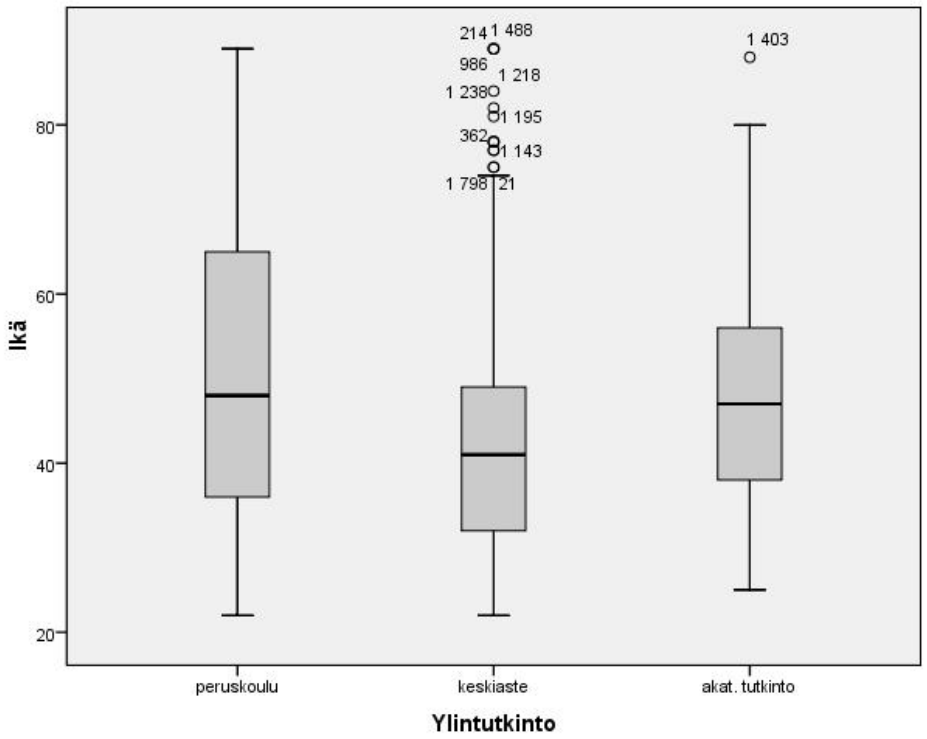
4. UUDEMPAA MALLINNUSTA VARIANSSIANALYYSIN TUEKSI

Varianssianalyysi (analysis of variance, ANOVA) on itse asiassa joukko erilaisia menetelmiä, joiden avulla yleensä tarkastellaan ryhmien välisiä eroja testimuuttujien osalta. Yksinkertaisin tilanne on sellainen, että tarkastelemme tiettyjen ehtojen vallitessa mitta-asteikollista testimuuttujaa, jonka arvot on luokiteltu kahteen tai useampaan luokkaan jonkin luokittelevan muuttujan perusteella. Pyrimme sitten selvittämään, onko tämä luokitus (eli ”käsittelyt”, treatments) tuottanut samanlaisia luokkia testimuuttujan osalta.

Käytännössä me tavallisesti vertaamme näiden luokkien keskiarvoja toisiinsa, ja jos keskiarvot ovat likimain samoja, luokkien tulkitaan olevan testimuuttujan osalta samanlaisia. Tässä on itse asiassa kyse yksisuuntaisesta (one-way) varianssianalyysistä, koska käytämme mallissamme vain yhtä luokittelevaa muuttujaa.

4.1. Yksisuuntainen ANOVA

Voter-aineistomme osalta testimuuttujana voi olla vaikkapa ikä ja luokittelevana muuttujana koulutustaso. Tavoitteena on tällöin selvittää, ovatko äänestäjät keskimäärin samanikäisiä kaikissa koulutustasoryhmissä. Jos tutkimme aluksi kuvan 4.1 box-plot –piirrosta, keskiasteen koulutuksen saaneet näyttävät keskimäärin olevan nuorimpia ja vain peruskoulun tason suorittaneet vanhimpia (suorakulmioiden paksut poikkiviivat ovat mediaaneja ja suorakulmioiden ala- ja ylärajat ovat vastaavasti ykkös- ja kolmoskvartiilit):



Kuva 4.1. Voter-aineiston äänestäjien iät koulutusryhmittäin (box-plot –piirros, numeroidut pisteet ovat ryhmistä selvästi poikkeavia tapauksia).

Samalta näyttää tilanne kuvan 4.2 keskiarvopiirroksen mukaan ja ne on esitetty myös taulussa 4.1.

Taulu 4.1. Voter-aineiston ikien keskiarvot koulutus-tasojen mukaan.

	N	Mean	Std. Deviation
peruskoulu	1136	50,46	17,570
keskiaste	518	42,20	13,033
akat. tutkinto	193	48,19	12,509
Total	1847	47,91	16,334

Varianssianalyysin avulla voidaan nyt selvittää, tulkitaanko nämä ryhmien keskiarvot yhtä suuriksi (nollahypoteesi), vai luovutaanko tästä yhtäsuuruuden olettamuksesta (vaihtoehtoinen hypoteesi).

Varianssianalyysi tehdään parametrisella testillä, jos ryhmien väliset varianssit ovat yhtä suuret tai ryhmien havaintoarvot ovat normaalisti jakautuneet (tai aineisto ei ole pieni). Vaikka nämä ehdot eivät tässä täytykään (ainakaan Levenen varianssitestin ja Shapiro-Wilkin normaalisuustestin perusteella), pedagogisista syistä johtuen käytämme kuitenkin ensin SPSS-ohjelman parametrista testiä, joka tuottaa seuraavan ANOVA-taulun (Analyze – Compare means – One-way ANOVA):

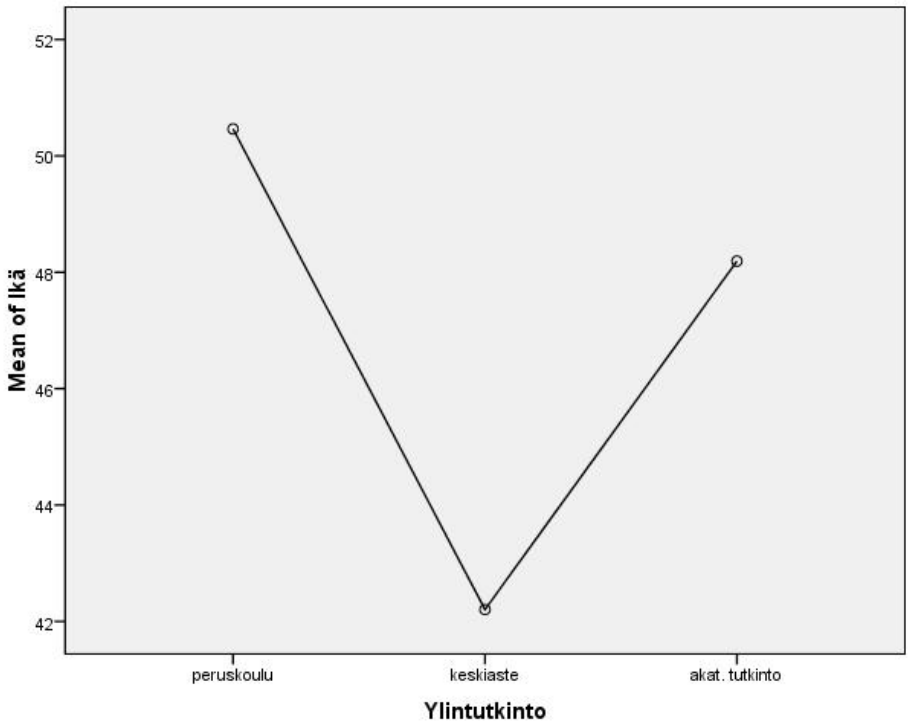
Taulu 4.3. Voter-aineiston ANOVA-taulu

Ikä

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	24281,238	2	12140,619	47,813	0,000
Within Groups	468228,663	1844	253,920		
Total	492509,900	1846			

ANOVA-taulun yleistestin perusteella p-arvo (Sig.) on niin pieni, että kaikkia keskiarvoja ei voida pitää yhtä suurina. Asian vahvistaa myös Brownin ja Forsythen testi, jota voidaan käyttää kun osajoukkojen varianssit eivät ole yhtä suuret. Niinpä ainakin pienin ja suurin keskiarvo, eli iät ryhmissä ”peruskoulu” ja ”keskiaste”, poikkeaisivat yleistestin perusteella toisistaan.

Koska ANOVA-yleistestin nollahypoteesi hylättiin, voimme selvittää ryhmien välisiä eroja tarkemmin vaikkapa Bonferronin ja Tamhanen parittaisten vertailujen testien avulla (pairwise comparison). Edellistä käytetään luokkien yhtä suurien varianssien tapauksissa, ja muulloin jälkimmäinen soveltuu paremmin (muitakin menetelmiä näihin on tarjolla). Tällöin todetaan, että 5 % merkitsevyystasolla parien ”peruskoulu” – ”keskiaste” ja ”keskiaste” – ”akateeminen tutkinto” tapauksessa keskiarvoja ei voida pitää yhtä suurina. Toisin sanoen, korkeintaan keskiasteen koulutuksen saaneet vastaajat olivat muita nuorempia kun taas kahden muun ryhmän vastaajat olivat käytännössä keskenään samanikäisiä (kuva 4.2).



Kuva 4.2. Voter-aineiston äänestäjien ikien aritmeettiset keskiarvot koulutusryhmittäin.

ANOVA-yleistesti suoritettiin myös SPSS:n Kruskall-Wallisin ei-parametrisella varianssianalyysillä, koska edellä mainitut havaintojen normaalisuus- ja varianssien yhtäsuuruusehdot eivät täyttyneet. Tämä testi tuotti kuitenkin saman johtopäätöksen kuin parametrinenkin yleistestimme:

Ranks

	Ylintutkinto	N	Mean Rank
Ikä	peruskoulu	1136	997,53
	keskiaste	518	745,27
	akat. tutkinto	193	970,90
	Total	1847	

Test Statistics[a,b]

	Ikä
Chi-Square	81,291
df	2
Asymp. Sig.	0,000

a. Kruskal Wallis Test

b. Grouping Variable: Ylin tutkinto

Yksisuuntaisessa varianssianalyysissä edetään tavallisesti siis edellä kuvatulla tavalla. Toisaalta tämä tarkastelu voidaan kuvitella myös mallinnukseksi, jossa luokitteleva muuttuja on syy- eli selittävä ja testimuuttuja seuraus- eli selitettävä muuttuja:

Koulutustaso → Ikä

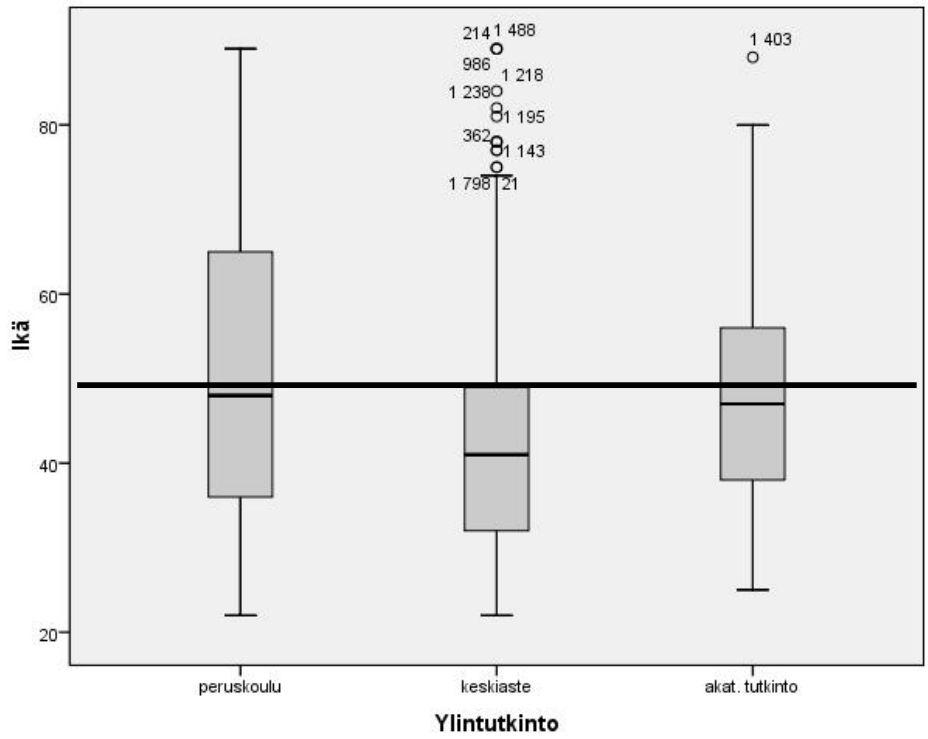
Jos ryhmien keskiarvojen katsotaan olevan samat (ANOVA:n yleistestin nollahypoteesin tilanne), kaikkien ryhmien osalta voidaan käyttää koko aineiston keskiarvoa, ja silloin mallimme lähtökohtana on yhtälö

Ikä = ikien keskiarvo + virhetermi

Jos taas mallimme perustuu ikien keskiarvoihin eri koulutustasoryhmissä (kun ryhmien ikien välillä on todettu olevan eroa), sovellamme jokaisen ryhmän osalta yhtälöä

Ikä = koulutustasoryhmän ikien keskiarvo + virhetermi

Kuvassa 4.3 on käytetty koko aineiston ikien keskiarvoa (47,9 vuotta) edustamaan kaikkia koulutustasoja (siis nollahypoteesin tilanne). Tällainen sovite-suora, jossa suoran kulmakerroin on nolla (eli vaakasuora viiva), ei käytännössä kuitenkaan ole toivottu ratkaisu ainakaan regressiomallinnuksessa.

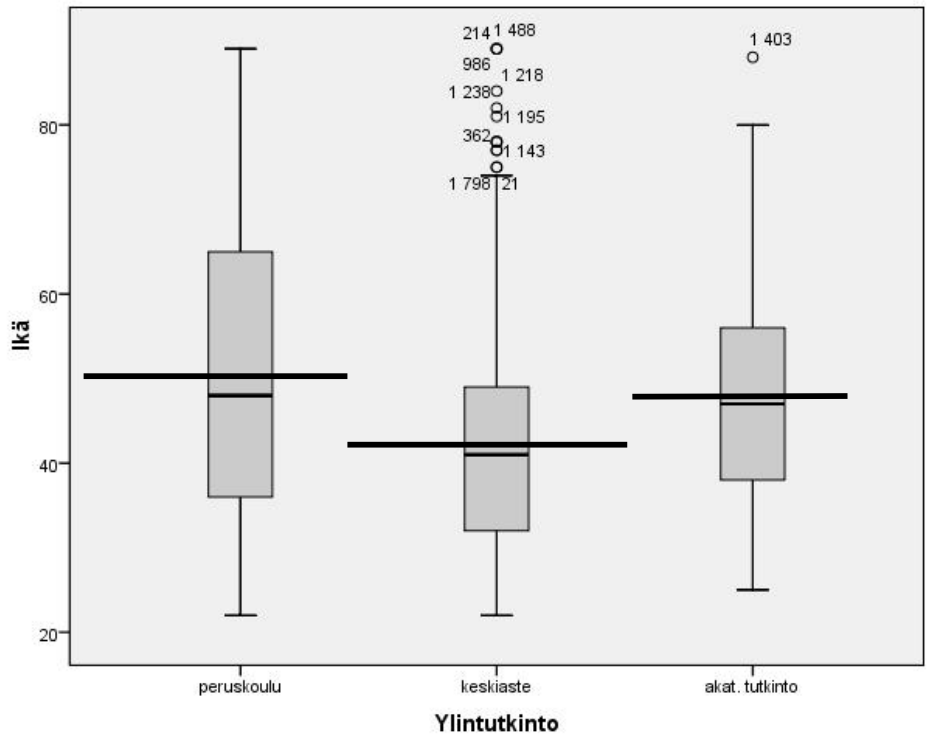


Kuva 4.3. Voter-aineiston mallin ”sovite” (vaakasuora viiva) kun selitettävänä on ikä ja selittäjänä koulutustaso. Mallissa käytetty jokaisen koulutustason osalta koko aineiston ikien keskiarvoa.

Parempi sovite saadaan käyttämällä koulutustasoluokkien keskiarvoja, ja näin tehdään erityisesti kun näiden keskiarvojen välillä on todettu tilastollisesti merkittäviä eroja. Niinpä mallimme rakennus perustuu nyt sääntöihin:

1. Jos on korkeintaan peruskoulutason käynyt , ikä on 50,5 vuotta.
2. Jos on korkeintaan keskiasteen tutkinto, ikä on 42,2 vuotta.
3. Jos on akateeminen tutkinto, ikä on 48,2 vuotta.

Vastaava sovitteemme on silloin kuvan 4.4 mukainen.

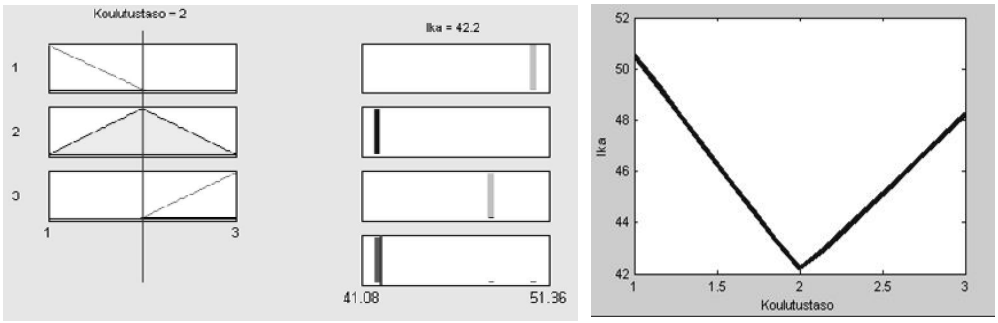


Kuva 4.4. Voter-aineiston mallin ”sovite” (vaakasuorat viivat) kun selitettävänä on ikä ja selittäjänä koulutustaso. Mallissa käytetty ikien ryhmäkeskiarvoja koulutustasoittain.

Olemme nyt siis itse asiassa olettaneet, että ikiä edustavat eri koulutustasoryhmissä niiden ryhmäkeskiarvot, eli meillä on kolme ikäklusteria, joiden keskuksia edustavat niiden keskiarvot. Niinpä tämä lähestymistapa, kun varianssianalyysi ajatellaan näin mallin rakentamiseksi, muistuttaakin jo paljon sumeiden sääntöjen avulla tehtyjä malleja. Me voimmekin käyttää edellä olevia sääntöjä suoraan vastaavassa sumeassa systeemissä, eli silloin mallimme rakennus perustuu sumeisiin sääntöihin

1. Jos on korkeintaan peruskoulun käynyt (=1), ikä on noin 50,5 vuotta.
2. Jos on korkeintaan keskiasteen tutkinto (=2), ikä on noin 42,2 vuotta.
3. Jos on akateeminen tutkinto (=3), ikä on noin 48,2 vuotta.

Kuvassa 4.5 on luonnosteltu vastaavaa sumeaa systeemiä kun on käytetty edellä mainittuja sääntöjä ja olemme soveltaneet Takagi-Sugenon 0. kertaluvun menetelmää (myös Mamdani-menetelmää olisi tietenkin voitu käyttää).

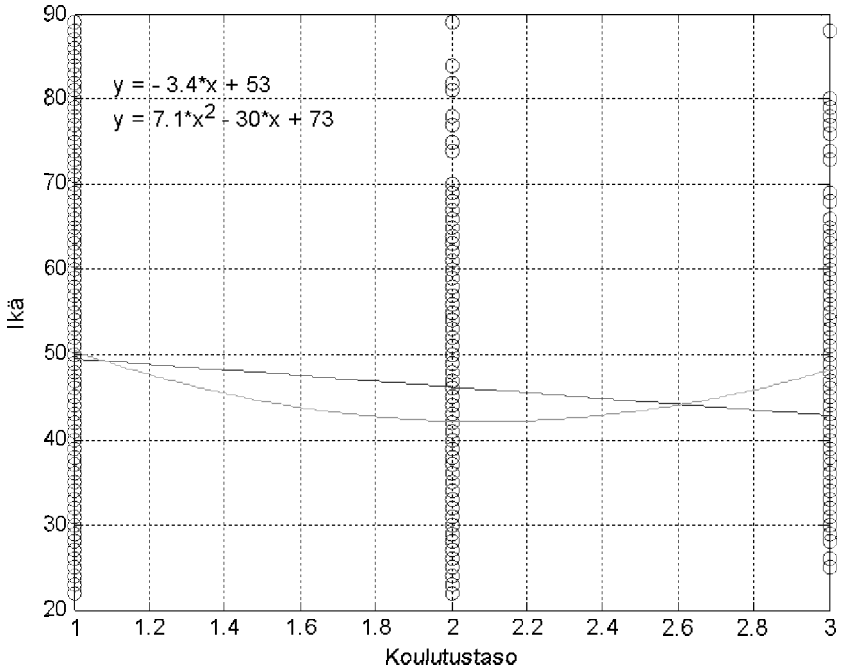


Kuva 4.5. Iän ryhmäkeskiarvojen perusteella tehty sumea mallinnus Voter-aineistosta kun koulutustaso on luokitteleva muuttuja.

Tässä sumeassa systeemissä on ikien ryhmäkeskuksina käytetty niiden keskiarvoja, mutta voimme rakentaa vastaavan mallin suoraankin kyseisistä muuttujista alkuperäisen havaintoaineiston perusteella, jolloin ryhmien keskukset voidaan periaatteessa määrittää muullakin tavalla (kuten sumealla ryhmittelyanalyysillä). Tässä tapauksessa saadaan ainakin Takagi-Sugenon 1. kertaluvun mallilla kolmea sääntöä käyttäen sama tulos kuin ylläkin, vaikka ryhmien keskukset onkin etsitty s*ubclust*-menetelmällä.

Jos taas teemme puhtaan matemaattisen mallin suoraan havaintoaineiston perusteella, saamme kuvassa 4.6 esitetyt sovitteet kun olemme käyttäneet lineaarista mallia (suora) ja toisen asteen polynomia. Kyseiset funktiot ovat vastaavasti (kun koulutustasot on koodattu 1-3).

$$\text{Ikä} = -3{,}4 \cdot \text{koulutustaso} + 53$$
$$\text{Ikä} = 7{,}1 \cdot \text{koulutustaso}^2 + 30 \cdot \text{koulutustaso} + 73$$



Kuva 4.6. Voter-aineiston matemaattisten mallien sovitteet kun selitettävänä on ikä ja selittäjänä koulutustaso (suora on lineaarisen ja käyrä ei-lineaarisen mallin sovite).

Edellä kuvailtu lähestymistapa auttaa meitä tarkastelemaan, miten luokitteleva muuttuja vaikuttaa testimuuttujaan. Varsinkin sumean mallin avulla voidaan helposti operoida silloinkin kun parametrisen ANOVA:n ehdot eivät täytykään. Tämä ajattelutapa auttaa myös kytkemään varianssianalyysin, kovarianssianalyysin ja regressioanalyysin toisiinsa, sillä tietyssä mielessä ensimmäisessä ovat selittäjinä luokittelumuuttujat, toisessa sekä luokittelu- että jatkuvat muuttujat ja kolmannessa jatkuvat muuttujat.

4.2. Kaksisuuntainen ANOVA

Yksisuuntaisen ANOVA:n tavoin voimme tuottaa kaksisuuntaisen varianssianalyysin (two-way ANOVA) malleja. Voter-aineiston tapauksessa me voimme selittää ikää vaikkapa koulutustason ja sukupuolen perusteella, jolloin saamme mallin

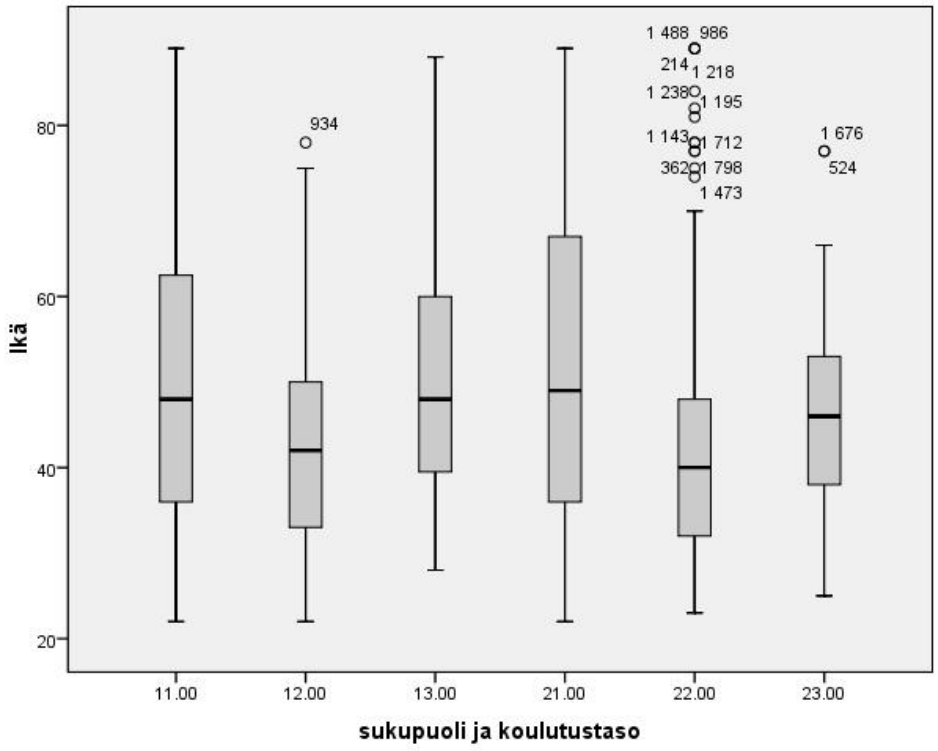
Koulutustaso →

Ikä

Sukupuoli →

Silloin ryhmämme ovat homogeenisempia kuin yksisuuntaisen ANOVA-mallin tapauksessa, ja meillä on kolmen ryhmän sijasta 2·3 = 6 ryhmää. Niinpä meillä on mahdollista saada luotettavampia tuloksia.

Jos sovellamme nyt parametristä ANOVA-mallinnusta, perinteinen tilastollinen mallinnus perustuu näihin kuuteen ryhmäkeskiarvoon siten, että vertailemme keskenään kahta sukupuolen ja kolmea koulutustasoluokan keskiarvoa. Jos lisäksi tutkimme näiden muuttujien yhdysvaikutusta (interaction), vertailemme keskenään kaikkia kuutta ryhmäkeskiarvoa yhtä aikaa. Kuvassa 4.7 on näiden kuuden ryhmän ”klusterit” box-plot –piirroksena, ja suorakulmiot eivät siinä näytä olevan aivan samalla tasolla vaakasuorassa suunnassa. Ovatko nämä erot sitten merkittäviä, selvitetään seuraavaksi testaamalla.



Kuva 4.7. Voter-aineiston äänestäjien iät sukupuolen mukaan ja koulutustasoluokittain.

Taulussa 4.4 ovat ikien keskiarvot näissä ryhmissä, ja näiden lukujen eroja siis tilastollisesti testataan.

Taulu 4.4. Dependent Variable: Ikä

RESPONPON-DENTS SEX	Ylintutkinto	Mean	Std. Deviation	N
mies	peruskoulu	49,40	16,670	488
	keskiaste	42,52	12,048	216
	akat. tutkinto	50,38	13,826	100
	Total	47,68	15,518	804
nainen	peruskoulu	51,26	18,190	648
	keskiaste	41,98	13,709	302
	akat. tutkinto	45,84	10,489	93
	Total	48,09	16,941	1043
Total	peruskoulu	50,46	17,570	1136
	keskiaste	42,20	13,033	518
	akat. tutkinto	48,19	12,509	193
	Total	47,91	16,334	1847

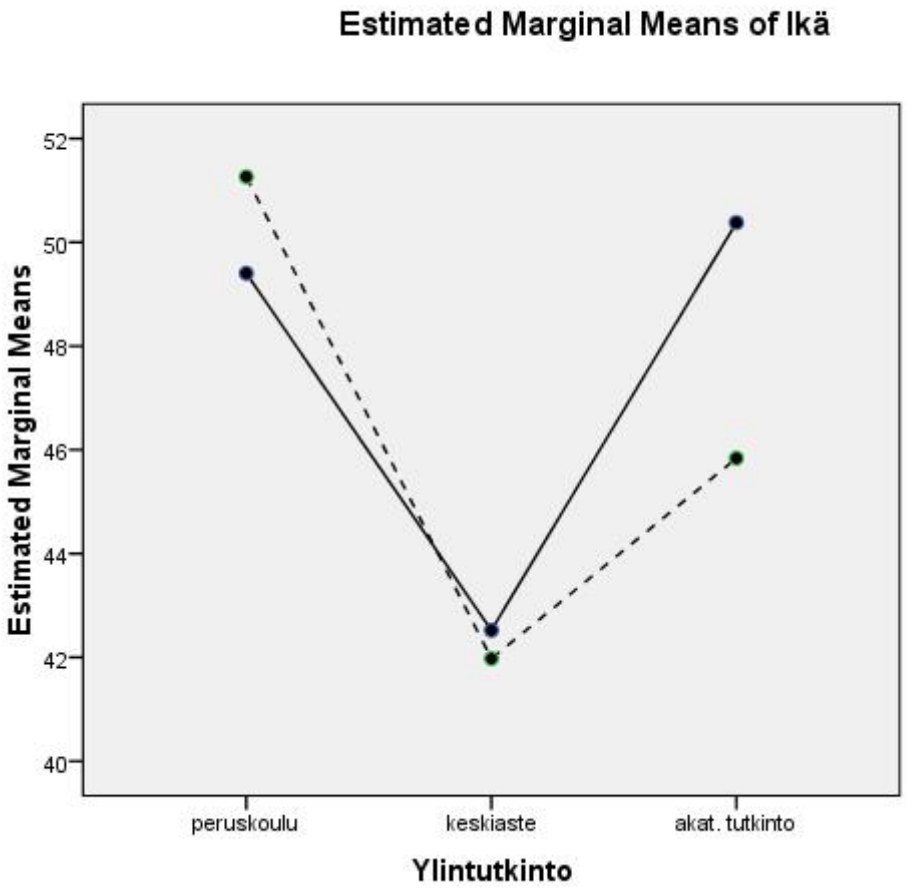
Parametrinen kaksisuuntainen varianssianalyysi tuottaa sukupuolen, koulutustason ja yhdysvaikutuksen osalta Taulussa 4.5 olevat merkitsevyydet, eli 5 % merkitsevyystasolla sukupuolen tapauksessa ryhmäkeskiarvot eivät poikkea toisistaan, kun taas koulutustason kaikki keskiarvot eivät ole samoja (Analyze – General linear model – Univariate). Sukupuolen ja koulutustason välillä on myös yhdysvaikutusta, joten ikien keskiarvojen erot eivät ole samanlaisia eri ryhmissä (kuva 4.8).

Taulu 4.5. Tests of Between-Subjects Effects.

Dependent Variable:Ikä

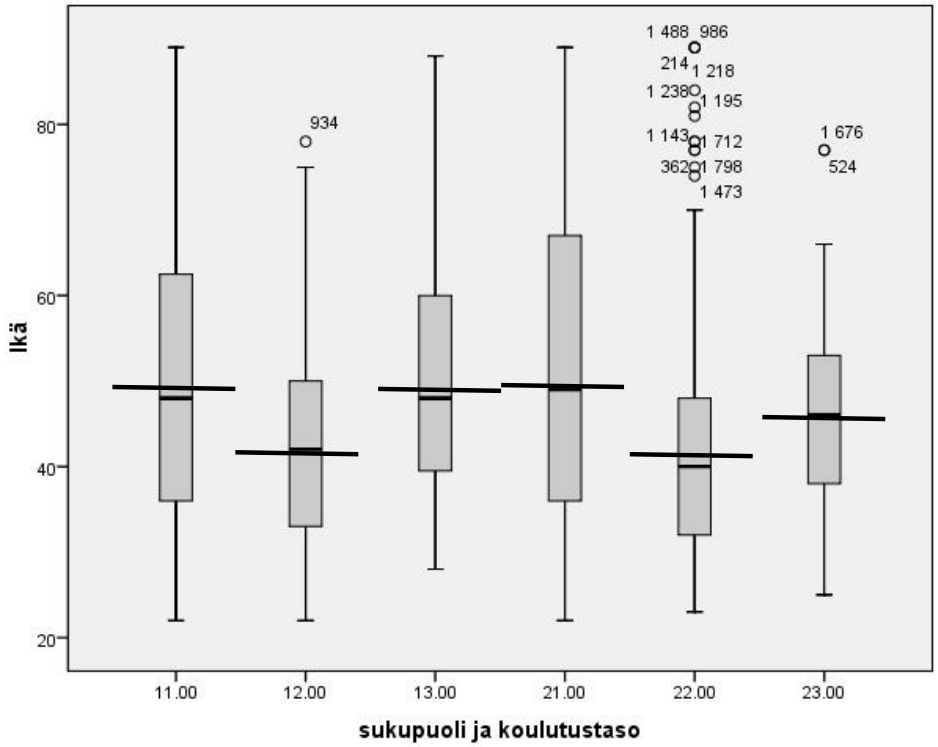
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	26276,358[a]	5	5255,272	20,751	,000
Intercept	2452341,148	1	2452341,148	9683,473	,000
Sukup	322,523	1	322,523	1,274	0,259
Ylintutkinto	22669,007	2	11334,504	44,756	0,000
Sukup * Ylintutkinto	1877,875	2	938,937	3,708	0,025
Error	466233,543	1841	253,250		
Total	4731981,000	1847			
Corrected Total	492509,900	1846			

a. R Squared = .053 (Adjusted R Squared = .051)



Kuva 4.8. Ikien keskiarvopiirros Voter-aineistosta kaksisuuntaisessa varianssianalyysissä.

Ryhmien keskiarvojen perusteella saadaan iälle sovite, joka on esitetty kuvassa 4.9 (esim. mies keskiasteella = 12, naisella akateeminen koulutus = 23) .



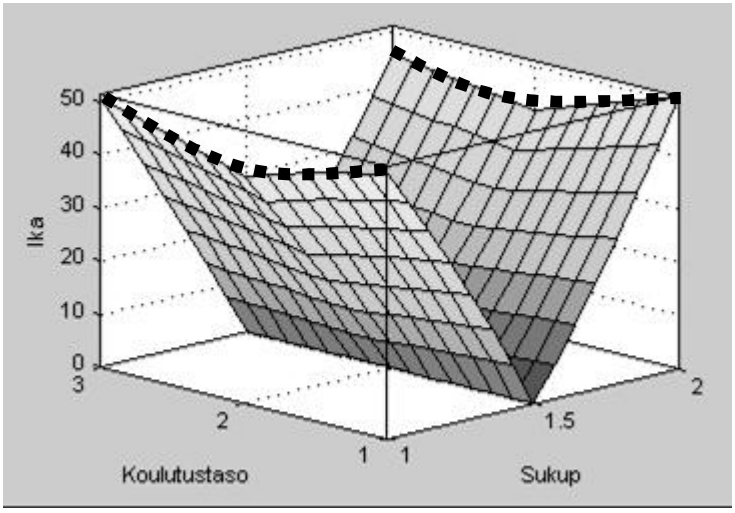
Kuva 4.9. Voter-aineiston mallin ”sovite” (vaakasuorat viivat) kun selitettävänä on ikä ja selittäjinä sukupuoli ja koulutustaso. Mallissa käytetty ikien ryhmäkeskiarvoja näissä luokissa.

Taulun 4.5 tiedot voidaan esittää kielellisinä ja sumeina sääntöinä seuraavasti:

1. Jos on mies (=1) ja korkeintaan peruskoulun käynyt (=1), ikä on noin 49,4 vuotta.
2. Jos on mies ja korkeintaan keskiasteen tutkinto (=2), ikä on noin 42,5 vuotta.
3. Jos on mies ja akateeminen tutkinto (=3), ikä on noin 50,4 vuotta.
4. Jos on nainen (=2) ja korkeintaan peruskoulun käynyt, ikä on noin 51,3 vuotta.
5. Jos on nainen ja korkeintaan keskiasteen tutkinto, ikä on noin 42,0 vuotta.
6. Jos on nainen ja akateeminen tutkinto, ikä on noin 45,9 vuotta.

Tämäntyyppisten sääntöjen avulla voidaan sitten tarvittaessa rakentaa myös sumea malli yksisuuntaisen ANOVA:n esimerkin mukaisesti, ja ku-

vassa 4.10 on yksi ehdotelma tällaisen mallin tuottamasta sovitteesta (sukupuoli on koodattu arvoilla 1 - 2 ja koulutustaso 1 – 3).



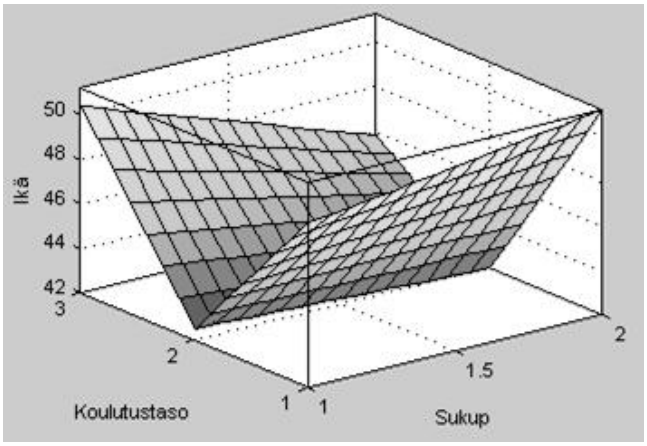
Kuva 4.10. Sumea mallinnus iästä ryhmäkeskiarvojen perusteella kun selittäjinä ovat sukupuoli ja koulutustaso.

Mallin hyvyyttä voidaan sitten arvioida vaikkapa selitysasteen avulla, eli laskemalla ryhmien välisen varianssin (SS between) suhde kokonaisvarianssiin (total). Tällaisten mallien avulla me voimme sitten selittää tai ennustaa, kuinka sukupuoli ja koulutustaso liittyvät ikään.

Jos muodostamme näistä muuttujista suoraan sumeita sääntöjä ilman alkuperäisiä ryhmäkeskiarvoja, niin esimerkiksi *subclust*-menetelmä tuottaa vaikuttavuussäteellä (range of influence) 0,6 taulun 4.6. Ehdotelma sumean mallin sovitepinnaksi on kuvassa 4.11, ja se on samantyyppinen kuvan 4.8 kanssa.

Taulu 4.6. Ryhmien keskukset s*ubclust*-menetelmällä kun muuttujina sukupuoli, koulutustaso ja ikä.

Sääntö	**Sukupuoli**	**Koulutustaso**	**Ikä noin (v)**
	Jos		**niin**
1	mies	peruskoulu	49
2	nainen	peruskoulu	51
3	mies	keskiaste	43
4	nainen	keskiaste	42
5	mies	akat.	50
6	nainen	akat.	46

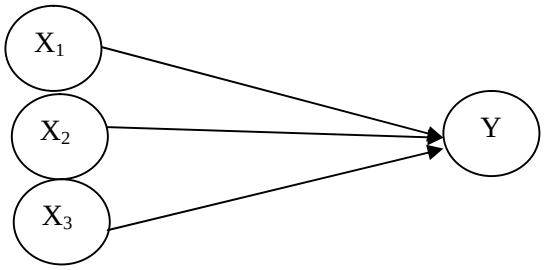


Kuva 4.11. Sumea mallinnus iästä suoraan aineiston perusteella kun selittäjinä ovat sukupuoli ja koulutustaso.

Samoin kuin luvussa 4.1, sumeat mallit, ja vastaavat sumeat säännöt, voivat helpottaa tarkasteluamme, koska silloin voimme ymmärtää paremmin luokittelevien muuttujien ja testimuuttujien välisen yhteyden. Lisäksi sumeita malleja voidaan käyttää ei-parametrisesti ilman jakauma- tai varianssioletuksia, ja ne voivat tuottaa myös ei-lineaarisia ratkaisuja. Sumea mallinnus pääsee kuitenkin tämäntyyppisissä asetelmissa selvemmin oikeuksiinsa silloin, kun käytämme jatkuvia muuttujia, ja seuraavaksi tarkastelemmekin regressiomalleja.

5. REGRESSIOMALLEJA PERINTEISESTI JA NEURO-SUMEASTI

Regressioanalyysi on ehkä tyypillisin tilastollisen tutkimuksen sovellusalue ryhmittely- ja erotteluanalyysin lisäksi, jossa neuro-sumeaa ja geneettis-sumeaa mallinnusta voidaan hyödyntää. Tällöin mallissamme on yksi selitettävä muuttuja (riippuva muuttuja, dependent variable) ja yksi tai useampia selittäviä eli riippumattomia muuttujia (independent variable). Tavoitteena on yleensä selittää tai laatia ennusteita riippuvasta muuttujasta riippumattomien muuttujien perusteella. Ongelmanasettelu voidaan esittää kuvan 5.1 mukaisena kausaalimallina, jossa Y on riippumaton muuttuja.



Kuva 5.1. Regressioanalyysin malli, tässä tapauksessa kolme selittävää muuttujaa.

Kuten luvussa 4 todettiin, varianssi- ja kovarianssianalyysiä voidaan tietyssä mielessä pitää sellaisena regressioanalyysin esivaiheena, jossa kaikki selittäjät, tai osa niistä, ovat pelkästään epäjatkuvia luokittelumuuttujia.

Eksploratiivisessa analyysissä me pyrimme löytämään sopivat muuttujat malliimme, jotta se olisi mahdollisimman hyvä, kun taas konfirmatorisessa analyysissä tiedämme mallin muuttujat esimerkiksi taustateoriamme perusteella ja tavoitteena on verrata malliamme muihin vastaaviin malleihin.

Mallin muuttujien pitäisi olla mitta-asteikon jatkuvia muuttujia, mutta siinä voidaan käyttää selittäjinä jopa luokitteluasteikon muuttujia, tarvittaessa tietyin erityisjärjestelyin (esim. dummy-muuttujat). Jos erityisesti selitettävä muuttuja on luokitteluasteikollinen, sitä varten on tarjolla ”tavallisen” regressioanalyysin sijasta logistinen (logistic, dikotomiselle muuttujalle) ja multinomiaalinen (multinomial) regressioanalyysi.

Tavallisesti regressioanalyysi perustuu lineaarisiin malleihin, eli me pyrimme selittäjien avulla muodostamaan sellaisen suoran tai tason yhtälön, siis sovitteen, jonka kuvaajan pisteet ovat mahdollisimman lähellä selitettävän muuttujan arvoja (esimerkiksi minimoimalla muuttujan arvojen ja sovitteen välisen etäisyyden neliöiden summaa tai *rmse*:n arvoja). Kyse on siis

optimoinnista, jossa regressiosuoran tai –tason yhtälölle pyritään määrittämään matemaattisesti sopivat kertoimet.

Älykkäiden järjestelmien piirissä taas tämä optimointi, samoin kuin koko mallinnus, voidaan tehdä esimerkiksi sumeiden systeemien, neuroverkkojen tai geneettisten algoritmien avulla, ja nämä voivat tuottaa myös ei-lineaarisia malleja ilman tilastollisia jakaumaoletuksia.

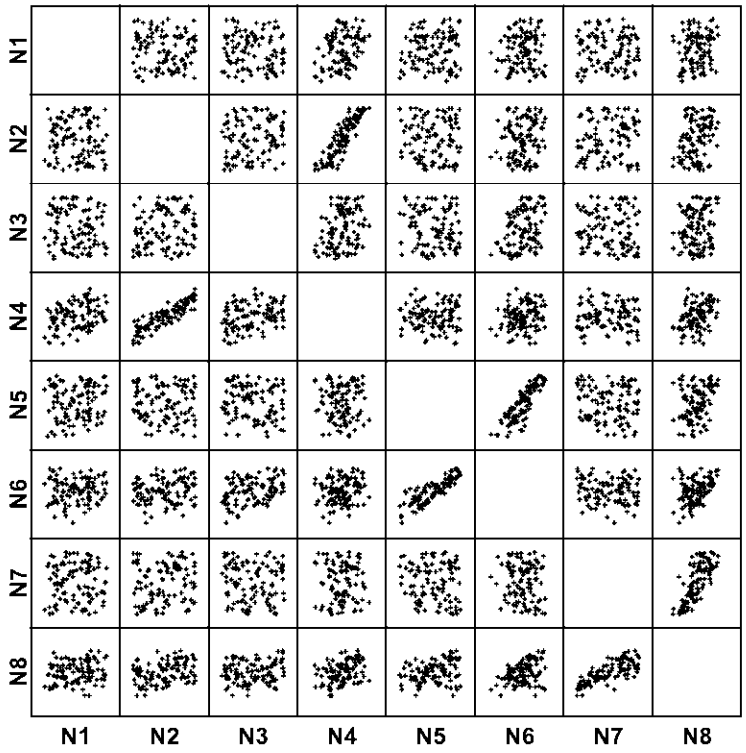
Tässä yhteydessä me keskitymme eksporatiiviseen analyysiin ja käytämme kahdeksan keinotekoisesti tuotetun muuttujan arvoja (muuttujat N1 – N8, vrt. myös luku 7, jossa käytämme vielä samaa aineistoa), jotka saavat arvoja nollan ja ykkösen välillä. Muuttujien tunnuslukuja on taulussa 5.1 (emme voi nyt hyödyntää Voter-aineistoa, koska siinä ei ole oikein sopivia muuttujia).

Valitsemme selitettäväksi muuttujaksi muuttujan N6 ja yritämme löytää sopivat selittäjät. Lineaarisissa malleissa sopivia selittäjiä ovat yleensä ne muuttujat, joilla on tarpeeksi korkea lineaarinen korrelaatio selitettävän muuttujan kanssa. Toisaalta selittäjät eivät saisi korreloida keskenään. Vielä pitäisi pyrkiä myös mahdollisimman pieneen selittäjien määrään mallin hyvyyden siitä kuitenkaan pahemmin kärsimättä.

Kuvassa 5.2 on esitetty sirontakuviot muuttujien välisistä yhteyksistä, ja merkitsevät lineaariset korrelaatiot muuttujalla N6 on korrelaatioanalyysin perusteella muuttujien N3 (0,347), N5 (0,857) ja N8 (0,400) kanssa, joten nämä ovat ilmeisesti sopivia selittäjäkandidaatteja. Toisaalta N5 korreloi myös N8:n kanssa, joten molempia ei kannattane ottaa malliin mukaan.

Taulu 5.1. Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
N1	100	1,00	,00	1,00	,5088	,26700
N2	100	,99	,00	1,00	,5034	,30459
N3	100	1,00	,00	1,00	,4861	,29876
N4	100	,88	,07	,94	,5033	,19836
N5	100	,97	,00	,98	,5407	,26947
N6	100	,87	,04	,92	,5590	,19224
N7	100	,99	,00	1,00	,5126	,29815
N8	100	,72	,12	,84	,4938	,16739
Valid N (listwise)	100					



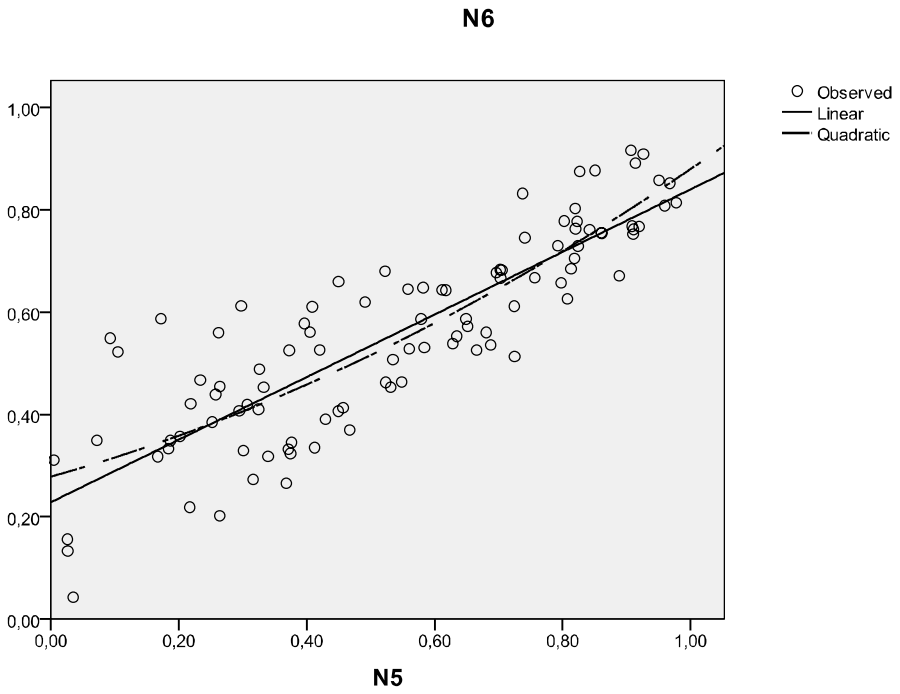
Kuva 5.2. Aineiston muuttujien väliset suhteet sirontakuvioina.

Tarkastellaan ensin yhden selittäjän lineaarista regressiomallia. Nyt meidän pitää löytää sopivin selittäjä muuttujalle N6, ja me hyödynnämme tässä SPSS:n regressioanalyysin stepwise-vaihtoehtoa (Analyze – Regression – Linear, method = stepwise), joka etsii meille ”automaattisesti” parhaat selittäjät selitysasteen perusteella poistamalla ja lisäämällä sopivasti muuttujia. Lopputuloksena saadaan selittäjät, jotka eivät korreloi keskenään, eli siis mallissa ei esiinny multikollineaarisuutta. Tällä perusteella paras selittäjä olisi N5, jolloin saamme regressioyhtälön

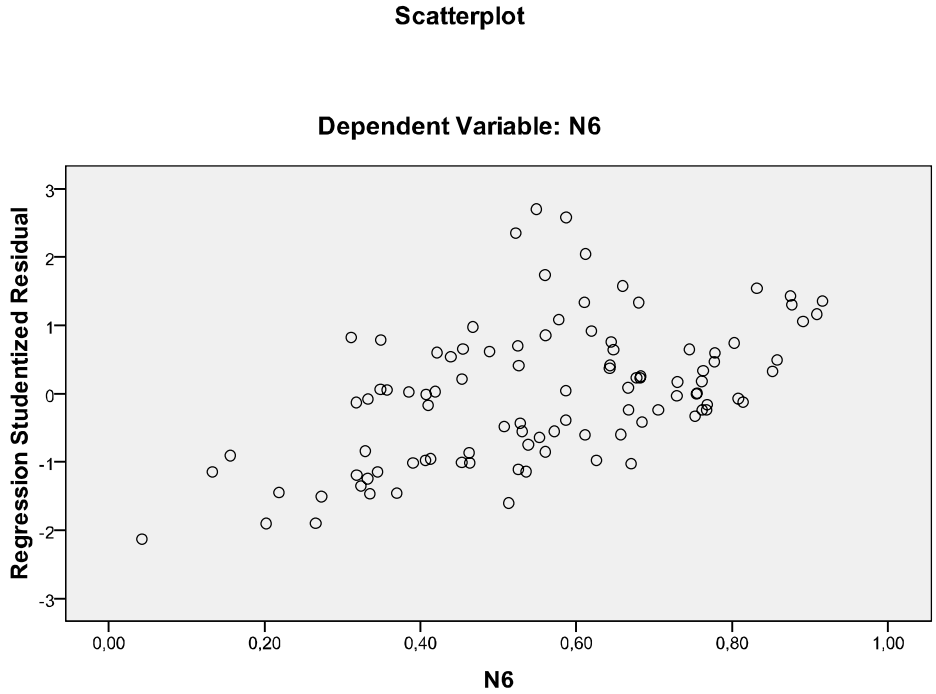
$$N6 = 0{,}611 \cdot N5 + 0{,}228$$

ja selitysasteeksi 0,734 (*rmse* = 0,099), mikä on jo melko korkea. Kuvassa 5.3 on esitetty kyseinen regressiosuora eli sovite. Kuvasta havaitaan, että havaintopisteet eivät ole asettuneet aivan suoraviivaisesti, joten lineaarinen malli ei liene aivan paras mahdollinen tähän tilanteeseen. Tätä olettamusta vahvistaa myös kuvan 5.4 residuaalipiirros, jossa studentisoidut (Studenti-

zed) residuaalit (virhetermit, residuals), eli selitettävän muuttujan arvojen pystysuorat etäisyydet regressiosuorasta, jotka sitten on jaettu residuaalien keskihajonnalla, eivät ole asettuneet symmetrisesti suorakulmion muotoon nolla-tason kummallekin puolelle.



Kuva 5.3. Lineaarinen ja ei-lineaarinen matemaattinen sovite yhden selittäjän mallissa.



Kuva 5.4. Sirontakuvio studentisoiduista residuaaleista yhden selittäjän mallissa.

Kuvassa 5.3 tulee selvästi myös esille regressiomallinnuksen, ja yleisemmin monen muunkin matemaattisen mallinnuksen, perusidea, nimittäin meidän on määritettävä sellainen matemaattinen funktio, jonka arvot (so. vastaavan kuvaajan pisteet) ovat mahdollisimman lähellä selitettävän muuttujan arvoja. Yleensä tämä etsintä käytännössä perustuu juuri residuaalien neliöiden summan tai *rmse*:n minimointiin. Sama menetelmä on tavallisesti käytössä myös älykkäissä järjestelmissä.

Lisäksi kuvassa 5.3 on esitetty yksi ei-lineaarinen sovite, nimittäin toisen asteen polynomi (SPSS:n Analyze – Regression – Curve estimation)

$$N6 = 0{,}252 \cdot N5^2 + 0{,}349 \cdot N5 + 0{,}278$$

jolloin selitysaste on hieman parempi (0,742, *rmse* = 0,097).

Ei-lineaaristen mallien tapauksessa meillä on kuitenkin ääretön määrä erilaisia regressiofunktioita tarjolla, ja emme useinkaan tiedä, minkätyyppinen funktio antaisi parhaimman tuloksen, joten sopivan funktion löytäminen on vaikeaa. Lisäksi ei-lineaariset mallit ovat yleensä matemaattisesti raskaita ja käyttäjän kannalta vaikeasti ymmärrettäviä. Joskus aineistoa linearisoi-

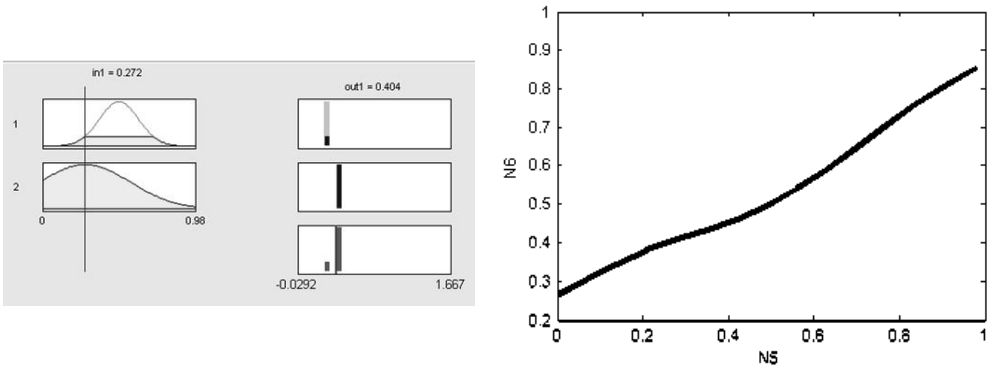
daan sopivien matemaattisten muunnoksien avulla, mutta tämä voi auttaa vain tietyissä tapauksissa.

Niinpä käytännössä, varsinkin tilastotieteen opetuksessa, mallinnus keskittyy lineaarisiin malleihin, vaikka elämä itse on yleensä epistä ja ei-lineaarista. Sumeiden, ja muiden älykkäiden järjestelmien, mallien avulla voidaan taaskin tuottaa käyttökelpoisia ei-lineaarisia malleja useimpiin tilanteisiin, ja ne ovat yleensä myös käyttäjäystävällisempiä.

Kuvassa 5.5 on esimerkki sumeista säännöistä ja neuro-sumean mallin sovitteesta kun on käytetty *anfisedit*-työkalua, 1.kertaluvun Takagi-Sugeno –mallia ja ryvästekniikkaa. Jo tämän mallin *rmse* = 0,096, eli se on kilpailukykyinen edellä mainittujen mallien kanssa. Tämä malli perustuu alustaviin sumeisiin sääntöihin (rypäiden keskuksiin)

1. Jos N5 on noin 0,32, niin N6 on noin 0,41.
2. Jos N5 on noin 0,70, niin N6 on noin 0,67.

ja nämä antavat jo jonkinlaisen yleiskuvan muuttujien N5 ja N6 välisestä suhteesta. Sääntöjä lisäämällä saadaan hieman parempia malleja.



Kuva 5.5. Kahden säännön (vas.) regressiomallin sovite yhden selittäjän tapauksessa, 1.kertaluvun Takagi-Sugeno –malli.

Teemme myös muuttujalle N6 kahden selittäjän mallin, ja lineaarinen versiomme perustuu jälleen SPSS:n edellä mainitulla *stepwise*-menetelmällä valittuihin ”parhaisiin” selittäjiin, jotka sitten ovat N3 ja N5. Tällöin selitysaste on jo 0,931 (*rmse* = 0,050), ja taulun 5.2 mukaan selittäjät eivät keskenään korreloi, koska toleranssit ja VIF lähellä ykkösiä.

Taulu 5.2. Coefficients[a]

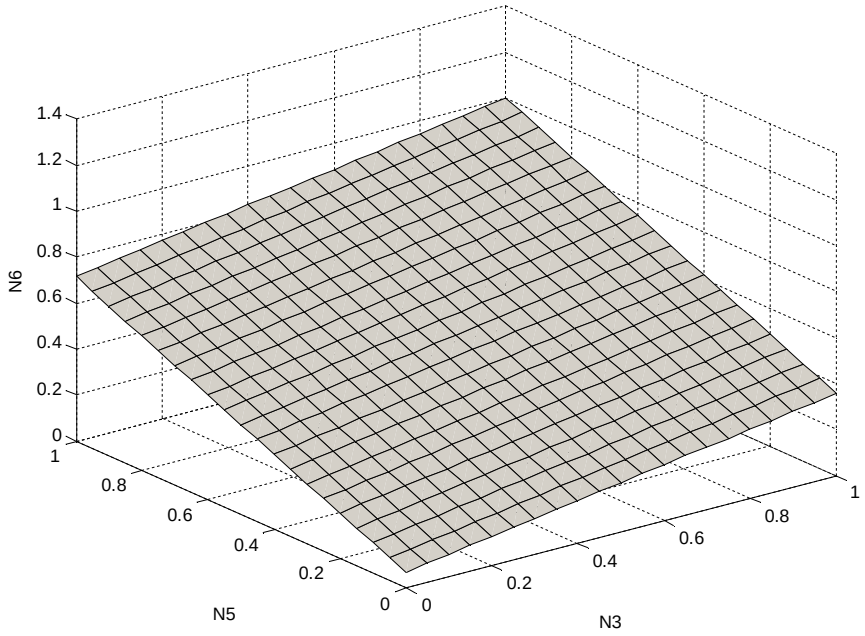
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,070	,015		4,698	,000		
	N3	,287	,017	,446	16,616	,000	,988	1,012
	N5	,646	,019	,906	33,733	,000	,988	1,012

a. Dependent Variable: N6

Regressioyhtälö on siis nyt

$$N6 = 0{,}287 \cdot N3 + 0{,}646 \cdot N5 + 0{,}070$$

ja beta-kertoimien mukaan N5 on se ”tärkeämpi” selittäjä. Kuvassa 5.6 on vastaava regressiotaso.

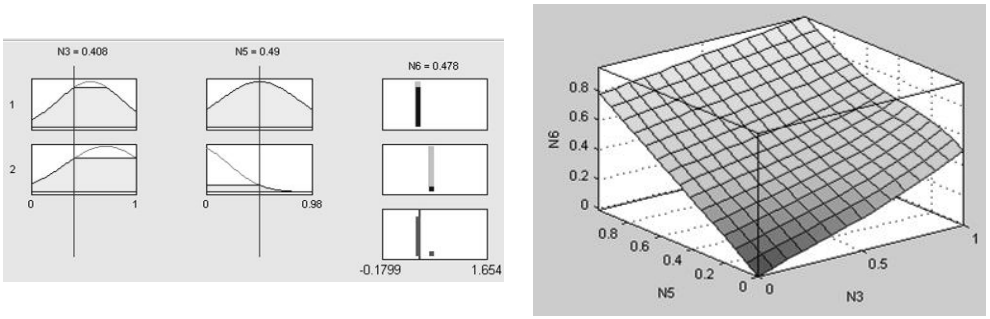


Kuva 5.6. Lineaarisen regressiomallin sovite kahden selittäjän tapauksessa.

Kahden selittäjän ja kahden säännön sumea malli on kuvassa 5.7 kun on käytetty *anfisedit*-työkalua, 1.kertaluvun Takagi-Sugeno -mallia ja ryväs-

tekniikkaa. Tällöin *rmse* = 0,037, eli hieman parempi kuin edellä, ja sääntöjä lisäämällä saadaan vieläkin parempia, joskin myös monimutkaisempia malleja. Alustavat sumeat säännöt ovat tässä tapauksessa

1. Jos N3 on noin 0,48 ja N5 noin 0,58, niin N6 on noin 0,59
2. Jos N3 on noin 0,73 ja N5 noin 0, niin N6 on noin 0,31



Kuva 5.7. Kahden säännön (vas.) regressiomallin sovite kahden selittäjän tapauksessa, 1.kertaluvun Takagi-Sugeno –malli.

Edellä sumeiden mallien viritys on perustunut neuroverkkoihin, mutta nykyään näiden tilalla käytetään optimoinnissa yhä enemmän geneettisiä algoritmeja, eli siis geneettis-sumeita malleja (genetic-fuzzy, [18] [23]), koska ne näyttävät antavan parempia tuloksia. Matlab'issa geneettis-sumeiden algoritmien käyttö edellyttää vielä hieman omaa ohjelmointityötä. Joka tapauksessa kaikki nämä älykkäät järjestelmät ovat erityisen hyvin sovellettavissa juuri regressiomalleihin.

Mallien residuaaleja voidaan tarkastella perinteisin tilastollisin menetelmin. Niinpä voimme vaikkapa tutkia, ovatko residuaalit normaalisti jakautuneet nollan ympärille, onko niissä liian poikkeavia tapauksia (outliers), ja onko niiden keskiarvo nolla. Voimme myös vertailla eri menetelmin tuotettuja residuaaleja siten, että kunkin mallin residuaalit muodostavat yhden ryhmän, ja sitten ANOVA:n avulla testaamme, onko näiden ryhmien keskiarvojen välillä eroa. Samanlainen analyysi voidaan tehdä myös mallien vastearvojen osalta.

Jos käytämme koko aineistomme mallin rakennukseen, malli voi olla tämän aineiston osalta hyvä, mutta muiden samasta perusjoukosta poimittujen aineistojen osalta huono (ylideterminoituminen). Mallimme on niin sanotusti "oppinut ulkoa" aineistomme. Mikäli regressiomallille halutaan myös yleistyskykyä, tai ainakin yleistyskykyä halutaan arvioida, voimme jakaa aineistomme kahteen osaan, opetus- ja vertailuaineistoon (training

data ja control data). Silloin malli rakennetaan opetusaineiston kanssa, mutta sen hyvyyttä arvioidaan tai yleistyskelpoisuutta viritetään vertailuaineiston perusteella.

Tämä menettelytapa, eli ristiinvalidointi (cross validation), on mahdollista kun aineisto on riittävän suuri, ja silloin alkuperäisestä aineistosta poimitaan ensin satunnaisesti muutama kymmenen prosenttia havainnoista vertailuaineistoon ja loput ovat opetusaineistoa. Näin molempien aineistojen tapauksessa havainnot yleensä peittävät syötemuuttujien virittämässä avaruudessa saman alueen (kuten pitääkin). Vertailuaineistona voi tietysti olla muukin, esimerkiksi erikseen kerätty tai jokin jo olemassaoleva samaa perusjoukkoa edustava aineisto. Ristiinvalidoinnin periaate soveltuu hyvin konfirmatooriseen analyysiin.

Anfisedit-mallinnus tuottaa ristiinvalidoinnissa sellaisen mallin, jossa *rmse* on pienimmillään vertailuaineiston osalta (checking data) kun alkuperäistä mallia viritetään opetusaineiston ja neuroverkon avulla. Näin mallin yleistyskyky on parhaimmillaan. Ristiinvalidointia käsitellään vielä luvussa 7.

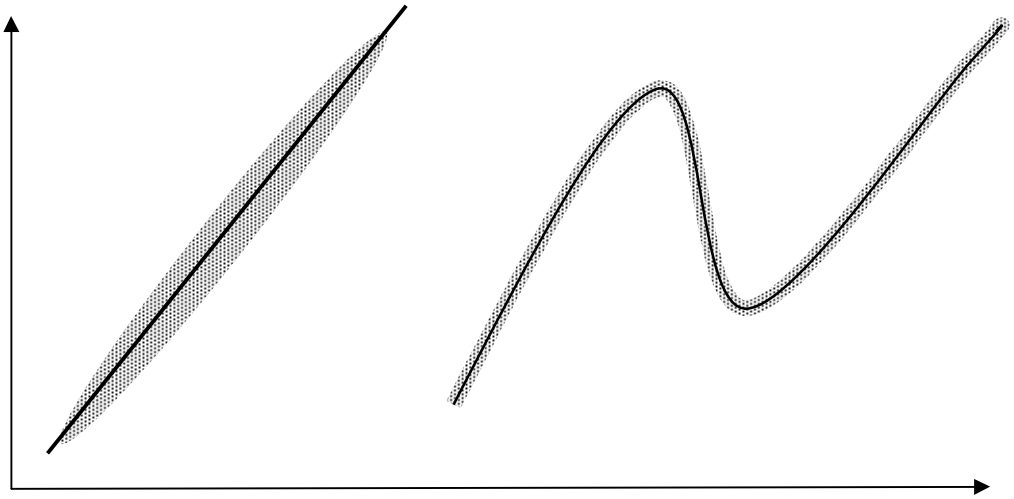
Ei-lineaariseen mallinnukseen voidaan soveltaa myös yleistettyä keskiarvoa ([17] [38], vrt. myös luku 7), ja tällöin tarvittaessa muuttujien arvot skaalataan ensin välille nollasta ykköseen. Tämä menetelmä on hyödyksi muun muassa silloin, kun ei tiedetä, millaista matemaattista ei-lineaarista mallia olisi käytettävä, mutta kuitenkin halutaan käyttää matemaattista mallia. Toinen matemaattinen ei-lineaarinen lähestymistapa ovat neuroverkot, jotka kykenevät tuottamaan hyviä malleja, mutta tällöin muuttujien väliset yhteydet ovat hyvin vaikeasti ymmärrettäviä, koska neuroverkon laskenta perustuu lukuisiin funktioihin ja painokertoimiin.

Exploratiivisessa regressioanalyysissä voidaan myös tutkia yksitellen kaikki mahdolliset selittäjäyhdistelmät parhaan mallin löytämiseksi, mutta se on työlästä, jos muuttujia on paljon. Edellä käytetty SPSS:n *stepwise*-menetelmä voi tällöin tuoda ratkaisun perinteisessä mallinnuksessa, mutta näin ei välttämättä saada parasta ratkaisua. Sumeiden systeemien osalta taas yhdeksi menetelmäksi on ehdotettu, että lasketaan *rmse* erikseen jokaisen mahdollisen selittäjän osalta kun vain yksi selittäjä on mallissa mukana. Tämän jälkeen selittäjien tärkeys on kääntäen verrannollinen *rmse*:n arvoon eli tärkeimmällä on pienin ja vähiten tärkeällä suurin *rmse*-arvo. Tämä lähestymistapa on itse asiassa analoginen askeltavien regressiomallien kanssa. Luvussa 7 esitellään sitten vaihtoehtoinen, yleistettyyn keskiarvoon ja kognitiivisiin karttoihin perustuva menetelmä. Kyseisessä luvussa käsitellään

myös varsinaisesti monimuuttujaregression ja kanonisen korrelaation sovelusta.

Selittäjien määrää voidaan myös yrittää vähentää käyttämällä summamuuttujia, jolloin hyviä apuvälineitä ovat pääkomponentti- ja faktorianalyysi sekä osioanalyysi. Nämäkin nojautuvat kuitenkin keskeisesti lineaarisuuteen, erityisesti muuttujien välisiin lineaarisiin korrelaatioihin, joten niiden sovellusalue on rajallinen. Sumeiden systeemien puolella ei vielä ole varsinaista omaa ratkaisua summamuuttujien luomiseen, joten tältä osin kaivataan lisätutkimusta.

Mainittakoon myös vielä uudestaan jo luvussa 1.3 kuvailtu switching regression –menetelmä, jossa aineiston jokaiselle erillisille havaintoarvojoukolle (eli siis klustereille) tuotetaan oma sovitteensa. Nämä sovitteet voivat sitten perustua vaikkapa matemaattisiin funktioihin tai sumeisiin malleihin (kuva 5.8).



Kuva 5.8. Jokaiselle aineiston havaintoarvojen ryhmälle voidaan tuottaa myös oma sovitteensa matemaattisten funktioiden tai sumeiden mallien avulla.

Tässä luvussa on kuvailtu regressiomallinnuksen perusidea sumeiden systeemien avulla, ja tämä lähestymistapa on jo sinänsä käyttökelpoinen monissa sovelluksissa. Tätä mallinnusperiaatetta voidaan hyödyntää myös monimutkaisemmissa systeemeissä, ja näitä asioita tarkastellaan luvuissa 6 ja 7. Seuraavaksi hyödynnämmekin edellä mainittua tekniikkaa erotteluanalyysissä.

6. EROTTELUANALYYSISTÄ OHJATTUUN OPPIMISEEN

Edellä tarkastelluissa ryhmittelyanalyyseissä pyrimme löytämään aineistosta havaintopisteiden rypäitä, ja nämä menetelmät saavat etsiä kyseisiä ryhmiä melko vapaasti. Tätä lähestymistapaa kutsuttiin ei-ohjatuksi oppimiseksi (unsupervised learning), sillä tutkija ei silloin yleensä etukäteen tiedä, millaisia ryhmiä saadaan tulokseksi. Kun olemme löytäneet aineistostamme sopivat ryhmät, voimme käyttää tätä tietoa hyväksemme myös muiden samantyyppisten aineistojen osalta. Samoin voidaan tietenkin hyödyntää toisten tutkijoiden tuloksia tai tutkittavaan ilmiöön liittyvää teoriaa. Muodollisemmin sanottuna, hyvän ryhmittelyn pitäisi olla käyttökelpoinen muissakin samasta perusjoukosta poimituissa aineistoissa.

Kun sitten haluamme tuottaa onnistuneen ryhmittelyn perusteella aineistollemme käyttökelpoisen ”luokitteluohjeiston”, perinteisesti sovelletaan erotteluanalyysiä (discriminant analysis). Vastaavia uusia menetelmiä ovat taas sellaiset ohjatun oppimisen (supervised learning) sovellukset kuten sumean päättelyn mallit tai neuroverkkojen LVQ-tekniikat (learning vector quantization). Näitä uusia menetelmiä sovelletaan erityisesti hahmontunnistukseen.

Näissä kaikissa on tavoitteena malli, jonka avulla aineiston havainnot voidaan sijoittaa tiettyjen kriteerien tai piirteiden perusteella ennalta määritettyihin ryhmiin. Mitä paremmin tämä sijoittelu sitten onnistuu, sitä parempi mallimme on. Lisäksi voidaan pyrkiä mahdollisuuksien mukaan löytämään luokittelun kannalta oleelliset piirteet ja toisaalta jättämään luokittelusta pois epäoleellisia piirremuuttujia mallin yksinkertaistamiseksi.

Oletetaan, että olemme tutkineet joukkoa hedelmiä ja löytäneet ryhmittelyanalyysin avulla tästä joukosta tiettyjen piirteiden perusteella sellaisia ryhmiä kuten omenat, appelsiinit, banaanit, persikat ja luumut. Tämän jälkeen voimme tulevaa hahmontunnistusta varten yrittää tuottaa ”säännöt”, joiden perusteella voimme aina sijoittaa aineistosta poimitun hedelmän joihinkin näistä ryhmistä.

Tieteellisempiä sovellusesimerkkejä ovat varastetun luottokortin tunnistaminen sillä tehtyjen, tavallisuudesta poikkeavien ostosten perusteella tai pankin asiakkaiden ryhmittely luottoriskin perusteella. Muita paljon käytettyjä sovelluksia ovat puheen- ja käsialantunnistus sekä röntgenkuvien tai Internetin verkkosivujen tulkinta (ks. esim. verkosta WEBSOM-sovellus, joka perustuu akateemikko Teuvo Kohosen itseorganisoituviin karttoihin).

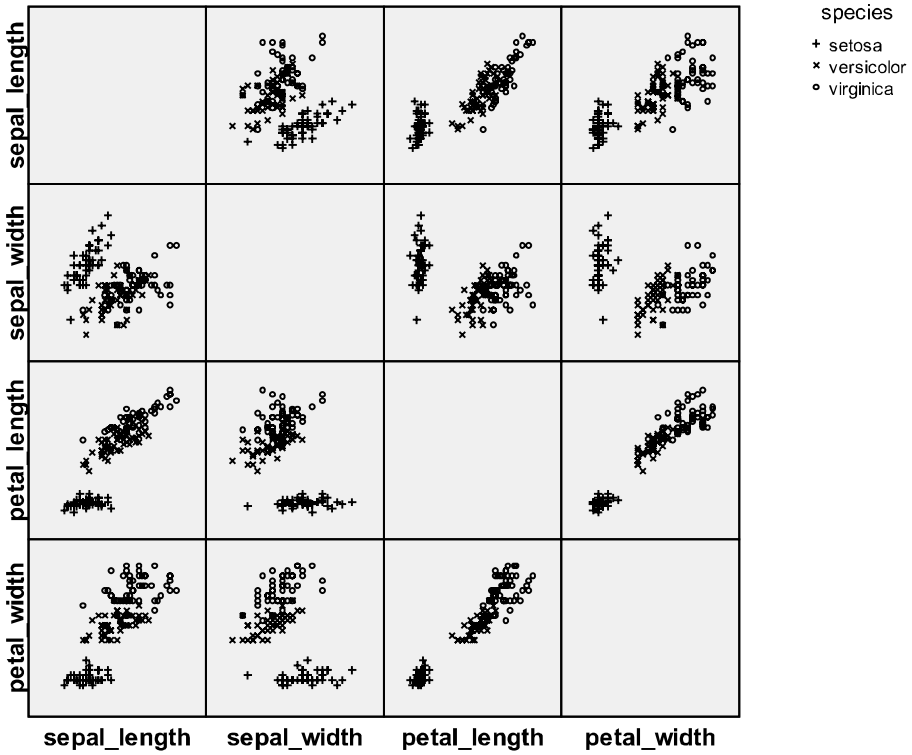
Erotteluanalyysin luokittelumuuttujina on tilastotieteessä perinteisesti käytetty matemaattisia funktioita, erottelufunktioita (discriminant functions), ja näiden funktioiden määrittäminen on tämän analyysin keskeinen tehtävä. Meillä on aluksi toisaalta joukko olioiden piirteitä kuvaavia (yleensä jatkuvia mitta-asteikollisia) muuttujia ja toisaalta luokitusmuuttuja. Analyysin tuloksena saadut erottelufunktiot ovat sitten piirremuuttujista muodostettuja, näiden piirteiden tärkeyden perusteella painotetttuja summamuuttujia (lineaarikombinaatioita), jotka pyrkivät tuottamaan mahdollisimman suuret erot luokitusmuuttujan määrittämien luokkien välille. Niinpä tässä mielessä kyse on yksisuuntaiselle varianssianalyysille analogisesta tilanteesta kun jokaisen erottelufunktion osalta pyrimme siihen, että sen tuottamien luokkien sisäiset varianssit ovat mahdollisimman pieniä ja luokkien väliset varianssit taas mahdollisimman suuria.

Regressioanalyysiä erotteluanalyysi muistuttaa taas siltä osin, että tässäkin tapauksessa pyritään löytämään piirremuuttujien joukosta hyvät ja oleelliset selittäjät luokitukselle. Uutena lähestymistapana tarkastelemme tässäkin yhteydessä erityisesti sumeaan päättelyyn perustuvaa menetelmää.

Käytämme tässä yhteydessä R. Fisherin kuuluisaa ryhmittely- ja erotteluanalyysin testiaineistoa, eli Iiris-aineistoa, jossa 150 iiristä (siis kukkaa) on luokiteltu kolmeen ryhmään (Iris Setosa, Iris Versicolor, Iris Virginica) neljän kriteerin, nimittäin verholehden pituus (sepal lenght), verholehden leveys (sepal width), terälehden pituus (petal lenght) ja terälehden leveys (petal width) perusteella. Meillä on siis neljä syöte- eli piirremuuttujaa ja yksi luokitus- eli vastemuuttuja (kuva 6.1). Onnistuneessa lopputuloksessa kuhunkin ryhmään on sijoitettu oikein 50 iiristä.

Vaikka tämä aineisto ei olekaan varsinaisesti ihmistieteisiin kuuluva, sen käyttö on perusteltua laajan tunnettavuutensa perusteella. Voimme tarvittaessa unohtaa muuttujien alkuperäisen sisällön ja keskittyä vain analyysin tekniseen puoleen, jos se lukijan tarkastelua helpottaa.

Perinteinen erotteluanalyysi edellyttää, että piirremuuttujat ovat normaalisti jakautuneita, ja että varianssit (tai kovarianssit) ovat yhtä suuria luokittelumuuttujan määräämissä luokissa. Lisäksi piirremuuttujat eivät saisi regressioanalyysin ehtojen tapaan korreloida keskenään. Jos nämä alkuehdot eivät riittävästi toteudu, voidaan erotteluanalyysin sijasta käyttää logistista tai multinomiaalista regressioanalyysiä.



Kuva 6.1. Iiris-aineiston piirremuuttujien sirontakuviot.

Luonnostelemme aluksi SPSS:n erotteluanalyysin (Analyze – Classify – Discriminant) avulla erottelufunktiot kun käytämme Iiris-aineiston alkuperäisen ajatuksen mukaisesti mallissamme luokittelevan muuttujan (grouping variable) lisäksi kaikkia neljää piirremuuttuja. Teemme näin, vaikka Boxin M-testin mukaan kovarianssimatriisit eivät olekaan yhtä suuret (nollahypoteesi hylätään).

Analyysi tuottaa meille kaksi erottelufunktiota (eli kolmen luokan tapauksessa maksimimäärän), joista ominaisarvojen tarkastelun perusteella ensimmäisen suhteellinen erottelukyky on yksinään jo 99,1 % ratkaisun varianssista. Wilksin lambdojen perusteella taas voitiin todeta, että erottelufunktioiden määrittämät luokat todella erosivat toisistaan (nollahypoteesit hylättiin). Tauluissa 6.1 ja 6.2 ovat sitten standardoitujen piirremuuttujien painokertoimet erottelufunktioilla ja alkuperäisten piirremuuttujien korrelaatiot erottelufunktioiden, eli uusien ”summamuuttujien”, kanssa. Taulun 6.2 perusteella ensimmäinen funktio näyttäisi olevan verholehti-erottelija (sepal) ja toinen lehtien leveys –erottelija (width) funktioiden tärkeimpien komponenttien perusteella.

Taulu 6.1. Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
sepal_length	-,427	,012
sepal_width	-,521	,735
petal_length	,947	-,401
petal_width	,575	,581

Taulu 6.2. Structure Matrix

	Function	
	1	2
petal_length	,706*	,168
sepal_width	-,119	,864*
petal_width	,633	,737*
sepal_length	,223	,311*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

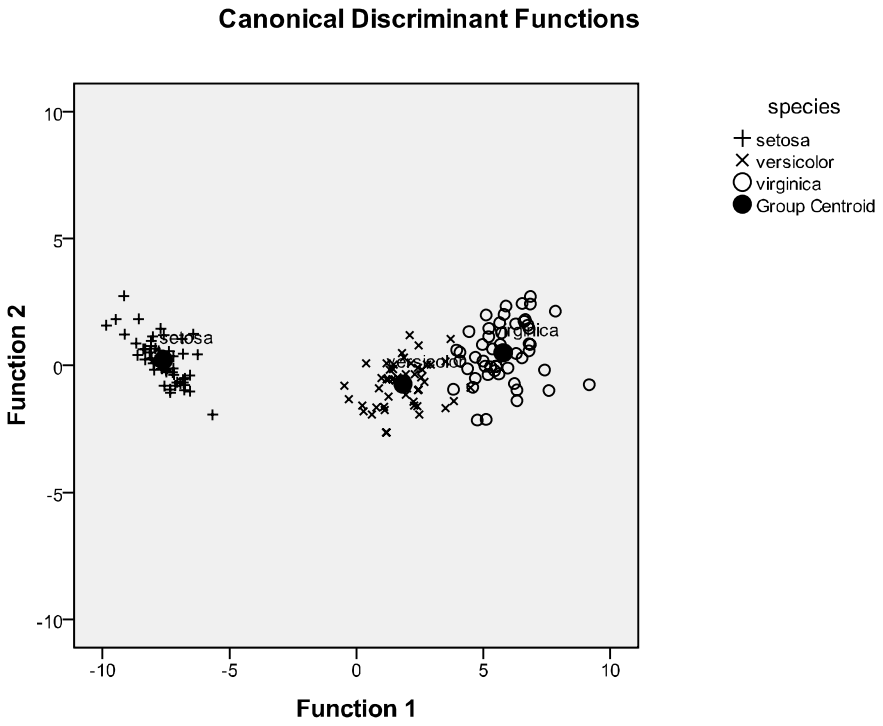
*. Largest absolute correlation between each variable and any discriminant function

Taulussa 6.3 on sitten esitetty erottelufunktioiden määrittämät ryhmien keskimääräiset erottelupisteet ja kuvassa 6.2 on esitetty sama asia graafisesti.

Taulu 6.3. Functions at Group Centroids

species	Function 1	Function 2
setosa	-7,608	,215
versicolor	1,825	-,728
virginica	5,783	,513

Unstandardized canonical discriminant functions evaluated at group means



Kuva 6.2. Erottelufunktioiden määrittämät erottelupisteet.

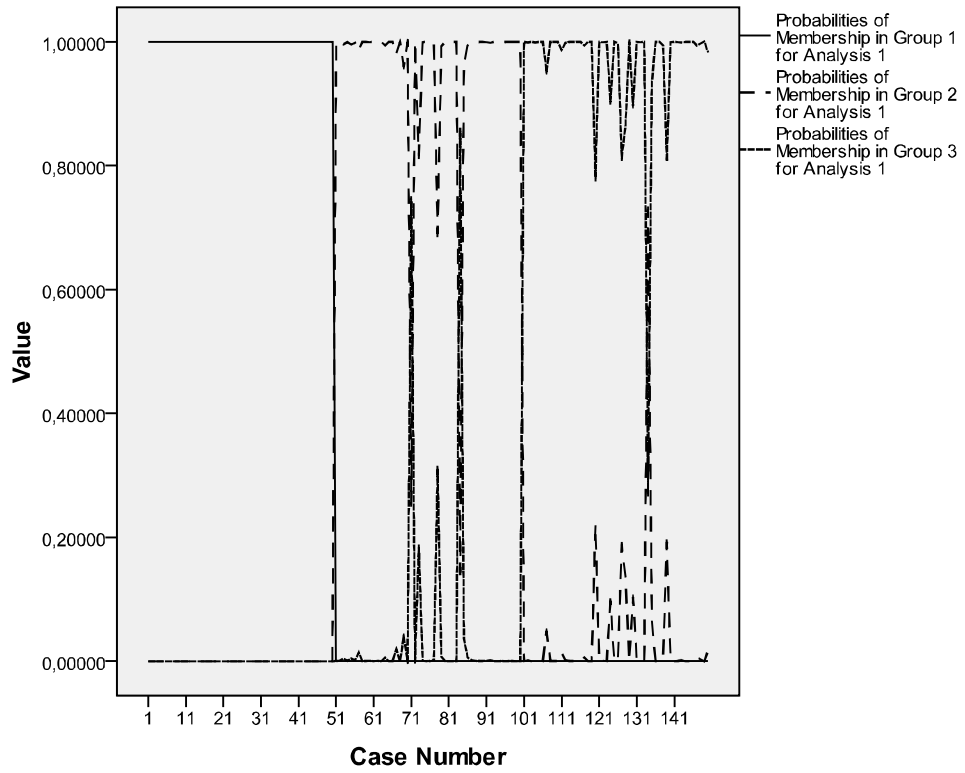
Lopulta taulussa 6.4 on esitetty, kuinka hyvin nämä funktiot luokittelevat Iiris-aineiston. Kuten todetaan, versicolor-luokasta jää uupumaan kaksi ja virginica-luokasta yksi kasvi, joten yhteensä kolme virheellistä luokitusta.

Taulu 6.4. Predicted Group for Analysis 1 * Species, Crosstabulation

Count

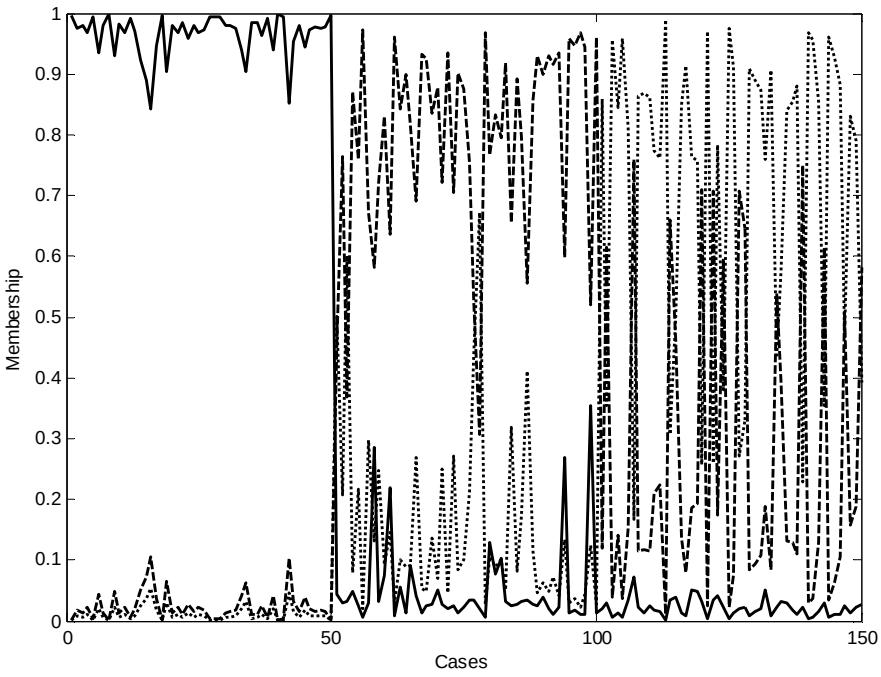
		Species			
		setosa	versicolor	virginica	Total
Predicted Group for Analysis 1	setosa	50	0	0	50
	versicolor	0	48	1	49
	virginica	0	2	49	51
Total		50	50	50	150

Perinteinen erotteluanalyysi tuottaa myös todennäköisyydet, joilla oliot kuuluvat ryhmiin. Esimerkiksi kuvassa 6.3 on esitetty yllä olevan analyysin tulokset. Toteamme, että ensimmäiseen ryhmään (setosa) kaikki oikeat tapaukset kuuluvat todennäköisyydellä yksi ja muut arvolla nolla, kun taas muissa ryhmissä näyttää olevan oikeitakin tapauksia alle yhden todennäköisyydellä. Luotettavamman kuvan erottelukyvystä saa, jos tehdään ristiinvalidointi eli arvioidaan luokittelun onnistumista sellaisen aineiston kanssa, jota ei ole käytetty mallin rakentamisessa.



Kuva 6.3. Iiristen todennäköisyydet kuulua eri luokkiin erotteluanalyysin perusteella (- setosa, - - versicolor, .. virginica).

Kuva 6.3 ja taulun 6.4 perusteella onkin helppo siirtyä sumeaan mallinnukseen ja ohjattuun oppimiseen. Jos vaikkapa käytämme *fuzzy c-means* –algoritmia kolmen ryhmän löytämiseksi pelkästään piirremuuttujien perusteella, Iiris-aineiston kukat tuottavat meille kuvan 6.4 mukaiset jäsenyysasteet (todennäköisyyksien sijasta sovellamme siis nyt jäsenyysasteita). Pyöristämällä jäsenyysasteet kokonaisluvuiksi ja koodaamalla ne arvoiksi 1 - 3 voisimme sitten määrittää luokituksen muodostamalla tällä tavoin uuden luokittelumuuttujan.



Kuva 6.4. Iiristen jäsenyysasteet eri luokkiin sumean ryhmittelyanalyysin perusteella (- setosa, - - versicolor, .. virginica).

Koska Iiris-aineistossa on jo luokittelumuuttujakin annettu, voimme tässä yhteydessä suoraan soveltaa ohjattua oppimista eli määrittää sumean ryhmittelyanalyysin perusteella ensin alustavat sumeat säännöt, ja sitten niiden perusteella luokituksen. Matlab'in *fcm*-algoritmi tuottaa jo kolmen säännön ratkaisuna säännöt, jotka antavat tiivistetyssä muodossa hyvän kuvan luokittelumekanismista (Matlab'in komennolla *irisfcm* käynnistyy tätä aineistoa koskeva demo). Sääntöjä, eli ryhmien määrää, lisäämällä luokittelua voidaan vielä tarkentaa:

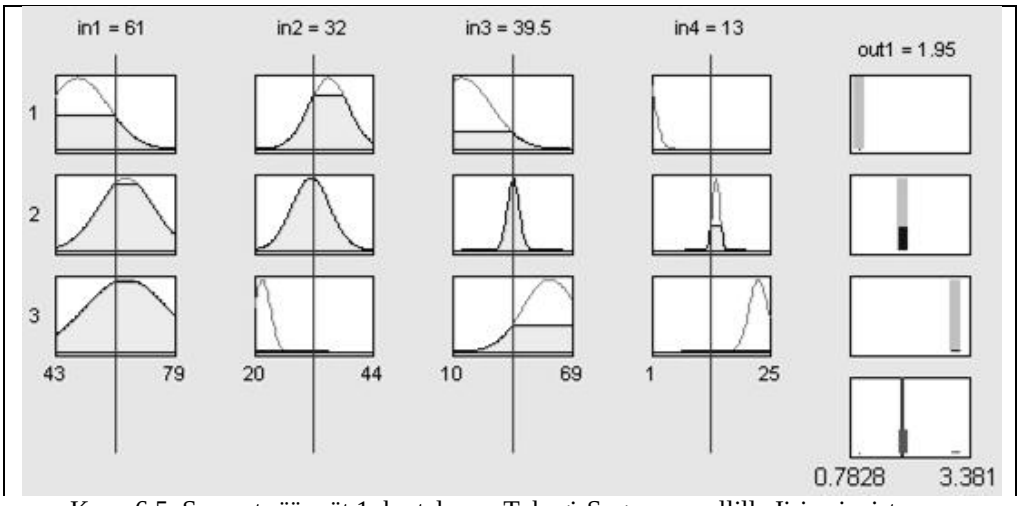
	Jos				niin
1.	sl = 50,04	sw = 34,14	pl = 14,83	pw = 2,54	class = 1,01
2.	sl = 58,89	sw = 27,61	pl = 43,63	pw = 13,97	class = 2,16
3.	sl = 67,74	sw = 30,52	pl = 56,46	pw = 20,53	class = 2,93

Subclust-algoritmi taas määrittää kolmen ryhmän ratkaisuna säännöt

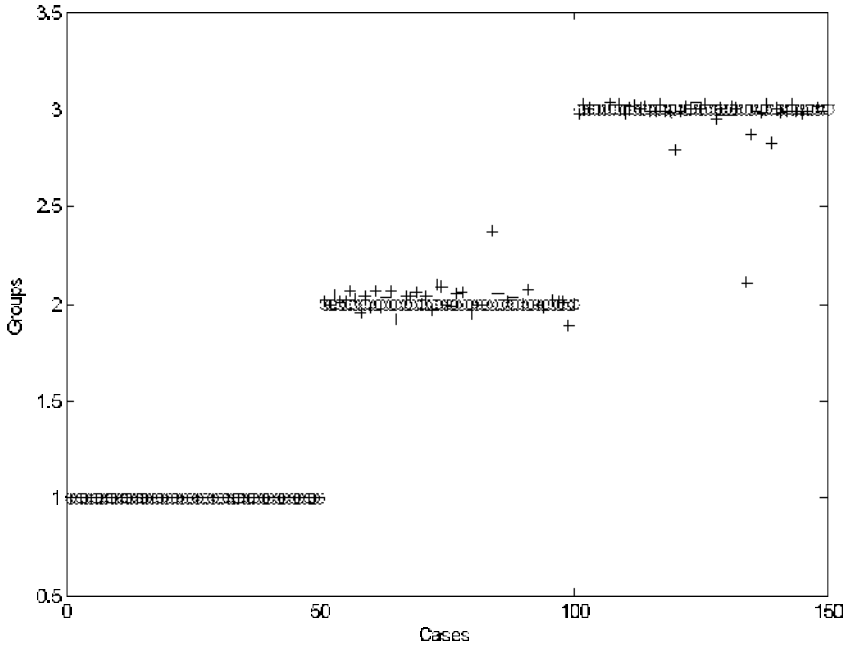
	Jos				niin
1.	sl = 50	sw = 34	pl = 15	pw = 2	class = 1
2.	sl = 60	sw = 29	pl = 45	pw = 15	class = 2
3.	sl = 68	sw = 30	pl = 55	pw = 21	class = 3

ja näihin perustuu sumea mallimmekin. Siis sen lisäksi, että voimme sumeiden sääntöjen avulla selvittää, kuinka luokittelu pelkistetysti tapahtuu, voimme rakentaa luokittelumallin vaikkapa tulevaa hahmontunnistusta varten.

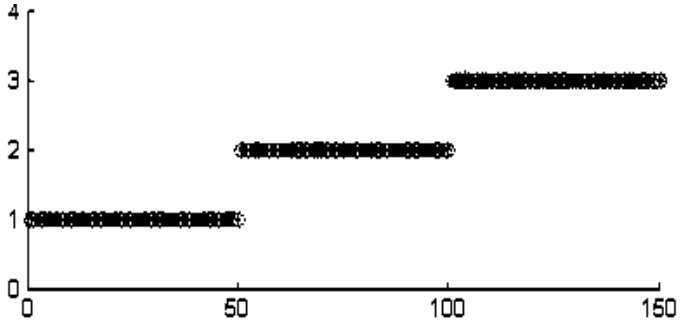
Kuvassa 6.5 ovat yhden tällaisen mallin sumeat säännöt, jotka on tuotettu Matlab'illa ryvästekniikalla (siis *subclust*-algoritmilla) 1. kertaluvun Takagi-Sugenon mallia varten. Kuvassa 6.6 ovat mallin tuottamat ja vastaavat todelliset arvot. Kuten havaitaan, ensimmäisen ryhmän osalta luokittelu on onnistunut täydellisesti, kun taas muiden luokkien osalta on joitakin eitoivottuja tapauksia. Jos aina pyöristämme mallin vastearvot kokonaisluvuiksi, opetusaineistomme tuottaa vain yhden virheluokituksen kuten taulussa 6.5 on esitetty. Täysin oikeaan luokitukseen päästään sääntöjä lisäämällä, ainakin opetusaineiston osalta, kuten kuvassa 6.7 esitetty kymmenen säännön neuro-sumean mallin sovite osoittaa.



Kuva 6.5. Sumeat säännöt 1. kertaluvun Takagi-Sugeno –mallille Iiris-aineistossa.



Kuva 6.6. Sumean mallin tuottama luokittelu Iiris-aineistolle (o havaitut, + mallin arvot), kolme sääntöä.



Kuva 6.7. Sumean mallin tuottama luokittelu Iiris-aineistolle (o havaitut, + mallin arvot), kymmenen sääntöä.

Taulu 6.5. Fuzzymodel * Species, Crosstabulation

Count

		Species			
		setosa	versicolor	virginica	Total
Fuzzymodel	1,00	50	0	0	50
	2,00	0	50	1	51
	3,00	0	0	49	49
Total		50	50	50	150

Sumeat mallit tarjoavat siis yksinkertaisen tavan tarkastella, millaisiin periaatteisiin tai sääntöihin tietty luokitus perustuu ja lisäksi voidaan sopivasti edustavan opetusaineiston perusteella rakentaa luokittelumalli muitakin vastaavia aineistoja varten. Myös neuroverkot tuottavat luokitukseen hyviä malleja, mutta tällöin luokituksen periaatteet jäävät yleensä käyttäjälle ”mustaksi laatikoksi” laskennan monimutkaisuuden takia. Älykkäiden järjestelmien mallit eivät edellytä mitään jakauma- eivätkä varianssioletuksia, ja niiden avulla voidaan tuottaa myös ei-lineaarisia malleja, joten niiden sovellusalue on laajempi kuin perinteisen erotteluanalyysin. Näissäkin tapauksissa ristiinvalidointi on suositeltavaa, mikäli mallille halutaan yleistyskykyä.

Kirjallisuudessa on useita artikkeleita, joissa on pyritty rakentamaan Iiris-aineistoa käyttäen mahdollisimman yksinkertaisia, mutta kuitenkin käyttökelpoisia neuro- ja geneettis-sumeita luokittelumalleja (esim. [24] [56]).

7. AIKASARJOJA TALON TAPAAN – SUMEAT KOGNITIIVISET KARTAT JA MUUTAKIN HELPPOA

Aikasarja-analyysissä (time series analysis) muuttujien osalta kerätään havaintoja tiettyinä ajankohtina ja tietyllä aikavälillä aikasarjaksi. Esimerkiksi Helsingin lämpötiloja voidaan mitata Kaisaniemessä aamuisin kymmenen vuoden aikana, jolloin olemme keränneet aikasarjan. Samat mittaukset voisi tietysti tehdä myös vaikkapa viikoittain tai kuukausittain. Aikasarjoissa on tavallisesti yksi havainto mittauskertaa kohti, ja niissä peräkkäiset havainnot ovat riippuvaisia toisistaan. Aikasarja-analyysin tavoitteet ovat samoja kuin edelläkin, eli pyrimme esimerkiksi kuvailemaan muuttujan vaihtelua tai selittämään, kuinka tietyt tekijät vaikuttavat aikasarjaamme. Voimme myös pyrkiä ennustamaan aikasarjamme tulevia arvoja.

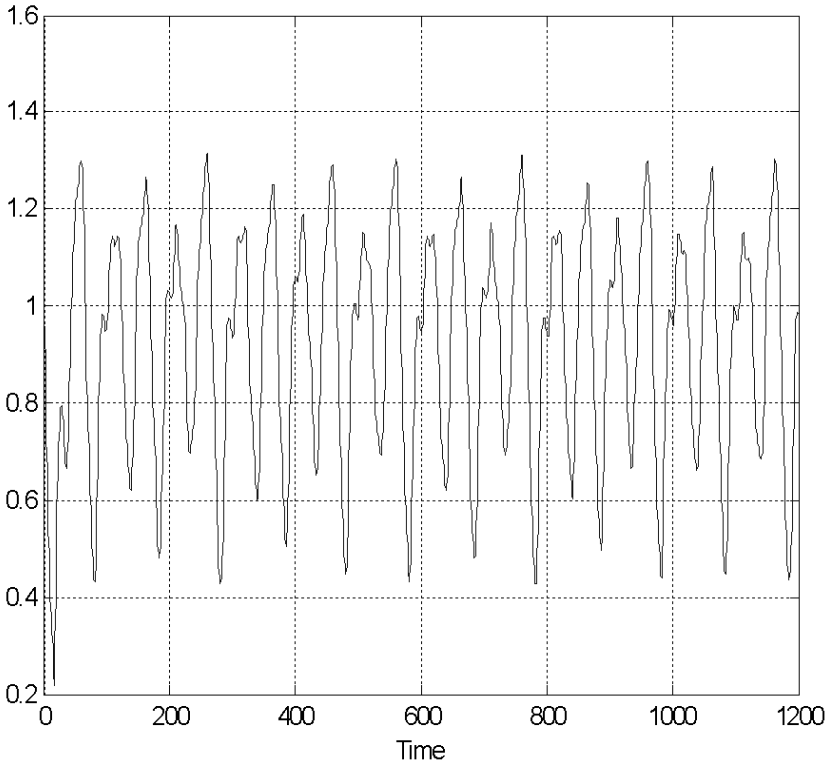
Tilastotieteessä aikasarjojen tarkasteluun on tarjolla valmiita malleja ja mallinnusohjelmistoja kuten ARMA ja ARIMA, ja SPSS-ohjelmassa ne löytyvät valikosta *Analyze - Forecasting*.

Me keskitymme vain kahteen helppoon mallinnustapaan, joita sumeiden systeemien tutkimuksessa on esitetty, ja varsinkin niistä jälkimmäinen, kognitiivinen kartta, on aika monipuolinen mallinnusväline.

7.1. Yhdestä muuttujasta tuotettu aikasarja

Aloitamme yksinkertaisella mallilla, jossa (aika-akselin lisäksi) on vain yksi muuttuja, ja tavoitteena on lähitulevaisuuden ennustaminen. Tällaisissa autoregressiivisissä malleissa vastearvot perustuvat yleensä muutamaan aika-akselilla aikaisemmin mitattuun muuttujan arvoon, ja näin ollen siis tavallaan aikasarjan menneisyyden arvojen avulla pyritään ennustamaan tulevaisuutta. Testiaineistona käytetään aluksi tässä yhteydessä tunnettua Mackey-Glassin funktioon perustuvaa kaoottista aikasarjaa (kuva 7.1, 1200 havaintoa, perustuu Matlab'in Fuzzy Toolbox –käsikirjassa olevaan esimerkkiin, joka käynnistyy komennolla *mgtsdemo*).

Tavoitteena on tuottaa ennustemallimme avulla kuvan x mukainen käyrä. Käytämme tästä aineistosta 500 ensimmäistä aikasarjan arvoa opetusaineistona ja seuraavat 500 arvoa käytetään ristiinvalidointiin eli mallin hyvyyden ja yleistyskelpoisuuden tarkasteluun.



Kuva 7.1. Mackey-Glassin kaoottinen aikasarja.

Yksi tavallinen tapa tällaisen mallin rakentamiseksi on poimia aina muutama arvo aika-akselin suunnassa tietyin välein ja ennustaa näillä saman välimatkan päässä tulevaisuudessa oleva arvo. Esimerkiksi

- aika-akselilla kohdissa 1, 5 ja 9 olevat arvot tuottavat ennusteen kohdan 13 arvolle,
- aika-akselilla kohdissa 2, 6 ja 10 olevat arvot tuottavat ennusteen kohdan 14 arvolle,
- jne.

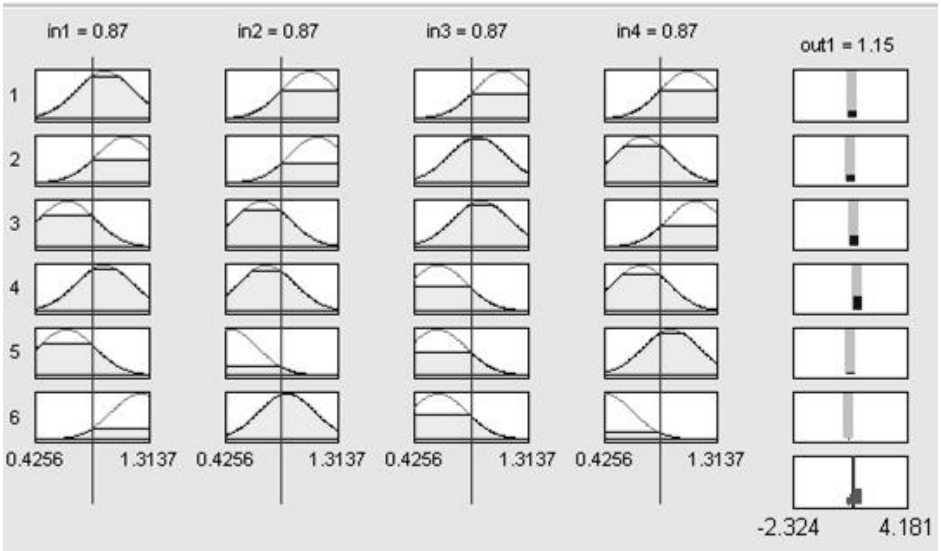
Edellä on siis aina kolme arvoa, jotka ovat neljän yksikön etäisyydellä toisistaan, syötearvoina, ja vastearvo on myös neljän yksikön päässä viimeisimmästä syötearvosta.

Kuvan 7.1 tapauksessa käytämme Matlab'in esimerkin tapaan neljää syötearvoa, jotka ovat kuuden yksikön etäisyydellä toisistaan. Siis opetusaineisto käsittää aika-akselilta *t* poimitut vektorit

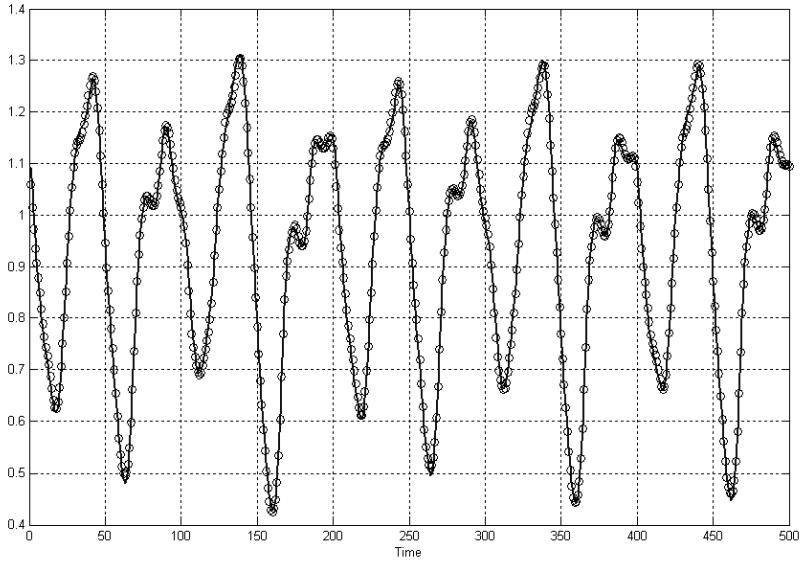
$(t_1,t_7,t_{13},t_{19},t_{25})$
$(t_2,t_8,t_{14},t_{20},t_{26})$
…

joissa vektorien viimeiset arvot ovat toivottuja tulosteita.

Meidän neuro-sumea mallimme perustuu nyt itse asiassa neljään syötemuuttujaan, ja kuvassa 7.2 on esitetty kuuden sumean säännön ratkaisu kun on käytetty Matlab'in *anfisedit*-mallinnuksen *Subclust*-menetelmää ja 1.kertaluvun Takagi-Sugeno –päättelyä (Matlabin vastaava demo käyttää 16 sääntöä). Mallin sovitteen *rmse* vertailuaineiston osalta on noin 0,01, eli varsin hyvä, joten mallimme vaikuttaa yleistyskelpoiselta eli se ei ole hyvä pelkästään opetusaineiston osalta. Vertailuaineiston todelliset ja sovitteen arvot ovat kuvassa 7.3.



Kuva 7.2. Kuuden säännön ratkaisu Mackey-Glassin aikasarjalle, Subclust-menetelmä, 1. kertaluvun Takagi-Sugeno –päättely.



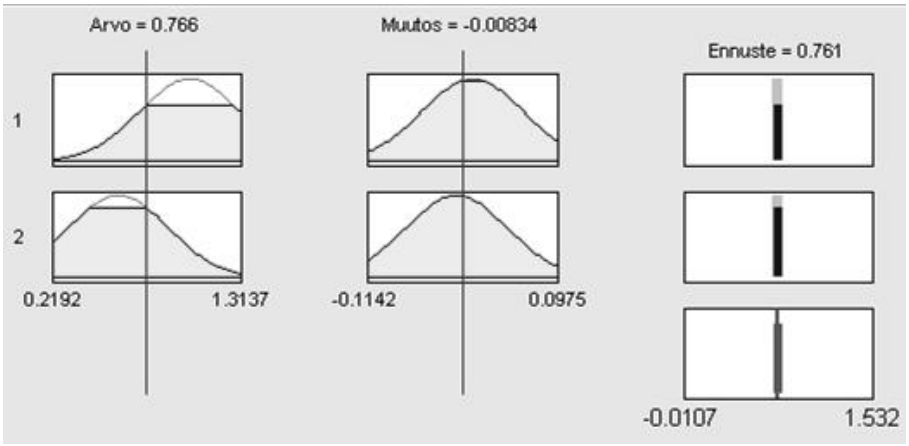
Kuva 7.3. Mackey-Glassin aikasarjan todelliset arvot (o) ja sovitteen arvot (kiinteä viiva) kun kuuden säännön neuro-sumeaa mallia on sovellettu vertailuaineistoon.

Edellä esitettyä aineistoa voidaan mallintaa toisellakin tavalla, jos sovelletaan Stachin ja Pedryczin esittämää menetelmää [61]. Tämä vaihtoehtoinen mallinnus käyttää syötteinä vain tarkasteltavan muuttujan arvoja ja niiden muutoksia tiettynä ajankohtana, ja vasteina ovat muuttujat arvot seuraavana ajankohtana. Jos sovellamme tätä ajatusta Mackeyn ja Glassin aineistoon, opetusaineistomme on tätä tyyppiä:

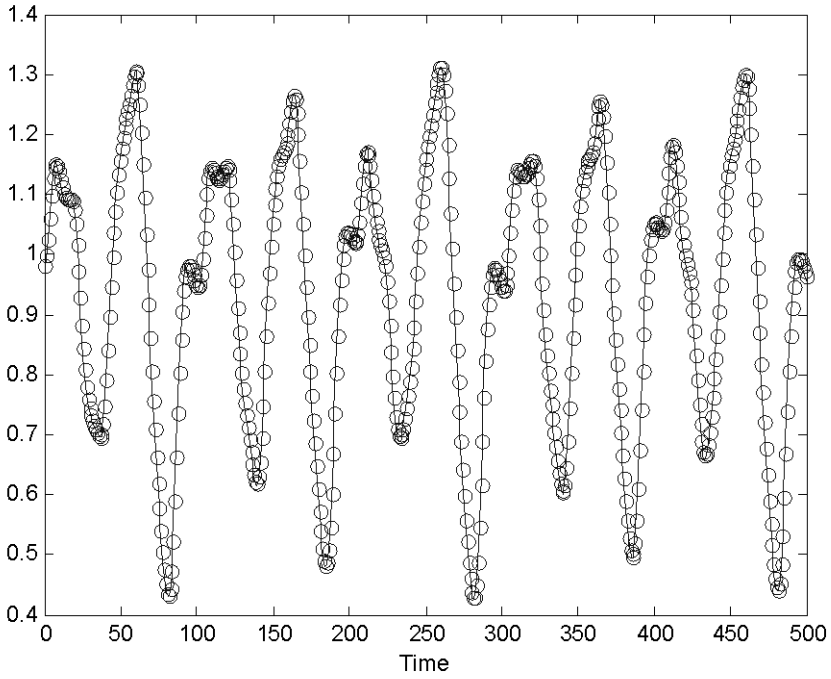
Aika	**Muuttujan arvo**	**Muutos edelliseen arvoon (nykyinen–edellinen)**	**Muuttujan seuraava arvo**
1	1,09	-0,11	**0,98**
2	**0,98**	-0,10	**0,89**
3	**0,89**	-0,09	0,80
4	0,80	-0,08	0,73
5	0,73	-0,08	0,66
6	0,66	-0,07	0,60
7	0,60	-0,06	0,54
8	0,54	-0,06	0,49

Tarkastellaan vaikkapa taulukon toista riviä (Aika=2). Muuttujan arvo on 0,98 ja sen erotus edellisen arvon (Aika=1) kanssa on -0,11. Nämä ovat tällä rivillä syötearvoja. Toivottu vastearvo on nyt muuttujan seuraava arvo 0,89 (Aika=3). Muuttujan alkuperäisten arvojen avulla on siis luotu kaksi muutakin saraketta, joskin tällöin menetämme muuttujan alkuperäisistä arvoista ensimmäisen ja viimeisen, koska näiden osalta ei erotusta tai ajallisesti seuraavaa arvoa voida laskea. Niinpä esimerkiksi ensimmäisellä aineistomme rivillä on itse asiassa vasta toinen muuttujan alkuperäisistä arvoista. Tällä lähestymistavalla ei onneksi ole käytännössä mallinnukselle merkitystä, jos aineistot ovat riittävän suuria.

Kun rakennamme neuro-sumean mallin tästä aineistosta, saamme jo kahdella säännöllä *Subclust*-menetelmää ja 1. kertaluvun Tagaki-Sugeno päättelyä käyttäen kuvien 7.4 ja 7.5 mukaiset säännöt ja vasteet, ja *rmse* on, samoin kuin edellä, vain 0,01.



Kuva 7.4. Kahden syötemuuttujan ja säännön ratkaisu Mackey-Glassin aikasarjalle, Subclust-menetelmä, 1. kertaluvun Takagi-Sugeno –päättely.



Kuva 7.5. Mackey-Glassin aikasarjan todelliset arvot (o) ja sovitteen arvot (kiinteä viiva) kun kahden säännön neuro-sumeaa mallia on sovellettu vertailuaineistoon.

Lineaarinen regressioanalyysikin tuottaa tässä yhteydessä hyvän tuloksen (selitysaste = 0,985 ja *rmse* = 0,03) ja regressioyhtälö on

Uusi arvo = 0,979·Arvo + 0,982·Muutos + 0,190

Edellä esitetty mallinnus sellaisenaan on käyttökelpoisuudeltaan aika rajallista, koska se perustuu lähinnä yhteen muuttujaan. Näyttävämpiä aikasarjamallinnuksia saadaan kun tarkastellaan yhtä aikaa useita eri tekijöitä ja niiden vuorovaikutuksia. Aikasarjoissakin meitä nimittäin usein esimerkiksi kiinnostaa, millaiset taustatekijät vaikuttavat tutkimuskohteena oleviin muuttujiin. Tällaisiin malleihin tutustutaan seuraavaksi.

7.2. Kaikki riippuu kaikesta - sumeat kognitiiviset kartat

Sumeat kognitiiviset kartat (fuzzy cognitive maps) on ensisijaisesti tarkoitettu dynaamisten mallien simulointiin tietokoneympäristössä. Nämä kartat perustuvat erityisesti Robert Axelrodin [4] esittämiin periaatteisiin, joiden

avulla monimutkaisiakin ilmiöitä voidaan usein mallintaa, joskin melko pelkistetysti. Myöhemmin Bart Kosko [36] esitti näistä kartoista ensimmäiset sumeat versiot, mutta ne olivat vielä numeerisia. Nykyään sumeat kielelliset kartat ovat uusinta uutta alan tutkimuksessa.

Aikasarjojen näkökulmasta on kyse regressiivisestä ennustusmenetelmästä, koska mallissa on mukana myös yksi tai useampi ulkoinen muuttuja. Käsittelemme ensin näitä karttoja Axelrodin ja Koskon esittämien ideoiden pohjalta, ja sen jälkeen alan viimeaikaisimpia tuloksia.

Kuvassa 7.6 on esitetty näiden karttojen perusidea eli me tarkastelemme muuttujien tai käsitteiden muodostaman verkoston (usein kausaalisia) keskinäisiä vuorovaikutuksia. Kartassa esitetään muuttujat tai käsitteet eli ”solmut” (concepts, nodes) laatikkoina ja niiden vaikutukset (edges, arcs) toisiin muuttujiin kuvataan nuolilla (jos ei kahden muuttujan välillä ole vuorovaikutusta, ei siis ole nuoltakaan).

Perinteisesti näiden karttojen avulla halutaan simuloida, kuinka muuttujien arvot muuttuvat tietyllä aikavälillä kun muuttujille on annettu tietyt alkuarvot. Muuttujien osalta esitetään entä-jos –kysymyksiä kuten

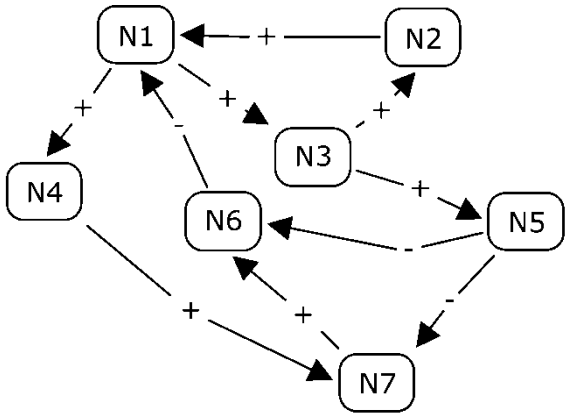
> ”jos muuttujan N2 arvo kasvaa ja muuttujan N7 arvo pienenee, mitä tapahtuu muuttujien N4 ja N6 arvoille?”

Kognitiivisia karttoja voidaan hyödyntää myös muuttujien välisiä vuorovaikutuksia määritettäessä, varsinkin jos empiiristä aineistoa on käytettävissä. Tällöin karttojen teko voi perustua vaikkapa korrelaatioihin tai regressioanalyyseihin. Tämä sovellustapa on analoginen esimerkiksi rakenneyhtälömallinnuksen (structural equation modelling) ja LISREL-ohjelmoinnin mallinnuksen kanssa. Älykkäiden järjestelmien piirissä tämäntyyppisiin tilanteisiin on sumeiden systeemien lisäksi käytetty Bayes-verkkoja (Bayesian networks), mutta niiden sovellusalue on suppeampi.

Ilmiöt voidaan suoraan kuvailla kognitiivisten karttojen avulla, mutta on mahdollista tehdä ensin alustavat versiot vaikkapa käsitekarttojen (concept map) avulla. Jälkimmäinen tapa voi olla helpompi, jos mallissa on kvalitatiivisia, epätäsmällisiä tai kielellisiä muuttujia.

Tällaisissa dynaamisissa malleissa muuttujilla ovat simuloinnin alussa siis tietyt alkuarvot, toisena ajanhetkenä muuttujien toisilleen alkuarvojen perusteella lasketut uudet arvot, kolmantena ajanhetkenä toisen ajanhetken perusteella lasketut arvot ja niin edelleen. Tietokoneympäristössä kyse on

toistoista eli iteroinneista (iteration): edellisen kierroksen muuttujien arvot ovat syötearvoina seuraavalle kierrokselle.



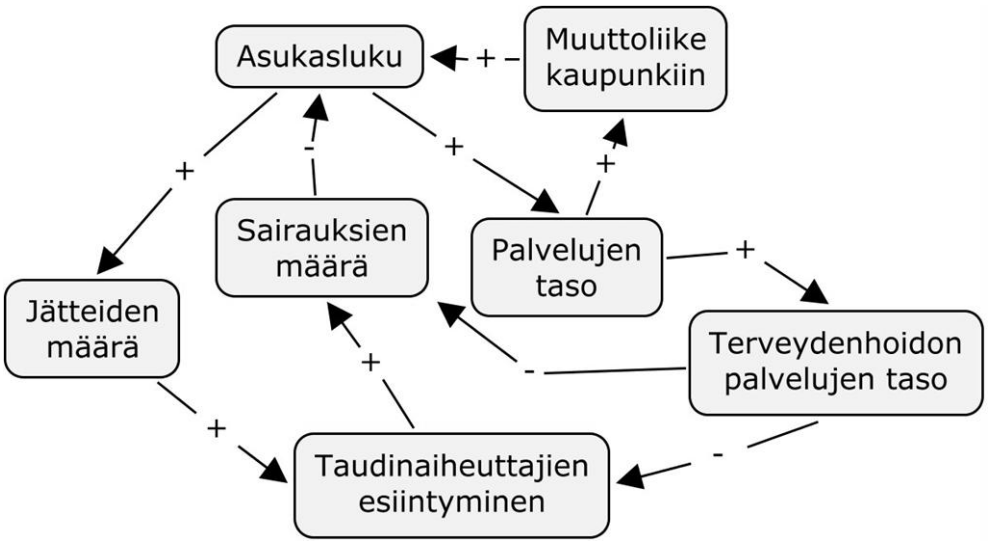
Kuva 7.6. Esimerkki kognitiivisesta kartasta.

Kuvan 7.6 esimerkissä meillä on seitsemän muuttujaa (N1-N7) ja nuolet kertovat, miten muuttujat vaikuttavat toisiinsa. Kuvassa ovat vain suuntaa-antavat yhteydet eli ”+” tarkoittaa lisäävää ja ”-” vähentävää vaikutusta. Toinen tapa olisi kuvitella nämä merkit vaikkapa positiivisiksi ja negatiivisiksi korrelaatioiksi. Niinpä esimerkiksi kun muuttujan N4 arvo kasvaa, myös muuttujan N7 arvo kasvaa. Kuten huomataan, myös takaisinkytkennät (palaute, feedback) ovat sallittuja, ja tämä piirre tekee kognitiivisista kartoista monipuolisempia, kuin eräisiin samantyyppisiin tilanteisiin sovelletut Bayesin verkot, joissa takaisinkytkentä ei ole mahdollista.

Kuinka sitten hyödynnämme tällaista mallia käytännössä? Kuvassa 7.7 laatikoihin on sijoitettu todelliset muuttujat erään mallinnusesimerkin perusteella [60], ja nyt voimme esimerkiksi todeta seuraavia vuorovaikutuksia mallin perusteella:

- Mitä suurempi muuttoliike kaupunkiin, sitä enemmän asukasluku kasvaa.
- Mitä enemmän asukasluku kasvaa, sitä enemmän asukkaista kertyy jätteitä.
- Mitä enemmän asukasluku kasvaa, sitä enemmän palveluja pitää kehittää.
- Mitä enemmän palveluja kehitetään, sitä enemmän myös terveydenhoitopalveluja kehitetään.
- Mitä enemmän terveydenhoitopalveluja kehitetään, sitä vähemmän esiintyy taudinaiheuttajia.

- Mitä enemmän terveydenhoitopalveluja kehitetään, sitä vähemmän esiintyy sairauksia.
- Mitä enemmän jätteitä kertyy, sitä enemmän esiintyy taudinaiheuttajia.
- Mitä enemmän esiintyy taudinaiheuttajia, sitä enemmän esiintyy sairauksia.
- Mitä enemmän sairauksia, sitä vähemmän asukasluku kasvaa.
- Mitä enemmän palveluja kehitetään, sitä enemmän se lisää muuttoliikettä kaupunkiin.



Kuva 7.7. Kaupunki-malli kognitiivisena karttana.

Axelrodin alkuperäisissä malleissa vaikutuksia kuvattiin vain arvoilla ”on vaikutus” tai ”ei vaikutusta” tai sitten kolmiarvoisesti ”vähentää”, ”ei vaikutusta”, ”kasvattaa”. Kosko puolestaan esitti yleistetymmän sumean version, jossa muuttujat saivat arvoja nollasta ykköseen ja vuorovaikutusten vaihteluväli oli -1 ja 1. Muuttujat siis vaikuttivat toisiinsa erilaisilla negatiivisilla ja positiivisilla intensiteeteillä (intensity), ja nämä arvot esitettiin tämäntyyppisenä ”solmu”- eli vuorovaikutusmatriisina (node matrix) [4]:

	Muuttuja 1	Muuttuja 2	Muuttuja 3	Muuttuja 4	Muuttuja 5	Muuttuja 6	Muuttuja 7
Muuttuja 1	0	0	0.6	0.9	0	0	0
Muuttuja 2	→ **0.5** ↑	0	0	0	0	0	0
Muuttuja 3	0	0.6	0	0	0.8	0	0
Muuttuja 4	0	0	0	0	0	0	1
Muuttuja 5	0	0	0	0	0	→ **-0.8** ↑	-0.9
Muuttuja 6	-0.3	0	0	0	0	0	0
Muuttuja 7	0	0	0	0	0	0.8	0

Syy-muuttujat valitaan riveiltä ja seuraus-muuttujat sarakkeilta. Niinpä esimerkiksi N2 vaikuttaa N1:een intensiteetillä 0,5, ja N5 N6:een intensiteetillä -0,8.

Muuttujien arvot voidaan puolestaan tulkita siten, että 0 tarkoittaa esimerkiksi pientä, huonoa tai alhaista, kun taas 1 on vastaavasti suuri, hyvä tai korkea. Arvo 0,5 edustaisi sitten näiden ääriarvojen puoliväliä (”keskinkertainen” tms.).

Kuten edellä todettiin, sumeita kognitiivisia karttoja on käytetty mallintamaan ilmiöiden muutoksia tietyllä aikavälillä. Tyypillisiä ovat erilaiset ”entä-jos” –simuloinnit: miten muuttujien arvot verkostossa muuttuvat tietyllä aikavälillä, jos vaikkapa muuttujan N1 arvo aluksi kasvaa ja muuttujan N5 arvo pienenee?

Käytännössä muuttujien vuorovaikutusten intensiteetit voidaan määrittää solmumatriisiin asiantuntijoiden näkemysten tai empiirisen aineiston perusteella. Mallin hyvyyttä arvioidaan sitten sen tuottamien tulosten perusteella eli kuinka oikeaan osuvasti se ajan mittaan tuottaa muuttujille uusia arvoja.

Varsinainen simulointi on sitten melko yksinkertaista: muuttujien alkuarvot annetaan vektorina, joka sitten kerrotaan matriisituloa käyttäen solmumatriisimme kanssa. Lopputuloksena on vektori, jossa ovat muuttujien arvot seuraavana ajankohtana. Tätä operaatiota toistetaan sitten haluttu määrä.

Jos esimerkiksi muuttujiemme alkuarvot on annettu vektorina

(0,04;0,85;0,93;0,68;0,76;0,74;0,39)

jossa nämä arvot ovat järjestyksessä N1:stä N7:ään, tämän vektorin ja edellä olevan matriisin tulona saamme seuraavan ajanhetken muuttujien arvoiksi vektorin

(0,20;0,56; 0,02; 0,03; 0,75; -0,29; -0,00)

Koska tämän vektorin komponenttien arvot eivät välttämättä ole matriisilaskennan jälkeen välillä nollasta ykköseen, käytämme sopivaa muunnosfunktiota (squashing function) näille arvoille. Muunnosfunktiona voi olla vaikkapa

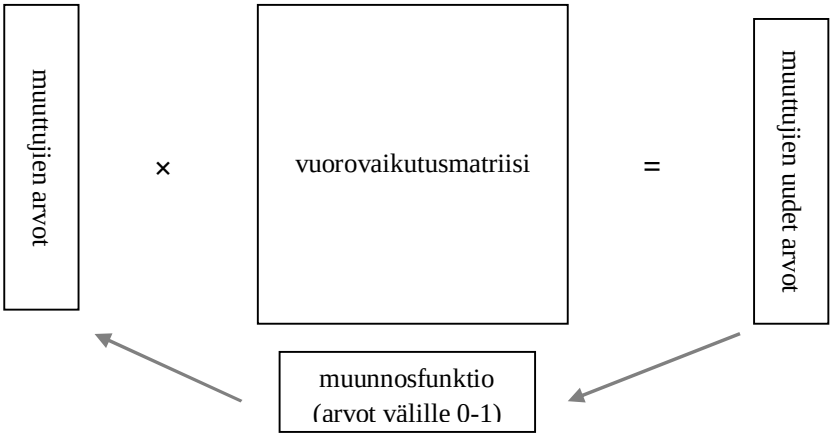
$$f(x) = 1/(1+\exp(-\alpha \cdot x)),$$

missä *x* on muuttujan arvo, a>0 (esim. 5) ja *exp* on eksponenttifunktio. Kun sovellamme tätä funktiota edellä olevaan tulovektoriimme, saamme toisen ajanhetken muuttujan arvoiksi vektorin

(0,73; 0,94;0,53; 0,54; 0,98; 0,19; 0,50)

eli N1 = 0,73, N2 = 0,94, …, N7 = 0,50. Tämä vektori kerrotaan nyt vuorostaan matriisimme kanssa, jolloin saamme, muunnoslaskun jälkeen, kolmannen ajanhetken muuttujien arvot, ja niin edelleen (kuva 7.8).

Jos käytämme tässä yhteydessä kuvan 7.7 muuttujia, niin esimerkiksi pieni asukasluku (=0,04) on muuttunut melko suureksi (=0,73). Vastaavasti kohtalaisen suuri jätteiden määrä (=0,68) on muuttunut hieman pienemmäksi (=0,54).



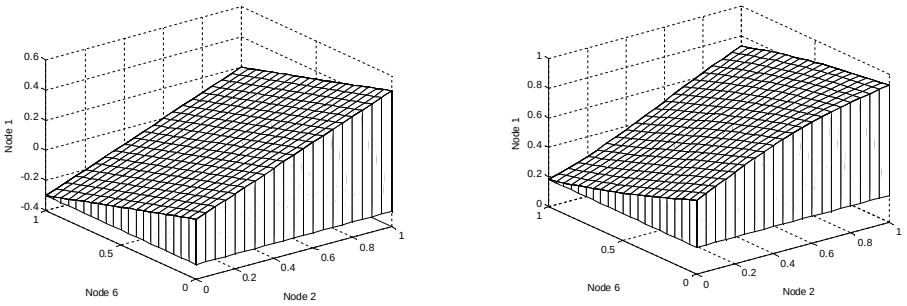
Kuva 7.8. Kognitiivisen kartan iteroinnin perusidea.

Kuvassa 7.10 on esitetty muuttujien historiakäyrät (history curves) eli tässä tapauksessa muuttujien arvot kymmenen iteraation jälkeen (aika-akselin nollakohdassa ovat alkuarvot). Tällä tavalla voimme seurata muuttujien arvojen muutoksia verkostossamme tietyllä aikavälillä. Joissakin tilanteissa, kun käytämme vaikkapa kvalitatiivisia muuttujia ja kielellisiä arvoja, käyrien suunnat kertovat meille vain trendeistä eli muuttujan arvojen muutoksien suunnasta, mutta tämäkin voi jo olla meille arvokasta informaatiota.

Matemaattiselta kannalta kognitiivisten karttojen osalta on kyse graafeista (graph) ja lineaarisista yhtälöistä kun muuttujien arvot on ensin skaalattu välille nollasta ykköseen. Voimme siis ajatella, että jokaisen muuttujan uusi arvo perustuu ”regressioyhtälöön”, jonka kertoimet on annettu vuorovaikutusmatriisissa ja selittäjien arvot ovat edellisen ajanhetken arvot. Esimerkiksi muuttujan N1 osalta sovellamme yhtälöä

$$N1 = 0.5 \cdot N2 - 0.3 \cdot N6$$

joten “selittäjinä” ovat tässä tapauksessa muuttujat N2 ja N6. Tämän jälkeen laskemme vielä muunnosfunktion avulla muunnetun arvon. Kuvassa 7.9 on esitetty näiden muuttujien alkuperäinen lineaarinen yhteys sekä sama yhteys muunnosfunktion soveltamisen jälkeen arvolla a=5 (jolloin itse asiassa yhteys on ei-lineaarinen).



Kuva 7.9. ”Regressiomallit” kognitiivisen kartan perusteella kun muuttuja N1 on selitettävä ja N2 sekä N6 selittäjiä. Ennen muunnosfunktion käyttöä (vas.) ja sen käytön jälkeen.

Edellä esitetty, matriisituloon perustuva laskentatapa, tuottaa matemaattisten ominaisuuksiensa vuoksi aina tietyn iterointimäärän jälkeen muuttujille uudelleen samoja arvoja, eli historiakäyrät itse asiassa aaltoilevat toistaen samoja arvoja (”oskilloivat”) tai muuttuvat suoriksi viivoiksi. Niinpä sitä ei valitettavasti voi soveltaa kovin pitkiin aikasarjoihin. Sumeita sääntöjä soveltavat kognitiiviset kartat voivat kuitenkin ratkaista tämän ongelman, mutta niitä tarkastellaan hieman myöhemmin.

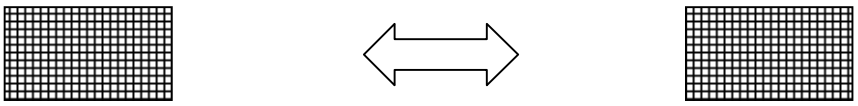
Kognitiivisia karttoja voidaan rakentaa myös osaverkosto kerrallaan, ja näitä osia sitten liitetään toisiinsa suuremmiksi kokonaisuuksiksi. Eri asiantuntijoiden tekemiä karttoja voidaan myös yhdistää Delfoi-tutkimuksen idean mukaisesti yhdeksi ”konsensus-kartaksi”. Voimme myös ensin tehdä käsitekartan (concept map), jonka pohjalta sitten vastaava kognitiivinen kartta muodostetaan. Tämä menettely on hyödyllistä erityisesti laadullisessa tutkimuksessa.

Kognitiivisissa kartoissa vuorovaikutusmatriisi voidaan tää ”asiantuntemuksen” tai empiirisen perusteella. Edellä tarkastelimme jo asiantuntemukseen perustuvaa, joskin melko kuvitteellista, matriisia. Toinen tapa on määrittää vuorovaikutusmatriisin arvot empiirisen, niin sanotun historia-aineiston perusteella (history data), ja silloin määritämme matriisin solujen arvot neuroverkkojen tai geneettisten algoritmien avulla. Silloin pyrimme tuottamaan kartan alkuarvojen ja optimoitavan vuorovaikutusmatriisin perusteella sellaisia muuttujien arvoja jokaiselle ajanhetkelle, jotka mahdollisimman paljon muistuttavat todellista historia-aineistoamme. Siis mallimme optimoitavia parametreja ovat vuorovaikutusmatriisin solujen arvot.

Tehdään edellä esitetty mallinnus soveltaen kyseistä optimointitekniikkaa. Olkoon meillä käytössä seitsemästä muuttujastamme N1 – N7 historia-aineisto, joka on kuvassa 7.10, ja tavoitteena on määrittää nämä arvot

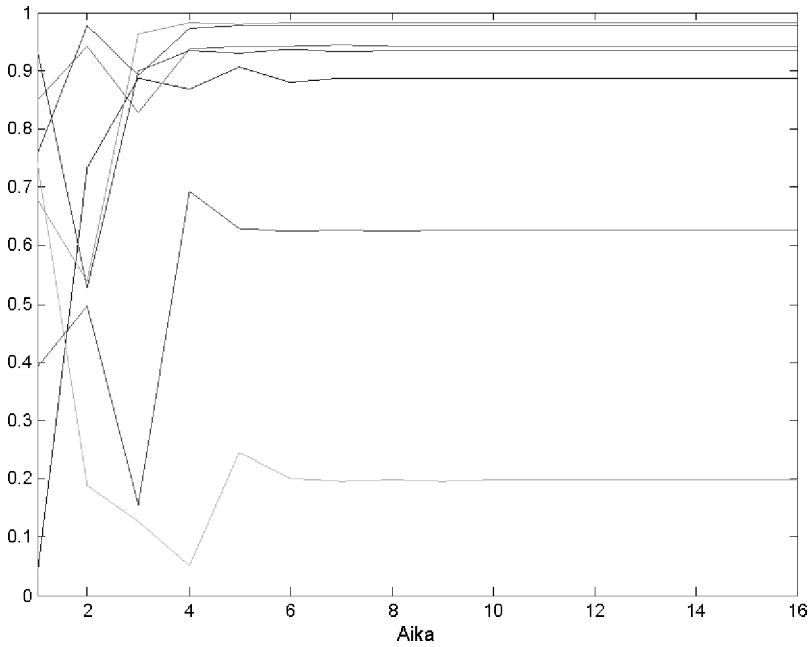
tuottava vuorovaikutusmatriisi. Aineistossamme on siis alkuarvojen lisäksi viidentoista muun ajanhetken arvot, eli havaintomatriisissamme on kuusitoista riviä ja seitsemän saraketta. Tämä aineisto on itse asiassa laskettu edellä olevan vuorovaikutusmatriisin perusteella käyttämällä muuttujille N1 – N7 alkuarvoja, jotka ovat kuvan 7.10 vasemmassa reunassa. Tavoitteena on siis määrittää sellainen vuorovaikutusmatriisi, joka tuottaa alkuarvoistamme mahdollisimman paljon historia-aineistoamme muistuttavia arvoja.

Käytämme vuorovaikutusmatriisin määrittämiseen Matlab'in geneettistä algoritmia *ga* siten, että vuorovaikutusmatriisin soluparametrien avulla (joita on tässä 7·7 = 49 kappaletta) lasketaan ensin alkuarvoista viisitoista muuta historia-aineistomme riviä, ja sitten laskemme, kuinka paljon meidän mallimme historia-aineiston arvot poikkeavat todellisista arvoista. Mallin hyvyys perustuu arvoon, joka saadaan laskemalla jokaisen havainnon osalta todellisen arvon ja mallimme tuottaman arvon erotuksen neliö, ja nämä neliöt sitten summataan (siis residuaalien neliöiden summa). Mitä pienempi tämä arvo on, sen parempi malli:



Todelliset havainnot, t_i — Minimoi tämä erotuksien neliöiden summa: $\Sigma(t_i-m_i)^2$ — Mallin tuottamat arvot, m_i

Pyrimme samalla kertaa myös löytämään optimaalisen arvon muunnosfunktion parametrille *a*, joten se on 50. optimoitava parametri. Optimoinnin voi tietenkin suorittaa muillakin menetelmillä, jopa Excel'in ratkaisimella (solver). Geneettisten algoritmien etuna on kuitenkin se, että ne pyrkivät löytämään globaalisti optimaalisen ratkaisun.



Kuva 7.10. Historia-aineisto vuorovaikutusmatriisiin määrittämiseksi (ensimmäiset arvot vasemmalla ovat muuttujien alkuarvot).

Matlab'in pientä, itse kirjoitettua aliohjelmaa (m-tiedosto) ja *ga*-algoritmia soveltaen saamme muunnosfunktion parametrin *a* arvoksi 5,02, jolloin tämä funktio on muotoa

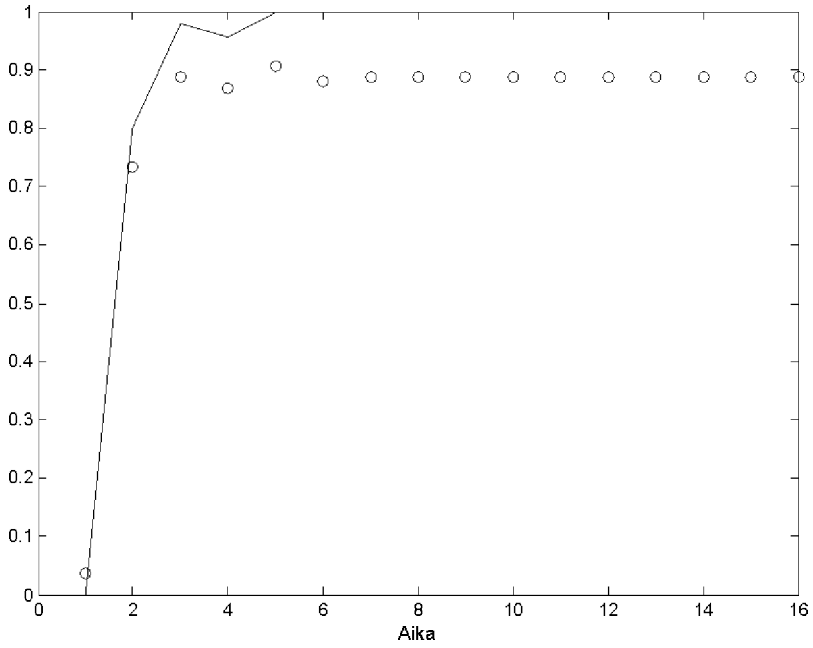
$$f(x) = 1/(1+\exp(-5{,}02 \cdot x)).$$

Tällöin saamme vuorovaikutusmatriisin

	N1	N2	N3	N4	N5	N6	N7
N1	-1,00	-0,96	1,00	0,81	1,00	-1,00	1,00
N2	0,82	-0,28	-1,00	-0,24	-1,00	0,51	-0,71
N3	1,00	1,00	-0,58	-0,38	1,00	0,03	0,57
N4	0,28	1,00	1,00	1,00	1,00	0,29	1,00
N5	0,87	-0,47	1,00	1,00	0,00	0,00	-1,00
N6	-1,00	-0,67	-0,44	-0,74	-0,17	-0,47	-0,46
N7	0,13	1,00	-0,41	-1,00	0,50	-0,01	-0,40

Tämä matriisi ei paljonkaan muistuta alkuperäistä, mutta mallin vastearvojen *rmse* on vain 0,01, joten annetun historia-aineiston näkökulmasta mallimme vaikuttaa hyvältä. Voisimme siis käyttää tätä matriisia tutkittavan

ilmiömme simulointiin. Kuvassa 7.11 on muuttujan N1 osalta todelliset arvot ja mallimme sovite.



Kuva 7.11. Historia-aineiston todelliset (o) ja optimoidun kognitiivisen kartan tuottaman sovitteen arvot muuttujan N1 osalta.

Sumeiden kognitiivisten karttojen yhtenä suurena etuna on, että voimme käyttää malleissamme suhteellisia arvoja, ja kuitenkin saamme käyttökelpoisia tuloksia. Esimerkiksi edellä meidän ei muuttujan *asukasluku* tapauksessa tarvitse käyttää absoluuttisia asukaslukuja simuloinneissamme, joten malli soveltuu kaiken kokoisten kaupunkien tarkasteluun. Näin myös mallit ovat käyttäjille helpommin ymmärrettävissä, ja toisaalta niiden rakentaminen ei vaadi niin spesifiä asiantuntemusta.

Nämä alkuperäiset, lineaarisiin yhtälöihin perustuvat numeeriset kartat ovat kuitenkin sovelluskelpoisuudeltaan vielä melko rajattuja, koska niillä voidaan mallintaa vain muuttujien välisiä monotonisia kausaalisia vuorovaikutuksia. Lisäksi niissä ei ole yleensä solmuja itse ilmiöiden muutoksille. Kognitiivisten karttojen käyttäjäystävällisyyttä ja käyttökelpoisuutta voidaan vielä lisätä hyödyntämällä kielellisiä arvoja sekä sumeita sääntöjä kuten seuraavaksi tulemme tekemään.

7.3. Kielelliset kognitiiviset kartat

Uusinta uutta sumeiden kognitiivisten karttojen osalta ovat kielelliset (linguistic) kartat, joissa muuttujat saavat kielellisiä arvoja ja vuorovaikutukset esitetään sumeina sääntöinä. Tällöin muuttujien arvot perustuvat tavallisesti Osgood- tai Likert-asteikkoon. Niinpä muuttuja voi saada numeeristen arvojen

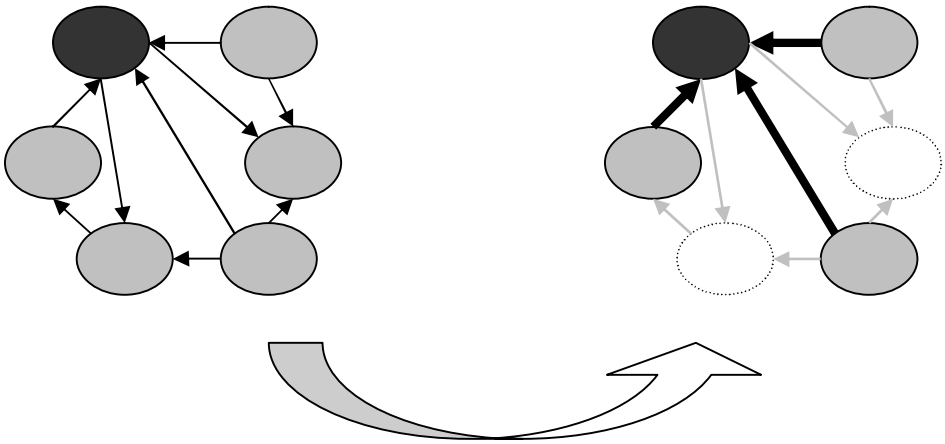
0 – 0,25 – 0,5 – 0,75 - 1

sijasta vaikkapa selvästi käyttäjäystävällisemmät arvot

huono – melko huono – keskinkertainen – melko hyvä – hyvä

Toisaalta mallin rakentaminen ei nyt käy yhtä helposti kuin edellä, koska jokaisen muuttujan osalta on erikseen määritettävä sumeat säännöt. Tämä lähestymistapa tarjoaa kuitenkin selvästi monipuolisemmat mahdollisuudet mallien rakentamiseen.

Tarkastellaan uudestaan kuvan 7.6 karttaa. Nyt voimme ajatella kyseessä olevan alun perin kanonisen korrelaation tapaisen asetelman, jonka me muutamme monimuuttujaregressio-asetelmiksi siten, että valitsemme aina yhden muuttujan kerrallaan ja tutkimme, mitkä ovat sen syy-muuttujia (ks. luvut 5 ja 7.5). Toisin sanoen, valitsemme yhden muuttujan kerrallaan ”selitettäväksi muuttujaksi”, ja sen jälkeen tutkimme, mitkä ovat sen ”selittäjiä” (kuva 7.12). Tämän jälkeen rakennamme tälle kokonaisuudelle päättelymallin.



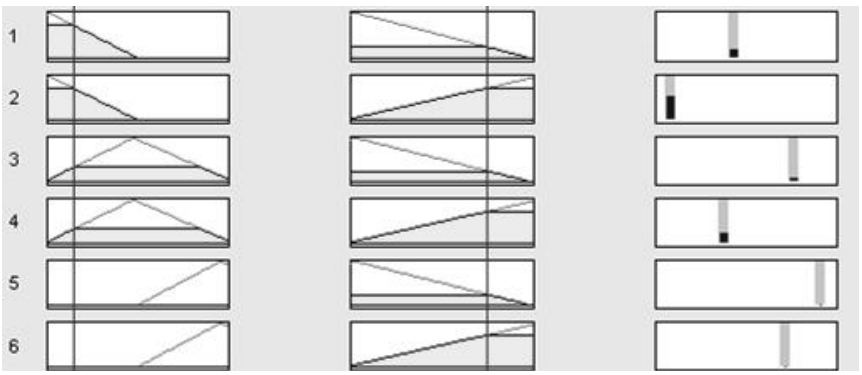
Kuva 7.12. Kielellisen kartan mallinnus. Jokaisen muuttujan osalta tehdään sumea ”regressiomalli” erikseen tämän muuttujan ja sen selittäjien osalta.

Jos esimerkiksi olisimme esittäneet kuvan 7.6 tapauksessa muuttujien N2 ja N6 yhteyden muuttujaan N1 funktion

$$N1 = 1/\exp(1 + (-5 \cdot (0.5 \cdot N2 - 0.3 \cdot N6)))$$

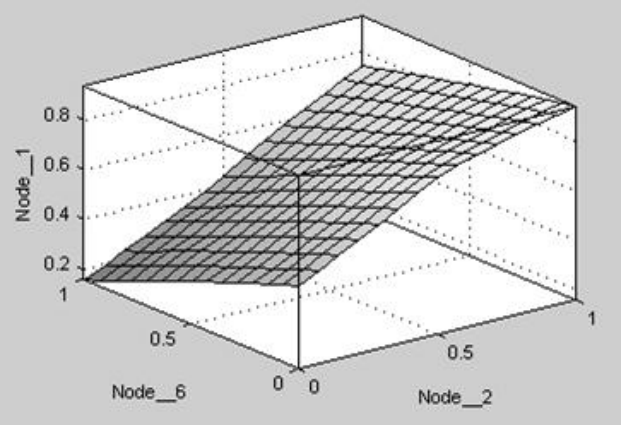
sijasta vastaavalla tavalla kielellisesti, olisimme voineet käyttää vaikkapa sääntöjä (kuva 7.13)

1. Jos N2 on pieni ja N6 on pieni, niin N1 on keskinkertainen.
2. Jos N2 on pieni ja N6 on suuri, niin N1 on melko pieni.
3. Jos N2 on keskinkertainen ja N6 on pieni, niin N1 on melko suuri.
4. Jos N2 on keskinkertainen ja N6 on suuri, niin N1 on keskinkertainen.
5. Jos N2 on suuri ja N6 on pieni, niin N1 on suuri.
6. Jos N2 on suuri ja N6 on suuri, niin N1 on melko suuri.



Kuva 7.13. Sumeat säännöt, jotka kuvaavat muuttujien N2 ja N6 yhteyttä muuttujaan N1.

Tässä olemme siis itse asiassa muodostaneet opetusaineiston edellä mainitun funktion perusteella, ja sitten olemme soveltaneet *anfisedit*-mallinnuksessa ruudukko-tekniikkaa (grid) ja 0. kertaluvun Takagi-Sugeno mallia. Vastaava sovite on kuvassa 7.14, ja se on odotetusti kuvan 7.9 ei-lineaarisen version mukainen. Todellisessa tilanteessahan sääntömme perustuisivat asiantuntemukseen ja empiiriseen aineistoon.

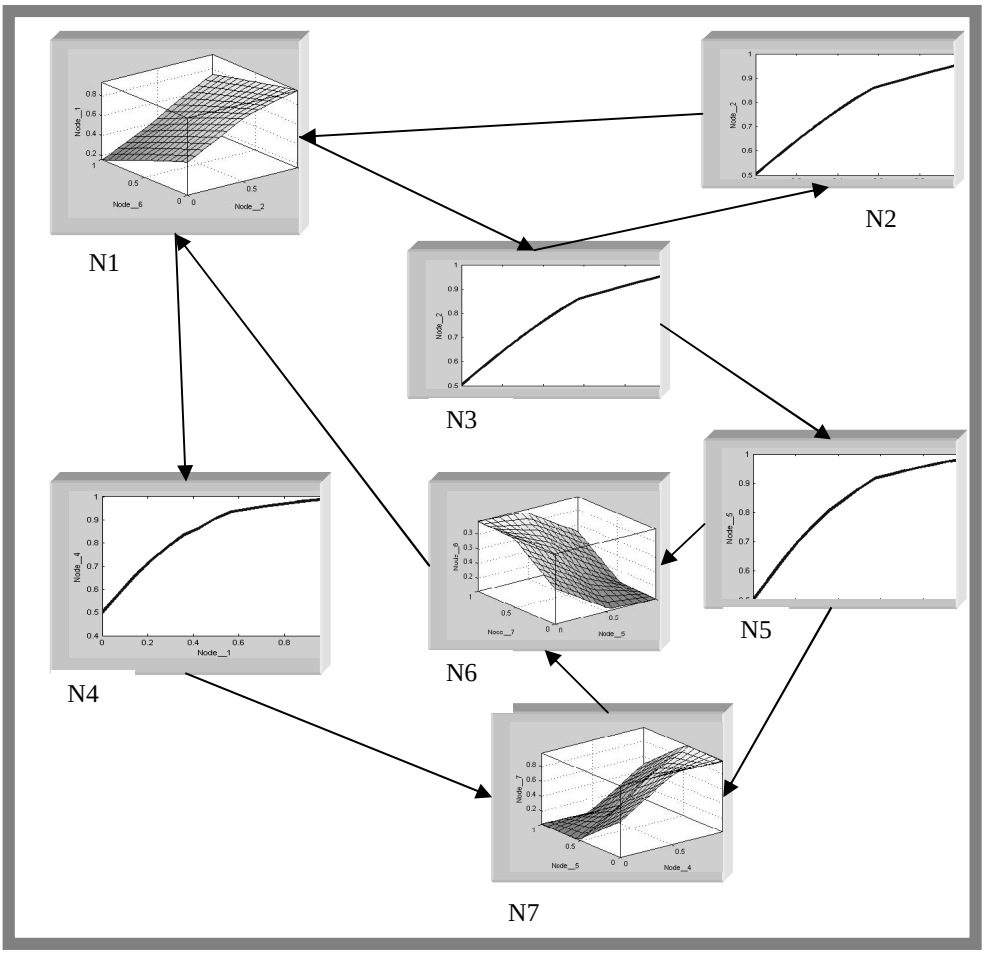


Kuva 7.14. Muuttujien N2 ja N6 yhteys muuttujaan N1 kielellisen mallin perusteella.

Jos taas sovellamme edellä mainittuja sääntöjämme kuvan 7.7 karttaan, voimme esittää ne vaikkapa muodossa

1. Jos MUUTTOLIIKE on pientä ja SAIRAUKSIEN MÄÄRÄ on pieni, niin ASUKASLUKU on keskimääräinen.
2. Jos MUUTTOLIIKE on pientä ja SAIRAUKSIEN MÄÄRÄ on suuri, niin ASUKASLUKU on melko pieni.
3. Jos MUUTTOLIIKE on keskinkertaista ja SAIRAUKSIEN MÄÄRÄ on pieni, niin ASUKASLUKU on melko suuri.
4. Jos MUUTTOLIIKE on keskinkertaista ja SAIRAUKSIEN MÄÄRÄ on suuri, niin ASUKASLUKU on keskimääräinen.
5. Jos MUUTTOLIIKE on suurta ja SAIRAUKSIEN MÄÄRÄ on pieni, niin ASUKASLUKU on suuri.
6. Jos MUUTTOLIIKE on suurta ja SAIRAUKSIEN MÄÄRÄ on suuri, niin ASUKASLUKU on melko suuri.

Kuvan 7.6 kartan kaikki ”regressiomallit” on luonnosteltu kuvassa 7.15 kun sumeiden mallien opetusaineistoina on kunkin muuttujan osalta käytetty vastaavia numeerisen kognitiivisen kartan intensiteetteihin perustuvia funktioita.



Kuva 7.15. Kuvan x muuttujien välisten yhteyksien regressiomallit sumeiden mallien avulla.

Vaikka kielelliset kognitiiviset kartat ovatkin käyttäjäystävällisempiä kuin puhtaasti numeeriset, Koskon ideoihin perustuvat vastineensa, niiden rakentaminen on työläämpää, jos kartassa on paljon muuttujia. Kielelliset sääntöpohjaiset kartat näyttävät kuitenkin monipuolisuutensa ansiosta paremmin vastaavan reaalimaailman ilmiöitä, sillä Kosko-kartat, jotka ovat oikeastaan vain neuroverkon soveltamista monimutkaisiin ilmiöihin, eivät voi esittää vuorovaikutuksia yhtä laaja-alaisesti.

Molempiin karttatyyppeihin liittyy vielä eräitä haasteita kuten erityisesti aikaviiveiden (time delay) mallinnus. Aikaviiveet tarkoittavat tapahtumia, jotka eivät toteudu joka iterointikerralla. Tähänkin kyllä on jo esitetty ratkaisuehdotuksia eräitä [12] [13]. Taulussa 7.1 on esitetty kumpaakin edellä mainittua tyyppiä olevien karttojen etuja ja haittoja.

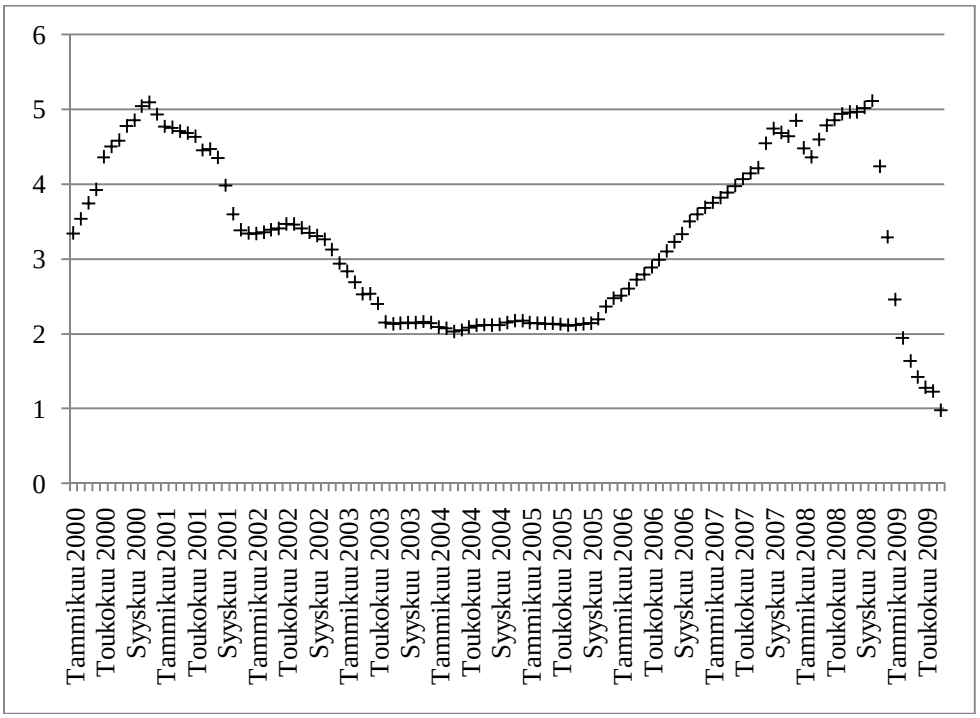
Taulu 7.1. Numeerisen ja kielellisen sumean kognitiivisen kartan vertailua.	
Numeerinen	**Kielellinen**
Melko hyvä monimutkaisten ilmiöiden mallinnuksessa.	Hyvä monimutkaisten ilmiöiden mallinnuksessa.
Matemaattiselta kannalta melko yksinkertainen.	Matemaattiselta kannalta melko yksinkertainen.
Kartta voi sisältää paljonkin muuttujia.	Kartta voi sisältää vain jonkin verran muuttujia.
Voi kuvata vain kausaalisia monotonisia suhteita muuttujien välillä.	Voi kuvata monipuolisesti muuttujien välisiä suhteita.
Pitää käyttää numeerisia arvoja ja intensiteettejä. Neuroverkkomainen.	Voidaan käyttää sekä numeerisia että kielellisiä arvoja ja intensiteettejä.
Voidaan helposti mallintaa myös palautetta (silmukat)	Palautteen mallintaminen hieman työläämpää kuin numeerisessa tapauksessa.
Kartan koon ja rakenteen muuttaminen helppoa.	Kartan koon ja rakenteen muuttaminen voi olla työlästä.
Useimmat kartat perustuvat vielä vain asiantuntemukseen eivätkä empiirisiin aineistoihin.	Useimmat kartat perustuvat vielä vain asiantuntemukseen eivätkä empiirisiin aineistoihin.
Aikaviiveiden mallinnus edellyttää lisäratkaisuja.	Aikaviiveiden mallinnus edellyttää lisäratkaisuja.
Muuttujien ja muuttujien välisten suhteiden tulkinta voi olla ilman havaintoaineistoa ongelmallista.	Muuttujien ja muuttujien välisten suhteiden tulkinta voi olla ilman havaintoaineistoa ongelmallista.

Edellä on tarkasteltu kognitiivisia karttoja seuraten pääasiassa Axelrodin ja Koskon alkuperäisiä ajatuksia. Nämä eivät kuitenkaan ota tarpeeksi huomioon mallien dynaamista luonnetta, eli sitä, että tarkastellaan nimenomaan ilmiöiden muutoksia tietyllä aikavälillä. Koskon mallit ovat myös liian neuroverkkomaisia, ja näin ollen sovellusalueeltaan rajallisia. Toisaalta nämä kartat ovat hyvin tuoneet esille tämän mallinnuksen toimintaperiaatteen. Seuraavaksi tarkastellaankin sellaisia kognitiivisia karttoja, joissa myös ilmiöiden muutos on otettu selvemmin huomioon.

7.4. Muutoskin aikasarjamalliin selvemmin mukaan

Luonnontieteiden dynaamisissa malleissa on jo pitkään otettu huomioon sekä muuttujien arvot että näiden arvojen muutokset. Ne ovat kuitenkin usein olleet matemaattisesti vaikeasti toteutettavissa ja vaatineet esimerkiksi raskasta differentiaalilaskentaa. Uusimmissa kognitiivisissa kartoissa on myös muutokset otettu selvästi paremmin huomioon, vaikka mallinnus on perustunut kielellisiin muuttujiin ja vuorovaikutuksiin.

Tarkastellaan yksinkertaista taloustieteellistä esimerkkiä, jonka osalta oletamme, että tutkittavaa ilmiötä vastaavan kognitiivisen kartan yhtenä muuttujana on euribor-korko, ja keskitymme vain tähän muuttujaan. Tämän koron osalta tarkastelemme nyt, kuinka tietyn ajanhetken korko sekä koron muutos edellisestä ajanhetkestä liittyvät seuraavan ajanhetken korkoon (tätä asetelmaa olemmekin itse asiassa jo soveltaneet Mackeyn ja Glassin aineistoon luvussa 7.1). Näin ollen itse korko-muuttujan lisäksi tarkastelemme myös koron muutosta. Aineistona käytämme euribor-koron kuukausittaisia arvoja tammikuun 2000 ja heinäkuun 2009 välisenä aikana (lähde: Suomen pankin verkkosivut), joka on esitetty kuvassa 7.16.



Kuva 7.16. Euribor-korko (keskimääräinen, %), tammikuu 2000 – heinäkuu 2009.

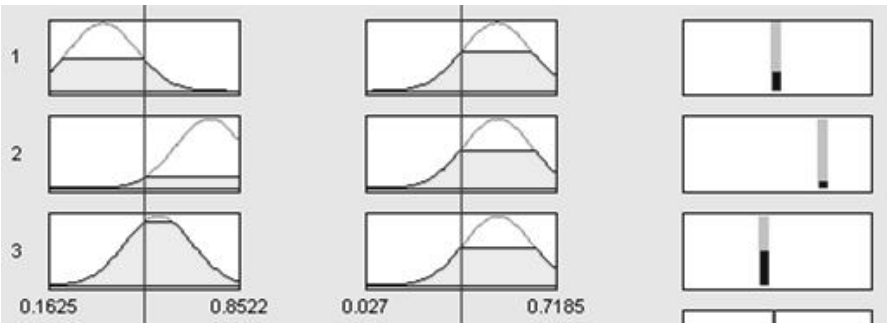
Teemme nyt kyseisen muuttujan osalta regressiomallin, jossa selittäjinä ovat kuukauden (keskimääräinen) euribor-korko (%) ja tämän koron muutos edellisestä kuukaudesta sekä selitettävänä seuraavan kuukauden korko (%). Käytännössä muodostamme opetusaineistomme niin, että esimerkiksi maaliskuun 2005 korkoennuste perustuu saman vuoden helmikuun korkoon sekä helmi- ja tammikuun korkojen erotukseen (maaliskuu-helmikuu). Lisäksi skaalaamme alkuperäisen aineiston arvot välille nollasta yhteen, jotta asteikkojen vaikutus saadaan eliminoitua. Oletamme tässä, että koron alkuperäinen vaihteluväli on nollasta kuuteen prosenttiin ja muutok-

sen miinus yhdestä yhteen. Tällöin vaihteluvälin alarajan arvo koodataan nollaksi ja ylärajan arvo ykköseksi. Tarkemmin, käytämme muunnosta

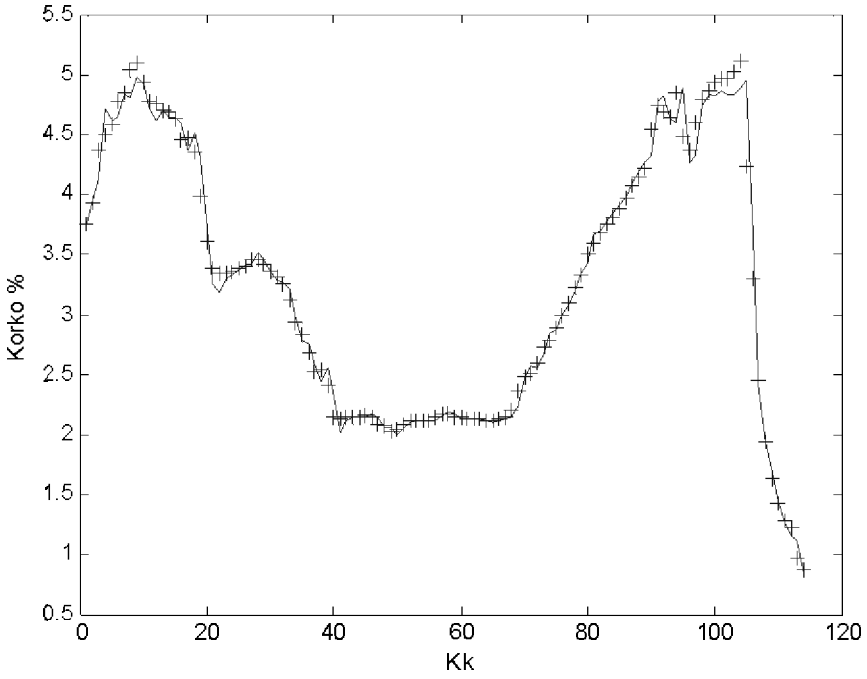
$$\| x - \min \| \; / \; \| \max - \min \|,$$

kun x on alkuperäinen havaintoarvo ja $\| . \|$ on etäisyysmitta.

Tällä tavoin *anfisedit*-mallinnus tuottaa jo kolmella sumealla säännöllä (kuva 7.17, ryväsmenetelmä (subtractive), 1. kertaluvun Takagi-Sugeno –päättely) kuvan 7.18 mukaisen sovitteen, jonka *rmse* = 0,02. Samalla olemme oppineet uuden ja yksinkertaisen tavan tuottaa aikasarjamalleja.



Kuva 7.17. Esimerkki sumeista säännöistä euribor-korkojen mallinnukseen.



Kuva 7.18. Euribor-koron todelliset arvot (+) ja sumean mallin tuottama sovite.

Tässä tapauksessa voidaan myös etsiä tyypillisiä piirteitä koron ”käyttäytymisestä” aineistossa sumeiden sääntöjen avulla. Esimerkiksi Matlab'in *subclust*-tekniikan avulla voimme tuottaa vaikkapa säännöt

1. Jos nykyinen korko on noin 2.14 % ja muutos edellisen kuun korkoon on noin -0.004 prosenttiyksikköä, niin ensi kuun korko on noin 2.16 %.
2. Jos nykyinen korko on noin 3.39 % ja muutos edellisen kuun korkoon on noin 0.034 prosenttiyksikköä, niin ensi kuun korko on noin 3.41 %.
3. Jos nykyinen korko on noin 3.98 % ja muutos edellisen kuun korkoon on noin 0.084 prosenttiyksikköä, niin ensi kuun korko on noin 4.07 %.
4. Jos nykyinen korko on noin 4.76 % ja muutos edellisen kuun korkoon on noin -0.015 prosenttiyksikköä, niin ensi kuun korko on noin 4.71 %.

Jos suosimme kielellisempää ja suhteellisempaa lähestymistapaa, kyseiset säännöt voisivat olla muotoa

1. Jos nykyinen korko on melko alhainen ja muutos edellisen kuun korkoon on hyvin pieni negatiivinen, niin ensi kuun korko melko alhainen.
2. Jos nykyinen korko on keskimääräinen ja muutos edellisen kuun korkoon on hyvin pieni positiivinen, niin ensi kuun korko keskimääräinen.
3. jne.

Samalla tavalla voisimme esittää säännöt tai jopa mallintaa euribor-koron tulevat muutokset koron ja koron aikaisemman muutoksen perusteella.

Tässä on käsitelty vain yhtä kognitiivisen kartan muuttujaa, mutta todellisuudessa samaa menetelmää sovelletaan yleensä kaikkiin kartan muuttujiin. Siis niiden osalta tulisi tarkastella sekä varsinaisia arvoja että arvojen muutosta. Käytännössä se vaatii hieman omaa ohjelmointityötä tai makrojen rakentelua, mikä esimerkiksi Matlab’in osalta edellyttää m-tiedostojen kirjoittamista. Matlab’in liitännäisenä oleva Simulink-ohjelma, jossa voidaan käyttää sumeitakin systeemejä, soveltuu myös hyvin kognitiivisten karttojen mallinnukseen.

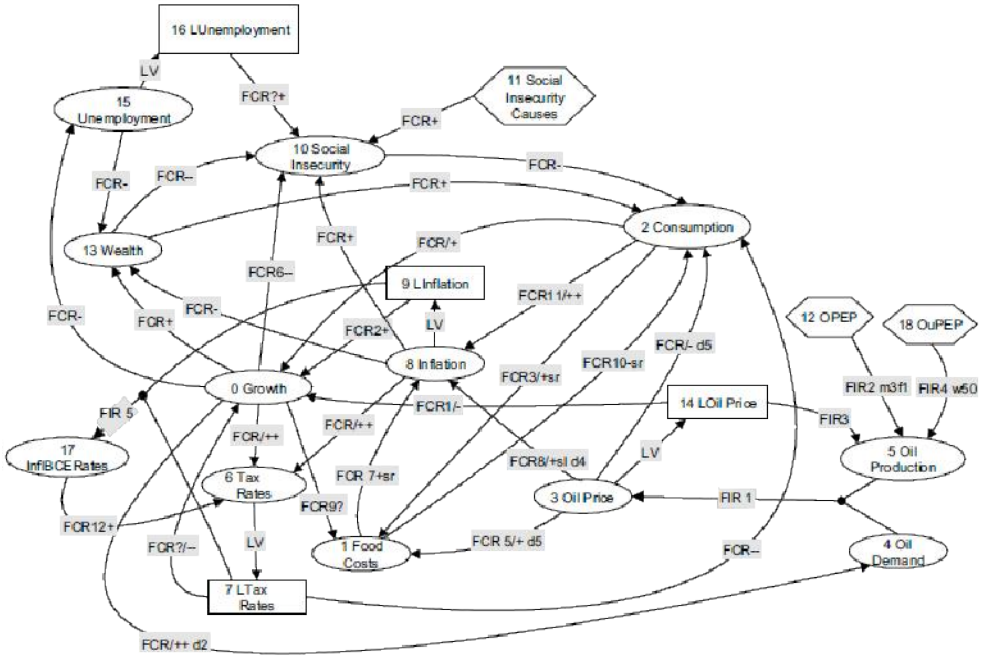
Tämän luvun loppukevennyksenä esitetään kognitiivinen, aikasarja-analyysiä käyttävä kartta, jonka Carvalho ja Tome tekivät jo muutama vuosi sitten eräiden taloustieteellisten julkaisujen perusteella eurooppalaisesta sosio-ekonomisesta systeemistä (kuva 7.19, [13] s. 1826). Kesällä 2009 Lissabonissa pidetyssä sumean logiikan maailmankongressissa he esittivät, kuinka kyseisenä vuonna vallinnut lama oli ennustettavissa, jopa heidän itsensä yllätykseksi, kyseisen mallin avulla (tarkemmat yksityiskohdat ker-

rotaan kyseisessä artikkelissa, jonka voi tarvittaessa pyytää tekijältä). He käyttivät mallissaan paljon suhteellisia arvoja, jotta se soveltuu paremmin erilaisiin olosuhteisiin. Mallin muuttujina olivat muun muassa öljyn kysyntä, öljyn tuotanto, öljyn hinta, ruoan hinta, veroaste, inflaatio, taloudellinen kasvu, varallisuus, kulutus, työttömyys ja sosiaalinen epävarmuus. Lisäksi muuttujina oli näiden tekijöiden muutoksia.

Esimerkiksi ruoan hinnan ja inflaation suhteesta he esittivät seuraavat, suhteellisia arvoja sisältävät säännöt (s. 1823):

1. If food cost decreases very much, then inflation has a large decrease.
2. If food cost decreases much, then inflation has a large decrease.
3. If food cost decreases, then inflation has a large decrease.
4. If food cost decreases few, then inflation decreases.
5. If food cost decreases very few, then inflation decreases.
6. If food cost maintains, then inflation decreases.
7. If food cost increases very few, then inflation has a small decrease.
8. If food cost increases few, then inflation has a very small decrease.
9. If food cost increases normally, then inflation maintains.
10. If food cost increases much, then inflation has a small increase.
11. If food cost increases very much, then inflation increases.

Niinpä nämä säännöt, mikäli ne ovat vuorovaikutuksiltaan riittävän yleistyskelpoisia, ovat suoraan siirrettävissä muihinkin sosio-ekonomisiin systeemeihin.

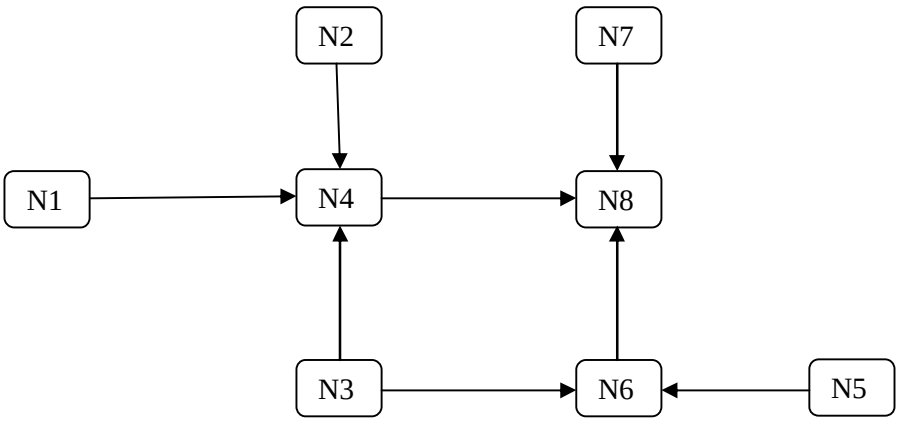


Kuva 7.19. Carvalho'n ja Tome'n makrotalouden malli.

7.5. Regressioanalyysin ja kanonisen korrelaation välimailla

Edellä esitettyä kognitiivisen kartan ideaa voidaan soveltaa myös regressiomalleihin, erityisesti monimuuttujaregressioon (multiple regression) ja kanoniseen korrelaatioon (multivariate regression). Jos nimittäin haluamme tarkastella koko havaintomatriisin osalta kerralla muuttujien välisiä yhteyksiä, voimme nyt määrittää vuorovaikutusmatriisin, jossa on esitetty vastaavat ”regressiokertoimet” tai muut yhteydet.

Tarkastellaan jo luvussa 5 esitettyä kahdeksan muuttujan (N1, …, N8) ja 100 havainnon aineistoa, jossa havainnot on tuotettu muuttujien N1, N2, N3, N5 ja N7 osalta välillä 0 – 1 olevista satunnaisluvuista. Muuttujien N4, N6 ja N8 arvot on taas määritetty siten, että ne perustuvat muiden muuttujien arvoihin kuvan 7.20 osoittamalla tavalla, eli ne ovat toivottuja selitettäviä muuttujia. Aineistostamme pitäisi siis löytyä ainakin kolme regressiomallia, joista muuttujan N4 ollessa selitettävä muuttuja malli on lineaarinen ja muiden osalta ei-lineaarinen (tämä sama rakenne on esitetty evoluutiolaskennan mallinnuksena, joskin eri aineistolla, Leungin artikkelissa [41]).

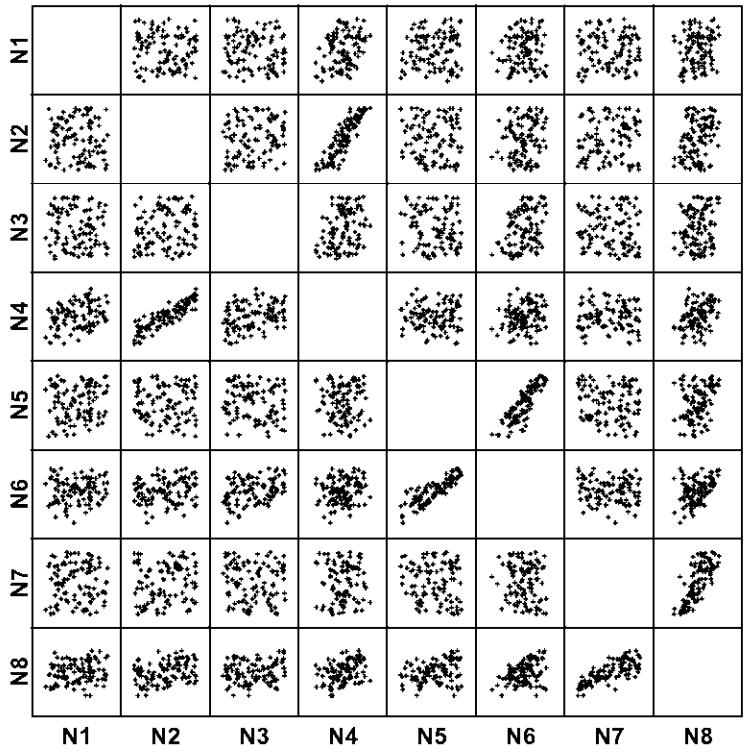


Kuva 7.20. Testiaineiston muuttujien väliset odotetut vuorovaikutukset.

Taulussa 7.2 ovat aineistomme perustunnusluvut (jotka jo olivat luvussa 5) ja kuvassa 7.21 muuttujaparien sirontakuviot ”graafisena korrelaatiomatriisina”. Taulussa 7.3 ovat muuttujien väliset (lineaariset) korrelaatiot. Shapiron ja Wilkin testin perusteella muuttujat N2, N3 ja N7 eivät olisi normaalisti jakautuneita.

Taulu 7.2. Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
N1	100	1,00	,00	1,00	,5088	,26700
N2	100	,99	,00	1,00	,5034	,30459
N3	100	1,00	,00	1,00	,4861	,29876
N4	100	,88	,07	,94	,5033	,19836
N5	100	,97	,00	,98	,5407	,26947
N6	100	,87	,04	,92	,5590	,19224
N7	100	,99	,00	1,00	,5126	,29815
N8	100	,72	,12	,84	,4938	,16739
Valid N (listwise)	100					



Kuva 7.21. Testiaineiston muuttujien väliset suhteet sirontakuvioina.

Sirontakuvioiden perusteella näyttää siltä, että melko selviä lineaarisia korrelaatioita olisi ainakin muuttujaparien N2-N4, N5-N6 ja N7-N8 välillä, ja nämä näkyvät korkeina korrelaatioina myös taulun x korrelaatiomatriisissa. Muitakin tilastollisesti merkitseviä korrelaatioita tästä taulukosta havaitaan.

Taulu 7.3. Correlations

		N1	N2	N3	N4	N5	N6	N7	N8
N1	Pearson Correlation	1	-,069	-,092	,322**	,090	,052	,002	,176
	Sig. (2-tailed)		,494	,363	,001	,375	,609	,988	,080
	N	100	100	100	100	100	100	100	100
N2	Pearson Correlation	-,069	1	,185	,886**	-,009	,102	,093	,337**
	Sig. (2-tailed)	,494		,065	,000	,932	,311	,358	,001
	N	100	100	100	100	100	100	100	100
N3	Pearson Correlation	-,092	,185	1	,297**	-,109	,347**	-,072	,165
	Sig. (2-tailed)	,363	,065		,003	,280	,000	,474	,100
	N	100	100	100	100	100	100	100	100
N4	Pearson Correlation	,322**	,886**	,297**	1	,027	,189	,026	,377**
	Sig. (2-tailed)	,001	,000	,003		,786	,059	,795	,000
	N	100	100	100	100	100	100	100	100
N5	Pearson Correlation	,090	-,009	-,109	,027	1	,857**	-,093	,309**
	Sig. (2-tailed)	,375	,932	,280	,786		,000	,357	,002
	N	100	100	100	100	100	100	100	100
N6	Pearson Correlation	,052	,102	,347**	,189	,857**	1	-,110	,400**
	Sig. (2-tailed)	,609	,311	,000	,059	,000		,277	,000
	N	100	100	100	100	100	100	100	100
N7	Pearson Correlation	,002	,093	-,072	,026	-,093	-,110	1	,763**
	Sig. (2-tailed)	,988	,358	,474	,795	,357	,277		,000
	N	100	100	100	100	100	100	100	100
N8	Pearson Correlation	,176	,337**	,165	,377**	,309**	,400**	,763**	1
	Sig. (2-tailed)	,080	,001	,100	,000	,002	,000	,000	
	N	100	100	100	100	100	100	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Sovellamme nyt kognitiivisten karttojen periaatetta löytääksemme aineistostamme yhdellä kertaa kaikki mahdolliset regressiomallit. Käytännössä jokainen muuttuja on vuorollaan selitettävä muuttuja ja selittäjiksi

valitaan aina muiden muuttujien joukosta relevantit muuttujat. Perinteisillä tavoilla tämä voidaan tehdä esimerkiksi polkuanalyysillä (path analysis), rakenneyhtälömalleilla (structural equation modelling, SEM) ja kanonisilla korrelaatioilla. Ohjelmistoina tähän on tarjolla esimerkiksi *Lisrel*™ ja SPSS:n kylkiäisenä tarjolla oleva *Amos* (tässä myös Bayesin verkkoja).

Voimme myös tehdä regressioanalyysit yksitellen jokaisen muuttujan osalta, joskin tämä on työlästä suuren muuttujajoukon tapauksessa. Usein nämä perinteiset menetelmät perustuvat kuitenkin lineaarisiin vuorovaikutuksiin ja tiettyihin muuttujien jakaumaolettamuksiin, joten niiden käyttöalue ja -kelpoisuus ovat rajallisia.

Me käytämme yksinkertaista, Matlab'in geneettistä algoritmia *ga* hyödyntävää laskentaa, joka pyrkii löytämään myös ei-lineaariset yhteydet. Esitämme muuttujat taulun 7.4 mukaisessa matriisissa, jossa selittäjät ovat riveinä ja selitettävät sarakkeina. Ei-lineaariset regressiomallimme perustuvat tässä tapauksessa aikaisemmin esitettyyn yleistettyyn keskiarvoon (generalized mean [17] [38]), jolloin optimoitavina parametreina ovat selittävien muuttujien painot *w* sekä potenssin *p* arvo. Niinpä esimerkiksi muuttujan N4 selittäjien painot, eli "regressiokertoimet", ovat neljännellä sarakkeella, ja tämän sarakkeen loppuun tulostuu myös potenssin arvo (jonka mukaan kyseessä likimäärin aritmeettinen keskiarvo). Luonnollisesti selitettävä ei itse ole mukana selittäjissä eli nämä solut ovat nollia. Tällä tavoin meidän pitää siis optimoida 7·7+8 = 57 parametria. Muuttujien arvot ovat jo valmiiksi välillä 0 – 1, muuten ne on tällaisiksi muunnettava.

Taulu 7.4. Yleistettyjen keskiarvojen muuttujien painot, potenssit (p), rmse-arvot ja selitysasteet.

	N1	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**
N1	0,000	0,000	0,000	0,294	0,000	0,003	0,000	0,044
N2	0,000	0,000	0,002	0,578	0,000	0,000	0,000	0,043
N3	0,000	0,000	0,000	0,112	0,000	0,286	0,000	0,000
N4	0,602	1,000	0,391	0,000	0,000	0,020	0,000	0,078
N5	0,148	0,000	0,000	0,010	0,000	0,674	0,000	0,000
N6	0,000	0,000	0,602	0,006	1,000	0,000	0,000	0,444
N7	0,000	0,000	0,001	0,000	0,000	0,000	0,000	0,390
N8	0,250	0,000	0,005	0,000	0,000	0,017	1,000	0,000
p	0,093	0,831	0,003	1,022	1,152	1,903	1,016	0,342
rmse	0,257	0,157	0,271	0,038	0,145	0,039	0,202	0,039
R^2				0,947		0,945		0,945

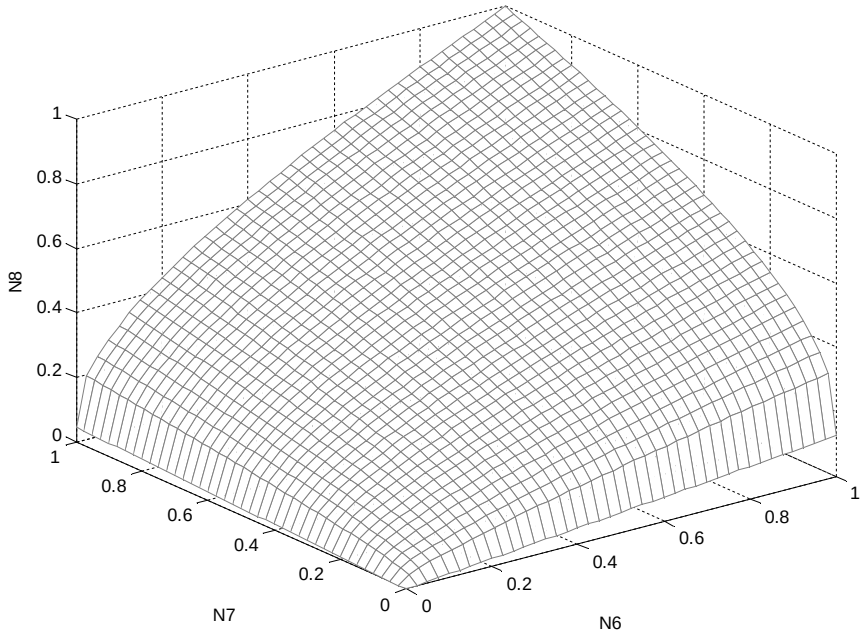
Pyrimme optimoinnin avulla löytämään relevantit selittäjät, ja käytännössä siis tuottamaan nolla- tai hyvin pieniä painoja muille selittäjille. Kun optimointi perustui vain yksinkertaisesti jokaisen selitettävän muuttujan

osalta regressiomallin *rmse*-arvon minimointiin havainto- ja mallin vastearvojen välisien residuaalien perusteella, saimme jo ihan käyttökelpoisia tuloksia. Kolme selvästi pienintä *rmse*-arvoa saatiin kuvan 7.20 mukaisesti selitettävien N4, N6 ja N8 osalta, ja myös kertoimet olivat aina selvästi suurimmat toivottujen selittäjien osalta. Näistä selitettävän muuttujan N4 tapauksessa malli oli miltei lineaarinen (p likimain 1 eli lähellä aritmeettista keskiarvoa) joten tämäkin tavoite miltei saavutettiin.

Jos nyt eliminoimme pienien painokertoimien perusteella vaikkapa muuttujan N8 epäoleelliset selittäjät (siis vain N6 ja N7 jäävät selittäjiksi), saamme yleistettyä keskiarvoa käyttäen, kun optimoidaan kertoimet uudestaan vain näiden muuttujien osalta, lopulliseksi regressioyhtälöksi

$$N8 = (0{,}566 \cdot N6^{0{,}273} + 0{,}434 \cdot N7^{0{,}273})^{(1/0{,}273)}$$

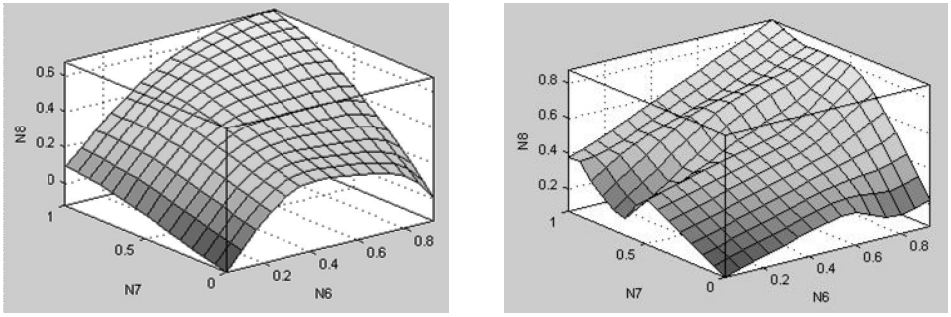
ja *rmse* = 0,059 (selitysaste = 0,876). Kuvassa 7.22 on esitetty tämän mallin ei-lineaarinen sovitepinta.



Kuva 7.22. Yleistettyyn keskiarvoon perustuva sovite muuttujalle N8 kun selittäjinä ovat vain muuttujat N6 ja N7.

Numeerisen ratkaisun lisäksi edellä käsitelty mallinnus voidaan tehdä sumeiden mallien avulla, jolloin yleistettyjen keskiarvojen sijasta jokainen

selitettävä muuttuja siis mallinnetaan sumeiden sääntöjen perusteella. Tällä tavoin vuorovaikutuksia voidaan esittää monipuolisemmin ja tarvittaessa myös kielellisesti, mutta optimoitavia parametreja on enemmän. Kuvassa 7.23 on luonnosteltu sumeat mallit *anfisedit*-mallinnuksella vastemuuttujan N8 tapauksessa siten, että toisessa on käytetty kaikkia muita muuttujia syötemuuttujina, kun taas toinen on tehty edellä olevan numeerisen mallinnuksen perustella vain syötemuuttujien N6 ja N7 perusteella. Tällä tavoin *rmse*-arvot ovat pienempiä (0,007 ja 0,048), joten tässä mielessä neuro-sumeat mallit näyttävät olevan parempia kuin edellä olevat matemaattiset mallit.



Kuva 7.23. Neuro-sumeisiin malleihin perustuvat sovitteet muuttujalle N8. Projektio kahdeksan säännön mallin sovitteesta kun kaikki muut muuttujat ovat selittäjiä (vas., ryväsmenetelmä, rmse=0,007). Neljän säännön malli kun vain N6 ja N7 ovat selittäjiä (oik., ryväsmenetelmä, rmse=0,048).

Vertailun vuoksi tehtiin myös lineaariset regressiomallit edellä mainittujen kolmen selitettävän muuttujan osalta (taulut 7.5-7.10), ja ne perustuivat SPSS:n askeltavaan (stepwise) menetelmään, jotta ohjelma saattoi vapaasti valita ”parhaimmat” selittäjät. Tällä tavoin valittujen mallien joukkoon kuuluivat myös edellä esitetyt ei-lineaariset mallit, joskin eivät aina parhaimpina lineaarisina malleina, mikä ilmeisesti johtui aineiston ei-lineaarisesta luonteesta.

Lineaariset mallit olivat tässä tapauksessa myös hyviä, koska aineistossa ei esiintynyt kovin paljon ei-lineaarisuutta.

Taulu 7.5. Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,886[a]	,786	,783	,09232
2	,966[b]	,933	,932	,05176
3	,980[c]	,961	,960	,03964
4	,981[d]	,963	,962	,03883

a. Predictors: (Constant), N2

b. Predictors: (Constant), N2, N1

c. Predictors: (Constant), N2, N1, N3

d. Predictors: (Constant), N2, N1, N3, N7

e. Dependent Variable: N4

Taulu 7.6. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,213	,018		11,889	,000		
	N2	,577	,030	,886	18,949	,000	1,000	1,000
2	(Constant)	,058	,015		4,016	,000		
	N2	,595	,017	,913	34,733	,000	,995	1,005
	N1	,286	,020	,385	14,656	,000	,995	1,005
3	(Constant)	,008	,013		,649	,518		
	N2	,575	,013	,882	43,109	,000	,963	1,038
	N1	,296	,015	,399	19,744	,000	,989	1,011
	N3	,113	,014	,171	8,326	,000	,960	1,042
4	(Constant)	,023	,014		1,645	,103		
	N2	,578	,013	,887	43,996	,000	,952	1,051
	N1	,296	,015	,399	20,159	,000	,989	1,011
	N3	,111	,013	,167	8,260	,000	,952	1,051
	N7	-,030	,013	-,045	-2,250	,027	,983	1,017

Taulu 7.6. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,213	,018		11,889	,000		
	N2	,577	,030	,886	18,949	,000	1,000	1,000
2	(Constant)	,058	,015		4,016	,000		
	N2	,595	,017	,913	34,733	,000	,995	1,005
	N1	,286	,020	,385	14,656	,000	,995	1,005
3	(Constant)	,008	,013		,649	,518		
	N2	,575	,013	,882	43,109	,000	,963	1,038
	N1	,296	,015	,399	19,744	,000	,989	1,011
	N3	,113	,014	,171	8,326	,000	,960	1,042
4	(Constant)	,023	,014		1,645	,103		
	N2	,578	,013	,887	43,996	,000	,952	1,051
	N1	,296	,015	,399	20,159	,000	,989	1,011
	N3	,111	,013	,167	8,260	,000	,952	1,051
	N7	-,030	,013	-,045	-2,250	,027	,983	1,017

a. Dependent Variable: N4

Taulu 7.7. Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,857[a]	,734	,732	,09960
2	,965[b]	,931	,929	,05104

a. Predictors: (Constant), N5

b. Predictors: (Constant), N5, N3

c. Dependent Variable: N6

Taulu 7.8. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,228	,022		10,189	,000		
	N5	,611	,037	,857	16,457	,000	1,000	1,000
2	(Constant)	,070	,015		4,698	,000		
	N5	,646	,019	,906	33,733	,000	,988	1,012
	N3	,287	,017	,446	16,616	,000	,988	1,012

a. Dependent Variable: N6

Taulu 7.9. Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,763[a]	,582	,578	,10878
2	,905[b]	,818	,815	,07205
3	,944[c]	,890	,887	,05626
4	,946[d]	,895	,891	,05536

a. Predictors: (Constant), N7

b. Predictors: (Constant), N7, N6

c. Predictors: (Constant), N7, N6, N4

d. Predictors: (Constant), N7, N6, N4, N1

e. Dependent Variable: N8

Taulu 7.10. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,274	,022		12,632	,000		
	N7	,428	,037	,763	11,680	,000	1,000	1,000
2	(Constant)	,021	,027		,773	,441		
	N7	,458	,024	,817	18,762	,000	,988	1,012
	N6	,426	,038	,489	11,242	,000	,988	1,012
3	(Constant)	-,066	,024		-2,794	,006		
	N7	,451	,019	,804	23,615	,000	,986	1,015
	N6	,380	,030	,436	12,587	,000	,951	1,052
	N4	,231	,029	,274	7,942	,000	,962	1,040
4	(Constant)	-,079	,024		-3,291	,001		
	N7	,451	,019	,804	24,017	,000	,986	1,015
	N6	,380	,030	,437	12,814	,000	,951	1,052
	N4	,211	,030	,250	7,001	,000	,864	1,157
	N1	,045	,022	,072	2,040	,044	,896	1,116

a. Dependent Variable: N8

Tässäkin yhteydessä olisi voinut tarvittaessa hyödyntää myös summamuuttujia selittäjien määrän vähentämiseksi ja näin ollen mallien yksinkertaistamiseksi, mutta kuten aikaisemmin on jo todettu, epäoleellisten selittäjien eliminointiin tai samantyyppisten muuttujien yhdistämiseen on käytössä pääasiassa vain perinteisiä, usein aineiston lineaarisuuteen perustuvia tekniikoita (esim. pääkomponentti- ja faktorianalyysi). Niinpä edelläkin oli turvauduttava sellaisiin optimointimenetelmiin, jotka pyrkivät myös ottamaan huomioon epäoleellisten selittäjien ongelman. Tältä osin kaivataan kuitenkin lisää metodologisia apuvälineitä.

Kognitiivisen kartan periaatteen soveltaminen tilastotieteeseen avaa mielenkiintoisia näkymiä, kunhan siihen saadaan kehitettyä sopivia algoritmeja. Edellä käsitellyn, monimuuttujaregression, kanonisen korrelaation ja rakenneyhtälömallin välimailla olevan, sovelluksen lisäksi tätä ajatusta voisi hyödyntää esimerkiksi sumeutetussa pääkomponentti- tai faktorianalyysissä

(R-tekniikka), ja tätä aihepiiriä tekijä tällä hetkellä pohtiikin. Laadullisessa tutkimuksessa taas kannattanee ehkä ensin tehdä ilmiöstä käsitekartta (concept map), joka sitten ”operationalisoidaan” eli muunnetaan kielelliseksi kognitiiviseksi kartaksi muun muassa muodostamalla erikseen muuttujat osailmiöistä ja niiden muutoksista. Tämän jälkeen voidaan tehdä simulointeja tietokoneympäristössä, mikä muuten on ollut laadullisessa tutkimuksessa ongelmallista.

8. LOPUKSI

Edellä olemme tarkastelleet eräitä sellaisia ihmistieteiden sovelluskohteita, joissa älykkäitä ja oppivia järjestelmiä voidaan hyödyntää sellaisenaan tai perinteisten menetelmien tukena. Kvantitatiivisessa tutkimuksessa, johon näitä järjestelmiä on tällä hetkellä enemmän sovellettu, voimme erityisesti tällä tavoin korvata tai täydentää perinteistä tilastollis-matemaattista mallinnusta. Kvalitatiivisessa tutkimuksessa taas voimme uusimuotoisten kielellisten malliemme avulla hyödyntää selvästi enemmän tietokoneympäristöä.

Molemmissa tutkimusperinteissä edellä käsitellyt mallinnukset edellyttävät sellaista uutta ihmisen ajattelua muistuttavaa ajattelutapaa, joka on usein luonteeltaan kielellistä ja jopa epätäsmällistä. Tällainen ajattelutavan muutos voi olla vaikeaa erityisesti kvantitatiivisessa tutkimusperinteessä, jossa tutkijat on opetettu muuntamaan ilmiöt matemaattisiksi funktioiksi. Ihanteellisena tilanteena kuitenkin olisi sellainen älykkäiden ja oppivien järjestelmien edesauttama ihmistieteiden *kaikkien* menetelmien rinnakkaiselo, että tutkijat ennakkoluulottomasti, mutta samalla terveen kriittisesti, valitsisivat käyttöönsä parhaimmat menetelmät.

Koska olemme edellä keskittyneet vain muutamiin tyypillisiin älykkäiden ja oppivien järjestelmien sovelluskohteisiin, lukijan on itsenäisesti hankittava lisätietoa omasta tutkimuskohteestaan. Tältä osin hänellä on onneksi runsaudenpula, sillä jo pelkästään sumeista systeemeistä on tarjolla kymmeniä tuhansia julkaisuja. Esimerkiksi Springer-kustantamolla on parikin kirjasarjaa, jotka keskittyvät pelkästään näihin uusiin menetelmiin. Lisäksi tästä aihepiiristä julkaistaan useita tieteellisiä julkaisusarjoja, joista tunnetuimmat ovat Fuzzy Sets and Systems sekä IEEE:n sarjat. Helpointa on kai aluksi hyödyntää Internetin kirjasto- ja hakupalveluja (NELLI, Google Scholar jne.).

Ihmistieteiden tutkimuksen perimmäinen tavoite on lisätä ihmiskunnan hyvinvointia ympäristöämme vahingoittamatta, ja näissä tutkimuskohteissa on vielä paljon haasteita. Toivottavasti älykkäät ja oppivat järjestelmät omalta osaltaan edesauttavat lukijoita tässä tärkeässä tehtävässä.

LÄHTEET

1. Adaptive and Intelligent Systems Applications 1994−1999. Tekes Technology Programme Report, Karisto Oy, Helsinki, 2002.
2. Alander, J.: Geneettisten algoritmien mahdollisuudet. Tekes, julkaisu 59/98, Paino-Center, Helsinki, 1998.
3. Alkula, T. ym.: Sosiaalitutkimuksen kvantitatiiviset menetelmät. WSOY, Juva, 1994.
4. Axelrod, R.: Structure of Decision: The Cognitive Maps of Political Elites. Princeton University Press, Princeton, 1976.
5. Bandemer, H. & Näther, W.: Fuzzy Data Analysis. Kluwer, Dordrecht, 1992.
6. Barabasi, A.-L.: Linked, Plume, New York, 2003.
7. Bertoluzza, C. ym. (toim.): Statistical Modeling, Analysis and Management of Fuzzy Data. Physica Verlag, Berlin, 2002.
8. Bezdek, J. ym.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer, Norwell, 1999.
9. Bezdek, J. ym. (toim.): Fuzzy Sets in Approximate Reasoning and Information Systems. Kluwer, Dordrecht, 1999.
10. Bezdek, J. ym.: Fuzzy Kohonen Clustering Networks. Proceedings of the IEEE International Conference on Fuzzy Systems, San Diego, 1992. ss. 1035-1043.
11. Castillo, O. ym. (toim.): Theoretical Advances and Applications of Fuzzy Logic and Soft Computing, Vol. 42, Springer, Heidelberg, 2007.
12. Carvalho, J. & Tome, J.: Qualitative Modelling of an Economic System Using Rule Based Fuzzy Cognitive Maps, FUZZ-IEEE 2004 - IEEE International Conference on Fuzzy Systems, 2004.
13. Carvalho, J. & Tome, J.: Rule based fuzzy cognitive maps in socio-economic systems. Proceedings of the IFSA-EUSFLAT Congress, Lisbon, 2009. ss. 1821-1826.
14. Chiu, S.: Fuzzy Model Identification Based on Cluster Estimation. Journal of Intelligent and Fuzzy Systems 2, 1994. ss. 267-278.
15. Cohen, L. & Manion, L.: Research Methods in Education. Routledge, London, 1989.
16. Dimitrov, V. & Hodge, B.: Social Fuzziology – Study of Fuzziness of Social Complexity. Physica Verlag, Heidelberg, 2002.
17. Dyckhoff, H. & Pedrycz, W.: Generalized Means as Model of Compensative Connectives. Fuzzy Sets and Systems 14, 1984. ss. 143-154.
18. Eiben, A.E. & Smith, J.E.: Introduction to Evolutionary Computing. Natural Computing Series Springer, Heidelberg, 2007.
19. Fullér, R.: Fuzzy Reasoning and Fuzzy Optimization. Turku Centre for Computer Science, TUCS General Publication 9, 1998.
20. Grzegorzewski, P. ym. (toim.): Soft Methods in Probability, Statistics and Data Analysis. Physica Verlag, Heidelberg, 2002.
21. Guilford, J. & Fruchter, B.: Fundamental Statistics in Psychology and Education. McGraw-Hill, London, 1978.
22. Gupta M. & Rao D.: On the Principles of Fuzzy Neural Networks. Fuzzy Set and Systems 61 (1), 1994. ss. 1-18.
23. Herrera, F. & Verdegay, J.: Genetic Algorithms and Soft Computing. Physica Verlag, Heidelberg, 1996.
24. Hong, T.-P. & Chen, J.-B.: Finding Relevant Attributes and Membership Functions. Fuzzy Sets & Systems, 103 (3), 1999. pp. 389-404.

25. Hruschka, E. R. ym.: A Survey of Evolutionary Algorithms for Clustering. IEEE Transactions on Systems, Man and Cybernetics. Part C: Applications and Reviews 39 (2), 2009. pp. 133-155.
26. Hyvönen, E. (toim.): Inhimillinen kone − konemainen ihminen. Yliopistopaino, Helsinki, 2001.
27. Hyvönen, E. (toim.): Tekoälyn ensyklopedia. Gaudeamus, Hämeenlinna, 1993.
28. Isomursu, P. ym.: Sumean logiikan mahdollisuudet. Tekes, julkaisu 34/93, PunaMusta, Helsinki, 1995. Uudistettu painos Tekesin Internet-sivulla http://www.tekes.fi/julkaisut/sumea/index.html.
29. Kacprzyk, J. and Fedrizzi, M. (toim.): Fuzzy Regression Analysis, Physica Verlag, Heidelberg, 1992.
30. Kohonen, T.: Self-Organization and Associative Memory. Springer Verlag. Berlin, 1987.
31. Koikkalainen, P. (toim.): Neurolaskennan mahdollisuudet. Tekes, julkaisu 43/94, Paino-Center, Helsinki, 1994.
32. Jang, R.: ANFIS: Adaptive Network-Based Fuzzy Inference System. IEEE Transactions on Systems, Man and Cybernetics 23 (3), 1993. ss. 665-685.
33. Jokivuori, P. & Hietala, R.: Määrällisiä tarinoita. Monimuuttujamenetelmien käyttö ja tulkinta. WSOY, Porvoo, 2007.
34. Kardaras, B. & Karakostas, B.: The Use of Fuzzy Cognitive Maps to Simulate the Information Systems Strategic Planning Process, Information and Software Technology 41, 1999. ss. 197-210.
35. Kim, S. & Lee, C.: Fuzzy Implications of Fuzzy Cognitive Map with Emphasis on Fuzzy Causal Relationship and Fuzzy Partially Causal Relationship, Fuzzy Sets and Systems 97 (3), 1998. ss. 303-313.
36. Kosko, B.: Fuzzy Engineering, Prentice-Hall, Upper Saddle River, New Jersey, 1997.
37. Kosko, B.: Sumea logiikka. Art House, Jyväskylä, 1993.
38. Krishnapuram, R. & Lee, J.: Fuzzy Connective-Based Hierarchical Aggregation Networks for Decision Making. Fuzzy Sets and Systems 1 (46), 1992. ss. 11-28.
39. Kruse, R. & Meyer, K.: Statistics with Vague Data. Reidel, Dordrecht, 1987.
40. Lee K. ym.: Strategic Planning Simulation Based on Fuzzy Cognitive Map Knowledge and Differential Game, Simulation, 71, 1998. ss. 316-327.
41. Leung, K.-S. ym.: Learning non-linear multiregression networks based on evolutionary computation. IEEE Transactions on Systems, Man and Cybernetics, Part B, 32 (5), 2002. ss. 630-644.
42. Mattila, J. K.: Sumean logiikan oppikirja. Art House, Vantaa, 2002.
43. Metsämuuronen, J.: Tutkimuksen tekemisen perusteet ihmistieteissä. International Methelp Ky, Gummerus, Jyväskylä, 2006.
44. Y. Miao, ym.: Dynamical Cognitive Network – an Extension of Fuzzy Cognitive Map, IEEE Transactions on Fuzzy Systems 9 (5), 2001. ss. 760-770.
45. Myllymäki, P. & Tirri H.: Bayes−verkkojen mahdollisudet. Tekes, julkaisu 58/98, Painocenter, Helsinki, 1998.
46. Niemi, A.: Johdatus sumeisiin joukkoihin ja sumeaan logiikkaan. Opetushallitus, Hakapaino Oy, Helsinki, 1996.
47. Niskanen, V. A.: Prospects for Integrating Analysis of Variance with Soft Computting, Information Sciences 134, 2001. ss. 135-166.

48. Niskanen, V. A.: Sumea logiikka, kirkasta älyä ja mallinnusta. WSOY, Porvoo, 2003.
49. Niskanen, V. A.: Soft Computing Methods in Human Sciences, Springer Verlag, Heidelberg, 2003.
50. Novak, J.: Learning, Creating and Using Knowledge, Concept Maps as Faciliative Tools in Schools and Corporations, Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.
51. Puolakka, H.: Sumea logiikka käytännön sovelluksissa. Opetushallitus, Hakapaino Oy, Helsinki, 1997.
52. Ragin, C.: Fuzzy-Set Social Science. University of Chicago Press, Chicago, 2000.
53. Ragin, C.: Redesigning Social Inquiry, Fuzzy Sets and Beyond. University of Chicago Press, Chicago, 2008.
54. Römer C. & Kandel, A.: Statistical Tests for Fuzzy Data. Fuzzy Sets and Systems 72 (1), 1995. ss. 1-26.
55. Schneider M. ym.: Automatic Construction of FCMs, Fuzzy Sets and Systems vol. 93 (2),1998. ss. 161-172.
56. Setnes, M. & Roubos, H.: GA-Fuzzy Modeling and Classification: Complexity and Performance. IEEE Transactions on Fuzzy Systems 5 (8), 2000. ss. 509-522.
57. Silverman, D.: Interpreting Qualitative Data. Sage, Thousand Oaks, 1993.
58. Smithson, M.: Fuzzy Set Analysis for Behavioural and Social Sciences. Springer Verlag, New York, 1987.
59. Ranta, E. ym.: Biometria. Yliopistopaino, Helsinki, 1991.
60. Stach, W. ym.: Genetic Learning of Fuzzy Cognitive Maps, Fuzzy Sets and Systems 153 (3), 2005. ss. 371-401.
61. Stach, W. ym.: Numerical and Linguistic Prediction of Time Series with the Use of Fuzzy Corgnitive Maps. IEEE Transactions on Fuzzy Systems, 1 (16), 2008. ss. 61-72.
62. Sneath, P. & Sokal, R.: Numerical Taxonomy. Freeman, San Francisco, 1973.
63. Stylios, C. ym.: Modeling Complex Systems Using Fuzzy Cognitive Maps, IEEE Transactions on Systems, Man and Cybernetics Part A 34 (1), 2004. ss. 155-162.
64. Stylios C. ym.: The Challenge of Modeling Supervisory Systems Using Fuzzy Cognitive Maps, J. of Intelligent Manufacturing 9, 1998. ss. 339-345.
65. Takagi, T. & Sugeno, M.: Fuzzy Identification of Systems and Its Applications to Modeling and Control. IEEE Transactions on Systems, Man and Cybernetics, 15 (1), 1986. ss. 116-132.
66. Yager, R. & Filev, D.: Generation of Fuzzy Rules by Mountain Clustering. Journal of Intelligent and Fuzzy Systems 2, 1994. ss. 209-219.
67. Zadeh, L. (1999): From Computing with Numbers to Computing with Words - From Manipulation of Measurements to Manipulation of Perceptions. IEEE Transactions on Circuits and Systems 45, 1999. ss. 105-119.
68. Zadeh, L.: From Search Engines to Question Answering Systems? The Problems of World Knowledge, Relevance, Deduction and Precisiation. In Sanchez, E. (toim.), Fuzzy Logic and the Semantic Web, Elsevier, Amsterdam, 2006.
69. Zadeh, L.: Fuzzy Logic and Approximate Reasoning. Synthese 30. 1975. ss. 407-428.
70. Zadeh, L.: Fuzzy Logic = Computing with Words, IEEE Transactions on Fuzzy Systems 2, 1996. ss. 103-111.
71. Zadeh, L.: Toward Extended Fuzzy Logic – A First Step. Fuzzy Sets and Systems 160. ss. 3175-3181.

72. Zadeh, L.: Toward a Perception-Based Theory of Probabilistic Reasoning with Imprecise Probabilities. Journal of Statistical Planning and Inference 105/2, 2002. ss. 233-264.
73. Zadeh, L.: Toward a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic, Fuzzy Sets and Systems 90 (2), 1997. ss. 111-127.
74. Zar, J.: Biostatistical Analysis. Prentice-Hall, Englewood Cliffs, New Jersey, 1984.