EPÄTÄYDELLISEN DATAN KÄSITTELY JA SEN VAIKUTUS ANALYYSIIN

Lappeenrannan-Lahden teknillinen yliopisto LUT

Laskennallisen tekniikan kandidaatintyöm

2022

Leevi Kämäräinen

Tarkastaja: dos. TkT Matylda Jablonska-Sabuka

TIIVISTELMÄ

Lappeenrannan-Lahden teknillinen yliopisto LUT
School of Engineering Science
Laskennallinen tekniikka

Leevi Kämäräinen

Epätäydellisen datan käsittely ja sen vaikutus analyysiin

Diplomityö
2022
27 sivua, 16 kuvaa, 1 taulukko
Tarkastaja: Matylda Jablonska-Sabuka

Hakusanat: epätäydellinen data, puuttuva data, poikkeama, väestönlaskenta, lähimmän naapurin menetelmä

Epätäydellistä dataa löytyy miltei kaikesta datasta. Se, kuinka paljon epätäydellisyyttä on datassa ja mitä tyyppiä se on, vaikuttaa paljon siihen, kuinka dataa voidaan hyödyntää. Tässä kandidaatintyössä tutkitaan eri epätäydellisyyden tyyppejä sekä käsittelytekniikoita MATLAB-ohjelmiston avulla. Työ toimii oppaana aiheen teoriaan ja pyrkii esimerkkien kautta tuomaan konkreettisesti esille sen, mitä on epätäydellinen data.

Työn tavoitteena oli tutkia yksinkertaisen lähimmän naapurin menetelmän sekä lineaariregression hyödyntämistä epätäydellisen datan korjaamiseen. Erityisesti tutkittiin epätäydellisellä datalla ennusteiden tekemistä ja sitä, kuinka paljon epätäydellistä dataa voi olla. Työn data-aineistona toimii WorldBank:in tarjoama väestönlaskentadata maailman väestöstä. Data-aineistosta keskityttiin tutkimaan eri maiden väkilukuja. Tuloksena saatiin korjattua epätäydellisyyttä siten, että sekä korjattu data että sillä tehdyt ennusteet mukailivat kohtuullisesti todellisen datan tilannetta. Vaikka data-aineistot olivat kooltaan pieniä, väestönlaskentadatan korjaaminen työn menetelmillä oli mahdollista ja erot todellisen dataan olivat tyydyttävällä tasolla.

Symboli- ja lyhenneluettelo

AR	Autoregressiivinen
ARMA	Autoregressiivinen liukuvan keskiarvon malli
MA	Liukuva keskiarvo
MAR	Satunnainen puuttuvuus
MCAR	Täysin satunnainen puuttuvuus
MNAR	Ei-satunnainen puuttuvuus
OLS	Pienimmän neliösumman menetelmä
R^2	Selitysaste

Sisällysluettelo

1 Johdanto **5**
1.1 Epätäydellinen data 5
1.2 Tutkimusongelma, tavoitteet ja rajaus 5
1.3 Työn toteutus 6
1.4 Työn rakenne 6

2 Epätäydellisen datan käsittely **7**
2.1 Puuttuvat arvot 7
2.2 Poikkeamat 9
2.3 Aikasarja-analyysi ja ARMA-malli 11
2.4 Korjausmenetelmät 12
2.5 Lineaariregressio 13

3 Aineisto ja ohjelmistot **15**

4 Epätäydellisyyden korjaaminen **16**
4.1 Aikasarja-analyysi 16
4.2 Puuttuvan datan määrä 18
4.3 Regressio mallintaminen 19
4.4 Poikkeamat 21

5 Keskustelu **23**

6 Yhteenveto **26**

LÄHTEET **27**

Taulukot **28**

Kuvat **29**

1 Johdanto

Tässä kandidaatintyössä tutkitaan epätäydellistä väestönlaskentadataa ja kuinka sitä voidaan hyödyntää analyysissä. Datan laatu on tärkeää, kun halutaan tehdä tilastollista analyysiä. Epätäydellisen datan käyttäminen analyysissä voi huonoimmassa tapauksessa antaa hyvinkin erilaisia tuloksia, mitä todellinen data antaisi. Tämä voi johtaa virheellisiin johtopäätöksiin, jonka takia analyysin teettäjälle voi koitua mittavia tappioita.

1.1 Epätäydellinen data

Epätäydellinen data on ongelma, joka joudutaan ottamaan huomioon, kun tehdään dataanalyysiä. Kun datasta löytyy puuttuvia arvoja, vastauskatoa tai muita poikkeamia, niin luokitellaan data epätäydelliseksi. Ongelmia tuottaa se, että monessa tilastollisen analyysin menetelmässä oletetaan, että käytetty data on virheetöntä. (Allison 2002.)

Puuttuvia arvoja syntyy jo datan keräysvaiheessa. Mittausvirheet, laitteistoviat sekä virheellisesti suunnitellut mittaukset ovat esimerkkitilanteita, joissa seurauksena voi olla puuttuvaa dataa. Yleinen syy puuttuvuudelle määrällisessä tutkimuksessa on se, että tutkimukseen osallistuvat jättävät syystä tai toisesta vastaamatta kyselyihin. Epätäydellistä dataa löytyy kuitenkin kaikesta soveltavan tutkimuksen alueista, jonka takia on hyvin tärkeää, että sitä osataan myös käsitellä. (Faisal 2018.)

1.2 Tutkimusongelma, tavoitteet ja rajaus

Tässä työssä tutkitaan, mitä on epätäydellinen data ja kuinka sitä voidaan käsitellä eri menetelmillä. Tavoitteena on vertailla, kuinka todellisella datalla ja korjatulla datalla tehdyt ennusteet eroavat toisistaan, sekä kuinka suuria määriä epätäydellistä dataa voi hyödyntää. Datassa keskitytään käsittelemään eri maiden populaatioiden datoja sekä pyritään löytämään populaatioihin vaikuttavia muita tekijöitä. Työssä hyödynnetään kahta eri menetelmää korjaamaan epätäydellistä dataa, mutta pääpainona on tutkia, kuinka korjattua dataa voidaan hyödyntää analyysissä, ja sitä kuinka paljon korjattu data eroaa todellisesta.

1.3 Työn toteutus

Työssä epätäydellisyys datassa luodaan muokkaamalla data-aineistoa MATLAB-ohjelmistolla. Tämä tapahtuu poistamalla satunnaisesti arvoja tai luomalla satunnaisia poikkeamia dataan, kuitenkin siten, että samaan aikaan säilytetään myös alkuperäinen data. Epätäydellisyys korjataan eri korjausmenetelmiä hyödyntäen, ja saatua korjattua dataa verrataan alkuperäiseen dataan. Korjatulla datalla sekä alkuperäisellä datalla tehdään ennusteita, ja niiden tuloksia verrataan keskenään.

1.4 Työn rakenne

Työn alussa kerrotaan lyhyesti tutkimusongelman taustasta ja työn tavoitteista. Luvussa 2 käsitellään epätäydellistä dataa, työssä käytettäviä korjausmenetelmiä ja aikasarjan sekä työssä käytettävän ARMA mallin teoriaa. Luvussa 3 esitellään lyhyesti työn aineisto sekä käytettävä ohjelmisto. Luvussa 4 esitellään työn tuloksia ja luvussa 5 analysoidaan niitä sekä pohditaan jatkotutkimuksia. Luvussa 6 on yhteenveto sekä tehdyt johtopäätökset.

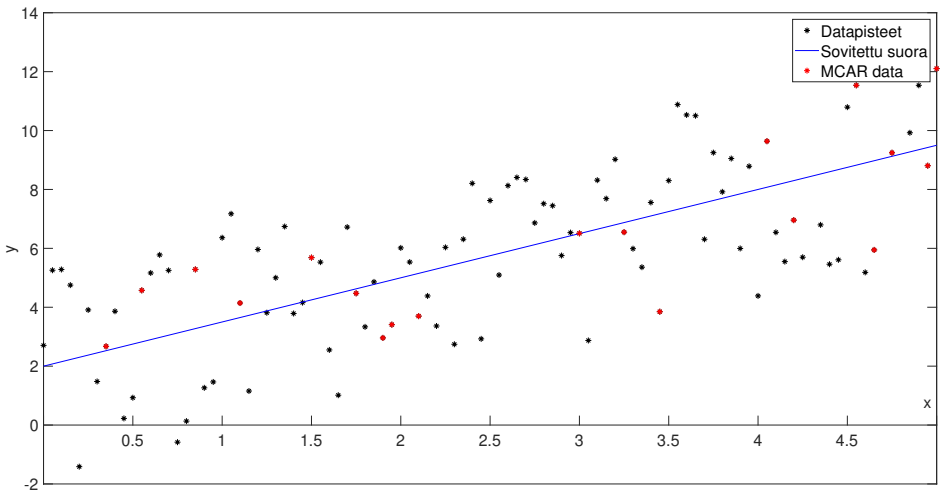
2 Epätäydellisen datan käsittely

2.1 Puuttuvat arvot

On hyvin yleistä, että data jota on saatavilla, ei ole kokonaista, vaan siitä usein puuttuu dataa jostain syystä. Se, kuinka paljon ja kuinka usein dataa puuttuu, riippuu paljon tieteenalasta, johon data-aineisto liittyy. Tieteenaloilla, missä datan kerääminen perustuu ihmisille tuotettuihin kyselyihin, kuten psykologian ja sosiaalisten tieteiden alat, on havaittavissa paljon puuttuvaa dataa. (Allison 2002.) Väestötutkimuksille yleinen syy puuttuvuudelle on vastauskato, eli tutkimukseen osallistuvat henkilöt jättävät syystä tai toisesta vastaamatta osaan tai kaikkiin kysymyksiin (Vriens et al. 2006). Tyypillisesti puuttuva data luokitellaan kolmeen eri kategoriaan: MCAR, MAR sekä MNAR tyyppeihin (Little et al. 2019).

- **MCAR - Täysin satunnainen puuttuvuus**

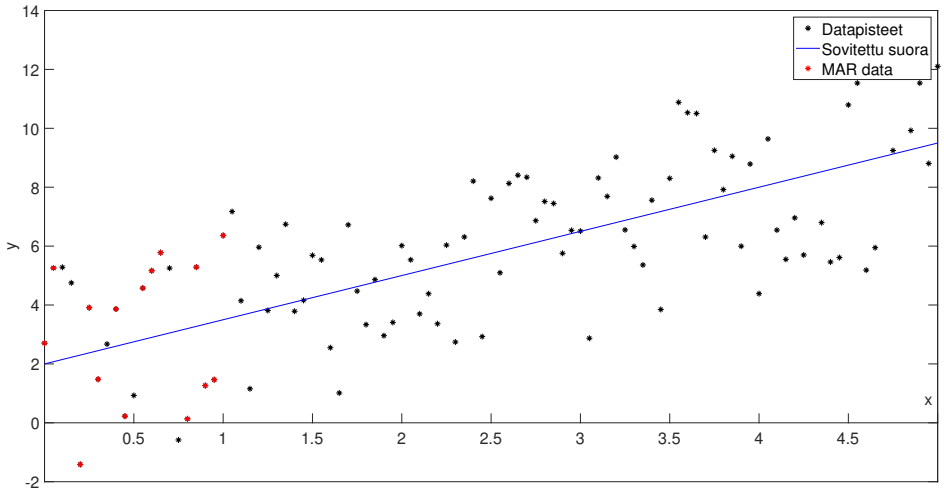
  Data on tyyppiä MCAR (*Missing completely at random*), kun todennäköisyys puuttuvalle datalle muuttujassa X ei ole kytkeytynyt muihin mitattuihin muuttujiin tai itse X:n arvoihin. Puuttuvuus on epäsystemaattista ja yksittäinen puuttuva arvo voidaan ajatella olevan täysin satunnaisesti tapahtunut. (Allison 2002.) Kuvassa 1 nähdään havainnollistettuna MCAR dataa. Kuvasta puuttuu atapisteitä satunnaisesti, eikä niissä ole havaittavissa mallia, jonka mukaan data puuttuisi.



Kuva 1. MCAR dataa. Kuvassa ei havaittavissa syytä datan puuttumiselle.

- **MAR- Satunnainen puuttuvuus**

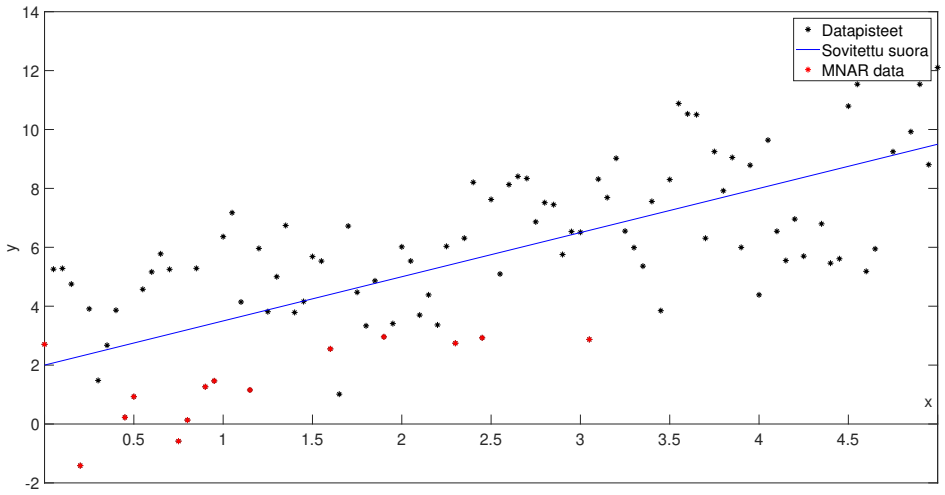
  Data luokitellaan tyypiksi MAR (*Missing at random*), jos puuttuvuus on riippuvainen muihin analyysin muuttujiin, mutta sillä ei ole riippuvuutta itse puuttuvien arvojen kanssa. Puuttuvuus siis ei ole nimensä mukaisesti satunnaista, vaan kyseessä on systemaattista puuttuvuutta, missä puuttuva data on korreloitunutta muiden tutkimuksen muuttujien kanssa. (Allison 2002). Kuvassa 2 havaitaan, että x:n arvoa 1 edeltävissä arvoissa havaitaan puuttuvuutta enemmän kuin muualla. Puuttuvuus kuvan tapauksessa voi esimerkiksi johtua siitä, että nuoremmat ihmiset eli alhaisemmat x:n arvot, jättävät vastaamatta kyselyihin useammin kuin muut ihmiset.



Kuva 2. MAR puuttuvuus. X-akselin arvot jotka ovat pienempiä kuin 1, puuttuvat todennäköisemmin.

- **MNAR - Ei-satunnainen puuttuvuus**

  MNAR (*Missing not at random*) datassa todennäköisyys puuttuvalle datalle on systemaattisesti riippuvainen puuttuvista arvoista. Käytännössä tämä tarkoittaa sitä, että syy datan arvon puuttumiselle liittyy jollain tavalla kyseisen datan arvoon. Kuvassa 3 havaitaan että y:n arvoa 3 pienemmissä arvoissa havaitaan puuttuvuutta enemmän kuin muualla datassa. Jos esimerkiksi y-akselilla on rahan käyttö harrastuksiin, ja x-akselilla on jokin sitä selittävä muuttuja, niin puuttuvuus voi johtua siitä, että ihmiset joilla on alhaisemmat tulot, jättävät vastaamatta kyselyihin.



Kuva 3. MNAR puuttuvuus. Y-akselin arvot, jotka ovat pienempiä kuin 3, puuttuvat todennäköisemmin.

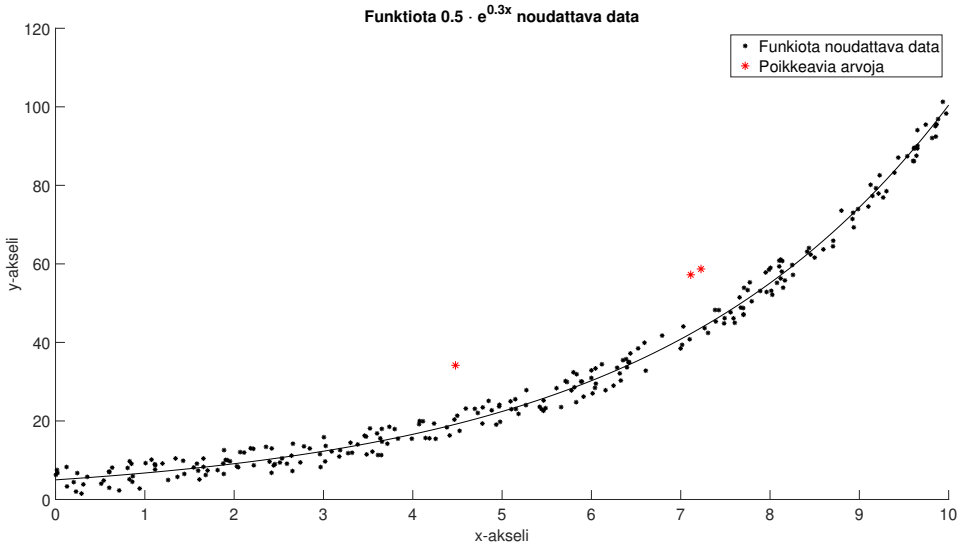
2.2 Poikkeamat

Poikkeama (*outlier*) on datapiste, joka poikkeaa merkittävästi muusta datasta. Analyysissä datalle on olemassa malli, jota pidetään normaalina. Tällöin poikkeama on datapiste, jonka epäillään poikkeavan tästä mallista. (Aggarwal 2017.)

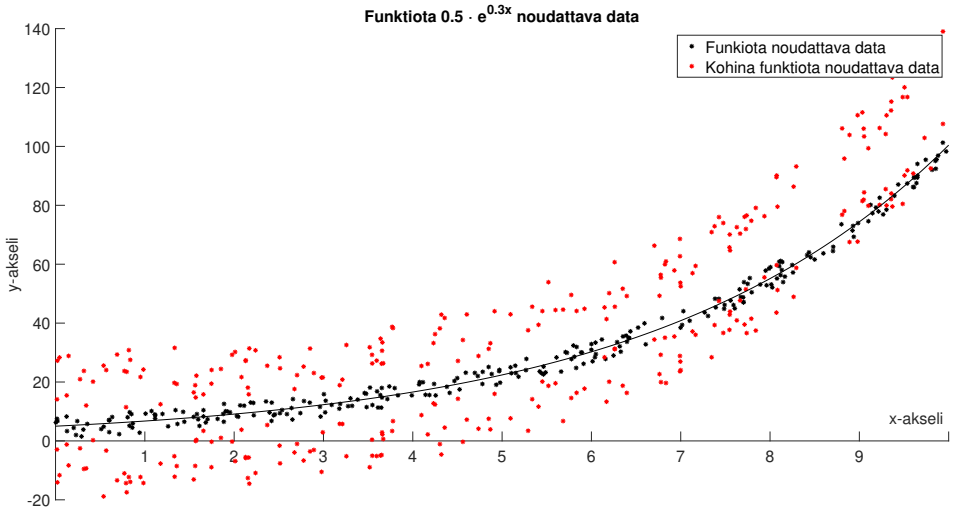
Poikkeamat usein luokitellaan erikseen heikkoihin ja vahvoihin poikkeamiin. Heikko poikkeama voi olla satunnaista kohinaa, jota dataan on kertynyt. Vahvat poikkeamat ovat taas yksittäisiä poikkeavia arvoja, jotka yleensä poikkeavat suhteessa enemmän normaalista mallista, kuin satunnainen kohina poikkeaa. (Aggarwal 2017.) Kuvassa 4 on tapauksena datapisteitä, mitkä luokitellaan vahvaksi poikkeamaksi, ja kuvasssa 5 on funktion ympärillä kohinaista dataa, mikä luokitellaan heikoksi poikkeamaksi.

Poikkeamien havaitsemiseen on monia eri keinoja. Tässä työssä esitellään teorian kannalta kaksi menetelmää niiden havaitsemiseen, mutta niiden korjaamisen tutkimisessa on tehty oletus, että tiedämme valmiiksi, mitkä arvot ovat poikkeamia.

Yksi menetelmä poikkeamien havaitsemiseen yksiuloitteisessa datassa on hyödyntää tilastollista Z-testiä. Z-testin tarkoitus poikkeamien havaitsemisessa on verrata, kuinka paljon yksittäisen datapisteen arvo eroaa datajoukon keskihajonnasta ja keskiarvosta. Z-testi



Kuva 4. Esimerkkikuva vahvoista poikkeamista.



Kuva 5. Esimerkkikuva kohinasta eli heikosta poikkeamasta.

saa muodon

$$Z_i = \frac{|X_i - \mu|}{\sigma},$$

missä Z_i on Z-testin arvo yksittäiselle datapisteelle X_i, μ on datajoukon keskiarvo ja σ on datajoukon keskihajonta. Yleinen sääntö poikkeamisen havaitsemiselle Z-testillä on se, että kun $Z_i \leq 3$, niin voidaan epäillä datapisteen olevan poikkeama. Z-testillä on oletuksena, että datajoukko on normaalijakautunut, ja että sen keskiarvo sekä keskihajonta voidaan arvioida tarkasti tai on muuta kautta tiedossa. Tämän takia usein Z-testi ei välttämättä sovi pienelle data-aineistolle, ja sen antamia tuloksia pitää tulkita tarkemmin. (Aggarwal 2017.)

Toinen menetelmä poikkeamien havaitsemiseen, jota voidaan hyödyntää erityisesti aikasarjoissa, on hyödyntää ikkunapohjaista analyysiä (Aggarwal 2017). Tämä tarkoittaa, että aikasarjadata jaetaan N määrään ikkunoita, jotka ovat kokoa W. Jokaiselle ikkunalle W lasketaan ikkunan sisäinen keskiarvo μ ja keskihajonta σ. Tällöin jos y_i on data-arvo ajanhetkellä i ja W_i on data-arvon ikkunan koko, niin voimme etsiä arvoja y_i jotka ovat eroavat yli kaksi kertaa ikkunan keskihajonnan σ_W verran ikkunan keskiarvosta μ_W. Toisin sanoen tarkastelemme onko arvo y_i välillä $\mu + 2\sigma$ ja $\mu - 2\sigma$. (Jablonska 2008.)

2.3 Aikasarja-analyysi ja ARMA-malli

Aikasarja on joukko tasaisesti jakautuneita havaintoja joltain aikaväliltä (Pickup 2015). Aikasarja-analyysissa usein tutkitaan, kuinka jokin ilmiö muuttuu ajan kuluessa. Tämän avulla analyytikot pystyvät tekemään ennustuksia tulevaisuuteen sekä arvioita siitä, kuinka ilmiö muuttuu, jos sen selittävät muuttujat muuttuvat. Perinteisiä aikasarja analyysin kohteita on sääilmiöiden ennustaminen ja ekonometria.

Aikasarjadata jaetaan usein neljään komponenttiin: trendi, syklinen, kausittainen ja epäsäännöllinen. Trendi kuvaa sitä, kuinka ajan kuluessa ilmiö joko kasvaa tai laskee arvoiltaan. Syklisyys ja kausittaisuus kuvaavat molemmat sitä, kuinka ilmiön datassa on selkeää vaihtelua. Erona on se että kausittainen komponentti on säännöllisesti toistuva, kuten esimerkiksi talvikausi, ja syklisyys taas on epäsäännöllistä vaihtelua. Epäsäännöllinen komponentti on se osa, mikä jää jäljelle dataan, kun siitä poistetaan kausittaisuus ja trendi. Tämä epäsäännöllinen komponentti kuvaa sitä kaikkea satunnaisuutta ja tiheitä muutoksia aikasarjassa. (Kendall 1973.)

ARMA malli on yksi monista tavoista selittää ja analysoida aikasarjadataa. Mallin nimi ARMA tulee siitä, että se koostuu kahdesta eri mallista: AR-malli eli autoregressiivinen malli sekä MA-malli eli liukuvan keskiarvon malli. ARMA on klassinen Box-Jenkins aikasarja-analyysin malli. (Box et al. 1994). Käytännössä mallin tarkoituksena on analysoida datasta löytyviä riippuvuuksia nykyisten ja historian arvojen välille.

ARMA mallissa yhdistyy siis AR- ja MA-mallit, jossa autoregressiivinen malli AR(r) kertoo nykyisten havaintojen yhteyden sen aikaisempiin arvoihin, ja liukuvan keskiarvon malli MA(m) kertoo nykyisten arvojen yhteyden datan historialliseen kohinaan. ARMA(r,m) malli saa siis muodon:

$$x_t = C + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots \phi_n x_{t-r} + \varphi_1 u_{t-2} + \varphi_2 u_{t-2} + \cdots + \varphi_n u_{t-m} + \epsilon_t,$$

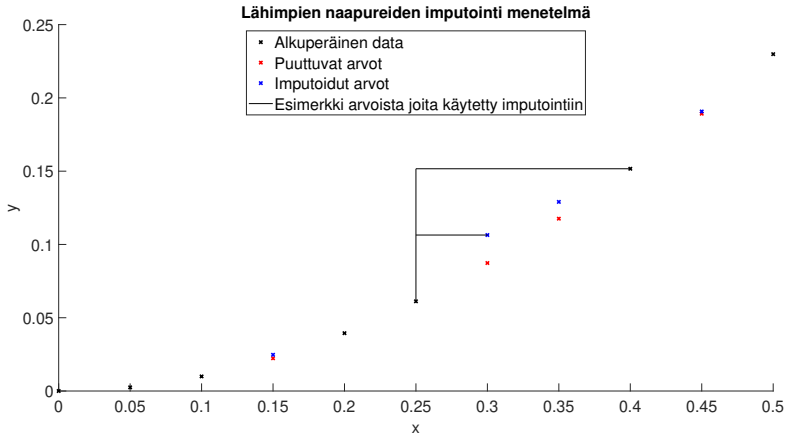
missä C on mallin vakio-osa, ϕ_n on autoregressiivisen mallin parametrejä, φ_n on liukuvan keskiarvon mallin parametrejä, ja ϵ_t on mallin kohinatermi. (Box et al. 1994).

ARMA mallissa on oletuksena, että mallin virhetermit (*residuaali*) ovat valkoista kohinaa eli normaalijakautuneita satunnaisarvoja, joiden keskiarvo on 0 ja varianssi on äärellinen. Arvot ovat myös tilastollisesti toisistaan riippumattomia. Tuomalla malliin r kappaletta havaintosarjojen viivettä ja m kappaletta valkoisen kohinan viivettä varmistetaan, että jäljelle jääneet virhetermit ovat täysin satunnaisia. (Box et al. 1994).

2.4 Korjausmenetelmät

Korjausmenetelmänä työssä käytetään lähimmän naapurin menetelmää (*Nearest neighbour imputation*), missä puuttuvat arvot korjataan lähimpien naapureiden keskiarvolla. Lähimmän naapurin menetelmästä on monia eri versiota, joita voidaan hyödyntää pieni- tai korkeaulotteiselle datalle (Faisal 2018). Työssä hyödynnetään kuitenkin yksinkertaistettua versiota lähimmän naapurin menetelmästä, koska käsiteltävä data on kaksiulotteista. Kaksiulotteista aikasarjadata-aineistoa käydään läpi vasemmalta oikealle siten, että puuttuva arvo paikataan aina edellisen sekä seuraavan löytyvän arvon keskiarvolla. Poikkeustapauksina on datan ensimmäinen ja viimeinen arvo, jolloin puuttuvat arvot imputoidaan lähimmällä olemassa olevalla arvolla. Kuvassa 6 havainnollistetaan sitä, mistä arvoista saadaan puuttuvan arvon korvaava arvo.

Eroja korjatun datan sekä todellisen datan välillä vertaillaan laskennallisesti absoluuttisel-



Kuva 6. Esimerkki lähimpien naapureiden imputointi menetelmän toiminnasta. Mustilla viivoilla kuvataan kahta lähintä saatavilla olevaa pistettä, jonka keskiarvolla on yksittäinen punaisella kuvattu arvo imputoitu.

la keskivirheellä (mean absolute error - MAE). Absoluuttinen keskivirhe saadaan kaavalla

$$MAE = \frac{|y_t - y_k|}{n},$$

missä y_t on alkuperäinen data, y_k on data, jonka epätäydellisyys on korjattu, ja n on datapisteiden määrä.

2.5 Lineaariregressio

Lineaariregressiossa luodaan malli kahdelle keskenään vahvasti korreloivalle satunnaismuuttujalle, mikä kertoo yhteyden muuttujien välillä. Korrelaatio kertoo, kuinka vahva lineaarinen yhteys muuttujien välillä on, ja se lasketaan kaavalla:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}},$$

missä $Cov(X,Y)$ on satunnaismuuttujien kovarianssi ja $Var(X)$ sekä $Var(Y)$ ovat muuttujien varianssit. Yksinkertainen lineaarisen regression malli on muotoa:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

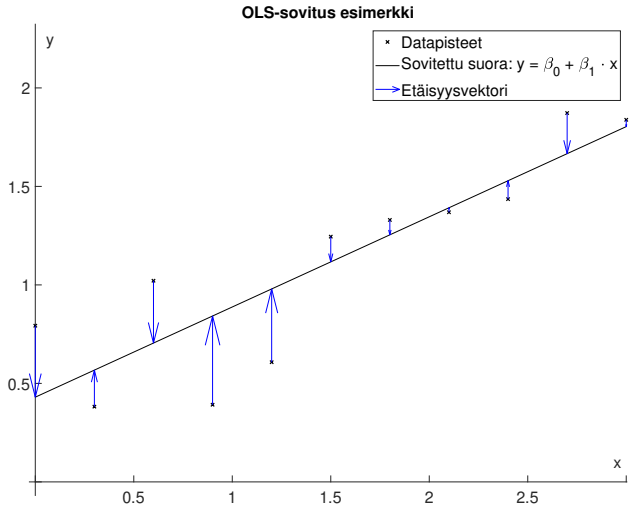
missä y_i on selitettävä muuttuja, x_i on selittävät muuttujat, β_0 ja β_1 ovat mallin kertoimet ja ϵ_i on virhetermi.

Lineaariregressiomallit usein, ja myös tässä työssä, sovitetaan hyödyntäen pienimmän

neliösumman menetelmää (*OLS - Ordinary Least Squares*). OLS:ssä on kyse siitä, että lineaarinen sovitus tehdään siten, että minimoidaan summa neliöityjä poikkeamia. OLS sovitus saa muodon:

$$\min S = \min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

missä y_i on alkuperäisen datan y:n, β_0 ja β_1 ovat sovitetun suoran kertoimet, ja x_i on alkuperäisen datan x:n arvot. OLS sovituksessa siis pyritään löytämään kertoimet β_0 ja β_1, joilla sovitettu suora saa pienimmät neliöidyt etäisyydet alkuperäisiin pisteisiin. Kuvassa 7 on esimerkki OLS sovituksesta, missä siniset nuolet datapisteistä kuvaavat etäisyyksiä sovitettuun suoraan.



Kuva 7. Esimerkki pienimmän neliösumman menetelmän suoran sovituksesta.

Sovituksen hyvyyttä eli selitysastetta mitataan työssä R^2-arvolla, mikä kertoo, kuinka hyvin malli selittää dataa ja mikä on sen ennustamisen vahvuus. R^2-arvo lasketaan kaavalla:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum (y_i - f(x_i, \hat{\beta}))^2}{\sum (y_i - \overline{y})},$$

missä y_i on alkuperäiset y:n arvot, $f(x_i, \hat{\beta})$ on sovituksen y:n arvot ja $\overline{y}$ on alkuperäisen y:n arvojen keskiarvo.

3 Aineisto ja ohjelmistot

Työssä käytetään aineistona World Bankin tarjoamaa väestönlaskentadataa maailman väestöstä sukupuolittain (WorldBank 2022). Kyseessä on aikasarjadataa vuosilta 1960-2020. Aineistosta keskitytään erityisesti eri maiden populaatioiden analysoimiseen, sillä populaatiodataa on parhaiten saatavilla suurimmasta osasta maista.

Työssä käytetään termiä *korjattu data*, kun puhutaan datasta mikä on muunneltu ensin epätäydelliseksi, minkä jälkeen siitä on korjattu epätäydellisyydet pois. Tilanteet missä dataan on luotu epätäydellisyyksiä ovat havainnollistusta varten, eivätkä välttämättä ole realistisia.

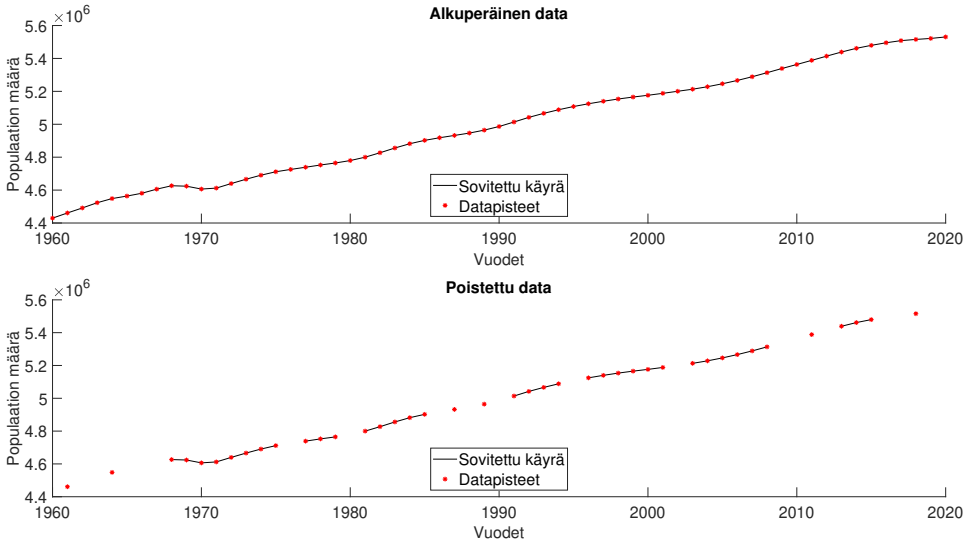
Väestönlaskentadataa käsitellään MATLAB-ohjelmiston avulla. MATLAB on ohjelmointikieli, mikä soveltuu hyvin matriisien ja numeeristen laskujen käsittelyyn. MATLABin System Identification ja Econometrics Toolbox ovat käytössä aikasarja-analyysissä.

4 Epätäydellisyyden korjaaminen

Data-aineistossa on eri maiden census dataa vuosilta 1960 - 2020. Työssä keskitytään eri maiden populaatiodatoihin, sekä pyritään löytämään korrelaatiota populaation ja muiden tekijöiden välillä.

4.1 Aikasarja-analyysi

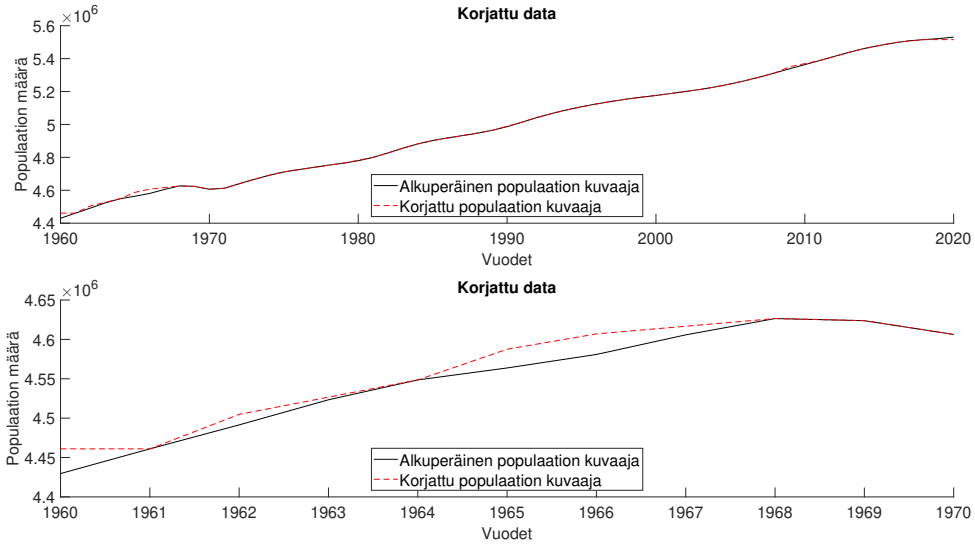
Kuvassa 8 käsitellään Suomen populaatiota. Kuvaajissa x-akselilla on aika, y-akselilla on populaation määrä sekä datapisteisiin on sovitettu käyrä. Alemmassa kuvaajassa datasta on poistettu täysin satunnaisesti 20 vuoden datapisteet, eli kyseessä on MCAR puuttuvuutta.



Kuva 8. Suomen populaatio vuosilta 1960 - 2020.

Poistettu data korjataan ympäröivien arvojen menetelmällä, missä yksittäinen puuttuva arvo paikataan edellisen ja seuraavan olemassa olevan arvon keskiarvolla, poikkeuksena ensimmäinen ja viimeinen arvo korvataan lähimmällä arvolla suoraan. Kuvassa 8 on korjatun datan sekä alkuperäisen datan kuvaajat piirrettynä. Kuvaajista havaitaan, että menetelmällä korjattu data mukailee alkuperäistä dataa. Välillä 1960 - 1970 havaitaan suurempia poikkeamia alkuperäisen kuvaajan kanssa.

Sekä alkuperäisestä datasta että sen korjatusta versiosta tehdään aikasarja-analyysin ja

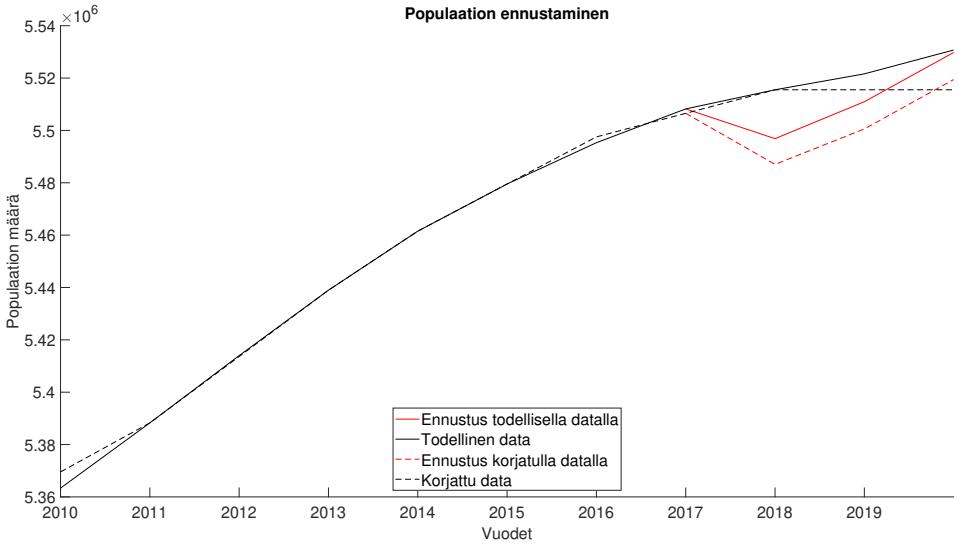


Kuva 9. Suomen korjatun datan kuvaajat vuosilta 1960 - 2020. Alemmassa kuvaajassa tarkennettuna vuosilta 1960-1970.

ARMA-mallin avulla ennustus viimeiselle kolmelle vuodelle oletuksena, että ne olisivat tuntemattomia. Tällöin vertailu ennusteiden sekä alkuperäisen datan välillä on mahdollista. Kuvassa 10 on tehty ennustus vuosille 2017 - 2020 siitä, mitä populaatiot olisivat korjatulla sekä alkuperäisellä datalla näinä vuosina. Taustalla on myös todelliset kuvaajat alkuperäiselle sekä korjatulle datalle. Vertailun kohteena on kuinka paljon molempien datojen ennusteet eroavat alkuperäisen datan kuvaajasta. Keskimäärin eroja todellisen datan arvoihin on alkuperäisen datan sovituksella $1.01 \cdot 10^4$ kpl ja korjatun datan sovituksella $2.03 \cdot 10^4$ kpl. Taulukossa 1 on ennusteiden prosentuaaliset erot sekä absoluuttiset erot todellisista arvoista vuosittain molemmille menetelmille.

Taulukko 1. Alkuperäisen datan ja korjatun datan ennusteiden prosentuaaliset sekä absoluuttiset erot vuosilta 2018-2020.

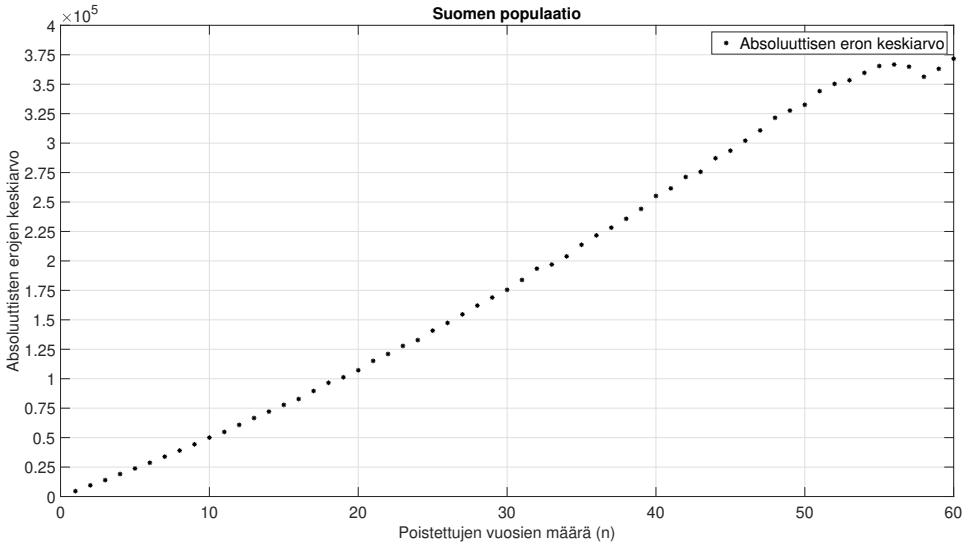
Data	2018	2019	2020
Alkuperäinen (%)	0.34 %	0.19 %	0.017 %
Alkuperäinen (kpl)	18671	10644	960
Korjattu (%)	0.52 %	0.38 %	0.20 %
Korjattu (kpl)	28492	21056	11289



Kuva 10. Suomen vuosien 2017-2020 populaation ennustaminen korjatulla sekä alkuperäisellä datalla.

4.2 Puuttuvan datan määrä

Kuvassa 11 simuloidaan, kuinka paljon puuttuvan datan määrä vaikuttaa lähimmän naapurin menetelmän korjauksen hyvyyteen. Simulaatiossa poistetaan Suomen populaatiodatasta vuosia 1-60 kappaletta ja korjataan ne lähimmän naapurin menetelmällä. Tämä simulaatio toistetaan 500 kertaa, jonka tuloksista otetaan keskiarvo kuvaajaa varten.

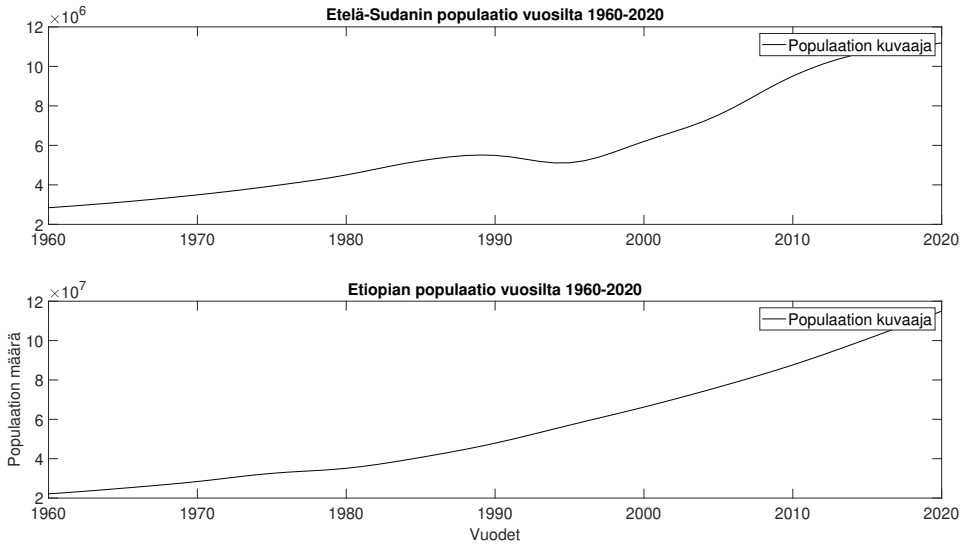


Kuva 11. Puuttuvan datan korjaamista suhteessa poistettujen vuosien määrään.

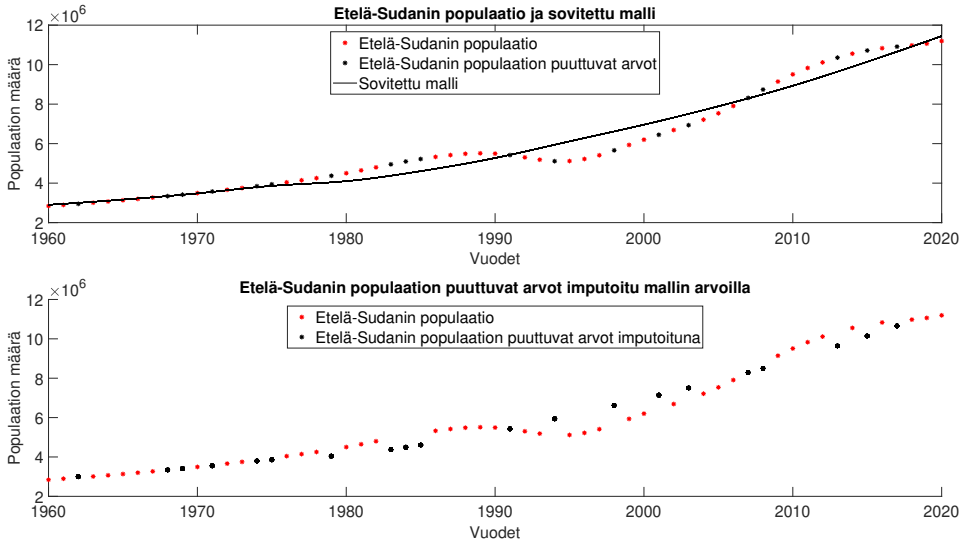
4.3 Regressio mallintaminen

Populaatioon määrään ja kasvuun on monia vaikuttavia tekijöitä. Tässä osiossa pyritään tutkimaan eri yhteyksiä populaation määrään, ja tekemään regressiomalli selittävien tekijöiden ja maan populaation välille. Regressiomalli toteutetaan tilanteessa, jossa selitettävän maan populaatiossa on puutteita datassa, ja puuttuvat arvot pyritään imputoimaan mallin avulla. Selittävissä muuttujissa ei ole puuttuvaa dataa. Regressiomallin luomisessa hyödynnetään siis naapurimaan väkiluvun arvoja vain niiltä vuosilta, mitkä löytyvät myös tutkittavassa maasta.

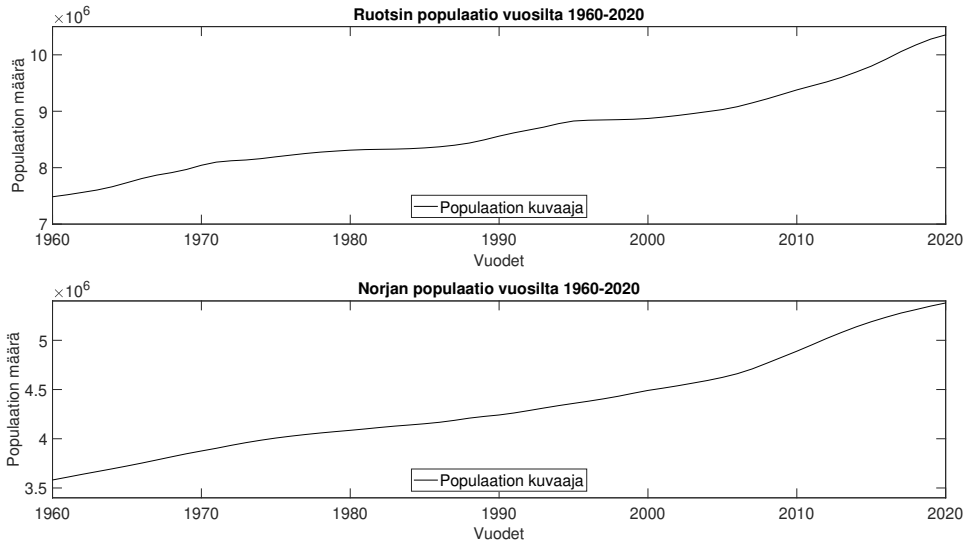
Kuvassa 12 on Etelä-Sudanin sekä sen naapurimaan Etiopian populaatiota vuosilta 1960-2020, ja kuvassa 14 on vastaava kuva Ruotsin ja Norjan populaatioille. Etelä-Sudanin sekä Ruotsin datasetistä poistetaan satunnaisesti 20 vuotta datapisteitä, jonka jälkeen datalle tehdään lineaariregressio, jossa selittävänä tekijänä on niiden naapurimaan populaatio. Etelä-Sudanin ja Etiopian datan välillä on havaittavissa 0.9820 korrelaatiokerroin, ja Ruotsin sekä Norjan välillä on havaittavissa 0.9951 korrelaatiokerroin. Kuvassa 13 on lineaariregressio Etelä-Sudanin populaatiolle, jonka sovituksen R^2 arvoksi saadaan 0.9650, ja kuvassa 15 on lineaariregressio Ruotsin populaatiolle, jonka sovituksen R^2 arvoksi saadaan 0.9899.



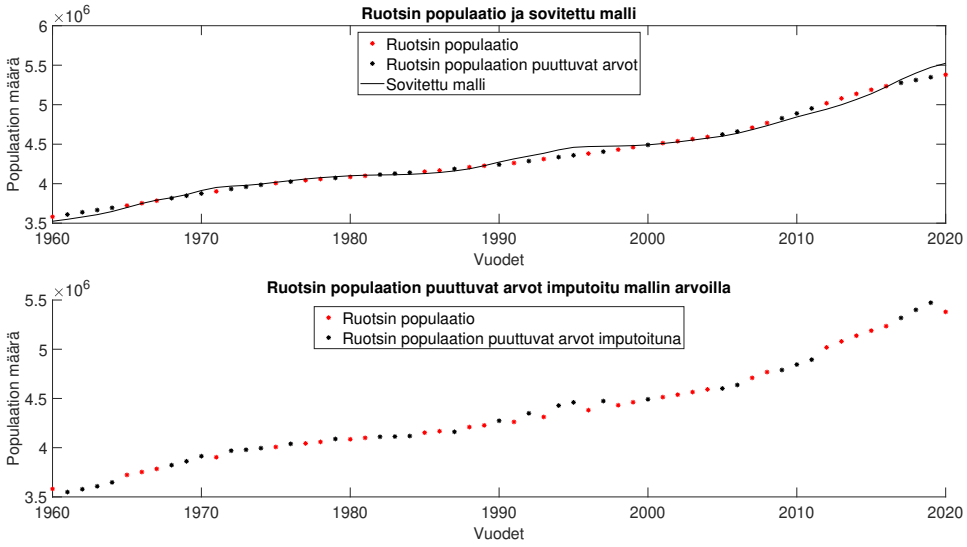
Kuva 12. Ylemmässä kuvassa Etelä-Sudanin populaatio ja alemmassa kuvassa Etiopian populaatio vuosilta 1960-2020.



Kuva 13. Etelä-Sudanin populaatiolle tehty lineaariregressio, jonka selittävän tekijänä on Etiopian populaatio. Ylemmässä kuvassa on Etelä-Sudanin populaation kuvaaja, josta puuttuu satunnaisesti arvoja, sekä regressiolla luotu malli. Alemmassa kuvassa puuttuvat arvot on imputoitu mallista saaduilla arvoilla.



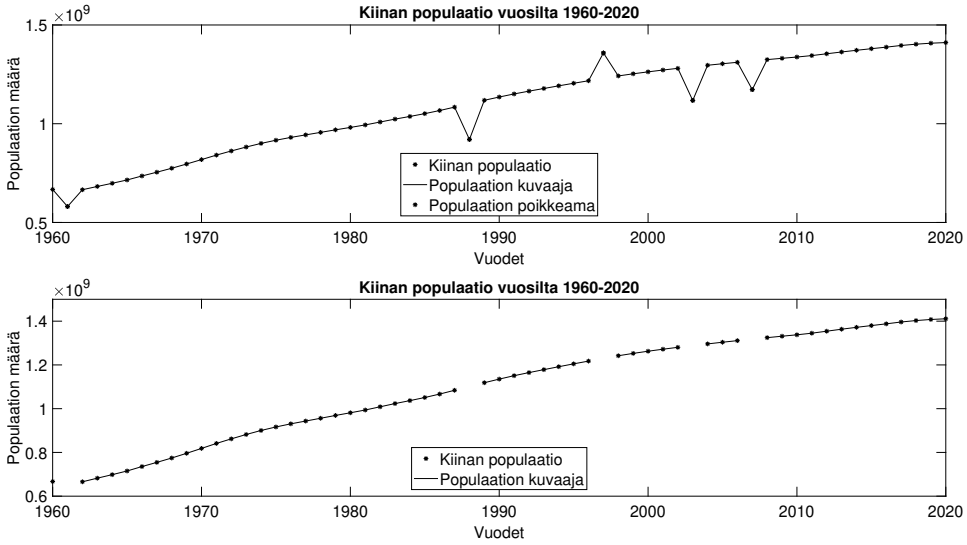
Kuva 14. Ylemmässä kuvassa Ruotsin populaatio ja alemmassa kuvassa Norjan populaatio vuosilta 1960-2020.



Kuva 15. Ruotsin populaatiolle tehty lineaariregressio, jonka selittävän tekijänä on Norjan populaatio. Ylemmässä kuvassa on Ruotsin populaation kuvaaja, josta on poistettu satunnaisesti arvoja, sekä regressiolla luotu malli. Alemmassa kuvassa puuttuvat arvot on imputoitu mallista saaduilla arvoilla.

4.4 Poikkeamat

Poikkeamia voidaan käsitellä vastaavilla tavoilla, millä voidaan käsitellä myös puuttuvia arvoja. Työssä ei erikseen esitellä poikkeamien käsittelyyn tarkoitettuja menetelmiä, vaan pyritään havainnollistamaan poikkeamien ja puuttuvien arvojen yhteyttä, sekä kuinka puuttuvien arvojen imputoinnin menetelmiä voidaan hyödyntää myös poikkeamien kanssa. Kuvassa 16 on Kiinan populaatiodata käsittelyssä, mihin on satunnaisesti luotu poikkeamia. Koska tiedämme poikkeamien sijainnit, voimme käsitellä niitä vapaasti. Poikkeamia voimme käsitellä kuvan 16 alemman kuvaajan mukaisesti eli yksinkertaisesti jättää ne huomioimatta. Tällöin voimme käsitellä poikkeamia samalla tavalla kuin käsittelisimme puuttuvia arvoja, esimerkiksi lähimmän naapurin menetelmällä.



Kuva 16. Kiinan populaation kuvaaja, mihin on luotu epätäydellisyyttä satunnaisilla poikkeamilla. Alemmassa kuvaajassa poikkeamat ovat poistettu, jolloin jäljelle jää vain puuttuvia arvoja.

5 Keskustelu

Aikasarja-analyysin tuloksista ja kuvasta 10 huomataan, että imputoimalla puuttuvat arvot, saadaan tehtyä ARMA-mallin mukainen ennuste, mikä on lähellä ennustetta, joka on tehty alkuperäisellä datalla. Alkuperäisen datan ennuste on kuitenkin selkeästi lähempänä todellista dataa, kuin korjatun datan ennuste. Suurimmat poikkeamat korjauksen ja alkuperäisen datan välillä kuvassa 9 on tilanteissa, joissa puuttuu useampi vuosi peräkkäin datasta.

Lähimmän naapurin menetelmän heikkoutena aikasarja-datan imputoimisessa on se, jos arvoja puuttuu paljon peräkkäin. Tällöin ympäröivät arvot eivät anna yhtä paljoa informaatiota siitä, mitä niiden välillä on tapahtunut. Tästä seuraa se, että jos puuttuva data on tyyppiä MNAR tai MAR, niin lähimmän naapurin menetelmällä ei saada yhtä tehokkaita tuloksia kuin MCAR puuttuvuuden kanssa. MAR- ja MNAR-puuttuvuuden ominaisuuksista johtuen niissä voi herkemmin tapahtua tilanteita, joissa arvoja puuttuu useampia peräkkäin.

On arvioitu, että jos puuttuvaa dataa on yli 15% koko datan määrästä, voi se vaikuttaa suuresti tuloksiin. Jos on $5-15\%$ puuttuvaa dataa, tarvitaan tällöin menetelmiä puuttuvuuden paikkaamiseen. Kuitenkin $1-5\%$ puuttuvuutta on helppo käsitellä, ja 1% on yleensä merkityksetöntä. (Acuna et al. 2004.) Kuvaajasta 11 huomataan, että keskimäärin poistettujen vuosien määrä vaikuttaa lineaarisesti imputoituun absoluuttiseen erojen keskiarvoon. Mitä lähemmäs päästään poistetuissa vuosissa n datan vuosien määrään N, sitä heikompia tuloksia menetelmä antaa. Jos poistettuja vuosia on kolme kappaletta eli noin viisi prosenttia koko datan vuosista, keskimäärin lähimmän naapurin menetelmässä on absoluuttinen ero noin 14000 kpl. Suomen väkiluvun nettovirhe on arvioitu olevan muutamia tuhansia henkilöitä (Nieminen 2014), minkä arvon tämä menetelmä keskimäärin ylittää, kun puuttuvaa dataa on viisi prosenttia. Kuitenkin on arvioitu, että maailman laajuisesti keskimääräinen virhe nykyhetken populaatiolle on noin kolme prosenttia (Council 2000). Vuoden 2020 Suomen väkiluvulle (noin 5.6 miljoonaa) kolmen prosentin nettovirhe tarkoittaisi noin 168000 henkilön virhettä, mikä on reilusti enemmän, mitä se todellisuudessa on.

Keskimääräistä virhettä maiden väestönmäärissä kasvattaa suuresti kehittyvät maat. Kehittyvissä maissa saatavilla oleva väestönlaskentadata on usein enemmän rajoittunutta ja epäluotettavampaa suuren kuolleisuuden ja syntyvyyden takia (Council 2000). Kehittyvissä maissa väestönlaskentaan hankaloittaa myös levottomuudet, kuten sota ja sairaudet, huono tietoliikenneinfrastruktuuri sekä ennakkoluulot valtiota kohtaan (Hoogeveen et al.

2020). Huomioitavaa siis on, että vaikka lähimmän naapurin menetelmä voi antaa heikkoja tuloksia maihin, missä väestönlaskenta on muutenkin hyvin tarkkaa, voi menetelmä olla hyödynnettävissä kehittyvien maiden väestönlaskennan tarkastelussa.

On arvioitu, että väkiluvun ennustaminen viiden vuoden päähän tuottaa alle viiden prosentin keskivirhettä, ja että jokaisen viiden vuoden lisääminen ennusteeseen lisää keskivirhettä noin 2.5 prosenttia (Council 2000). Taulukosta 1 huomataan, että sekä alkuperäisellä datalla että korjatulla datalla prosentuaaliset keskivirheet ovat alle yksi prosenttia kolmen vuoden ennusteelle. Saadaan siis ennusteet, minkä keskivirheet ovat reilusti alle yleisen arvion ennusteiden keskivirheiden suuruudesta. Vaikka Suomen populaatiodatasta puuttuu 20 vuotta dataa eli noin 33 prosenttia, niin saadaan imputoimalla puuttuvat arvot luotua ARMAn avulla ennuste, mikä vastaa hyvin paljon alkuperäisellä datalla luotua ennustetta. Vaikka käsittelyssä on maa, jonka väestön lukumäärä on hyvin tarkasti selvillä, tulosten perusteella menetelmää voisi hyödyntää maissa, joissa väestön lukumäärässä vuosien varrelta löytyy puutteita.

Regression tuloksista huomataan, että hyödyntämällä naapurimaan populaatiodataa, pystytään imputoimaan toisen maan puuttuvia arvoja muodostetulla mallilla. Vaikka Etelä-Sudanin ja Etiopian välille luotu lineaarinen regressiomalli saa hyvän R^2 arvon, niin imputoidut arvot, erityisesti 1980-2005 vuosien välillä, poikkeavat enemmän todellisista arvoista. Tämä johtuu siitä, että maiden väkilukujen välillä ei ole täysin lineaarista yhteyttä. Etelä-Sudanin väkiluku lähtee laskuun 1990 luvulla maan sisäisen sisällissodan takia. On arvioitu, että tämän sisällissodan aikana ihmisiä Sudanissa kuoli noin kaksi miljoonaa (Johnson 2018). Etiopian väkiluku on sen sijaan kasvanut hyvin lineaaristi 1980-luvulta lähtien. Vaikka maiden väkilukujen välillä on suurta korrelaatiota, on tärkeää tietää, kuinka samanlaista väkilukujen kasvu on ollut vuosien aikana.

Data, jota työssä hyödynnettiin, oli kooltaan pientä, sekä usein yksinkertaista laadultaan. Eri maiden väkiluvuissa, joita työssä tarkasteltiin, oli usein havaittavissa lineaarisesti kasvavaa trendiä. Datapisteiden vähäinen määrä aiheuttaa ongelmia siinä, että jo muutamalle vuodelle epätäydellisyyden luominen kattaa prosentuaalisesti suuren osan kokonaisvuosista. Esimerkiksi Etelä-Sudanin ja Ruotsin populaatioiden käsittelyssä poistettiin satunnaisesti vuosia 20 kappaletta, mikä vastaa noin kolmasosaa kaikista datapisteistä. Tässä huomataan selvästi se, että jos puuttuva data on selkeästi tyyppiä MCAR ja datassa on havaittavissa selkeä trendi, niin datan korjaaminen onnistuu, vaikka dataa puuttuisikin paljon.

Tutkimusta voitaisiin jatkaa erityisesti tilanteessa, jossa käsiteltäisiin jotain laajempaa da-

tasettiä. Lähimmän naapurin menetelmää voisi tutkia aikasarjalle, jossa data on tiheämpää, jolloin voitaisiin esimerkiksi soveltaa menetelmää ottamaan keskiarvot lähimmiltä vastaavilta päiviltä. Menetelmässä myös voisi tutkia, saataisiinko parempia tuloksia, jos otettaisiin esimerkiksi kolmen tai useamman lähimmän pisteen keskiarvo eikä pelkästään lähimmän kahden pisteen keskiarvo. Työssä oleva datan vähäinen määrä sekä vuoden aikaväli yksittäisten arvojen välillä, rajoitti epätäydellisyyden määrän luomista sekä menetelmien hyödyntämistä epätäydellisyyden korjaamiseksi. Suurempaa data-aineistoa tutkimalla saataisiin vertailtua, kuinka hyviä tuloksia menetelmät antavat, jos poistetaan prosentuaalisesti saman verran dataa kuin tässä työssä on tehty.

Väkiluvulle voitaisiin myös tutkia muita selittäviä tekijöitä kuin naapurimaiden väkiluvut. Esimerkiksi tiedetään, että väkiluvun määrään vaikuttaa suuresti kuolleisuus, koulutustaso ja syntyvyys. Näitä tutkimalla voitaisiin luoda väkiluvulle useamman muuttujan malli ja tutkia, selittäisivätkö nämä väkiluvun kehitystä paremmin. Myös muiden menetelmien hyödyntämistä väestönlaskentadataan olisi hyödyllistä tutkia. Olisi hyvä vertailla minkälaisia tuloksia ennusteilla saadaan, jos puuttuvat arvot jätettäisiin kokonaan huomioimatta tai hyödynnettäisiin esimerkiksi multi-imputointia (Faisal 2018).

6 Yhteenveto

Tässä kandidaatintyössä esiteltiin epätäydellisen datan ongelma sekä tuotiin esille eri tapoja, miten sitä voidaan käsitellä. Työssä käsiteltiin WorldBankin (WorldBank 2022) tarjoamaa väestönlaskentadataa maailman väestöstä MATLAB-ohjelmiston avulla. Erityisesti keskityttiin käsittelemään Euroopan, Afrikan ja Aasian maiden väkilukuja ja luomaan niihin keinotekoista epätäydellisyyttä.

Työssä tutkittiin lähimmän naapurin menetelmää, missä epätäydelliset arvot imputoitiin ympäröivien arvojen keskiarvolla sekä luotiin lineaariregressiomalleja tietyn maan väkiluvulle. Epätäydellisyys luotiin olemaan täysin satunnaista, jolloin puuttuva data oli tyyppiä MCAR. Korjatun datan kuvaajia verrattiin alkuperäisen datan kuvaajiin sekä tehtiin laskennallista vertailua niiden välillä. Suomen populaation korjatulle datalle sekä alkuperäiselle datalle tehtiin ARMA-mallin mukainen ennuste vuosille 2017-2020 ja tutkittiin niiden eroja todellisiin arvoihin.

Lähimmän naapurin menetelmän keskimääräiset korjauksen absoluuttisen erot kasvavat lineaarisessa suhteessa poistettujen vuosien määrään. Kuitenkin jo 10 vuoden puuttuminen tarkoittaa, että keskimäärin yksi paikattu arvo eroaa 50000 kpl todellisesta arvosta. Työssä huomattiin, että imputointimenetelmien hyödyntäminen pienille data-aineistoille tuottaa hankaluuksia, ja tulokset voivat erota merkittävästi toisistaan simulaatioita tehdessä.

Korjatun datan ja alkuperäisen datan avulla tehdystä ARMA-ennusteesta havaitaan, että vaikka datassa olisi puuttuvia arvoja, voidaan silti datasettiä hyödyntää aikasarja-analyysissä. Korjaamalla epätäydellisyys pois saavutettiin alle 0.2 % erot alkuperäisen datan ennusteisiin verrattuna. Tulosten perusteella voidaan siis todeta, että myös epätäydellistä voidaan hyödyntää, jos halutaan tehdä väkiluvuille ennusteita. Tärkeää on kuitenkin tietää kyseisen maan väestönlaskennan historiaa, jotta pystytään asettamaan korjauksien sekä ennusteiden tarkkuuksille sopivat virhemarginaalit.

Väestönlaskentadatan käsittelyn suurimpia ongelmia on tarkkojen data-arvojen huono saatavuus sekä se, että data-arvoja harvoin löytyy tiheältä aikaväliltä. Tässä kandidaatintyössä kuitenkin huomattiin, että epätäydellisyyttä väestönlaskentadatassa pystytään käsittelemään yksinkertaisilla korjausmenetelmillä, mikä implikoi, että menetelmiä voitaisiin pystyä myös hyödyntämään muissakin data-aineistoissa. Menetelmien yksinkertaisuuden vuoksi myös niiden kehittäminen on tarpeiden mukaan mahdollista.

Lähteet

Acuna, E. ja C. Rodriguez (2004). "The treatment of missing values and its effect on classifier accuracy". Teoksessa: *Classification, clustering, and data mining applications*. Springer.

Aggarwal, C.C. (2017). *Outlier Analysis.* Springer International Publishing AG 2017.

Allison, P.D. (2002). *Introduction. In: Missing Data, Quantitative Applications in the Social Sciences.* Thousand Oaks, CA: SAGE Publications.

Box, G., G. Jenkins ja G. Reinsel (1994). *Time Series Analysis: forecasting and control*. Prentice-Hall, Englewood Cliffs.

Council, National Research (2000). *Beyond Six Billion: Forecasting the World's Population.* Cuvillier Verlag.

Faisal, S. (2018). *Nearest Neighbor Methods for the Imputation of Missing Values in Low and High-Dimensional Data*. Cuvillier Verlag.

Hoogeveen, J. ja U. Pape (2020). *Data Collection in Fragile States : Innovations from Africa and Beyond*. World Bank.

Jablonska, M. (2008). *Analysis of outliers in electricity spot prices with example of New England and New Zealand markets*. Lappeenranta.

Johnson, H.F. (2018). *South Sudan : the untold story from independence to civil war*. I.B. Tauris.

Kendall, M.G. (1973). *Time Series*. Charles Griffin.

Little, R. J. A. ja D. B. Rubin (2019). *Statistical analysis with missing data.* John Wiley & Sons, Inc.

Nieminen, J. (2014). "Kuinka monta meitä on? Väestön laskemisessa on omat haasteensa". *Hyvinvointikatsaus* 1.

Pickup, M. (2015). *Introduction to Time Series Analysis.* SAGE Publications, Inc.

WorldBank (2022). *Gender Statistics*. URL: `https://datacatalog.worldbank.org/search/dataset/0037654/Gender-Statistics`.

Vriens, M. ja S. Sinharay (2006). *Dealing with missing data in surveys and databases*. SAGE Publications, Inc.

Taulukot

1 Alkuperäisen datan ja korjatun datan ennusteiden prosentuaaliset sekä absoluuttiset erot vuosilta 2018-2020. 17

Kuvat

1 MCAR dataa. Kuvassa ei havaittavissa syytä datan puuttumiselle. 7

2 MAR puuttuvuus. X-akselin arvot jotka ovat pienempiä kuin 1, puuttuvat todennäköisemmin. 8

3 MNAR puuttuvuus. Y-akselin arvot, jotka ovat pienempiä kuin 3, puuttuvat todennäköisemmin. 9

4 Esimerkkikuva vahvoista poikkeamista. 10

5 Esimerkkikuva kohinasta eli heikosta poikkeamasta. 10

6 Esimerkki lähimpien naapureiden imputointi menetelmän toiminnasta. Mustilla viivoilla kuvataan kahta lähintä saatavilla olevaa pistettä, jonka keskiarvolla on yksittäinen punaisella kuvattu arvo imputoitu. 13

7 Esimerkki pienimmän neliösumman menetelmän suoran sovituksesta. . . 14

8 Suomen populaatio vuosilta 1960 - 2020. 16

9 Suomen korjatun datan kuvaajat vuosilta 1960 - 2020. Alemmassa kuvaajassa tarkennettuna vuosilta 1960-1970. 17

10 Suomen vuosien 2017-2020 populaation ennustaminen korjatulla sekä alkuperäisellä datalla. 18

11 Puuttuvan datan korjaamista suhteessa poistettujen vuosien määrään. . . . 18

12 Ylemmässä kuvassa Etelä-Sudanin populaatio ja alemmassa kuvassa Etiopian populaatio vuosilta 1960-2020. 19

13 Etelä-Sudanin populaatiolle tehty lineaariregressio, jonka selittävän tekijänä on Etiopian populaatio. Ylemmässä kuvassa on Etelä-Sudanin populaation kuvaaja, josta puuttuu satunnaisesti arvoja, sekä regressiolla luotu malli. Alemmassa kuvassa puuttuvat arvot on imputoitu mallista saaduilla arvoilla. 20

14 Ylemmässä kuvassa Ruotsin populaatio ja alemmassa kuvassa Norjan populaatio vuosilta 1960-2020. 20

15 Ruotsin populaatiolle tehty lineaariregressio, jonka selittävän tekijänä on Norjan populaatio. Ylemmässä kuvassa on Ruotsin populaation kuvaaja, josta on poistettu satunnaisesti arvoja, sekä regressiolla luotu malli. Alemmassa kuvassa puuttuvat arvot on imputoitu mallista saaduilla arvoilla. . . 21

16 Kiinan populaation kuvaaja, mihin on luotu epätäydellisyyttä satunnaisilla poikkeamilla. Alemmassa kuvaajassa poikkeamat ovat poistettu, jolloin jäljelle jää vain puuttuvia arvoja. 22