Lokaalin regressiomallin sovittaminen järjestysasteikoisille muuttujille hyödyntäen aineiston augmentaatiota

Jussi Tanskanen

Tilastotieteen pro gradu –tutkielma
Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
7. toukokuuta 2017

Tiivistelmä

Tanskanen, Jussi: *Lokaalin regressiomallin sovittaminen järjestysasteikoisille muuttujille hyödyntäen aineiston augmentaatiota.* Tilastotieteen pro gradu -tutkielma, 28 s. + liitteitä 5 s., Jyväskylän yliopisto, Matematiikan ja tilastotieteen laitos, maaliskuu 2017.

Ihmistieteissä tilastollinen tutkimusaineisto kerätään usein kyselylomakkeissa järjestysasteikollisten Likert-asteikollisten muuttujien avulla. Näin kerätty aineisto on karkeaa, eli todellinen havaintoavaruus on pyöristetty rajalliseen määrään kategorioita, ja tällainen diskreetti järjestysasteikollinen tilastoaineisto voi osoittautua ongelmalliseksi niissä tapauksissa, joissa käytettävät tilastolliset analyysimenetelmät on kehitetty jatkuville muuttujille. Yksi tällainen menetelmä on parametriton lokaali regressioanalyysi, jota voidaan hyödyntää esimerkiksi käyräviivaisten yhteyksien analysointiin.

Tutkielma ehdottaa ja testaa aineiston augmentaatioon perustuvaa menetelmää lokaalin regressioanalyysin toteuttamiseen järjestysasteikollisilla muuttujilla. Menetelmässä järjestysasteikolliset muuttujat augmentoidaan satunnaisesti tasajakaumasta arpoen jatkuviksi muuttujiksi, joille suoritetaan lokaali regressioanalyysi. Simuloimalla agmentointeja ja lokaalin regression sovituksia saadaan estimoitua keskimääräinen sovite ja sille luottamusväli. Menetelmää testataan kolmella erilaiselle generoiduille aineistoille, jotka kuvaavat erilaisia käyräviivaisia riippuvuuksia ja empiirisellä esimerkillä.

Simuloinnin tulokset osoittavat menetelmän tarkkuuden olevan kiinni eniten järjestysasteikollisten muuttujien luokkien lukumäärästä. Luokkien lukumäärän kasvaessa menetelmän tarkkuus paranee. Aineiston koko tai yhteyden funktionaalisen muodon monimutkaisuus eivät olleet niin merkittäviä tekijöitä. Empiirisen esimerkin perusteella voi päätellä, että suuri hajonta ja muuttujien vinous heikentävät myös analyysin tarkkuutta. Empiiriset aineistot voivat olla hyvin karkeita, hajonta isoa, jakaumat vinoja ja efektit pieniä. Aineiston augmentaatioon perustuva lokaalin regression menetelmä antaa kuitenkin useaan käyttötarpeeseen riittävän approksimaation käyräviivaisen yhteyden luonteesta jo viisi- tai seitsemän-luokkaisilla järjestysasteikoillisilla muuttujilla.

Sisältö

1. Johdanto

Ihmistieteissä tilastollinen tutkimusaineisto kerätään usein kyselylomakkeiden avulla. Usein kyselylomakkeissa vastaukset annetaan diskreetillä skaalalla, vaikka ilmiö itsessään olisikin jatkuva. Esimerkiksi tyytyväisyyttä elämään voidaan pyytää arvioimaan Likert-asteikolla: 1 = ei lainkaan tyytyväinen, 2 = melko tyytymätön, 3 = ei tyytyväinen tai tyytymätön, 4 = melko tyytyväinen ja 5 = täysin tyytyväinen. Vastaajan on valittava joku annetuista vaihtoehdoista, vaikka todellisuudessa tyytyväisyys elämään voisi olla jotain annettujen arvojen väliltä. Tällaista aineistoa voidaan kutsua karkeaksi aineistoksi (coarse data), koska todellinen jatkuva havaintoavaruus pyöristetään rajalliseen määrään kategorioita (Heitjan & Rubin, 1991). Pyöristämisen lisäksi karkeaa dataa luovat muun muassa aineiston pyöristäminen sopiviin numeroihin (digit preference), sensurointi ja intervallisensurointi. Likert-asteikko on järjestysasteikollinen psykometrinen mitta-asteikko, jota käytetään paljon varsinkin ihmistieteellisissä survey-kyselylomakkeissa. Diskreetti järjestysasteikollinen tilastoaineisto voi osoittautua ongelmalliseksi niissä tapauksissa, joissa käytettävät tilastolliset analyysimenetelmät on kehitetty jatkuville muuttujille. Yksi tällainen menetelmä on lokaali regressioanalyysi, jota voidaan hyödyntää esimerkiksi käyräviivaisten yhteyksien analysointiin. Tämä Pro gradu –tutkielma ehdottaa ja testaa aineiston augmentaatioon perustuvaa menetelmää parametrittoman lokaalin regressioanalyysin toteuttamiseen järjestysasteikollisilla muuttujilla.

Ihmistieteissä ilmiöiden välisiä yhteyksiä mallinnetaan oletusarvoisesti lineaarisilla menetelmillä, vaikka teoriat vain harvoin määrittävät yksiselitteisesti riippuvuuksien funktionaalisen muodon (Beck & Jackman, 1998). Usein lineaarisuusoletus tehdään implisiittisesti edes pohtimatta ilmiön luonnetta. Todellisuudessa kuitenkin monet ilmiöt ovat käyräviivaisia. Grant ja Schwartz (2011) ovat eritelleet kolme erilaista mekanismia, jotka selittävät käyräviivaisen yhteyden: 1) arvoristiriidat, 2) kynnysarvot ja 3) ei-monotoniset vaikutukset. Arvoristiriita ilmenee, kun esimerkiksi yhden asian tekeminen erinomaisesti vaatii paljon aikaa, joka on pois muiden asioiden tekemiseltä ja kokonaisuus kärsii. Kynnysarvot viittaavat tilanteeseen, jossa positiivinen tai negatiivinen vaikutus alkaa tai loppuu tietyn kynnysarvon kohdalla. Esimerkiksi raha ei enää tietyn pisteen jälkeen tuo onnellisuutta ja fyysisesti tai henkisesti vaativat työtehtävät eivät kohtuullisina määrinä aiheuta ongelmia, mutta tietyn pisteen jälkeen alkavat kasvattaa stressiä. Yleisesti sanotaankin, että hyvät asiat saturoituvat ja

pahat eskaloituvat. Ekonomit viittaavat kynnysarvoihin laskevan rajahyödyn käsitteen avulla (esim. Hirvonen & Mangeloja, 2006). Jotkut ilmiöt taas ovat luonnostaan ei-monotonisia ja epälineaarisia. Esimerkiksi aktivointiteoria (activation theory) ja Yerkes-Dodsonin laki ehdottavat, että ihmiset tarvitsevat tietyn määrän aktivointia ollakseen motivoituneita, mutta liiallinen stimulointi johtaa stressiin (esim. Grant and Schwartz, 2011). Arkisempi esimerkki käyräviivaisesta yhteydestä löytyy esimerkiksi hampurilaisten syömisestä. On helppoa kuvitella miten hyvältä muutama hampurilainen maistuisikaan ja miten niiden syöminen ilahduttaisi, mutta samaan aikaan vähintään yhtä helposti voisi kuvitella, miten pahalta ja väkinäiseltä tuntuisi syödä esimerkiksi kuudes hampurilainen.

Käyräviivaisia yhteyksiä voidaan mallintaa parametrisilla ja parametrittomilla menetelmillä. Parametrisilla menetelmillä aineistoon sovitetaan jokin tietty ennalta valittu käyräviivainen funktio kuten esimerkiksi toisen asteen polynomi tai logaritmi. Logaritmi- ja eksponenttifunktioilla voidaan mallintaa kynnysarvojen tilannetta ja polynomifunktiolla esimerkiksi U-kirjaimen (tai käännetyn U-kirjaimen) muotoista yhteyttä. Parametristen menetelmien käyttö edellyttää kuitenkin tutkijalta ennakkokäsitystä tutkittavan yhteyden funktionaalisesta muodosta. Toinen ongelma liittyy siihen, että käyräviivaisuus voi olla hyvinkin lokaali ominaisuus rajoittuen vain jollekin rajatulle välille, jolloin globaalit funktiot eivät ole päteviä työkaluja. Lisäksi parametrisiin käyräviivaisiin menetelmiin voi myös liittyä multikollineaarisuutta sekä ne voivat toimia heikosti datan ääripäissä, joissa havaintoja ei ole paljoa (Beck & Jackman, 1998). Parametrittomat menetelmät toimivat lokaalisti ja estimoivat riippuvuuden funktionaalisen muodon suoraan aineistosta ilman a priori oletusta riippuvuuden muodosta.

Tässä tutkielmassa tarkastellaan aineiston augmentaatioon, eli lisäämiseen, perustuvaa menetelmää Likert-asteikollisten ja muiden järjestysasteikollisten muuttujien analysoimiseksi lokaalilla regressioanalyysillä. Yksinkertaisuuden takia tutkielmassa käsitellään vain kahden muuttujan välisiä yhteyksiä. Ideana on augmentoida järjestysasteikollinen diskreetti aineisto jatkuvaksi, jolloin parametrittomat menetelmät toimivat varmasti. Jatkuva muuttuja luodaan arpomalla tasajakaumasta havaintoja väleille, joita järjestysasteikollisen muuttujan arvot kuvaavat. Esimerkiksi viisiportainen Likert-asteikollinen muuttuja jaetaan viiteen osaan ja jokaiseen osioon arvotaan tasajakaumasta arvoja alkuperäisen Likert-asteikollisen muuttujan saamien arvojen lukumäärän mukaan. Koska Likert-asteikolliset muuttujat eivät aina ole välimatka-asteikollisia, valitaan luokittelurajat jatkuvalle muuttujalle satunnaisesti. Varsinainen lokaali regressioanalyysi suoritetaan augmentoiduilla

jatkuvilla muuttujilla. Simuloimalla uusia luokittelurajoja ja jatkuvia muuttujia saadaan selville keskimääräinen lokaalin regression sovite, jonka pitäisi edustaa oikeaa funktionaalista muotoa. Simulointikokeita suoritetaan eri kokoisille aineistoilla ja eri määrille järjestysasteikon luokkia. Menetelmää testataan myös oikeaan aineistoon.

Tutkielman aluksi tarkastellaan järjestysasteikoista Likert-asteikkoa ja sen erityispiirteitä tarkemmin, jonka jälkeen esitellään regressioanalyysin teoriaa keskittyen erityisesti lokaaliin regressioon. Luokittelun vaikutusta parametrittomaan lokaaliin regressioon tarkastellaan niin simulointikokeilla kuin oikean tilastollisen aineistonkin tapauksessa. Pro gradu –tutkielma päättyy johtopäätöksiin, joissa pohditaan menetelmän soveltuvuutta ja käyttöä.

2. Likert-asteikko

Likert-asteikko on psykometrinen mitta-asteikko, jota käytetään varsinkin ihmistieteellisissä survey-kyselylomakkeissa. Likert-asteikkoa käytetään erityisesti asenne- ja motivaatiomittareissa, joissa koehenkilö arvioi omaa käsitystään väitteen tai kysymyksen sisällöstä (Bryman, 2012). Vastausvaihtoehdot riippuvat mittarin tyypistä. Väitteiksi muotoilluissa mittareissa, kuten esimerkiksi: *”Olen tyytyväinen elämääni”* pyydetään arvioimaan kuinka samaa tai eri mieltä on väitteen kanssa. Useimmiten käytetään viisiportaista Likert-asteikkoa, jonka vastausvaihtoehdot ovat:

1 = täysin eri mieltä
2 = eri mieltä
3 = ei samaa eikä eri mieltä
4 = samaa mieltä
5 = täysin samaa mieltä.

Myös seitsenportaisia mittareita ja toisinaan jopa neliportaista mittaria käytetään. Väitteiden lisäksi voidaan kysyä muunkinlaisia kysymyksiä, kuten esimerkiksi tapahtumien yleisyyttä. Tapahtumien yleisyyttä kysyessä vastausvaihtoehdot voivat olla tarkoituksella epämääräiset (esim. harvoin, joskus, usein) tai tarkemmin määritellyt (esim. kerran kuussa, kerran viikossa). Esimerkkinä kysymys: *”Kuinka usein tunnet itsesi onnelliseksi”*, johon vastataan seitsenportaisella Likert-asteikolla, jonka vastausvaihtoehdot voivat olla seuraavan kaltaiset:

1 = ei koskaan	/ ei koskaan
2 = tuskin koskaan	/ muutaman kerran vuodessa
3 = harvoin	/ kerran kuussa
4 = joskus	/ muutaman kerran kuussa
5 = usein	/ kerran viikossa
6 = hyvin usein	/ muutaman kerran viikossa
7 = aina	/ päivittäin.

Likert-asteikollinen muuttuja voi joissain tapauksissa olla välimatka-asteikollinen, mutta joissain tapauksissa se ei sitä selvästikään ole, vaan on vain järjestysasteikollinen. Esimerkiksi tapahtuman yleisyyttä kysyttäessä vastausvaihtoehdot päivittäin, viikoittain tai kuukausittain eivät selvästikään ole välimatka-asteikon mukaisia. Myös samaa mieltä – eri mieltä –asteikon muuttujat eivät ainakaan yksikäsitteisesti ole välimatka-asteikollisia, koska on hankala määrittää, onko esimerkiksi *täysin samaa mieltä* olemisen ja *samaa mieltä* olemisen välinen välimatka yhtä suuri kuin *samaa mieltä* olemisen ja *ei samaa eikä eri mieltä* olemisen välillä. Ei ole myöskään yksikäsitteistä, millaista väliä Likert-asteikon saamat arvot edustavat esimerkiksi ihmisen todelliselta jatkuvalta samaa mieltä – eri mieltä -asteikolta. Jos ajatellaan, että vastaajalla olisi olemassa reaalinen jatkuva asteikko samaa mieltä olemiselleen, joka saisi arvoja esimerkiksi välillä 0 (ei lainkaan samaa mieltä) ja 1 (täysin samaa mieltä). Likert-asteikon vastausvaihtoehdot eivät luultavasti ole tasavälisesti sijoittuneet jatkuvan muuttujan skaalalle ja riippuu paljon myös vastaajasta, että sijoittaisiko hän esimerkiksi 0,8 arvoisen hyväksyntänsä väitteelle Likert-asteikon kategoriaan *samaa mieltä* vai *täysin samaa mieltä*.

Usea Likert-asteikollinen muuttuja voidaan liittää yhteen summamuuttujaksi, joka muodostetaan laskemalla yhteen useiden muuttujien, eli osioiden, arvoja ja usein vielä keskiarvoistamalla summamuuttuja alkuperäisten muuttujien skaalalle. Summamuuttujan voidaan ajatella olevan osiensa summa tai niitä voidaan käyttää faktorianalyysin tapaan latenttien muuttujien estimointiin. Summamuuttuja voidaan myös tulkita esimerkiksi indikaattori- tai proxy–muuttujaksi. Siinä missä faktoripistemäärämuuttuja olisi aidosti jatkuva muuttuja, on taas keskiarvoistettu summamuuttuja skaalansa takia helpompi tulkita. Toisaalta, mitä useammasta osiosta (muuttujasta) summamuuttuja koostuu, sitä enemmän eri arvoja summamuuttuja saa. Summamuuttujan mahdollisten arvojen lukumäärä lasketaan kaavalla

$$LKM_T = k(LKM_Y - 1) + 1,$$

jossa LKM_T on summamuuttujan mahdollisten arvojen lukumäärä, k on summamuuttujan osioiden lukumäärä ja LKM_Y osioiden mahdollisten arvojen lukumäärä. Summamuuttujan mahdollisten arvojen lukumäärä (ks. Taulukko 1) kasvaa melko nopeasti osioiden lisääntyessä ja summamuuttuja alkaa muistuttaa yhä enemmän jatkuvaa muuttujaa.

Taulukko 1. Summamuuttujan mahdollisten arvojen lukumäärä

	Mittarin skaala		
Osioiden lukumäärä	**1-4**	**1-5**	**1-7**
1	4	5	7
2	7	9	13
3	10	13	19
4	13	17	25
5	16	21	31
6	19	25	37
7	22	29	43
10	31	41	61

3. Regressiomenetelmien teoriaa

Regressiomenetelmillä mallinnetaan selittävien muuttujien ja vastemuuttujan välistä yhteyttä. Regressioanalyysin tarkoituksena on estimoida regressiofunktio eli vastemuuttuja Y:n ehdollinen odotusarvo ehdolla, että selittävät muuttujat saavat arvot $X_1 = x_1,...,X_p = x_p$. Tätä merkitään

$$\mathbb{E}[Y|X_1 = x_1, \dots, X_p = x_p] = \mu(x_1, \dots, x_p).$$

Regressiomalli voi olla parametrinen, jolloin muuttujien yhteyttä, eli ehdollista odotusarvoa, kuvataan jollain ennalta määrätyllä funktiolla. Esimerkiksi lineaarisessa regressiossa oletetaan, että

$$\mu(x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

Parametrittomassa regressiossa selitettävien muuttujien ja vastemuuttujan yhteydestä, eli ehdollisen odotusarvon funktionaalisesta muodosta ei tehdä oletuksia vaan yhteys estimoidaan aineistosta (Loader, 2004).

3.1. Parametriton regressio

Pisteparvessa (x_i,y_i), $i = 1,...,n$ selittävän muuttujan x_i ja vastemuuttujan y_i välistä yhteyttä $y_i = \mu(x_i) + \varepsilon_i$ voidaan mallintaa niin parametrisilla kuin parametrittomilla regressiomenetelmillä. Tässä tapauksessa x_i on yksiulotteinen, koska tarkastelu rajataan yhden selittävän muuttujan tapaukseen, ja yksinkertaisuuden vuoksi oletetaan satunnaisvirheet ε_i samavarianssisiksi, riippumattomiksi ja normaalijakautuneiksi. Tehtävänä on usein estimoida tuntematon funktio μ, joka kuvaa selittäjän ja vasteen yhteyttä. Parametrisesti mallintaen oletetaan, että selittäjän ja vasteen välinen yhteys μ noudattaa jotain tiettyä funktionaalista muotoa, joka voidaan täysin määritellä rajallisella määrällä parametreja. Esimerkkinä parametrisesta menetelmästä on lineaarinen regressioanalyysi, jossa funktionaalinen muoto määritellään suoraksi. Aina kuitenkaan selittäjän ja vasteen välistä yhteyttä ei voida kuvata yksinkertaisten funktioiden avulla. Parametrittomissa menetelmissä selittävän muuttujan ja vastemuuttujan väliselle yhteydelle, eli vasteen ehdolliselle odotusarvolle ei oleteta mitään tiettyä funktionaalista muotoa vaan muoto estimoidaan aineistosta (Loader, 2004). Aineistosta

parametrittomasti estimoitua regressiokäyrää kutsutaan tasoitusfunktioksi (*smoothing function*). Selittävän muuttujan ja vastemuuttujan välisen mahdollisesti hyvin monimutkaisen yhteyden selvittäminen on Härdlen (1990) mukaan yksi parametrittoman regression pääkäyttötarkoituksista. Parametrittomia regressiomenetelmiä ovat esimerkiksi lokaali regressio (*local regression*) ja tasoitussplinit (*spline smoothing*).

Vaikka parametrittomat menetelmät eivät tarvitse ennakkotietoa estimoitavan yhteyden funktionaalisesta muodosta, niin paradoksaalisesti parametrittomat menetelmät sisältävät kuitenkin paljon oletuksia esimerkiksi tasoituksen sileydestä ja painotusfunktioista, jotka voivat merkittävästi vaikuttaa estimoituun sovitteeseen. Oikeastaan parametrittomissa malleissa on hyvinkin paljon parametreja, kuten esimerkiksi paikalliset regressiokertoimet lokaalin regression tapauksessa, mutta nämä parametrit kuitenkin riippuvat toisistaan.

3.2. Lokaali regressio

Lokaalin regression (*local regression, locally weighted scatterplot smoothing, LOESS, LOWESS*) menetelmää esiteltiin jo 1800-luvun loppupuolella ja tilastollisessa kirjallisuudessa 1970-luvun lopulla usean kirjoittajan voimin (Loader, 2004). Tässä esitetty teoria lokaalista regressiosta perustuu Clevelandin (1979) ja Loaderin (2004) artikkeleihin. Lokaali regressio perustuu ideaan, jonka mukaan tasoitusfunktiota voidaan approksimoida lokaalisti ensimmäisen tai toisen asteen polynomilla jokaisen pisteen x_i naapurustossa.

Lokaali approksimaatio voidaan sovittaa aineistoon käyttäen lokaalia painotettua pienimmän nelisumman menetelmää. Painotus suoritetaan painofunktiolla W, jonka valinnassa tulee huomioida seuraavat vaatimukset:

1. $W(x) > 0$, kun $|x| < 1$.
2. $W(x) = 0$, kun $|x| \geq 1$.
3. $W(x) = W(-x)$.
4. $W(x)$ on vähenevä, kun $x \geq 0$.

Usein painofunktioksi valitaan

$$W(x) = \begin{cases} (1 - |x|^3)^3, & kun\ |x| < 1 \\ 0, & kun\ |x| \geq 1 \end{cases}$$

Tämä painofunktio painottaa eniten pistettä x_i lähellä olevia pisteitä. Jokaiselle x_i määritetään painot $w_k(x_i)$ siten että, vain pisteen x_i naapurustoon kuuluvat pisteet saavat nollasta eroavia arvoja.

$$w_k = W\left(\frac{x_k - x_i}{h_i}\right),\ k = 1,\ldots,n,$$

jossa h_i on etäisyys pisteen x_i ja kauimmaisen naapurin välillä. Naapuruston kokoa voidaan kuvata parametrilla r, joka kuvaa havaintojen lukumäärää naapurustossa. Tasoitusfunktion tasaisuutta kontrolloidaan tasoitusparametrilla f = r/n, joka kertoo, kuinka suuri osa havainnosta otetaan mukaan naapurustoon. Tasoitusta kasvatetaan ottamalla suurempi osa havainnosta naapurustoon, eli kasvattamalla tasoitusparametrin f arvoa.

Lokaalin regression tasoitusfunktio (loess-tasoituskäyrä) estimoidaan käyttäen painotettua pienimmän neliösumman menetelmää, jokaisessa pisteessä x_i. Minimoitavana on

$$\sum_{k=1}^{n} w_k(x_i)\left(y_i - \beta_0 - \beta_1 x_k - \ldots - \beta_d x_k^d\right)^2,$$

jossa d on polynomin asteluku. Sovitteet tehdään erikseen kaikille pisteille x_i, ja näiden sovitteiden mukaisesti muodostetaan loess-tasoituskäyrä. Oletusarvoisesti sovitteet lasketaan havaintopisteille, mutta tämä ei ole välttämätöntä ja tässä tutkimuksessa sovitteet lasketaan 50 tasaväliselle pisteelle välillä 0–1, koska simuloinnin kannalta on tärkeää, että pisteet ovat kaikilla iteraatiokierroksella samat.

Lokaalia regressiota käytettäessä pitää valita sopiva naapuruston koko (f), painofunktio (W) sekä paikallisen regressiomallin aste (d). Cleveland (1979) tarjoaa ohjeita näiden valintaan. Tasoitusparametri f valitaan usein silmämääräisesti kokeiluja tekemällä. Usein sopiva tasoituksen

määrä löytyy väliltä f=[0.2,0.8]. Myöhemmin on kehitetty myös työkaluja, kuten esimerkiksi ristiinvalidointi, joilla tasoitusparametri voidaan estimoida suoraan aineistosta (Loader, 2004). Esitelty ja käytetty painofunktio W täyttää hyvin painofunktiolle esitetyt vaatimukset ja lisäksi se antaa hyvän approksimaation virhevarianssin estimaattorille. Lisäksi esitetyn painofunktion W pitäisi taata riittävä tasoite lähes joka tilanteessa. Lokaalin käyrän astelukua d=1 pidetään usein riittävänä ja Cleveland (1979) piti astelukua d=2 laskennallisesti vaativana. Nykyään kuitenkin tietokoneiden laskentateho ei ole ongelma ja toisen asteen paikallinen regressiomalli tuo enemmän joustavuutta datan mallintamiseen. Tässä tutkimuksessa käytetään selkeyden takia samaa tasoitusparametria (f=2/3) kaikissa tilanteissa. Painofunktiona käytetään esiteltyä painofunktiota W ja paikallisen käyrän asteluvuksi asetetaan d=2. Nämä parametrien arvot ovat myös oletusarvoisia R-ohjelmistossa. Aineistojen generointi, simulointi ja oikean empiirisen aineiston analysointi on suoritettu R (3.3.1.) ohjelmistolla (R Core Team, 2016). Lokaalit regressiomallit sovitettiin hyödyntäen funktioita *loess* ja *loess.smooth*.

4. Simulointikokeita

Simulointikokeilla luokitellun aineiston laajentamista jatkuvaksi voidaan tarkastella helposti. Luokkien määrä ja aineiston koko voidaan ottaa huomioon ja tutkia niiden vaikutusta luokitellun aineiston laajentamiseen. Simulointi suoritettiin seuraavasti:

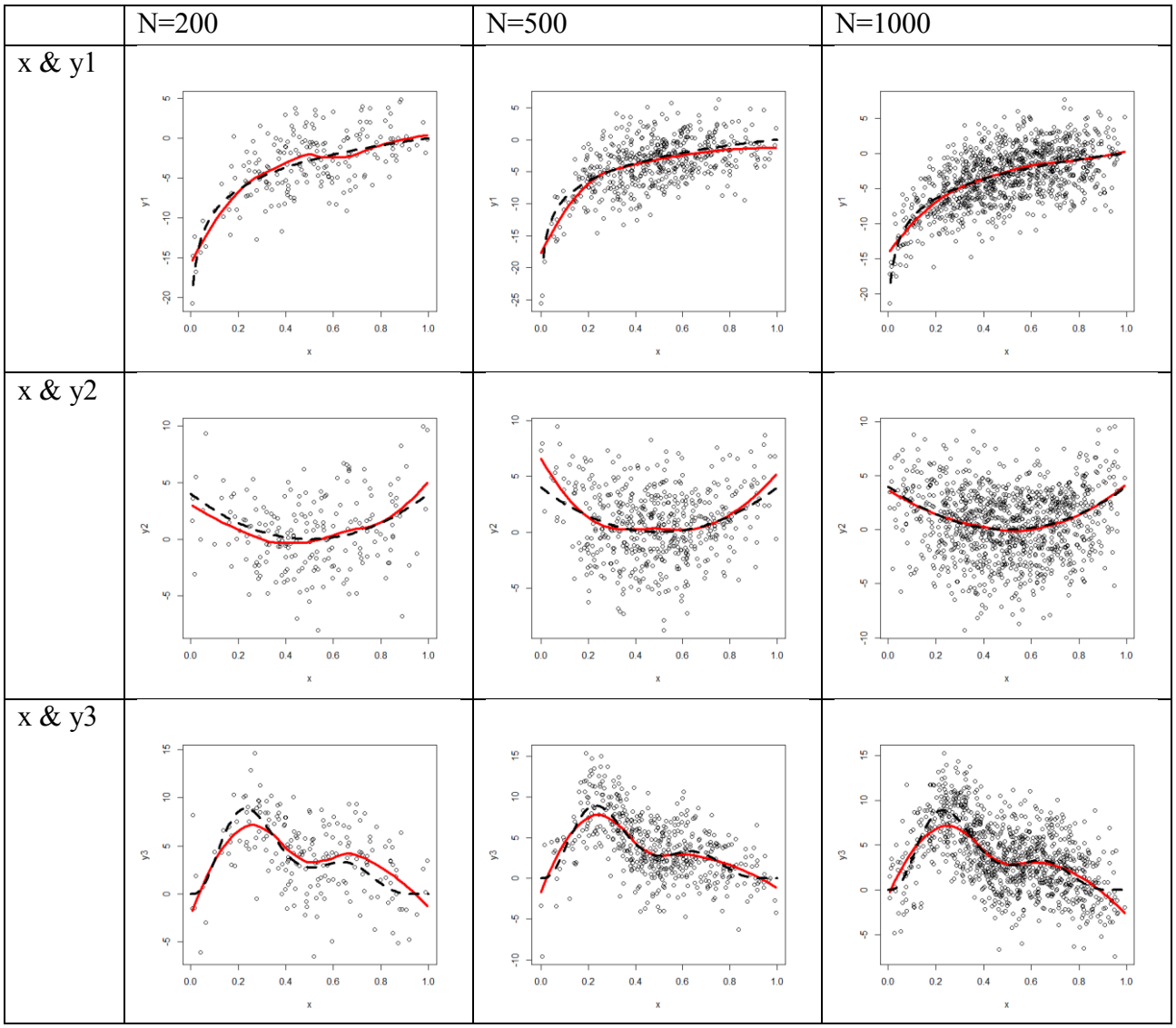
1. Simuloidaan jatkuva aineisto.
2. Luokitellaan jatkuva aineisto valitsemalla luokittelurajat arpomalla tasajakaumasta sekä tälle verrokiksi valitaan tasaväliset luokittelurajat.
3. Valitaan uudet luokittelurajat arpomalla ne tasajakaumasta.
4. Laajennetaan luokiteltu aineisto uudelleen jatkuvaksi, arpomalla datapisteiden sijainti tasajakaumista.
5. Suoritetaan lokaalin regression sovitus.
6. Toistetaan kohtia 3, 4 ja 5.

Simulaatiokokeita varten generoidaan dataa edustamaan kolmea erilaista käyräviivaista tilannetta. Ensimmäisessä tilanteessa x- ja y-muuttujien välisen yhteyden funktionaalinen muoto on logaritminen, toisessa tilanteessa U-kirjaimen muotoinen, eli toisen asteen yhteys, ja kolmannessa tilanteessa tarkastellaan monimutkaista käyrää. Simulaatioissa kokeillaan kolmea erikokoista aineistoa, jotta voidaan selvittää miten aineiston koko vaikuttaa analyysiin. Aineistojen koot (N=200, N=500 ja N=1000) on valittu edustamaan ihmistieteille tyypillisiä aineistoja.

X-selittäjä on jokaisessa tapauksessa sama ja se generoidaan katkaistusta normaalijakaumasta välille 0–1 keskiarvolla 0,5 ja keskihajonnalla 0,25. Kolme erilaista vastemuuttujaa (Y) luodaan yleisesti käytetyllä tavalla lisäämällä sopiviin funktioihin normaalijakaumasta N(0,3) satunnaisesti arvottu luku kuvaamaan vaihtelua. Vastemuuttujat generoidaan seuraavalla tavalla:

$$y1 = 4 * \log(x) + N(0,3)$$
$$y2 = (4 * (x - 0{,}5))^2 + N(0,3)$$
$$y3 = (0{,}2 * x)^{11} * \left(10 * (1 - x)\right)^6 + 10 * (10 * x)^3 + (1 - x)^{10} + N(0,3)$$

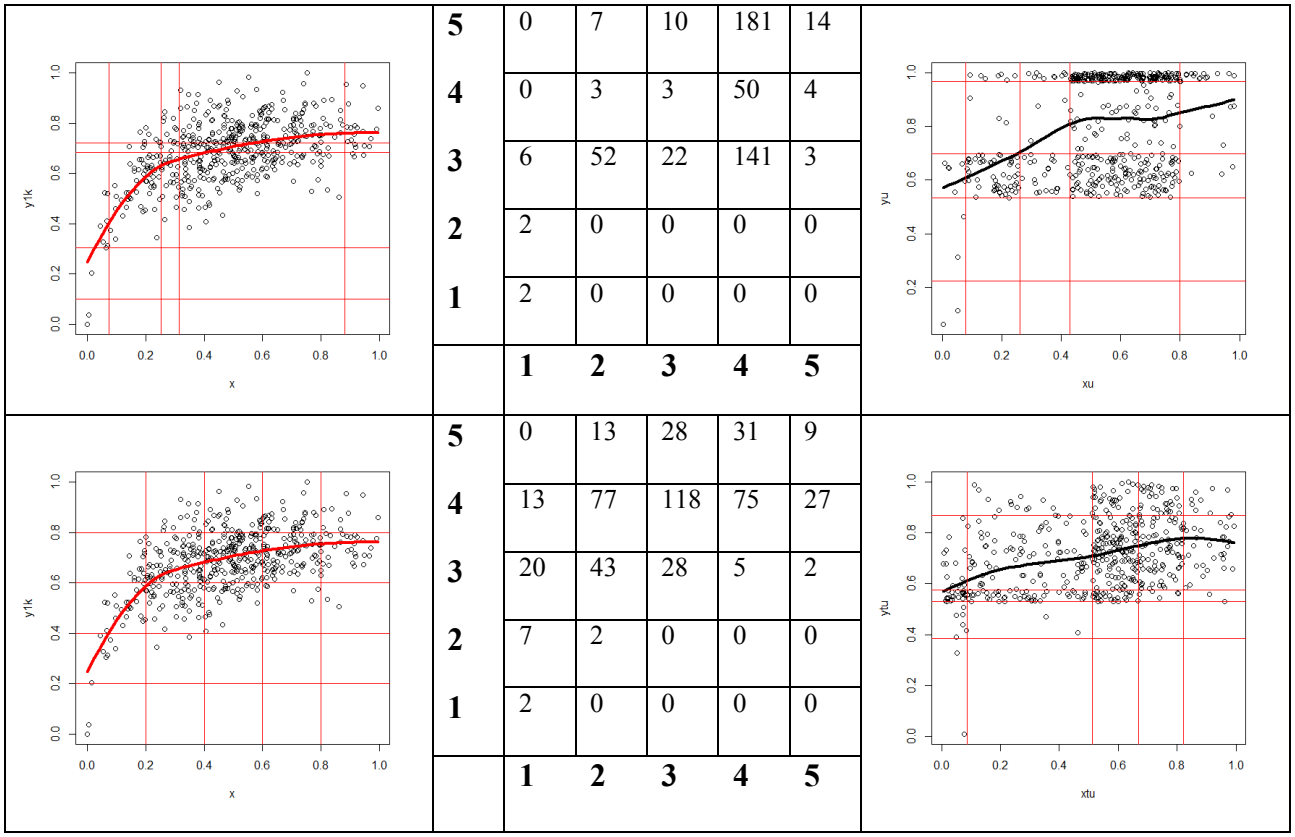
Kuvio 1 esittää simulointikokeissa käytetyt generoidut aineistot eri havaintomäärillä. Kuvioissa näkyvät havaintojen lisäksi datan generoimiseen käytetty funktio mustalla katkoviivalla sekä lokaalin regression sovite punaisella. Analyysissä selkeyden vuoksi myös y-muuttujat skaalataan välille 0–1.



Kuvio 1. Generoidut datat. Kuviossa generoidut datapisteet. Punainen viiva kuvaa lokaalin regression sovitetta ja musta katkoviiva funktionaalista muotoa.

Varsinainen simulointiprosessi alkaa valitsemalla jatkuvalla muuttujalle luokittelurajat. Luokittelurajat saadaan k luokkaiselle muuttujalle valitsemalla ja järjestämällä suuruusjärjestykseen k-1 kappaletta satunnaislukuja tasajakaumasta väliltä 0–1. Aineiston luokittelu satunnaisia rajoja käyttäen voi vaikuttaa ratkaisevasti menetelmän toimivuuteen. Tämän vuoksi tarkistellaan myös

tilannetta, jossa aineisto luokitellaan tasavälisesti, joka kuvaa keskimääräistä luokittelua. Jatkuva muuttuja luokitellaan k luokkaiseksi luokitelluksi muuttujaksi luokittelurajojen mukaisesti. Luokitelluille x ja y muuttujille lasketaan solufrekvenssit, joiden perusteella sama määrä havaintoja arvotaan uusien arvottujen luokittelurajojen määrittämälle alueelle ja näin laajennetaan aineisto taas jatkuvaksi. Tälle uudelle jatkuvalle aineistolle estimoidaan lokaalin regression sovite. Nämä simuloinnin vaiheet havainnollistetaan Kuviossa 2. Vaiheita toistetaan 1000 kertaa, joista voidaan laskea keskimääräinen lokaalin regression sovite ja tälle 95% luottamusväli.



5	0	7	10	181	14
4	0	3	3	50	4
3	6	52	22	141	3
2	2	0	0	0	0
1	2	0	0	0	0
	1	**2**	**3**	**4**	**5**

5	0	13	28	31	9
4	13	77	118	75	27
3	20	43	28	5	2
2	7	2	0	0	0
1	2	0	0	0	0
	1	**2**	**3**	**4**	**5**

Kuviossa datapisteet, punainen käyrä kuvaa lokaalin regression sovitetta ja punaiset viivat luokittelurajoja.

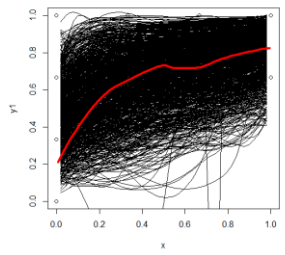
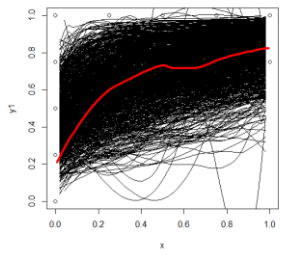
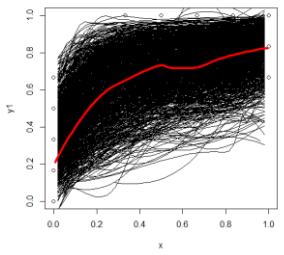
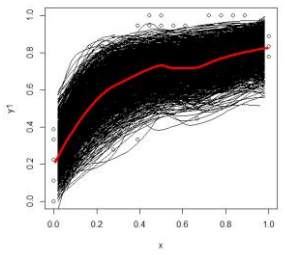
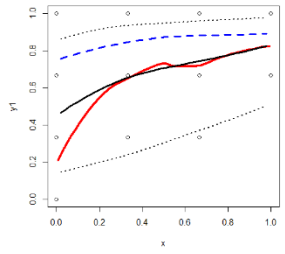
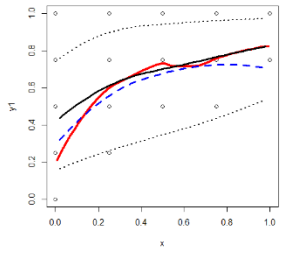
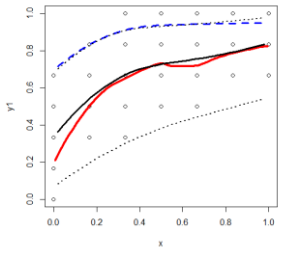
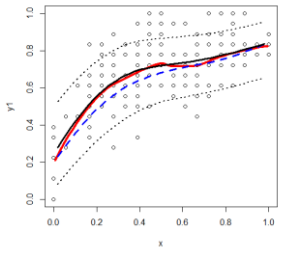
Kuvio 2. Jatkuvan aineiston satunnainen ja tasavälinen luokittelu ja sen augmentoiminen uudestaan jatkuvaksi sekä lokaalin regression sovitteet

Kuviossa 2 esitetään simulointiprosessia tilanteessa, jossa aineistoa luokitellaan viiteen luokkaan. Vasemman puoleisissa kuvissa esitetään alkuperäinen generoitu aineisto ja punaisella paksulla viivalla lokaalin regression sovite. Punaisella ohuella viivalla on merkitty arvotut luokittelurajat ylemmässä kuvassa ja alemmassa kuvassa luokittelurajat ovat tasavälisiä. Kuvion 2 keskellä on

esitetty solufrekvenssit luokitelluille muuttujille. Oikean puoleisimmissa kuvissa taas data on laajennettu uudestaan jatkuvaksi arpomalla tasajakaumista havaintoja uusien arvottujen luokittelurajojen määrittämien solujen sisään. Musta viiva kuvaa uuden arvotun aineiston lokaalin regression sovitetta.

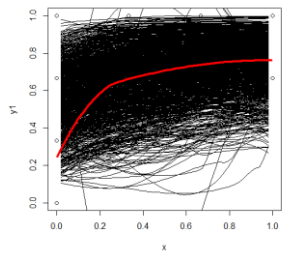
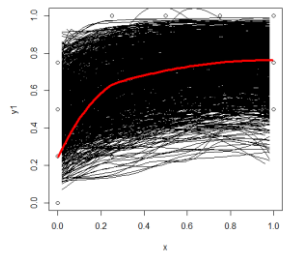
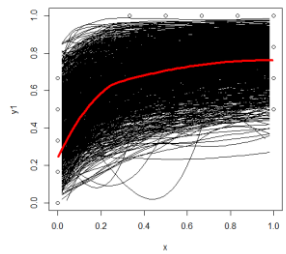
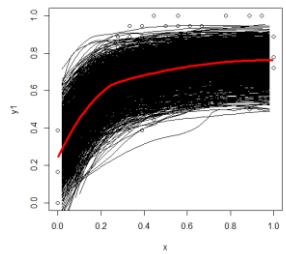
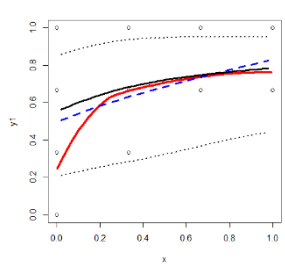
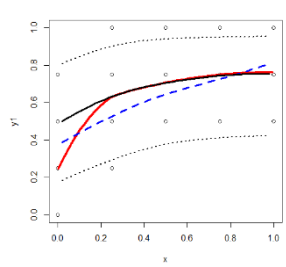
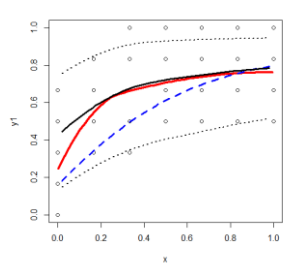
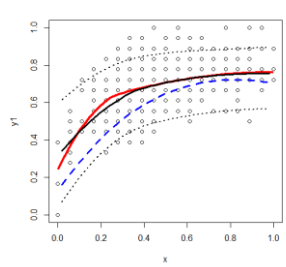
Simulaatiokokeita suoritetaan vaihtelemalla aineiston kokoa sekä luokitellun muuttujan luokkien määrää. Analyysejä tehdään 200, 500 ja 1000 havainnoin aineistoilla, koska nämä edustavat melko hyvin aineistoja joita kerätään ihmistieteissä. Luokitellun muuttujan luokkamääristä tarkastelussa ovat 4, 5, 7 ja 19. Kolme ensimmäistä edustavat tyypillisiä Likert-asteikkoja ja 19 luokkainen muuttuja taas esimerkiksi tilannetta jossa kolme seitsenportaista Likert-asteikollista muuttujaa muodostaa summamuuttujan.

Kuviot 3–11 esittävät simulaatiokokeiden tulokset. Yhdessä kuviossa esitetään tulokset koskien tietynmuotoista funktionaalista yhteyttä eri luokkamäärillä tietyn kokoisella aineistolla. Kuvioissa 3–5 esitetään tulokset koskien eri kokoisia aineistoja logaritmisen funktionaalisen muodon tapauksessa. Kuvioissa 6–8 tarkastellaan U-kirjaimen muotoista yhteyttä ja Kuvioissa 9–11 monimutkaista funktionaalista muotoa. Kuvioihin on piirretty yläriville mustalla värillä 1000 lokaalin regression sovitetta ja punaisella on merkitty alkuperäisen aineiston lokaalin regression sovite. Kuvioiden alemman rivin kuvissa punainen käyrä kuvaa edelleen alkuperäisen aineiston lokaalin regression sovitetta. Sinisellä katkoviivalla kuvataan augmentoitujen aineistojen keskimääräinen lokaalin regression sovitetta tilanteessa, jossa aineisto on alun perin luokiteltu satunnaisilla luokittelurajoilla. Mustalla on merkitty luokiteltujen ja uudelleen jatkuvaksi augmentoitujen aineistojen keskimääräinen lokaalin regression sovite ja mustalla katkoviivalla 95% luottamusväli näille sovitteilla tilanteessa, jossa alkuperäinen aineisto luokiteltiin tasavälisesti. Keskimääräinen sovite on määritelty simuloitujen sovitteiden mediaaniksi kussakin tarkastelupisteessä ja luottamusväliä varten on poimittu 2,5% ja 97,5% kvantiilit. Keskimääräinen sovite ja luottamusvälit on tämän jälkeen vielä tasoitettu lokaalilla regressiolla (f=2/3, d=2).

4 luokkaa	5 luokkaa	7 luokkaa	19 luokkaa
			
			

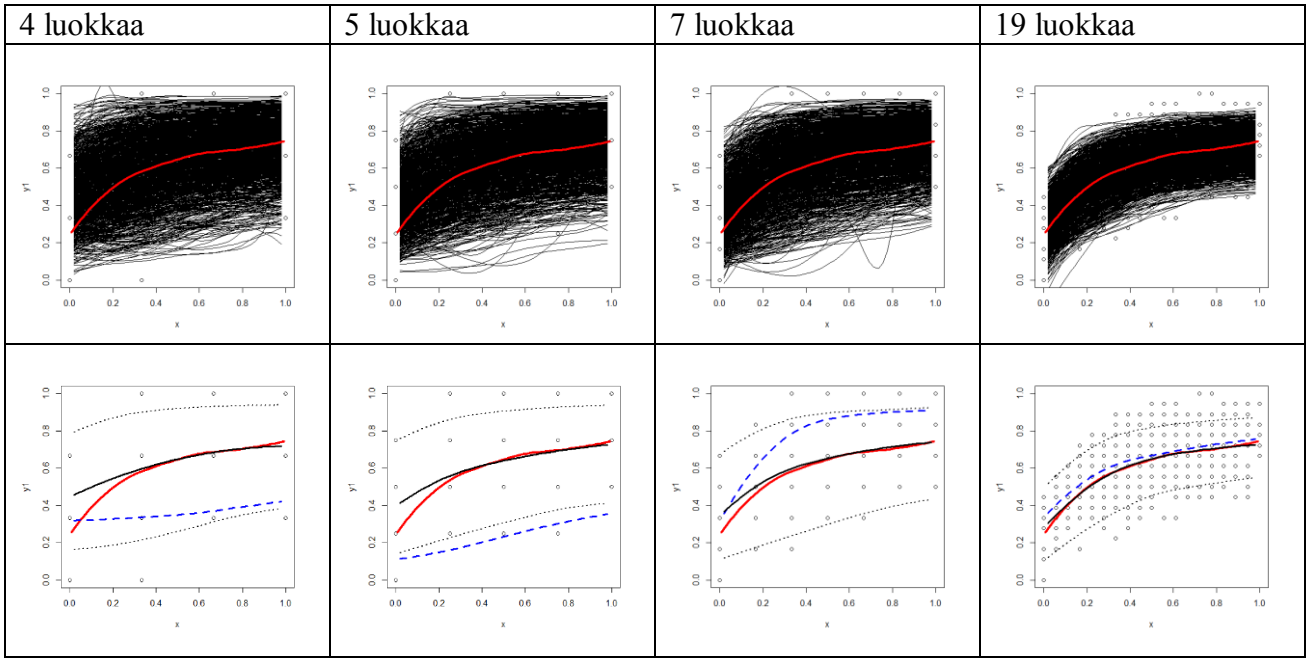
Kuvion merkinnät selitetty tekstissä.

Kuvio 3. Simulointikokeet (x & y1), n=200

4 luokkaa	5 luokkaa	7 luokkaa	19 luokkaa
			
			

Kuvion merkinnät selitetty tekstissä.

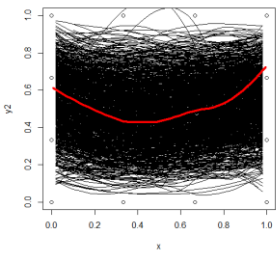
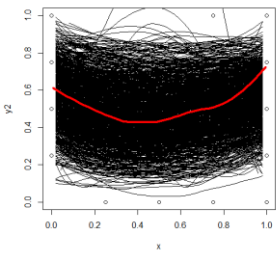
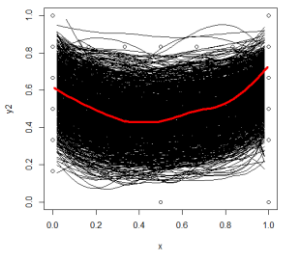
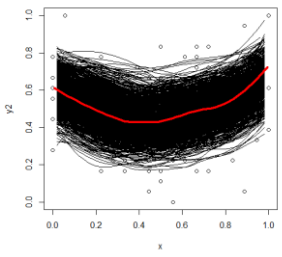
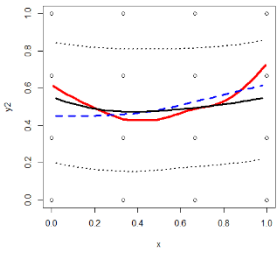
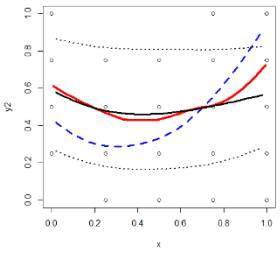
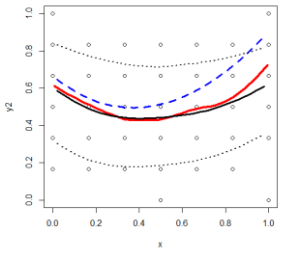
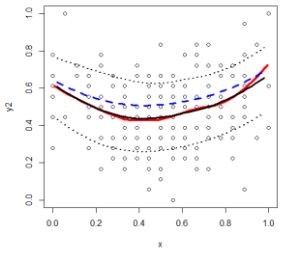
Kuvio 4. Simulointikokeet (x & y1), n=500



Kuvion merkinnät selitetty tekstissä.

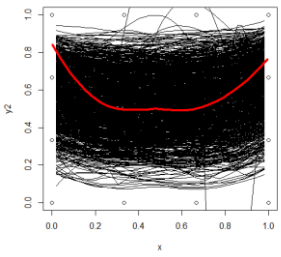
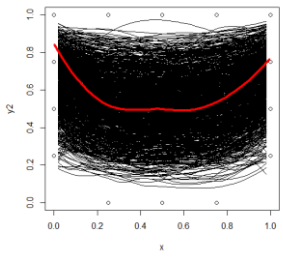
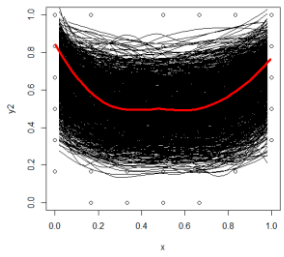
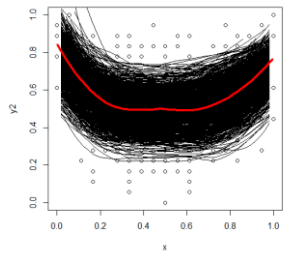
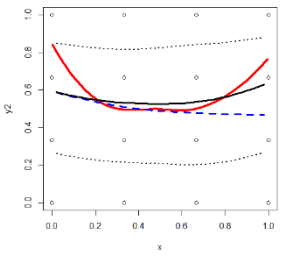
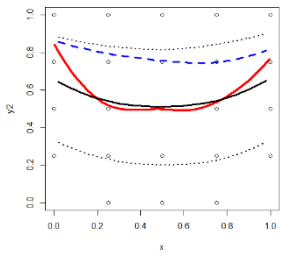
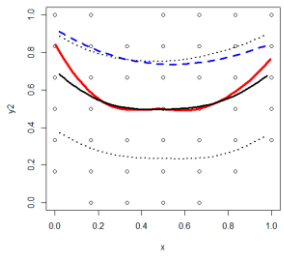
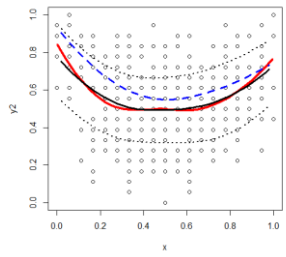
Kuvio 5. Simulointikokeet (x & y1), n=1000

Kuviot 3–5 osoittavat ensinnäkin, että alkuperäisen aineiston luokittelulla on suuri vaikutus tuloksiin. Joissakin tapauksissa satunnainen luokittelu on luonut tilanteita, joissa suurin osa havainnoista on luokiteltu esimerkiksi vain muutamaan pienimpään tai suurimpaan luokkaan, jolloin aineiston augmentaatioon perustuvan menetelmän sovite on selvästi liian korkealla tai matalalla tasolla. Jatkuvan muuttujan luokittelulla on siis suuri merkitys tuloksiin. Tasavälinen luokittelu kuvaa keskimääräistä luokittelua ja sen antamat tulokset antavat siten yleistettävämpiä tuloksia, joiden tulkintaan jatkossa keskitytään enemmän. Tuloksista huomataan, että luokitellun muuttujan luokkien lukumäärä on aineiston kokoa tärkeämpi tekijä. Aineiston koko ei tunnu vaikuttavan lainkaan tai vain vähän lopulliseen simuloinnin tuloksena saatuun keskimääräiseen lokaalin regression sovitteen tarkkuuteen. Sen sijaan luokitellun muuttujan luokkien lukumäärä vaikuttaa merkittävästi keskimääräisen sovitteen tarkkuuteen. Neljäluokkainen luokitus ei löydä käyräviivaista muotoa, mutta viisi- ja seitsenluokkainen toimivat paremmin. 19-luokkainen muuttuja löytää funktionaalisen muodon hyvin. Ongelma analyysissä voi olla, että aineistoa on jopa 1000 havainnon tapauksessa vain vähän siellä missä analyysin kannalta olisi mielenkiintoista. Usein käyräviivaiset ilmiöt tapahtuvat juuri skaalojen ääripäädyissä, joissa havaintoja on usein vain vähän.

4 luokkaa	5 luokkaa	7 luokkaa	19 luokkaa
			
			

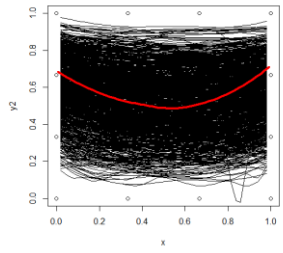
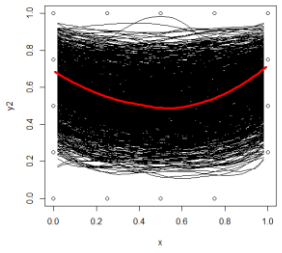
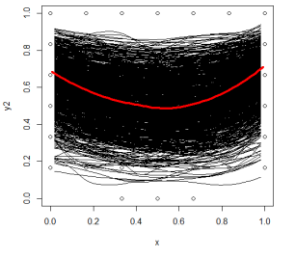
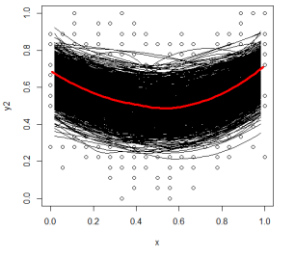
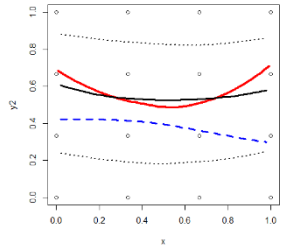
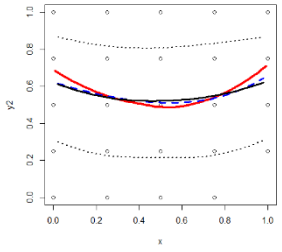
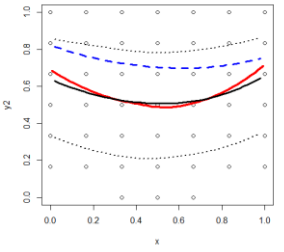
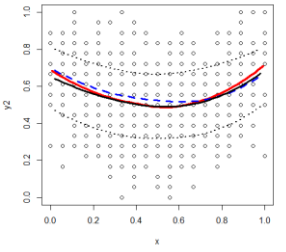
Kuvion merkinnät selitetty tekstissä.

Kuvio 6. Simulointikokeet (x & y2), n=200

4 luokkaa	5 luokkaa	7 luokkaa	19 luokkaa
			
			

Kuvion merkinnät selitetty tekstissä.

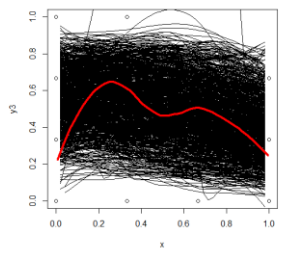
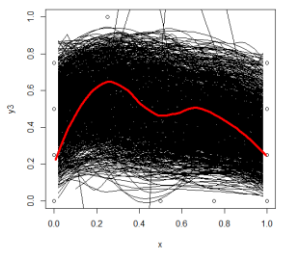
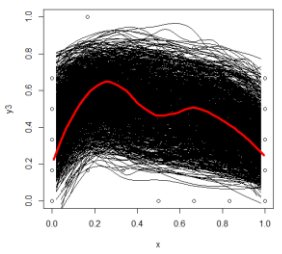
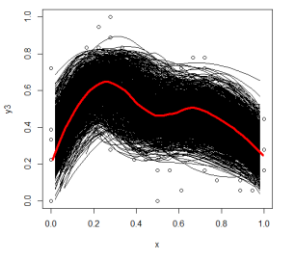
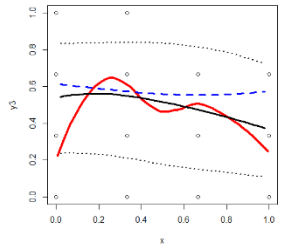
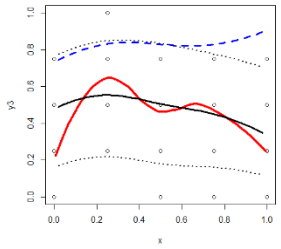
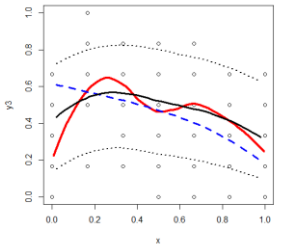
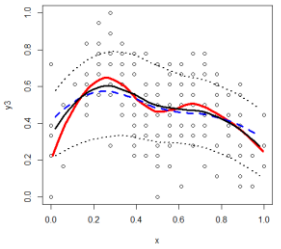
Kuvio 7. Simulointikokeet (x & y2), n=500

4 luokkaa	5 luokkaa	7 luokkaa	19 luokkaa
			
			

Kuvion merkinnät selitetty tekstissä.

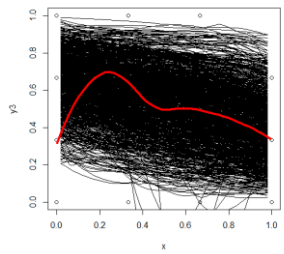
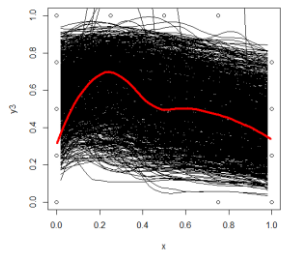
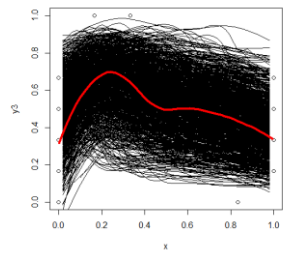
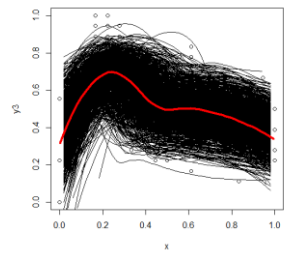
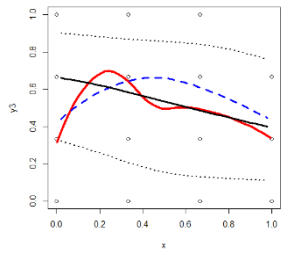
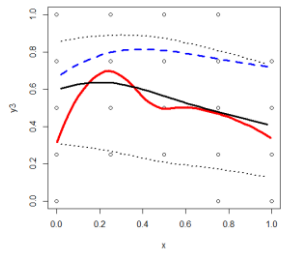
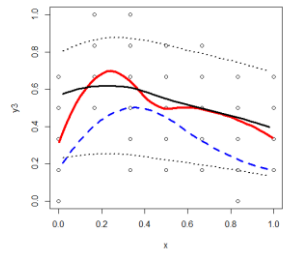
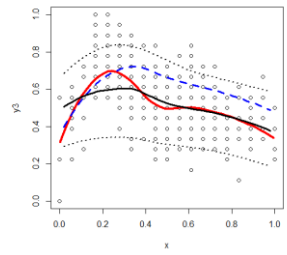
Kuvio 8. Simulointikokeet (x & y2), n=1000

U-kirjaimen muotoinen yhteys löytyy paremmin kuin logaritminen, joka johtuu varmastikin siitä, että käyräviivaisuus sijoittuu x-muuttujalla skaalan puoliväliin, jossa on paljon havaintoja. Jo viisi- ja seitsenportaiset luokittelevat muuttujat osuvat lähelle oikeaa funktionaalista muotoa ja vähintään viittaavat U-kirjaimen muotoon ja 19 luokkainen muuttuja on käytännössä identtinen alkuperäisen sovitteen kanssa.

4 luokkaa	5 luokkaa	7 luokkaa	19 luokkaa
			
			

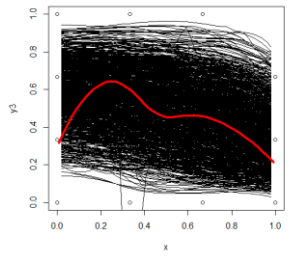
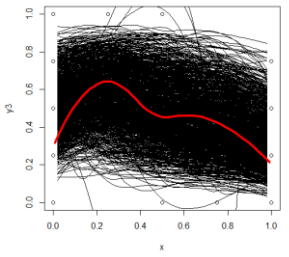
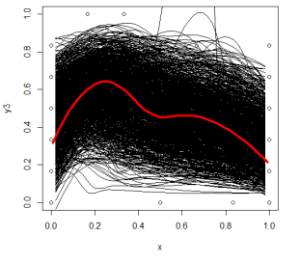
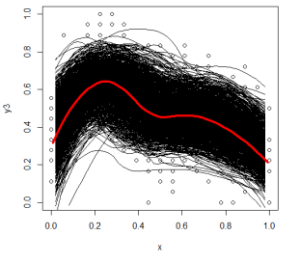
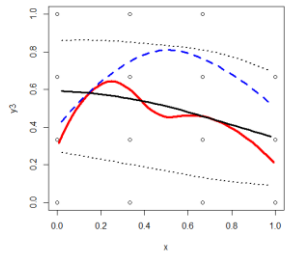
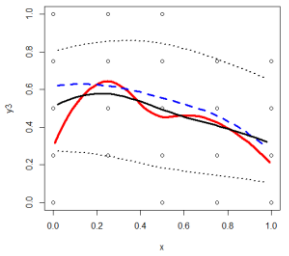
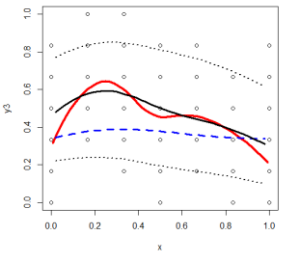
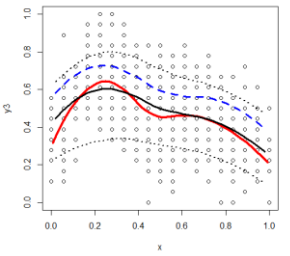
Kuvion merkinnät selitetty tekstissä.

Kuvio 9. Simulointikokeet (x & y3), n=200

4 luokkaa	5 luokkaa	7 luokkaa	19 luokkaa
			
			

Kuvion merkinnät selitetty tekstissä.

Kuvio 10. Simulointikokeet (x & y3), n=500

4 luokkaa	5 luokkaa	7 luokkaa	19 luokkaa
			
			

Kuvion merkinnät selitetty tekstissä.

Kuvio 11. Simulointikokeet (x & y3), n=1000

Kuten oli oletettavaakin, monimutkaisin funktionaalinen muoto osoittautui menetelmälle haastavimmaksi. Edes 19 luokkainen muuttuja ei löydä kunnolla funktionaalista muotoa.

5. Empiirinen esimerkki

Empiirisessä esimerkissä tarkastellaan kahden Likert-asteikolla mitatun muuttujan välistä yhteyttä. Analyysin kohteena on työhyvinvoinnin ja –motivaation tutkimus ja tässä tapauksessa työn kyynistymisen ja työhön uppoutumisen välinen yhteys. Tässä yhteydessä tarkastellaan kahden yksittäisen osion välistä yhteyttä lokaalia regressioanalyysiä hyödyntäen. Vastemuuttujana toimii muuttuja: *”Kun työskentelen, työ vie minut mukanaan”*, joka osa suomenkielistä validoitua työn imun mittaria (Utrecht Work Engagement Scale) ja kuuluu tarkemmin sanottuna sen uppoutumisen aladimensioon (Seppälä ym., 2009). Työn imu on suhteellisen uusi työhyvinvoinnin käsite, jonka avulla mitataan aitoa positiivista työhyvinvointia eikä vain pahoinvoinnin puutetta.

Selittävänä muuttuja analyysissä on: *”Huomaan, että minun on vaikea eläytyä asiakkaitteni tarpeisiin tai muiden työni kohteena olevien ihmisten ongelmiin”*, joka on taas yksi osio kyynistymisen kysymyspatteristosta, joka on osa laajempaa validoitua suomenkielistä työuupumuksen mittaria (Bergen Burnout Inventory) (Näätänen ym., 2003). Työn imun ja työuupumuksen ajatellaan olevan toisiinsa liittyviä ilmiöitä, jotka eivät kuitenkaan ole toistensa vastakohtia (Hakanen, 2009). Onkin teoreettisesti järkevää olettaa kyynistymisen vähentävän työn imun kokemusta.

Muuttujaa *”Kun työskentelen, työ vie minut mukanaan”* mitattiin seitsenportaisella Likert-asteikolla:

0 = ei koskaan
1 = muutaman kerran vuodessa
2 = kerran kuussa
3 = muutaman kerran kuussa
4 = kerran viikossa
5 = muutaman kerran viikossa
6 = päivittäin.

Vastaajia pyydettiin arvioida muuttujaa *”Huomaan, että minun on vaikea eläytyä asiakkaitteni tarpeisiin tai muiden työni kohteena olevien ihmisten ongelmiin”* kuusiportaisella Likert-asteikolla:

1 = täysin eri mieltä

2 = eri mieltä

3 = osittain eri mieltä

4 = osittain samaa mieltä

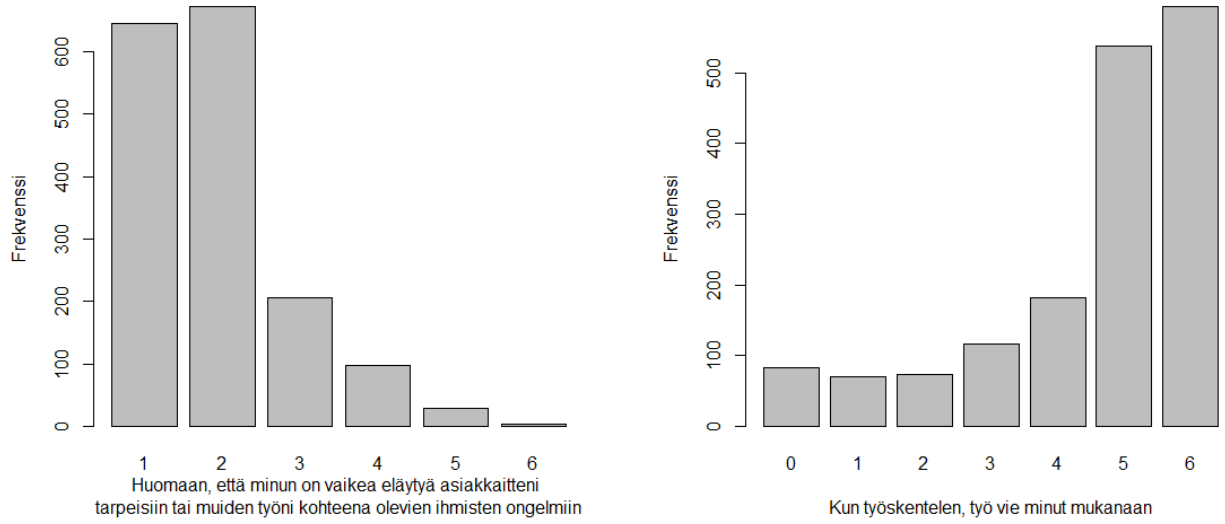
5 = samaa mieltä

6 = täysin samaa mieltä.

5.1. Aineiston kuvaus

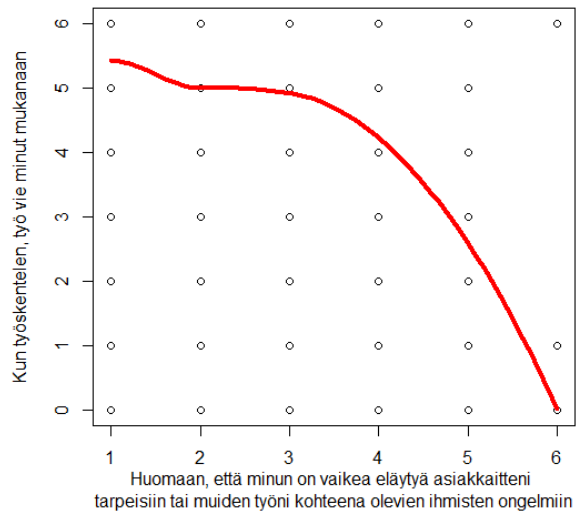
Tutkimusaineisto on kerätty Vaasan yliopiston LÄIKE-tutkimushankkeen yhteydessä vuosina 2011 ja 2012. Tutkimushankkeen teemana oli tarkastella lähijohtamista, sen kehittämistä ja sen yhteyttä työntekijöiden hyvinvointiin sekä tuloksellisuuteen. Määrällistä tutkimusaineistoa kerättiin palvelusektorilta viidestä suuresta organisaatiosta sekä PK-yrityksistä. Kysely toteutettiin internet- ja paperikyselylomakkeiden muodossa ja koska työntekijöitä pyydettiin arvioimaan esimiehiään, vastausten luottamuksellisuutta korostettiin. Aineisto muodostui seuraavasti eri organisaatioiden vastaajista: vakuutusala (N=334, 19,6 %), kunnallinen varhaiskasvatus (N=364, 21,4 %), logistiikka (N=488, 28,7 %), PK-sektori (N=129, 7,6 %), kaupan ala (N=175, 10,3 %) ja rahoitusala (N=211, 12,4 %). Kokonaisuudessaan kyselyyn vastasi 1701 työntekijää. Työntekijöistä suurin osa (81,4 %) oli vakituisessa työsuhteessa ja keskimääräinen työskentelyaika nykyisessä työpaikassa oli 10,96 vuotta (kh=11,02 vuotta). Vastaajien keski-ikä oli 41,83 vuotta (kh=12,14 vuotta) ja suurin osa vastaajista oli naisia (68,3 %), joka selittyy palvelusektorin ja varsinkin varhaiskasvatuksen naisvaltaisuudella. Kyselylomakkeessa oli kysymyksiä mm. esimiehen toiminnasta, työn piirteistä, omasta hyvinvoinnista sekä suoriutumisesta.

Aineistosta poistettiin puuttuva tieto (N=48) ja tarkasteluun otettiin vain 1653 havaintoa, joissa niin vastemuuttujalla kuin selittäjälläkin oli havaittu arvo. Muuttujien havaitut frekvenssit näkyvät Kuviossa 11. Molemmat muuttujat ovat vinoja ja varsinkin selittävän muuttujan arvossa 6 on hyvin vähän havaintoja, vain 4 kappaletta. X-muuttujan (*Huomaan, että minun on vaikea eläytyä asiakkaitteni tarpeisiin tai muiden työni kohteena olevien ihmisten ongelmiin*) keskiarvo on 1,91 ja keskihajonta 0,97. Y-muuttujan (*Kun työskentelen, työ vie minut mukanaan*) keskiarvo taas on 4,56 ja keskihajonta 1,69.



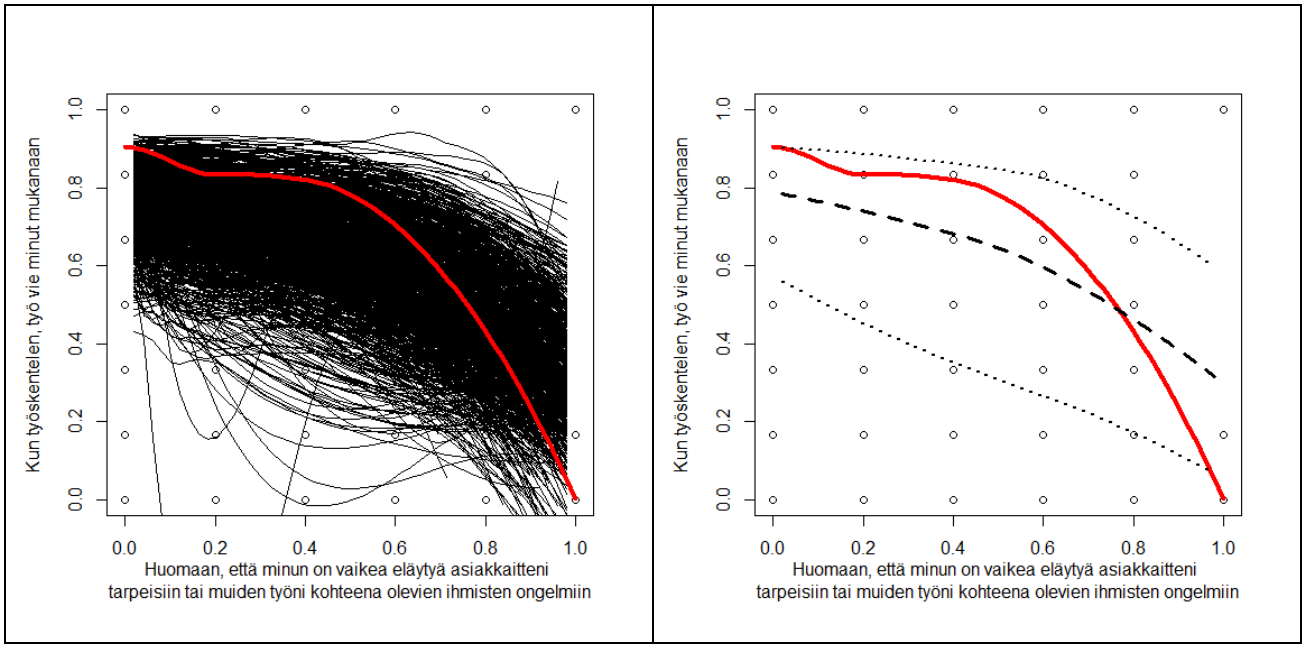
Kuvio 12. Muuttujien frekvenssit

5.2. Luokitellun aineiston analyysi



6	317	193	58	22	2	1
5	188	250	66	28	5	0
4	46	92	25	16	2	0
3	34	55	17	7	4	0
2	22	28	16	5	2	0
1	12	29	13	9	6	1
0	27	25	10	11	7	2
	1	**2**	**3**	**4**	**5**	**6**

Kuvio 13. Lokaalin regressioanalyysin sovite alkuperäisille luokitelluille muuttujille sekä havaintojen lukumäärät luokiteltujen muuttujien soluissa



Kuvio 14. Luokiteltujen muuttujien analyysi

Aineiston augmentaatioon perustuva menetelmä ehdottaa melko suoraviivaista yhteyttä muuttujien välillä, mutta toisaalta viitteitä kynnysarvoon on. Analyysi osoittaa, että kun kyynisyysoire kasvavat riittävän isoiksi, eli työntekijän on vaikea eläytyä asiakkaiden tarpeisiin ja työn kohteena olevien ihmisten ongelmiin, niin on työ ei enää myöskään vie työntekijää mukanaan, eli työn imun voidaan todeta laskevan. Ajoittaiset ja pienissä määrin ilmenevät työhön kyynistyminen ei näytä vaikuttavan työn imun kokemukseen vaan vasta suuri määrä kyynistymistä alentaa työn imevyyttä.

Tulosten tarkkuuteen vaikuttaa varmastikin datassa oleva suuri hajonta, jakaumien vinous ja havaintojen pieni määrä tietyissä luokissa. Simulointiin perustuva menetelmä antaa jossain määrin saman kaltaisen sovitteen muuttujien väliselle yhteydelle kuin alkuperäisten luokiteltujen muuttujine käyttäminen, jonka sovite on piirretty punaisella värillä Kuvioihin 13 ja 14.

6. Johtopäätökset

Tutkimuksessa esiteltiin datan augmentaatioon perustuva menetelmä sovittaa lokaali regressioanalyysi järjestysasteikollisille muuttujille. Simulointikokeita suoritettiin kolmen erilaisen funktionaalisen muodon estimoimiseksi. Muuttujien alkuperäisellä luokittelulla on suuri vaikutus tuloksiin ja jos luokittelun seurauksena esimerkiksi hyvin suuri osa havainnoista luokitellaan vain yhteen luokkaan ei augmentaatioon perustuva menetelmä vaikuta toimivan ainakaan systemaattisesti hyvin. Keskimäärin, eli tasavälisellä luokittelulla, tulokset olivat kuitenkin kohtuullisen hyviä. Simuloinnin tulokset osoittivat menetelmän tarkkuuden olevan kiinni eniten luokiteltujen muuttujien luokkien lukumäärästä. Aineiston koko tai yhteyden funktionaalisen muodon monimutkaisuus eivät olleet niin vaikuttavia tekijöitä, vaikka hyvin monimutkaiseen funktionaaliseen muotoon menetelmä ei antanut hyvää sovitetta. Toisaalta usein käyräviivaiset yhteydet ovat varsinkin ihmistieteissä aika yksinkertaisia, joten menetelmän pitäisi usein toimia. Viisi- ja varsinkin seitsemän-luokkainen järjestysasteikollinen muuttuja antoi useaan käyttötarpeeseen riittävän hyvän approksimaation yhteyden funktionaalisesta muodosta ja ylipäätään tarkkuuden kannalta on parempi mitä enemmän luokkia on. Likert-asteikolliset muuttujat sisältävät usein juurikin viisi tai seitsemän luokkaa, joten menetelmää voidaan käyttää useassa ihmistieteen sovelluksessa. Summamuuttujat pitävät tyypillisesti sisällään jo niin monta luokkaa, että estimaatin tarkkuus on jo hyvin suuri.

Empiirinen esimerkki osoitti, että suuri hajonta aineistoissa ja muuttujien vinous heikentää analyysin tarkkuutta. Nyt aineistoa on vain vähän sillä alueella, jolla käyräviivaisuus tapahtuu, joka heikentää analyysin onnistumista. Silti funktionaalinen muoto kuvaa käyräviivaista trendiä, joka näkyy myös alkuperäisille järjestysasteikollisille muuttujille tehdyssä analyysissä. Nämä tulokset tukevat toisiaan.

Empiirinen aineisto on ihmistieteissä usein sotkuista ja karkeaa. Hajonta voi olla hyvin isoa ja efektit pieniä. Myös aineisto voi olla hyvin epätasaisesti jakautunut ja sitä voi olla jakaumien hännissä vain vähän, jossa käyräviivaisuus näyttäytyy. Vaikka voidaan olettaa, ettei aineiston augmentaatioon perustuva menetelmä lokaalin regression sovitukseen yllä tarkkoihin estimaatteihin realistisilla empiirisillä aineistoilla, niin menetelmällä saavutetaan kuitenkin riittävä tarkkuus erottamaan ensinnäkin onko yhteys lineaarinen vai käyräviivainen ja myös antamaan karkean kuvauksen

mahdollisesta käyräviivaisesta muodosta.

Tämä tutkimus huomioi simulaatiokokeissa menetelmän toimivuuden erilaisissa tilanteissa riippuen riippuvuuden funktionaalisen muodosta, aineiston koosta ja järjestysasteikollisen muuttujan luokkien lukumäärästä. Tutkimuksen vahvuutena voidaan pitää varsinkin erilaisten funktionaalisten muotojen tutkimista. Kaksi muodoista ovat tyypillisiä ihmistieteiden teorioille ja kolmas monimutkainen muoto taas laajentaa ymmärrystä yleisempään tapaukseen. Kuitenkin jatkotutkimuksissa olisi hyvä ottaa huomioon enemmän erilaisia tekijöitä ja selvittää menetelmän toimivuuden reunaehtoja. Empiirinen esimerkki viittaa, että muuttujien suurella hajonnalla, epätasaisella jakaumalla ja varsinkin havaintojen puutteella siinä data-avaruuden alueella, jossa käyräviivaisuus tapahtuu, on suuri vaikutus analyysin tarkkuuteen. Simulointeja olisi hyvä tehdä myös erikokoisille efekteille. Aineiston augmentoinnissa voisi käyttää muita kuin tasajakaumia realistisemman tuloksen saamiseksi. Lisäksi muita lokaalin regression parametrien vaikutusta, kuten naapuruston kokoa, painofunktiota ja lokaakin regression astetta, voisi tutkia simulointikokeilla.

7. Lähteet

Beck, Nathaniel & Jackman, Simon (1998) Beyond Linearity by Default: Generalized Additive Models. *American Journal of Political Science*, 42(2), 596–627.

Bryman, Alan (2012) Social Research Methods, 4th Edition. Oxford: Oxford University Press.

Cleveland, William S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74 (368), 829–836.

Grant, Adam M., & Schwartz, Barry (2011) Too Much of a Good Thing: The Challenge and opportunity of the Inverted U. *Perspectives on Psychological Science*, 6(1), 61–76.

Hakanen, Jari (2009) Työn imua, tuottavuutta ja kukoistavia työpaikkoja? - kohti laadukasta työelämää. Työterveyslaitos.

Heitjan, Daniel F. & Rubin, Donald B. (1991) Ignorability and Coarse Data. *The Annals of Statistics*, 19(4), 2244–2253.

Hirvonen, Tatu & Mangeloja, Esas (2006) Miksi kolmas hampurilainen ei tee onnelliseksi? Jyväskylä: Atena.

Härdle, Wolfgang (1990) Applied nonparametric regression. Cambridge: Cambridge University Press.

Loader, Catherine (2004) Smoothing: Local Regression Tehniques. Teoksessa James E. Gentle, Wolfgang Härdle & Yuichi Mori (toim.) Handbook of Computational Statistics: Concepts and Methods. New York: Springer.

Näätänen, Petri, Aro, Antti, Matthiesen, Stig & Salmela-Aro, Katariina (2003) Bergen Burnout Indicator 15. Helsinki: Edita.

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Seppälä, Piia, Mauno, Saija, Feldt, Taru, Hakanen, Jari, Kinnunen, Ulla, Tolvanen, Asko & Schaufeli, Wilmar (2009) The construct validity of the Utrecht work engagement scale: Multisample and longitudinal evidence. Journal of Happiness Studies, 10, 459-481.

Liite A: R-koodi generoiduille aineistoille

Funktion parametrit

sim = iteraatioiden määrä (oletusarvoisesti 1000)
xl = satunnaisesti luokiteltu x-muuttuja
yl = satunnaisesti luokiteltu y-muuttuja
xlt = tasavälisesti luokiteltu x-muuttuja
ylt = tasavälisesti luokiteltu y-muuttuja
x = jatkuva x-muuttuja (pitää olla välillä 0–1)
y = jatkuva y-muuttuja (pitää olla välillä 0–1)
xl_min = valittu alaraja luokitellulle x-muuttujalle (esim. 1 tai 0)
xl_max = valittu yläraja luokitellulle x-muuttujalle (esim. 4, 5 tai 7)
yl_min = valittu alaraja luokitellulle y-muuttujalle (esim. 1 tai 0)
yl_max = valittu alaraja luokitellulle y-muuttujalle (esim. 4, 5 tai 7)
xnimi = kuvioihin x-akselille nimi
ynimi = kuvioihin y-akselille nimi
mnimi = kuvioihin otsikko

```
augmentaatio <- function(sim, xl,yl, xlt, ylt, x, y, xl_min, xl_max, yl_min, yl_max,
xnimi="x",ynimi="y",mnimi=" ") {
xpred<-seq(0,1,length=50)                    # pisteet joihin tehdään ennusteet
kayraty<-array(0,dim=c(sim,50))
kayratyt<-array(0,dim=c(sim,50))
xl_lkm <- xl_max - xl_min + 1                # x-muuttujan luokkien lkm
yl_lkm <- yl_max - yl_min + 1                # y-muuttujan luokkien lkm


# satunnaisesti luokitellut muuttujat
xlf <- factor(xl, levels=c(xl_min:xl_max))
ylf <- factor(yl, levels=c(yl_min:yl_max))
lkm<-table(xlf,ylf)                          # havaintojen lukumäärät eri soluissa
xl01 <- (xl - xl_min) / (xl_max - xl_min)    # laitetaan luokitellut muuttujat 0-1 välille
yl01 <- (yl - yl_min) / (yl_max - yl_min)
```

```
#tasavälisesti luokitellut muuttujat
xltf <- factor(xlt, levels=c(xl_min:xl_max))
yltf <- factor(ylt, levels=c(yl_min:yl_max))
lkmt<-table(xltf,yltf)
xlt01 <- (xlt - xl_min) / (xl_max - xl_min)          # laitetaan luokitellut muuttujat 0-1 välille
ylt01 <- (ylt - yl_min) / (yl_max - yl_min)

x11()
plot(xlt01,ylt01,main="",xlab=xnimi, ylab=ynimi)

for(k in 1:sim) {                                    #simulointi
x_rajat <- c(0, sort(runif(xl_lkm-1, 0,1)), 1)       # arvotaan augmentaatiorajat
y_rajat <- c(0, sort(runif(yl_lkm-1, 0,1)), 1)

# satunnaisesti luokiteltu data
xu<-NULL                                             #datan augmentointi
yu<-NULL
for(i in 1:xl_lkm){
for(j in 1:yl_lkm){
x0<-runif(lkm[i,j], min=x_rajat[i], max= x_rajat[i+1])
y0<-runif(lkm[i,j], min=y_rajat[j], max= y_rajat[j+1])
xu<-c(xu,x0)
yu<-c(yu,y0)
}
}
m <- loess(yu ~ xu,degree=2,span=2/3)                # LOESS-sovite augmentoidulle aineistolle
a<-predict(m, data.frame(xu=xpred))                  # ennusteet
kayraty[k,]<-a                                       #y-ennusteet talteen


# tasavälisesti luokiteltu data
xtu<-NULL
ytu<-NULL
```

```
for(i in 1:xl_lkm){
for(j in 1:yl_lkm){
xt0<-runif(lkmt[i,j], min=x_rajat[i], max= x_rajat[i+1])
yt0<-runif(lkmt[i,j], min=y_rajat[j], max= y_rajat[j+1])
xtu<-c(xtu,xt0)
ytu<-c(ytu,yt0)
}
}
mt <- loess(ytu ~ xtu,degree=2,span=2/3)
at<-predict(mt, data.frame(xtu=xpred))
lines(xpred,at)
kayratyt[k,]<-at
}

lines(loess.smooth(x,y,degree=2),col=2,lwd=4)          #alkuperäisen jatkuvan aineiston sovite

yq<-apply(kayraty, 2, quantile, probs=c(0.025,0.5,0.975),na.rm=TRUE)
yqt<-apply(kayratyt, 2, quantile, probs=c(0.025,0.5,0.975),na.rm=TRUE)

x11()
plot(xlt01,ylt01,main=mnimi,xlab=xnimi, ylab=ynimi)
lines(loess.smooth(x,y,degree=2),col=2,lwd=4)          #alkuperäisen jatkuvan aineiston sovite
#satunnaisesti luokiteltu data
#lines(loess.smooth(xpred,yq[1,],degree=2),lty=2,col=4,lwd=2)          #alaraja
#lines(loess.smooth(xpred,yq[3,],degree=2),lty=2,col=4,lwd=2)          #ylärajar
lines(loess.smooth(xpred,yq[2,],degree=2),lty=2,col=4,lwd=3)          #mediaani
#tasavälisesti luokiteltu data
lines(loess.smooth(xpred,yqt[1,],degree=2),lty=3,col=1,lwd=2)         #alaraja
lines(loess.smooth(xpred,yqt[3,],degree=2),lty=3,col=1,lwd=2)         #ylärajar
lines(loess.smooth(xpred,yqt[2,],degree=2),col=1,lwd=3)               #mediaani
}
```

Liite B: R-koodi empiiristen järjestysasteikollisten muuttujien analysointiin

sim = iteraatioiden määrä (oletusarvoisesti 1000)
xl = järjestysasteikollinen x-muuttuja
yl = järjestysasteikollinen y-muuttuja
xl_min = järjestysasteikollinen x-muuttujan alaraja
xl_max = järjestysasteikollinen x-muuttujan alaraja
yl_min = järjestysasteikollinen y-muuttujan alaraja
yl_max = järjestysasteikollinen y-muuttujan alaraja
xnimi = kuvioihin x-akselille nimi
ynimi = kuvioihin y-akselille nimi
mnimi = kuvioihin otsikko

```
augmentaatio <- function(sim, xl,yl, xl_min, xl_max, yl_min, yl_max,
xnimi="x",ynimi="y",mnimi=" ") {
xpred<-seq(0,1,length=50)                    # pisteet joihin tehdään ennusteet
kayraty<-array(0,dim=c(sim,50))
kayratyt<-array(0,dim=c(sim,50))
xl_lkm <- xl_max - xl_min + 1
yl_lkm <- yl_max - yl_min + 1

xlf <- factor(xl, levels=c(xl_min:xl_max))
ylf <- factor(yl, levels=c(yl_min:yl_max))
lkm<-table(xlf,ylf)                          # havaintojen lukumäärä eri soluissa
xl01 <- (xl - xl_min) / (xl_max - xl_min)    # laitetaan luokitellut muuttujat 0-1 välille
yl01 <- (yl - yl_min) / (yl_max - yl_min)

x11()
```

```
plot(xl01,yl01,main="",xlab=xnimi, ylab=ynimi)


for(k in 1:sim) {                                   #simulointi
x_rajat <- c(0, sort(runif(xl_lkm-1, 0,1)), 1)      # arvotaan augmentaatiorajat
y_rajat <- c(0, sort(runif(yl_lkm-1, 0,1)), 1)

xu<-NULL                                            #datan augmentointi
yu<-NULL
for(i in 1:xl_lkm){
for(j in 1:yl_lkm){
x0<-runif(lkm[i,j], min=x_rajat[i], max= x_rajat[i+1])
y0<-runif(lkm[i,j], min=y_rajat[j], max= y_rajat[j+1])
xu<-c(xu,x0)
yu<-c(yu,y0)
}
}
m <- loess(yu ~ xu,degree=2,span=2/3)               # LOESS-sovite augmentoidulle aineistolle
a<-predict(m, data.frame(xu=xpred))                 # ennusteet valituissa pisteissä
lines(xpred,a)                                      # piirtää sovitteen
kayraty[k,]<-a                                      #y-ennusteet talteen
}

lines(loess.smooth(xl01,yl01,degree=2),col=2,lwd=4)          #luokitellun aineiston sovite

yq<-apply(kayraty, 2, quantile, probs=c(0.025,0.5,0.975),na.rm=TRUE)    #luottamusväli
x11()
plot(xl01,yl01,main=mnimi,xlab=xnimi, ylab=ynimi)
lines(loess.smooth(xl01,yl01,degree=2),col=2,lwd=4)          #luokitellun aineiston sovite
lines(loess.smooth(xpred,yq[1,],degree=2),lty=3,col=1,lwd=2)          #alaraja
lines(loess.smooth(xpred,yq[3,],degree=2),lty=3,col=1,lwd=2)          #ylärajar
lines(loess.smooth(xpred,yq[2,],degree=2),lty=2,col=1,lwd=3)          #mediaani
}
```