaiheessa jo kirjoittamaan tutkielmaa. Suosin lähteissä uusimpia julkaisuja ja kirjoja lyhyiden artikkeleiden ja vanhojen tutkimusten yli. Monet vanhat aineistot ovat säilyttäneet pitkälti arvonsa, niin käytin myös vanhempia lähteitä. Alussa oli haasteita sopivien lähteiden löytämisessä, sillä eri julkaisuissa on suuria eroja siinä miten käytännönläheisiä ja teoreettisia ne ovat. En myöskään juuri tuntenut aihealuetta tai tilastollista sanastoa ennalta niin alussa oli aika jyrkkä oppimiskäyrä.

On olemassa monia muita tekniikoita puuttuvuuden käsittelyyn kuten apumuuttujien käyttö sekä puuttuvuuden huomioivien mallien soveltaminen, kuten suurimman uskottavuuden menetelmä. Nämä menetelmät mainitaan tosin vain pikaisesti ja on monia muita menetelmiä sekä strategioita puuttuvuuden käsittelyyn, joita en ole käynyt läpi lainkaan, sillä valitsin vain tärkeimmät ja geneerisimmät menetelmät tutkielman rajatun aihealueen huomioiden.

Monissa erilaisissa käytännön sovelluksissa joko MCAR tai MAR mekanismi on hyvä oletus ja puuttuvuuden korjaaminen on myös paljon yksinkertaisempi prosessi, joten syvennyn näihin tilanteisiin ja mainitsen vain lyhyesti menetelmiä, joita kannattaa käyttää puuttuvuuden ollessa MNAR. Tutkielma keskittyy klassisiin tilastollisiin menetelmiin eikä uudempiin koneoppimiseen pohjautuviin tekniikoihin, sillä edellisessä on usein monia etuja, kuten matalampi prosessointiteho ja ajan vaatimus [3]. Tutkielman tilastolliset menetelmät ovat usein myös sekä teknisesti että konseptuaalisesti yksinkertaisempia ja helpompia ymmärtää, joten väärinkäsityksistä johtuva virhealttius näiden implementoinnissa pienenee.

2 Mekanismit

2.1 Puuttuvan datan luokittelu mekanismeilla

Datan kerääminen voi johtaa puuttuvan datan erilaiseen jakaumaan, frekvenssiin ja rakenteeseen, joten sen tyypin arvioiminen ja luokittelu on tärkeätä ennen varsinaista puuttuvan datan käsittelyä. Tämä auttaa sopivan imputointimenetelmän valinnassa sekä muita raportin tai tutkimuksen lukijoita ymmärtämään tutkijan motiiveja eri päätöksille. Rubin ja tämän kollegat Little & Rubin keksivät luokittelutavan, joka on edelleen aktiivisessa käytössä tänään: MCAR, MAR ja MNAR [19]. Nämä mekanismit kuvaavat suhteita havaittujen ja mitattujen muuttujien sekä puuttuvan datan todennäköisyyden välillä. Mekanismeilla on tarkat matemaattiset määritelmät, mutta sisimmissään ne ovat kolme eri selitystä puuttuvalle datalle. Käytännön näkökulmasta, mekanismit ovat oletuksia, jotka vaikuttavat valitun puuttuvan datan käsittelyyn otetun tekniikan suorituskykyyn. Tämä johtuu siitä, että eri menetelmät olettavat erilaisia premissejä datasta toimiakseen ja voivat täten antaa hyvin vääristyneitä tuloksia niiden ollessa virheellisiä. Pitää silti muistaa, että valittaessa eri menetelmiä tiedämme harvoin varmuudella puuttumisen syytä. [17]

2.2 Kolme tyypillistä mekanismia

Missing Not At Random (MNAR), eli data ei puutu sattumanvaraisesti, sillä puuttuminen liittyy itse hypoteettiseen puuttuvaan datan arvoon eikä havaittuihin arvoihin. Kyselyssä saatetaan esimerkiksi kysyä, että käytätkö kannabista ja tähän saatetaan jättää vastaamatta riippuen maan laeista, kulttuurista ja siitä miten avoin vastaaja on valmis olemaan aiheesta. Jokin asia puuttuvan datan arvoissa, eli tässä tilanteessa myöntymys päihteen käyttöön, johtaa siis sen puuttumiseen. [14]

Missing At Random (MAR), harhaanjohtavasta nimestään huolimatta datan puuttuminen ei ole satunnaista vaan on tässä mekanismissa riippuvainen muista huomioiduista arvoista, mutta ei itsestään, eli hypoteettisista arvoista, jos data olisi ollut täydellistä. MAR esimerkkinä voidaan pitää tilannetta, jossa koulu jakaa opiskelijoille soveltuvuuskokeen ja vain ne, jotka ylittävät tietyn pisterajan osallistuvat vaativampaan matematiikan kurssiin. Tällöin vaativan matematiikan arvosanat ovat MAR, sillä niiden puuttuvuus on täysin riippuvainen soveltuvuuskokeen tuloksista. [6]

Missing Completely At Random (MCAR), eli data puuttuu täysin sattumanvaraisesti ja ei ole mitenkään riippuvainen muista huomioiduista arvoista tai puuttuvasta datasta itsessään. Puuttuvuus on siis täysin epäsystemaattista ja havaitun tiedon voidaan ajatella olevan satunnaisotos hypoteettisesti täydellisestä datasta. Puuttuvat arvot ovat MCAR esimerkiksi silloin, jos ihminen muuttaa toiseen kaupunkiin kesken tutkimuksen ja ei tämän vuoksi voi enää osallistua kyselyihin ja testeihin, ja muuton syyn ollessa täysin riippumaton muista muuttujista datasetissä. Muutto on myös riippumaton hypoteettisista arvoista, jotka ihminen olisi antanut, jos ei olisi muuttanut pois. [6]

On olemassa vielä neljäs luokitus nimeltä rakenteellisesti puuttuva data (SMD). Data puuttuu tarkoituksella, joten on yleensä helppoa selvittää tekijä, joka sen aiheutti. Esimerkiksi kyselyssä voi olla kysymys omaatko lapsia ja tämän jälkeen voi olla toinen kysymys: kuinka monta lasta sinulla on. Jos henkilöllä ei ole lapsia niin tämä ei pystyisi totuudenmukaisesti vastaamaan jälkimmäiseen kysymykseen, joten tämä luultavasti jättäisi kokonaan vastaamatta ja näin syntyisi rakenteellisesti puuttuvaa dataa. Joidenkin lähteiden mukaan tämä mekanismi on vain MNAR luokan aliluokka, sillä erolla, että SMD on helpompi havaita ja analysoida. SMD eroaa kuitenkin tästä ja muista mekanismeista sekä käsitteellisesti, että puuttuvan datan käsittelyyn käytettyjen tekniikoiden suhteen sen verran, että ansaitsee joidenkin lähteiden mukaan oman erillisen luokkansa [16].

2.3 Mekanismien rajoitukset ja testaaminen

Luokittelusta huolimatta puuttuvaa dataa voi harvoin selittää puhtaasti yhden mekanismin avulla, sillä dataa puuttuu monesta eri syystä. Mekanismit ovat siksi vain ohjenuoria siitä millaista puuttuva data on tyypiltään suurin piirtein. Datamekanismit eivät kuvaa koko datasetin piirteitä vaan vain oletuksia, jotka kohdistuvat tiettyyn analyysiin. Sama data voi siis johtaa kaikkiin luokitteluihin MAR, MNAR tai MCAR riippuen siitä mitä muuttujia ja datasetin osia on sisällytetty.

Monien metodologien mielestä MCAR oletus on niin tiukka, että se toteutuu erityisen harvoin käytännössä ja muissa kuin datan keinotekoisissa simulaatiotilanteissa [18]. Tämä käy järkeen, sillä tutkija tai muu datan kokooja on yleensä halunnut selittää jotain ilmiötä datalla, jossa eri muuttujat vaikuttavat toistensa arvoihin. MAR ja MNAR mekanismien oletukset ovat siksi vähemmän tiukkoja ja vaativia, sillä ne eivät sulje tätä mahdollisuutta kokonaan pois.

Kolmesta datamekanismista vain MCAR on mahdollista testata empiirisesti. Metodologit ovat esittäneet monia eri tekniikoita tähän, mutta valitettavasti ne yleensä kärsivät matalasta tilastollisesta voimasta. Eräät tilastolliset testit MCAR mekanismin oletuksen tukemiseen ovat t-testi, chi-square, logistinen regressio ja Little's MCAR testi [20][15]. Ensimmäiset kaksi testiä käytetään yhden muuttujan analyysiin ja jälkimmäisiä sovelletaan monimuuttuja-analyysissä. Näiden menetelmien pääidea on verrata puuttuvaa dataa sisältävät rivit niihin, jotka ovat täydellisiä ja jos erot ovat tarpeeksi suuret niin havainto tukee MCAR olettamusta. On kuitenkin kyseenalaista miten hyödyllisiä nämä testit ovat käytännössä, sillä monet hyvät menetelmät kuten moni-imputointi toimivat ongelmitta sekä MAR että MCAR oletuksen pohjalta [22].

MCAR mekanismia ei ole kuitenkaan mahdollista aukottomasti todistaa, sillä tämä vaatisi meitä näyttämään, että ei ole mitään mahdollista tapaa ennustaa puuttuvia arvoja havaittujen arvojen avulla. Voimme silti usein todistaa, että MCAR ei päde siten, että näyttää puuttuvuuden olevan jollain tavalla ennustettavissa havaituista arvoista. Eli muuttujien välillä on tietynlaisia riippuvuuksia, kuten lineaarinen suhde. [14]

MAR JA MNAR mekanismien tilastollinen verifiointi on taas mahdotonta, sillä ne perustuvat puuttuvaan dataan. Täten MI ja ML metodien perusolettamaa MAR ei voi siis juuri testata. On mahdollista kuitenkin muilla tavoin saada mahdollisimman hyvät oletukset mekanismin tyypistä. [7]

Saadessa käsiimme osajoukon puuttuvista arvoista niin voimme tehdä jonkinlaisia johtopäätöksiä muidenkin uupuvien arvojen mekanismista. Voisimme esimerkiksi soittaa kyselyyn vastanneille pyytäen tarkennusta puuttuviin arvoihin, mutta tällainen tai vastaava ei silti yleensä ole vaihtoehto. On siis hyvin tärkeää käyttää hyväksi datan lisäksi alaan liittyvää tietämystä kuten jo valmiiksi tunnettuja suhteita eri muuttujien välillä, syitä puuttuvaan dataan ja käytettyjä datankeräysmenetelmiä. Nämä keinot voivat auttaa meitä tunnistamaan MNAR mekanismin ja välttää harhan syntyä. [27]

3 Strategioita ja syitä imputointiin

3.1 Puuttuvuuden syntytavat

Tilastollista analyysia tekevät kohtaavat usein ongelmia puuttuvan datan vuoksi. Kyselyissä ei usein esimerkiksi haluta kertoa omaa palkkaa. Tarkoituksella vastaamatta jättämisen lisäksi on monta muuta syytä puuttuvaan dataan. Itsetäytetyissä kyselyissä saatetaan olla vastaamatta huolimattomuuden vuoksi tai koska ei ymmärretä kysymystä. Vastaaja ei joskus yksinkertaisesti tiedä vastausta, ei pääse käsiksi tietoon sillä hetkellä tai ei koe kysymyksen edes pätevän itseensä. On esimerkiksi hankalaa arvioida kyselyssä puolisoaan, jos ei sellaista omaa. Pitkittäistutkimuksessa voi syntyä ongelmia, jos aiemmin haastatellut ihmiset kuolevat tai muuttavat pois ennen seuraavaa kyselyä. Lisäksi todellista dataa saatetaan menettää sen keräämisen, tallentamisen, siirron tai virheellisen imputoinnin vuoksi, joka on johtanut datan puuttumiseen tai korruptoitumiseen. Puuttuvaa tai muuten huonolaatuista dataa on näistä syistä yleensä mahdotonta välttää kokonaan. [13]

3.2 Syyt imputoinnille

Imputointi saatetaan haluta suorittaa monista syistä. Arvoa ei ole ehkä lainkaan olemassa tai olemassa oleva arvo puuttuu osittain arvojen ollessa intervalleja. Epätarkka arvo halutaan usein vaihtaa tarkempaan uniikkiin arvoon, jotta saadaan esimerkiksi parempi ennuste jakaumasta. Toinen syy imputoinnille voisi olla, että arvo ei vaikuta olevan oikea, joten halutaan vaihtaa se totuudenmukaisempaan arvoon. Nykyinen arvo voi olla myös liian henkilökohtainen, jos jonkun henkilöllisyys voi paljastua sen kautta. Henkilökohtaisen arvon korvaaminen imputoidulla arvolla poistaisi monia ongelmia, mutta arvo olisi myös vähemmän tarkka. Joskus halutaan suorittaa erilaisia imputointeja erilaisia ennustuksia varten. Voi siis olla, että jotkut ennustukset tehdään tiettyjen imputoitujen arvojen avulla ja toiset ei. Tämä johtuu siitä, että tietynlainen imputointi ei aina paranna ennustusten tuloksia ja voi jopa huonontaa niitä. Imputointi kannattaa siis suorittaa vain niille arvoille, joista on hyötyä suoritettavien ennustusten suhteen. [9]

3.3 Strategioita imputointiin

Herkkyysanalyysi viittaa analyyseihin, jotka tehdään eri tavalla kuin pääanalyysissä. Täten voidaan helpommin arvioida miten pääanalyysistä poikkeavat oletukset, kuten MNAR mekanismi MAR oletuksen sijasta, saattavat vaikuttaa saatuihin tuloksiin. Herkkyysanalyysi antaa arvion siitä millainen vaihteluväli ja epävarmuus pääanalyysin tuloksilla on. Tämä kannattaa suorittaa jokaiselle puuttuvuutta sisältävälle muuttujalle ollessamme epävarmoja puuttuvien arvojen mahdollisesta vaikutuksesta analyysin tuloksiin. Se voitaisiin toteuttaa esimerkiksi siten, että luodaan simuloitu datasetti, poistetaan siitä dataa tietyn mekanismin mukaisesti ja lopuksi suoritetaan samanlainen analyysi kuin alkuperäiselle datalle ja vertaillaan näiden tuloksia keskenään. [25]

On tärkeätä ottaa huomioon puuttuvan datan prosentuaalinen määrä ennen kuin päättää mitä imputointitekniikkaa käyttää [7]. Mitä enemmän puuttuvaa dataa on, niin sitä tärkeämmäksi muodostuu hyvän menetelmän valinta. Esimerkiksi kun se on noin 15–20 % koko datasta niin se voi vaikuttaa parametrien ennustuksiin merkittävästi, jos dataa taas puuttuu vähemmän kuin noin 10 % niin sillä ei monissa tilanteissa ole niin paljon merkitystä mitä menetelmää käyttää puuttuvuuden käsittelyyn. [8] Puuttuvuuden ollessa suurta kuten yli 50 % niin datan laatu on yleensä liian huono, jos puuttuva dataa sisältävän rivit poistetaan. Joidenkin lähteiden mukaan moni-imputoinnista voi olla hyötyä jopa silloin, kun puuttuvuus on 90 %, kunhan mekanismi on MAR ja apumuuttujia on hyödynnetty [23][24]. Vaikka puuttuvuus olisi pientä kuten alle 5 %, niin jos on saatu selville, että tärkeämmät vastaajat puuttuvat niin kaikki mahdollinen kannattaa tehdä datan laadun parantamiseksi. Esimerkiksi tulokyselyssä, jossa kaikkien todella rikkaiden henkilöiden vastaukset puuttuvat voi vääristää tuloksia merkittävästi, vaikka puuttuvuus olisi pientä. Imputointi on silti joskus ainoa vaihtoehto.

Ennen kuin toteuttaa imputoinnin kannattaa selkeästi määrittää tämän tavoitteet. Eräs harvoin toteutuva tavoite on, kun imputoidut arvot halutaan jokainen mahdollisimman lähelle todellisia arvoja, eli onnistuminen yksilöllisellä tasolla. On kuitenkin yleensä hankala tietää miten hyviä imputoinnit ovat. Tämä on yleensä liian vaativa tavoite ja vaikea toteuttaa käytännössä. Toinen tavoite on pyrkimys saada imputoitujen arvojen jakauma mahdollisimman lähelle oikeiden arvojen jakaumaa. Tätä on myös hankala tarkistaa, mutta helpompaa kuin ensimmäisessä tavoitteessa. Kolmas tavoite on onnistuminen aggregaatti tasolla, joka voi olla tarpeeksi hyvä tulos. Esimerkkejä tästä ovat keskiarvo, mediaani, suhde, summa ja standardipoikkeama. Nämä voidaan tarkistaa jossain määrin, varsinkin jos kyselyt toistetaan. Neljäs ja vähiten vaativa tavoite on muuttujien välisten suhteiden säilyttäminen, kuten korrelaatio ja kovarianssi, joka on myös tärkeätä monissa tutkimuksissa. Käytännössä toinen ja kolmas mainituista tavoitteista ovat suosituimpia, sillä indikaattorit niiden suorituskyvylle ovat usein löydettävissä. Huonoin strategia on imputointi ilman, että kunnolla huomioi sen laatua tai dokumentoi mitkä puuttuvat arvot on täydennetty. Käytetyt imputointitekniikat ja strategiat pitäisi aina dokumentoida, jotta muut osaavat arvioida miten paljon dataan voi luottaa. [9]

Puuttuvan datan kuvio ja puuttuvan datan mekanismi tarkoittavat hyvin erilaisia asioita, vaikka tutkijat saattavat käyttää niitä synonyymeina. Ensimmäinen viittaa tapaan, miten havaittu ja puuttuva data sijaitsee datasetissä ja toinen kuvailee mahdollisia suhteita havaitun datan ja puuttuvan datan todennäköisyyden välillä. Datan kuvio siis kuvailee vain millaisia ”koloja” datassa on ja ei selitä miksi data puuttuu. Huomionarvoista on, että datamekanismi ei myöskään selitä miksi vaan missä tilanteissa data puuttuu. Molemmat voivat kuitenkin myös auttaa analyytikkoa vastaamaan kysymykseen miksi data puuttuu. [22]

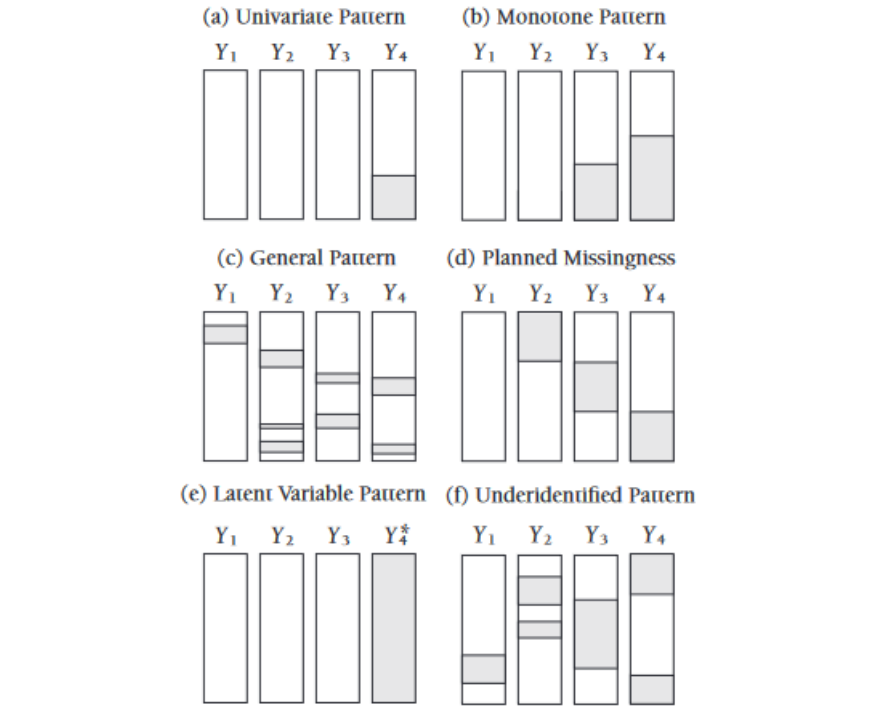


FIGURE 1.1. Six missing data patterns. The gray shaded areas of each bar represent missing observations.

Kuva 1. Lähde 22. Kuusi puuttuvan datan kuviota. Harmaaksi väritetyt alueet ovat puuttuvia arvoja.

Kuvassa 1.1 näkyy kuusi tyypillistä uupuvan datan kuviota, harmaat alueet osoittaen puuttuvien arvojen sijainnin datasetissä. A:ssa puuttuvat arvot ovat kohdistuneet vain yhteen muuttujaan, B:n tilanne syntyy yleensä pitkittäistutkimuksessa, jossa vastaajia lakkaa osallistumasta jossain vaiheessa ja eivät koskaan palaa takaisin. C on ehkä yleisin kuvio, jossa puuttuvat arvot sijaitsevat satunnaiselta vaikuttavalla tavalla datasetissä. Mielivaltainen kuvio on usein harhaanjohtava, sillä arvot voivat silti olla systemaattisesti uupuvia. Taulu D on esimerkki puuttuvan datan kuviosta, joka on suunniteltu. Vaikka se voi kuulostaa intuition vastaiselta niin joissakin tilanteissa on hyvä idea, että tutkija luo tarkoituksella epätäydellistä dataa. Suunniteltu datan puuttuvuus mahdollistaa esimerkiksi suuren määrän kysymyksiä kyselyssä niin, että vastaajien kokema kuormittavuus pysyy samalla hallinnassa. Kuviossa E kaikki data puuttuu yhdestä piilevästä muuttujasta. Nimensä mukaisesti muuttujan arvot ovat olleet aiemmin piilossa ja ne on saatu ja päätelty epäsuorasti matemaattisen mallin avulla, joka on hyödyntänyt muita suoraan havaittavissa olevia datasetin muuttujia. [1]

Kuvioiden analysoiminen ei ole yhtä tärkeätä kuin aiemmin, mutta edelleen siitä voi olla hyötyä. Menneisyydessä tutkijat kehittivät analyyttisiä tekniikoita, jotka oli tarkoitettu tietyille puuttuvan datan kuvioille. Käytännössä eri kuvioiden välillä tunnistaminen ei ole enää niin oleellista puuttuvuuden menetelmän valitsemisessa, sillä moni-imputointi ja suurimman uskovuuden menetelmä toimivat hyvin melkein millä tahansa kuviolla. Näiden puuttuvan datan kuvioiden ja yleisesti vastaavien kuvioiden visuaalinen analyysi on silti edelleen hyödyllistä, sillä niiden avulla voi tunnistaa eri muuttujien välisiä suhteita. Puuttuvuuden kuviot voivat auttaa herättämään tutkijassa kysymyksiä, jotka eivät muuten tulisi mieleen. Ymmärtäessämme paremmin dataa ja eri yhteyksiä voimme muun muassa yrittää tehdä tulevista kyselyistä parempilaatuisia siten, että vähemmän oleellista dataa tai muuttujia puuttuu. Tai kuvion D tapauksessa voimme puuttuvuuden kuvioita huomioimalla arvioida millä tavalla meidän kannattaa tarkoituksella poistaa dataa siten, että se luo mahdollisimman vähän harhaa ja tilastollisen voiman heikkenemistä. Analysoimalla kuvioita voimme myös helpottaa oikean datamekanismin selvittämistä. Esimerkiksi puuttuvuuden näyttäessä kuviolta C, eli täysin satunnaiselta, voimme olla tietyissä tilanteissa hieman varmempia MCAR oletuksen oikeellisuudesta. [1][22]

On olemassa kuitenkin tietynlaisia puuttuvuuden kuvioita, joiden kohdalla jopa modernit puuttuvan datan menetelmät voivat antaa huonoja ennustuksia. Esimerkkinä tästä on kuvio F, jossa data puuttuu kahdesta muuttujasta Y3 ja Y4 siten, että on hyvin vähän tai ei ollenkaan molempien muuttujien arvot sisältäviä rivejä. Tällöin olisi vaikeata tai jopa mahdotonta ennustaa kahden muuttujan välisiä suhteita. On siis tärkeätä seuloa tämän kuvion varalta aikaisessa vaiheessa analyysiä. [22]

4 Perinteiset menetelmät

4.1 Täydellisten havaintorivien analyysi

Täydellisten havaintorivien analyysi viittaa analyysiin, joka suoritetaan täydellisellä datalla ilman yhtään puuttuvia arvoja ja tämä saavutetaan usein poistomenetelmillä. Joidenkin lähteiden mielestä datan poistamiseen perustuvaa tekniikkaa ei kannata käyttää koskaan, vaikka sillä kuinka säästäisi aikaa ja resursseja helpon implementoinnin vuoksi. Useimpien mielestä sitä voi ja jopa kannattaa käyttää joskus, jos puuttuvan datan määrä on tarpeeksi pieni suhteessa kokonaisuuteen ja/tai puuttuvan datan mekanismi on MCAR [10]. Rivien poisto, jos rivistä puuttuu edes yksi arvo voi olla paras ratkaisu myös mekanismien MAR ja MNAR ollessa tosi tietyissä tilanteissa [22][21].

Koko sarakkeen poisto, eli muuttujan poisto kaikista havainnoista on harvoin hyvä idea. Kuitenkin jos olemme varmoja, ettei muuttuja ole hyödyllinen analyysillemme tai sarakkeen puuttuvan datan suhde on todella suuri niin sarakkeen poisto voi olla pienempi paha suhteessa aikaan ja vaivaan mitä sen sisältämän datan hyödyntämiseen menisi. Hyvä puoli tässä on se, ettei tämä luo harhaa dataamme mekanismista huolimatta, sillä muuttujan arvojen mahdolliset suhteet muihin muuttujiin poistuu myös. [10]

4.2 Yksittäisimputointi

Suurin osa perinteisistä yksittäisimputoinnin tai poistamisen menetelmistä tuottaa tarkkoja tuloksia vain silloin, jos puuttuva data on MCAR tyypiltään [1]. Käytännössä datan puuttuvuus on kuitenkin harvoin täysin satunnaista. Yksittäisimputointi johtaa usein harhan lisäksi ennustettujen parametrien keskivirheen aliarvioimiseen, sillä menetelmä ei ota huomioon puuttuvan datan epävarmuutta. Toisin sanoen vain yksi puuttuvuuden ennustus, vaikka se olisi miten hyvä tahansa, on todennäköisesti erilainen kuin todellinen arvo, jos siihen päästäisiin jotenkin käsiksi. [5]

Toinen alkeellinen tekniikka on keskiarvon imputointi, jossa yksinkertaisesti lasketaan kaikkien tietyssä sarakkeessa olevien arvojen keskiarvo ja täydennetään puuttuva data tällä. Tämä menetelmä johtaa lähes aina virheellisiin johtopäätöksiin harhan takia ja tilastollisen voiman heikkenemiseen. Seuraavassa osiossa käytetään pientä keinotekoista datasettiä ilmentämään millaista harhaa tämä ja muut alkeellisemmat menetelmät voivat aiheuttaa.

Toisaalta ML ja MI menetelmät tarjoavat ennustuksia, jotka ovat vapaita harhasta datan ollessa MCAR tai MAR ja täten paremmat mallien parametrien ennustukset. Nämä modernit tekniikat kohtaavat kuitenkin yleensä ongelmia datan ollessa MNAR, mutta useimmiten ne suoriutuvat paremmin kuin yksinkertaisemmat tekniikat. [5]

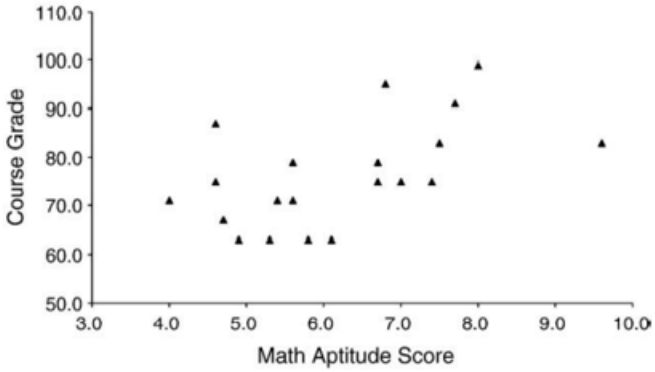
Table 1
Math performance data set.

Complete data		Observed data	Mean imputation	Regression imputation [a]	Stochastic regression imputation [a]	
Math aptitude	Course grade	Course grade	Course grade	Course grade	Random error	Course grade
4.0	71.00	–	81.80	65.26	7.16	72.42
4.6	87.00	–	81.80	68.22	0.73	68.95
4.6	74.00	–	81.80	68.22	12.01	80.23
4.7	67.00	–	81.80	68.71	−7.91	60.81
4.9	63.00	–	81.80	69.70	−4.07	65.63
5.3	63.00	–	81.80	71.68	27.41	99.09
5.4	71.00	–	81.80	72.17	25.76	97.93
5.6	71.00	–	81.80	73.16	2.76	75.92
5.6	79.00	–	81.80	73.16	−11.77	61.39
5.8	63.00	–	81.80	74.15	−0.56	73.59
6.1	63.00	63.00	63.00	63.00	–	63.00
6.7	75.00	75.00	75.00	75.00	–	75.00
6.7	79.00	79.00	79.00	79.00	–	79.00
6.8	95.00	95.00	95.00	95.00	–	95.00
7.0	75.00	75.00	75.00	75.00	–	75.00
7.4	75.00	75.00	75.00	75.00	–	75.00
7.5	83.00	83.00	83.00	83.00	–	83.00
7.7	91.00	91.00	91.00	91.00	–	91.00
8.0	99.00	99.00	99.00	99.00	–	99.00
9.6	83.00	83.00	83.00	83.00	–	83.00
Mean	76.35	81.80	81.80	76.12		78.70
Std. Dev.	10.73	10.84	7.46	9.67		12.36

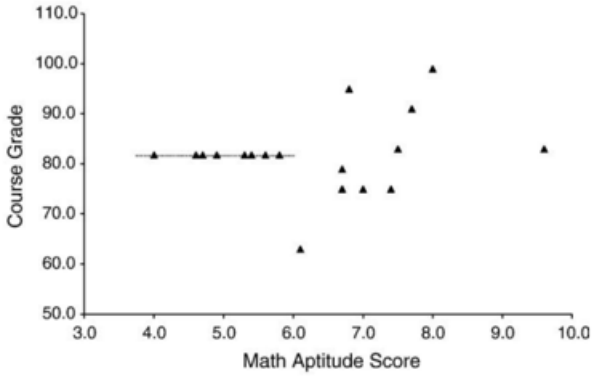
[a] Imputation regression equation: $\hat{Y} = 45.506 + 4.938(\text{Aptitude})$.

Kuva 2. Lähde 1, taulukosta 1 näkyy opiskelijoiden suoriutuminen kurssista ja matematiikan taitotestistä sekä eri imputointimenetelmien ennustukset.

Taulukossa 1 näkyy opiskelijoiden matematiikan taitotestin ja matematiikan kurssin kaikki todelliset arvot. Näiden rivien vieressä on vielä esitettynä ne kurssin arvosanat, jotka on keinotekoisesti poistettu eri menetelmien suorituskyvyn testaamista varten. Viimeiset sarakkeet taulukossa esittävät rivi riviltä eri imputointitekniikoiden puuttuvien matematiikan kurssin arvosanojen ennustukset. Näitä puuttuvia arvoja yritetään ennustaa keskiarvon, lineaarisen regression ja stokastisen lineaarisen regression imputoinneilla. Jälkimmäiset kaksi menetelmää käyttävät hyväkseen matematiikan taitotestin muuttujaa kurssin arvosanan puuttuvien arvojen ennustamisessa, toisin kuin keskiarvon imputointi, joka käyttää vain kurssin arvosanojen arvoja. Taulukossa on esitettynä myös lineaarisen regression yhtälö, jolla jokainen tämän menetelmän imputoitava arvo ennustettiin. Huomioi, että vain simuloimme tässä mahdollista datan ja puuttuvuuden tilannetta, joten pystymme poikkeuksellisesti täydellä varmuudella vertailemaan eri menetelmien suorituskykyä, sillä tiedämme jo oikeat arvot.

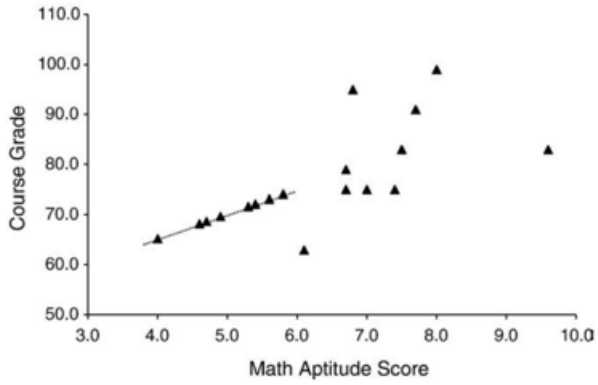


Kuva 3. Lähde 1, pistekuvio täydellisestä datasta, eli ennen kuin taulukon 1 kurssin arvosanoista poistettiin puolet. Vaakatasolla on x-akseli matematiikan taitotestin arvoille ja pystysuoralla y-akselilla on toisen matematiikan kurssin arvosanat, joista puolet aiotaan poistaa.



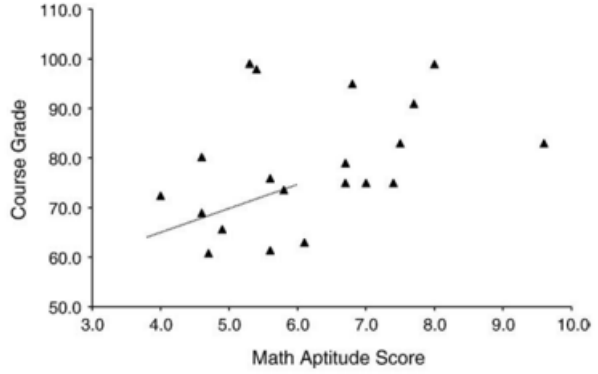
Kuva 4. Lähde 1, taulukon 1 dataan perustuva pistekuvio keskiarvolla imputoinnin ennustuksista

Kuten kuvasta 4. näkyy, ennustetuissa arvoissa ei ole lainkaan variaatiota vaan kaikki arvot ovat 81.80. Tämä imputoinnin arvo saatiin, kun laskettiin kaikkien kurssin arvosanojen havaittujen arvojen keskiarvo. Keskiarvolla imputointi loi harhaa, eli systemaattisesti loi ennusteita, jotka ovat liian pieniä tai suuria ottaen huomioon toisen muuttujan arvon, tässä tapauksessa saman opiskelijan matematiikan taitotestin arvon.



Kuva 5. Lähde 1, taulukon 1 dataan perustuva pistekuvio lineaarisen imputoinnin ennustuksista

Kuvasta 5 näkee miten tällä kertaa imputoidut arvot kuvaavat paremmin todellisia arvoja, sillä ne vaihtelevat välillä 65.26–74.15. Nämä arvot saatiin, kun puuttuvan datan arvot ennustettiin kurssin arvosanojen sekä matematiikan taitotestin muuttujan avulla. Lineaarisessa regressiossa malli, jota käytetään puuttuvien arvojen ennustamisessa, luodaan siten, että etsitään se suora tunnettujen pisteiden läpi, jonka summattu pisteiden absoluuttinen etäisyys suorasta on pienin mahdollinen. Tämä metodi ei kuitenkaan ota huomioon puuttuvien arvojen epävarmuutta ja siksi ennustetut arvot ovat epärealistisesti kaikki tismalleen mallin suoralla.



Kuva 6. Lähde 1, taulukon 1 dataan perustuva pistekuvio stokastisen lineaarisen imputoinnin ennustuksista

Kuvasta 6 näemme miten lisäämällä mallin funktioon satunnaista epävarmuutta muuttujalla e, saimme paljon paremmat ennustukset puuttuvuudelle, sillä kaikki ennustetut arvot eivät ole enää samassa linjassa suoralla ja näin ennustusten variaatio säilyy paremmin. Menetelmä ei luo harhaa puuttuvuuden ollessa MCAR tai MAR, mutta kärsii silti tilastollisen voiman heikkenemisestä MAR tapauksessa. Stokastinen regressio ei ole siten yleensä paras

mahdollinen tapa imputoida, sillä se ei ota huomioon yksittäisten imputointien epävarmuutta. Tässä tapauksessa se on kuitenkin parempi vaihtoehto kuin aiemmin esitetyt menetelmät [11].

5 Moni-imputointi

5.1 Moni-imputointi

Moni-imputointi on eräs modernimpi tekniikka puuttuvan datan täydentämiseen. Itse imputointi suoritetaan yleensä regressiota soveltaen, mutta menetelmän voi valita vapaasti. On erilaisia tapoja suorittaa moni-imputointi, mutta yhteistä kaikille tehokasta moni-imputointia soveltaville prosesseille on neljä vaihetta: ensin luo m imputointia puuttuville arvoille käyttäen jotain hyvää imputointimenetelmää, joka käyttää hyväkseen tietoa muista muuttujista ja sisällyttää satunnaisen komponentin. Tuloksena on m täydellistä datasettiä, jotka ovat kaikki hieman erilaisia imputoitujen arvojen suhteen satunnaisen komponentin vuoksi. Seuraavaksi analysoi jokainen valmis datasetti, jokaisen datasetin parametrien ennustukset poikkeavat toisistaan hieman, sillä ennustuksiin käytetty data on erilaista. Lopuksi yhdistä tulokset ja laske variaatio parametrien ennustuksissa. [11][7]

Moni-imputointi luo siis useita eri kopioita datasetistä, joissa jokaisessa on erilaiset imputoidut arvot. Analyysi suoritetaan jokaiselle datasetille käyttäen samoja menetelmiä kuin silloin, jos data olisi ollut täydellinen ja ilman puuttuvia arvoja. Jokaisen datasetin analysoiminen erikseen johtaa useisiin eri keskivirheisiin ja parametrien ennusteisiin. Nämä arvot yhdistetään lopuksi yhdeksi tulokseksi. Yllättäen vaadittu imputointien määrä m voi olla niin pieni kuin 5–10, vaikka se riippuu puuttuvan datan osuudesta. [11]

5.2 MICE-algoritmi

Multiple Imputation by Chained Equations (MICE) on yksi suosittu imputoinnin menetelmä. Moni-imputointi osuus algoritmista on se, kun luomme alussa monta kopiota alkuperäisestä datasta uupuvine arvoineen ja suoritamme MICE-algoritmin kaikki vaiheet erikseen jokaiselle kopiolle. Nämä vaiheet esitetään seuraavien taulukkojen avulla. On tärkeätä, että lisää lineaariseen regressioon satunnaisen komponentin. Tällä varmistetaan kaikkien datasettien erilaisuus. Suoritamme seuraavaksi moni-imputoinnin prosessin mukaisesti halutun analyysin jokaiselle datasetin kopiolle erikseen ja lopuksi yhdistämme näiden analyysien tulokset. MICE-algoritmia käytetään usein monimuuttuja-analyysissä, jossa käytetään useampia muuttujia hyväksi ennustaessa puuttuvia arvoja ja täten tulokset ovat yleensä paljon parempia. MICE olettaa datan olevan MAR, eli pystymme tekemään valaistuneen arvauksen puuttuvista arvoista käyttämällä muiden muuttujien arvoja hyväksi. [26][27]

ikä	kokemus	palkka
25	1	50
27	3	80
29	5	110
31	7	140
33	9	170
35	11	200

Kuva 7. Muuttujien nimet ovat ikä, kokemus ja palkka. Rivit ovat erilaisia havaintoja muuttujista. Taulukossa on todelliset arvot, joihin voimme verrata MICE-algoritmin tuloksia. Oranssilla on väritetty arvot, jotka tulevat myöhemmin keinotekoisesti puuttumaan ja joita yritetään ennustaa.

vaiheet	1			2			3			4			5		
	ikä	kokemus	palkka	ikä	kokemus	palkka	ikä	kokemus	palkka	ikä	kokemus	palkka	ikä	kokemus	palkka
	25	7	50	25	7	50	25	7	50	25		50	25	1.8538	50
	27	3	134	27	3	134	27	3	134	27	3	134	27	3	
	29	5	110	29	5	110	29	5	110	29	5	110	29	5	110
	31	7	140	31	7	140	31	7	140	31	7	140	31	7	140
	33	9	170	33	9	170	33	9	170	33	9	170	33	9	170
	29	11	200		11	200		11	200	36.2532	11	200	36.2532	11	200

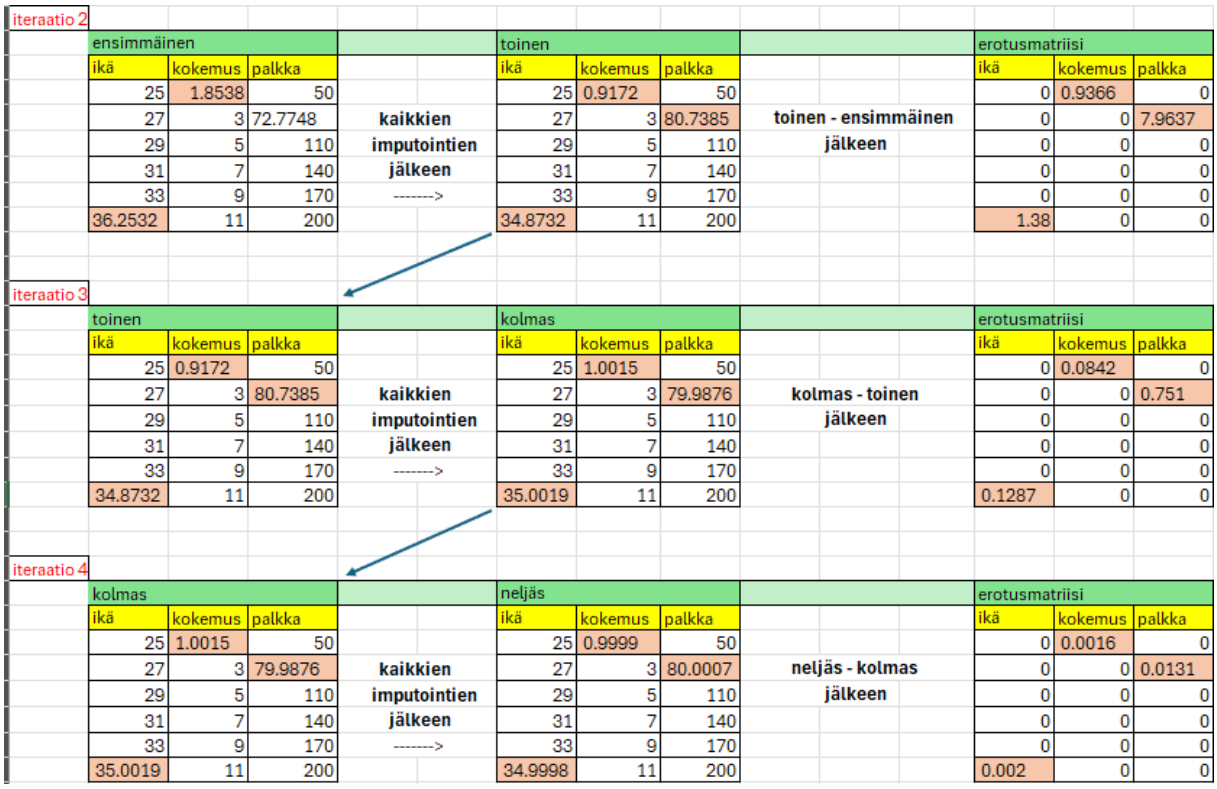
Kuva 8. MICE-algoritmin vaiheet iteraatiossa 1.

Seuraavaksi esitetään MICE-algoritmin eri vaiheet iteraatiossa yksi. Liikumme taulukossa vasemmalta oikealle keskiarvolla imputoituja arvoja (oranssilla väritetyt vaiheessa 1) ja poistamme ne vuorotellen väliaikaisesti, seuraavaksi käytämme kaikkia muita taulukon arvoja (harmaalla väritetty) paitsi samalla rivillä olevia ennustamaan uupuvan arvon, joka on kohdemuuttuja. Lineaarinen regressio antoi tässä puuttuvalle ikä arvolle ennustuksen 36.2532 ja hyödynnämme tätä kokemus sarakkeen uupuvan arvon ennustamiseen vaiheessa 4. Toistamme saman prosessin ennustaessamme puuttuvan palkka sarakkeen arvoa vaiheessa 5.

vaiheet	6				7				8		
	ikä	kokemus	palkka		ikä	kokemus	palkka		ikä	kokemus	palkka
	25	1.8538	50		25	7	50		0	-5.1462	0
	27	3	72.7748		27	3	134		0	0	-61.2252
	29	5	110	**miinus**	29	5	110	**=**	0	0	0
	31	7	140		31	7	140		0	0	0
	33	9	170		33	9	170		0	0	0
	36.2532	11	200		29	11	200		7.2532	0	0

Kuva 9. MICE-algoritmin vaiheet iteraatiossa 1.

Lopuksi, kun olemme imputoineet regressiolla kaikki puuttuvat arvot vaiheessa 6 niin laskemme näiden erotuksen alustavasta taulukostamme vaiheessa 7, jossa käytettiin yksinkertaista keskiarvolla imputointia. Vaiheessa 8 näemme iteraation 1 erotusmatriisin, joka antaa meille osviittaa siitä miten kaukana ennustetut arvot ovat todellisista arvoista.



iteraatio 2

ensimmäinen				toinen				erotusmatriisi		
ikä	kokemus	palkka		ikä	kokemus	palkka		ikä	kokemus	palkka
25	1.8538	50		25	0.9172	50		0	0.9366	0
27	3	72.7748	kaikkien	27	3	80.7385	toinen - ensimmäinen	0	0	7.9637
29	5	110	imputointien	29	5	110	jälkeen	0	0	0
31	7	140	jälkeen	31	7	140		0	0	0
33	9	170	------->	33	9	170		0	0	0
36.2532	11	200		34.8732	11	200		1.38	0	0

iteraatio 3

toinen				kolmas				erotusmatriisi		
ikä	kokemus	palkka		ikä	kokemus	palkka		ikä	kokemus	palkka
25	0.9172	50		25	1.0015	50		0	0.0842	0
27	3	80.7385	kaikkien	27	3	79.9876	kolmas - toinen	0	0	0.751
29	5	110	imputointien	29	5	110	jälkeen	0	0	0
31	7	140	jälkeen	31	7	140		0	0	0
33	9	170	------->	33	9	170		0	0	0
34.8732	11	200		35.0019	11	200		0.1287	0	0

iteraatio 4

kolmas				neljäs				erotusmatriisi		
ikä	kokemus	palkka		ikä	kokemus	palkka		ikä	kokemus	palkka
25	1.0015	50		25	0.9999	50		0	0.0016	0
27	3	79.9876	kaikkien	27	3	80.0007	neljäs - kolmas	0	0	0.0131
29	5	110	imputointien	29	5	110	jälkeen	0	0	0
31	7	140	jälkeen	31	7	140		0	0	0
33	9	170	------->	33	9	170		0	0	0
35.0019	11	200		34.9998	11	200		0.002	0	0

Kuva 10. MICE-algoritmin eri iteraatioiden tulokset.

Seuraavissa iteraatioissa käytetään alustavana taulukkona edellisen iteraation kaikkien imputointien lopputulosta ja tätä havainnollistetaan kuvassa 10 sinisillä nuolilla. Kuvan 10 ensimmäinen taulukko on siis sama kuin iteraation 1 vaiheen 6 taulukko. Saman kuvan toinen taulukko kuvastaa lopputulosta, kun vaiheet 1–8 on toistettu uudestaan siten, että alustavana taulukkona vaiheessa 1 oli ensimmäinen taulukko. Toistetaan vaiheet 1–8 niin monta iteraatiota kuin on tarpeen, kunnes erotusmatriisin arvot ovat tarpeeksi pieniä ja lähellä nollaa tai suoritit ennalta määrätyn määrän iteraatioita. [15]

Huomataan, että jo iteraation 4 jälkeen ennustetut arvot ovat hyvin lähellä todellisia kuvassa 7 esitettyjä arvoja. MICE-algoritmi ei yleensä onnistu luomaan näin hyviä ja tarkkoja ennustuksia, vaan data on luotu siten, että se parhaiten havainnollistaa algoritmin toimintaa. Lopussa ennustetut arvot ovat paljon parempia verrattuna uupuvien arvojen alustaviin keskiarvolla imputoituihin arvoihin. Monet tutkimukset ovat osoittaneet muun muassa erilaisilla simuloiduilla dataseteillä, miten moni-imputointi peittoaa myös stokastista regressiota soveltavan yksittäisimputoinnin [22]. Yleisesti ottaen moni-imputointi on siten suositeltavaa pelkän yksittäisimputoinnin sijasta ja MICE-algoritmi on hyvä valinta moni-imputoinnin toteuttamiseen. MICE käytti tässä esimerkissä lineaarista regressiota, mutta se

voisi käyttää mitä vaan muuta mallia puuttuvien arvojen ennustamiseen. Datasettien ollessa todella suuret regressiomenetelmät kuten MICE voivat vaatia hyvin paljon prosessointitehoa, joten tällöin voi olla aiheellista harkita vaihtoehtoista menetelmää.

6 Yhteenveto

Monet tutkijat ovat soveltaneet ja edelleen soveltavat vanhentuneita tai puutteellisia strategioita ja tekniikoita puuttuvan datan ongelman korjaamiseen. Yksittäisimputointia soveltavista menetelmistä suurin osa tarjoaa yleensä tai aina epäoptimaalisia tuloksia mekanismista huolimatta, mutta täydellisten havaintorivien analyysi, jossa poistamme puuttuvuutta sisältävät rivit, voi edelleen olla paras tai hyvä vaihtoehto. Nämä menetelmät vaativat tiukat oletukset puuttuvan datan luonteesta ja mekanismista, joten ne toimivat vain tietyissä tilanteissa. Toiset menetelmät eivät juuri koskaan toimi hyvin, kuten keskiarvon imputointi ja yksittäisimputointi regressiolla ilman stokastista termiä [22]. Näiden tekniikoiden käyttäminen voi johtaa virheelliseen keskivirheeseen, harhaan tai molempiin. Yksittäisimputointi johtaa usein huonompiin tuloksiin, erityisesti mekanismin ollessa jotain muuta kuin MCAR ja tällöinkin menetelmä voi antaa epäoptimaalisia tuloksia, sillä vaikka harhaa ei esiinny niin tilastollista voimaa voi olla vähemmän. Kehittyneemmät menetelmät, kuten moni-imputointi MICE-algoritmilla ja suurin uskottavuus ovat taas menestyksekkäästi käytössä monissa erilaisissa käytännön sovelluksissa, sillä MAR mekanismi on usein hyvä ja realistinen oletus, ja tekniikat toimivat hyvin myös MCAR oletuksella. Nämä menetelmät ovat nykyään vielä helposti kaikkien käytettävissä parempien ohjelmien ansioista. Erityisesti R ja Python ohjelmointikielillä löytyy monia laadukkaita paketteja, joilla voi toteuttaa kaikki tutkielmassa esitetyt imputointimenetelmät. Monet standardit tilastolliset ja koneoppimisen menetelmät eivät silti suoraan toimi puuttuvien arvojen kanssa, joten hieman vaivaa pitää nähdä puuttuvien arvojen oikaisemiseksi [5]. On tärkeätä harkita tarkkaan millaista menetelmää käyttää puuttuvuuden käsittelyyn ja tähän on olemassa monenlaisia strategioita kuten puuttuvuuden määrän ja kuvion huomioiminen, puuttuvuuden syntytapojen ja syiden arvioiminen, imputoinnin tai poiston tavoitteiden määrittäminen ja herkkyysanalyysi. Vaikka jotkut menetelmät puuttuvan datan käsittelyyn ovat yleensä tai aina selkeästi parempia kuin toiset kuten moni-imputointi regressiolla verrattuna esitettyihin perinteisiin menetelmiin niin mitään menetelmää ei voi oikein kuvailla erinomaiseksi. Jopa kehittyneimmät tekniikat pystyvät vain valistuneisiin arvauksiin ja ainoa todella hyvä ratkaisu on välttää kaikkea puuttuvaa dataa. Tämän ollessa käytännössä yleensä mahdotonta niin kannattaa pyrkiä puutteellisen datan vaikutuksen minimoimiseen. Esimerkiksi kyselylomakkeen tai datankeräysmenetelmän paremmalla suunnittelulla voi pyrkiä vähentämään puuttuvuuden syntyä ja jäljelle jäävien uupuvien arvojen kohdalla voi sitten soveltaa erilaisia imputointi- tai poistomenetelmiä.

Lähdeluettelo

[1] Baraldi, Amanda N., and Craig K. Enders, "An Introduction to Modern Missing Data Analyses.", *Journal of school psychology*, vol. 48.1, s. 5–37, 2010.

[2] Allison, Paul D, "Missing Data", *Thousand Oaks: SAGE Publications*, vol. 136, 2001.

[3] Rashid, Wajeeha, and Manoj Kumar Gupta, "A Perspective of Missing Value Imputation Approaches.", *Advances in Computational Intelligence and Communication Technology*, vol. 1086, s. 307–315, 2020.

[4] Lin, Wei-Chao, and Chih-Fong Tsai, "Missing Value Imputation: a Review and Analysis of the Literature (2006–2017).", *The Artificial intelligence review,* vol. 53.2, s. 1487–1509, 2020.

[5] WU, Lang, and Jin Qiu, "Applied Multivariate Statistical Analysis and Related Topics with R.", *EDP Sciences*, 2021.

[6] Petrazzini, Ben Omega, "Evaluation of Different Approaches for Missing Data Imputation on Features Associated to Genomic Data.", *BioData mining*, vol. 14.1, s. 1–44, 2021.

[7] Newman, Daniel A., "Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques.", *Organizational research methods,* vol. 6.3, s. 328–362, 2003.

[8] Roth, P. L., "Missing data: A conceptual review for applied psychologists.", *Personnel Psychology*, vol. 47, s. 537-560, 1994.

[9] Laaksonen, Seppo, "Survey Methodology and Missing Data: Tools and Techniques for Practitioners.", *Cham: Springer International Publishing AG*, 2018.

[10] Rachael A. Hughes, Jon Heron, Jonathan A. C. Sterne, Kate Tilling, "Accounting for missing data in statistical analyses: multiple imputation is not always the answer", *International Journal of Epidemiology*, Vol. 48, s. 1294–1304, 2019

[11] Li, Peng, Elizabeth A. Stuart, and David B. Allison, "Multiple Imputation: A Flexible Tool for Handling Missing Data." *JAMA: the journal of the American Medical Association,* vol. 314.18, s. 1966–1967, 2015.

[12] Khan, Faizan U. F., Kashan U. Z. Khan, and S. K. Singh, “Is Group Means Imputation Any Better Than Mean Imputation: A Study Using C5.0 Classifier.”, *Journal of Physics: Conference Series,* vol. 1060.1, 2018

[13] Waal, Ton de., Jeroen Pannekoek, and Sander Scholtus, “Handbook of Statistical Data Editing and Imputation.”, *Hoboken, N.J: Wiley*, 2011.

[14] Žitnik, Slavko and Štrumbelj, Erik, “Introduction to data science”, url: https://fri-datascience.github.io/course_ids/handbook/missing-data.html#introducing-bias ,“Käyty sivulla viimeksi 25.10.23”.

[15] W. Heymans, Martijn and Eekhout Iiris, ”Applied Missing Data Analysis With SPSS and (R)Studio”, url: https://bookdown.org/mwheymans/bookmi/multiple-imputation.html “Käyty sivulla viimeksi 25.10.23”.

[16] Little, Roderick J., James R. Carpenter, and Katherine J. Lee. “A Comparison of Three Popular Methods for Handling Missing Data: Complete-Case Analysis, Inverse Probability Weighting, and Multiple Imputation.”, *Sociological methods & research*, 2022.

[17] Rubin, Donald B., “Inference and Missing Data.”, *Biometrika*, vol. 63.3, s. 581–592, 1976.

[18] Raghunathan, Trivellore E., “What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data.”, *Annual review of public health,* vol. 25.1, s. 99–117, 2004.

[19] Little, Roderick J. A., and Donald B. Rubin, “Statistical Analysis with Missing Data.”, *Newy York: John Wiley & Sons*, 2002.

[20] Eekhout, Iiris, ”Missing data mechanisms”, url: https://www.iriseekhout.com/post/2022-06-28-missingdatamechanisms/ “Käyty sivulla viimeksi 25.10.23”.

[21] Pepinsky, Thomas B., “A Note on Listwise Deletion Versus Multiple Imputation.”, *Political analysis,* vol. 26.4, s. 480–488, 2018.

[22] Enders, Craig K., “Applied Missing Data Analysis.”, *New York: The Guilford Press*, 2022.

[23] Hughes, Rachael A., “Accounting for Missing Data in Statistical Analyses: Multiple Imputation Is Not Always the Answer.”, *International journal of epidemiology,* vol. 48.4, s. 1294–1304, 2019.

[24] Madley-Dowd, Paul, “The Proportion of Missing Data Should Not Be Used to Guide Decisions on Multiple Imputation.”, *Journal of clinical epidemiology,* vol. 110, s. 63–73, 2019.

[25] Jakobsen, Janus Christian, “When and How Should Multiple Imputation Be Used for Handling Missing Data in Randomised Clinical Trials - a Practical Guide with Flowcharts.”, *BMC medical research methodology*, vol. 17.1, s. 162–162, 2017.

[26] Golkhatmi, Nafiseh Seyyed Nezhad, and Mahboobeh Farzandi, “Enhancing Rainfall Data Consistency and Completeness: A Spatiotemporal Quality Control Approach and Missing Data Reconstruction Using MICE on Large Precipitation Datasets.”, *Water resources management,* vol. 38.3, s. 815–833, 2024.

[27] Van Buuren, Stef, and Karin Groothuis-Oudshoorn, “Mice: Multivariate Imputation by Chained Equations in R.”, *Journal of statistical software,* vol. 45.3, s. 1–67, 2011.

Liitteet

Liite 1. Käännökset ja lyhennykset

Statistical power = tilastollinen voima

Bias = harha

Maximum likelihood = suurimman uskottavuuden menetelmä (ML)

Multiple imputation = moni-imputointi (MI)

Standard error = keskivirhe

Single imputation = yksittäisimputointi (SI)

Missing value imputation = puuttuvan arvon imputointi (MVI)

Structurally missing data = rakenteellisesti puuttuva data (SMD)

Complete-case analysis = täydellisten havaintorivien analyysi (CCA)

Multivariate analysis = monimuuttuja-analyysi

Bivariate = Kahden muuttujan

Univariate = Yhden muuttujan

Auxiliary variable = apumuuttuja

Latent variable = piilevä muuttuja