Nina Hänninen
nhannine@student.uef.fi
10. lokakuuta 2011

Mikko Kolehmainen

Basics of Multivariate Methods 2011

Laskuharjoitus 4:
Itseorganisoituva kartta (SOM) ja Sammon-kartta

1 Johdanto

Itseorganisoituva kartta (SOM) ja Sammon-kartta ovat tehokkaita menetelmiä monidimensioisen datan analysoimiseen ja klusteroimiseen. SOM-menetelmässä datan muuttujien väliset suhteet esitetään yksinkertaisina geometrisina suhteina kaksiulotteisessa kartassa. SOM-karttaa voidaan kuvata myös U-matriisin avulla, jossa jokaiselle alkiolle lasketaan sen keskimääräinen etäisyys naapurialkioista. Sammon-kartta taas kuvaa datavektoreiden suhteellisia etäisyyksiä toisistaan.

Klusterointi on mahdollista sekä SOM- että Sammon-kartan perusteella. SOM-kartassa alkiot ovat järjestyneet siten, että samantyyppiset alkiot ovat vierekkäin ja niistä voidaan muodostaa klustereita. Klustereiden muodostamiseen voidaan lisäksi hyödyntää U-matriisin antamaa tietoa alkioiden välisistä etäisyyksistä. Sammon-kartasta klustereita voidaan etsiä lähekkäimmin toisiinsa nähden sijaitsevista alkioista.

Tässä työssä tutustuttiin itseorganisoituvaan karttaan (SOM) sekä Sammon-karttaan analysoimalla niiden avulla ilmanlaatumittauksesta saatua dataa. Datalle muodostettiin esikäsittelyn jälkeen SOM-kartta, jonka ominaisuuksia tarkasteltiin erilaisten kuvaajien ja U-matriisin avulla. SOM-kartan perusteella määritettiin myös Sammon-kartta, jonka jälkeen datasta pyrittiin näiden karttojen avulla löytämään klustereita. Lopuksi klustereille pyrittiin löytämään ilmanlaatua kuvaavat selitykset klustereista saatavien tietojen avulla.

2 Materiaalit ja menetelmät

2.1 Itseorganisoituva kartta (SOM)

Itseorganisoituva kartta (Self-organizing map, SOM) on tehokas väline monidimensioisen datan esittämiseen. Sen avulla monidimensioisen datan muuttujien monimutkaiset suhteet toisiinsa voidaan esittää yksinkertaisina geometrisina suhteina, esimerkiksi kaksiulotteisena karttana. [1]

SOM-kartta koostuu M kappaleesta neuroneita, jotka ovat järjestäytyneet kaksiulotteiseksi verkoksi. Jokaisella neuronilla on oma painovektorinsa $\mathbf{w_m} = (w_{m1}, w_{m2}, \ldots, w_{mp})$, $m = 1, \ldots, M$, jossa painojen lukumäärä on sama kuin mitattujen muuttujien lukumäärä p. Painovektorit alustetaan satunnaisilla arvoilla, jonka jälkeen jokaiselle datavektorille $\mathbf{x_i}$ etsitään sitä parhaiten vastaava neuroni (Best-Matching Unit, BMU), eli neuroni, joka on Euklidisen etäisyyden suhteen lähimpänä datavektoria:

$$c(\mathbf{x_i}, \mathbf{W}) = \arg \min_j \|\mathbf{x_i} - \mathbf{w_j}\|, \tag{1}$$

missä $\mathbf{W}$ sisältää kaikki painovektorit.

BMU:n ja sen naapurifunktion mukaan määritettävien naapurineuronien painot päivitetään seuraavan yhtälön mukaisesti kohti datavektoria $\mathbf{x_i}$:

$$\mathbf{w_m}(t+1) = \mathbf{w_m}(t) + h_{cm}(t)\left[\mathbf{x_i} - \mathbf{w_m}(t)\right], \tag{2}$$

missä t on iteraatiokierroksen järjestysluku, c BMU:n indeksi ja m päivitettävän neuronin indeksi. Naapurifunktio h_{cm} määritellään Gaussiselle funktiolle:

$$h_{cm}(t) = \alpha(t)\exp\left(-\frac{\|\mathbf{r_c} - \mathbf{r_m}\|^2}{2\sigma^2(t)}\right), \tag{3}$$

missä $\mathbf{r_c}$ ja $\mathbf{r_m}$ ovat vastaavien neuronien paikkavektorit, $\sigma(t)$ on kernelin leveys ja $\alpha(t)$ oppimiskykykerroin (learning rate factor).

Algoritmin konvergoitumisen jälkeen voidaan muodostaa kaksiulotteinen SOM-kartta. Kartan ominaisuutena on, että se on approksimaatio alkuperäisestä datasta, sillä painovektorit tiivistävät alkuperäisen datan informaatiota. Kartta on myös topologisesti järjestynyt, eli neuronin paikka verkossa vastaa tiettyä alkuperäisen datan ominaisuutta, joten lähekkäiset neuronit ovat keskenään samantyyppisiä. SOM-kartan muita piirteitä on se, että se heijastaa alkuperäisen datan jakaumaa ja että se osaa valita parhaat ominaisuudet datan jakauman approksimointiin, vaikka kyseessä olisi epälineaarinen jakauma. Tämä on etu verrattuna esimerkiksi PCA-algoritmiin, joka pystyy löytämään vain lineaariset jakaumat. [2]

2.2 U-matriisi

Eräs tapa kuvata SOM-karttaa on esittää se U-matriisina. U-matriisi (Unified Distance Matrix) kuvaa SOM-kartan neuronien välisiä suhteellisia etäisyyksiä. Neuronin etäisyys naapurineuroneihin lasketaan kaavalla:

$$d_{umat}(\mathbf{w_i}) = \frac{1}{N}\sum_{j=1}^{N}\|\mathbf{w_i} - \mathbf{w_j}\|, \ j \in S_{NN}, \tag{4}$$

missä N on naapurineuronien lukumäärä vektorille $\mathbf{w_i}$ ja S_{NN} kuvaa naapurineuronien joukkoa. Tuloksena saadaan jokaiselle neuronille etäisyyttä kuvaava lukuarvo, joka voidaan esittää harmaasävynä tai Z-akselin arvona kolmiulotteisessa kuvaajassa. Matalat vierekkäiset arvot (vaalea harmaasävy) kuvaavat lähekkäin olevia alueita, eli niistä voidaan muodostaa klustereita, kun taas suuret arvot toimivat klustereita erottavina alueina. [2, 3]

2.3 Sammon-kartta (Sammon's mapping)

Sammon-kartta on epälineaarinen kartta-algoritmi, jonka tarkoituksena on esittää p-dimensioisen avaruuden pisteet kaksiulotteisesti. Tämä suoritetaan niin, että alkuperäisten mittausdatavektoreiden rakenne säilyy mahdollisimman hyvin.

Sammon-kartta kuvaa datavektoreiden suhteellisia etäisyyksiä toisistaan, joten menetelmä on hyödyllinen klustereiden ja niiden välisten etäisyyksien määrittelemiseen. Sammon-kartta toimii samantapaisesti kuin SOM, mutta on laskennallisesti vaativampi. Suuren datan käsittelyssä onkin yleensä parasta yhdistää nämä kaksi menetelmää esimerkiksi siten, että Sammon-kartan alustamiseen käytetään SOM-algoritmilla saatuja painovektoreita. [2]

2.4 Työn toteutus

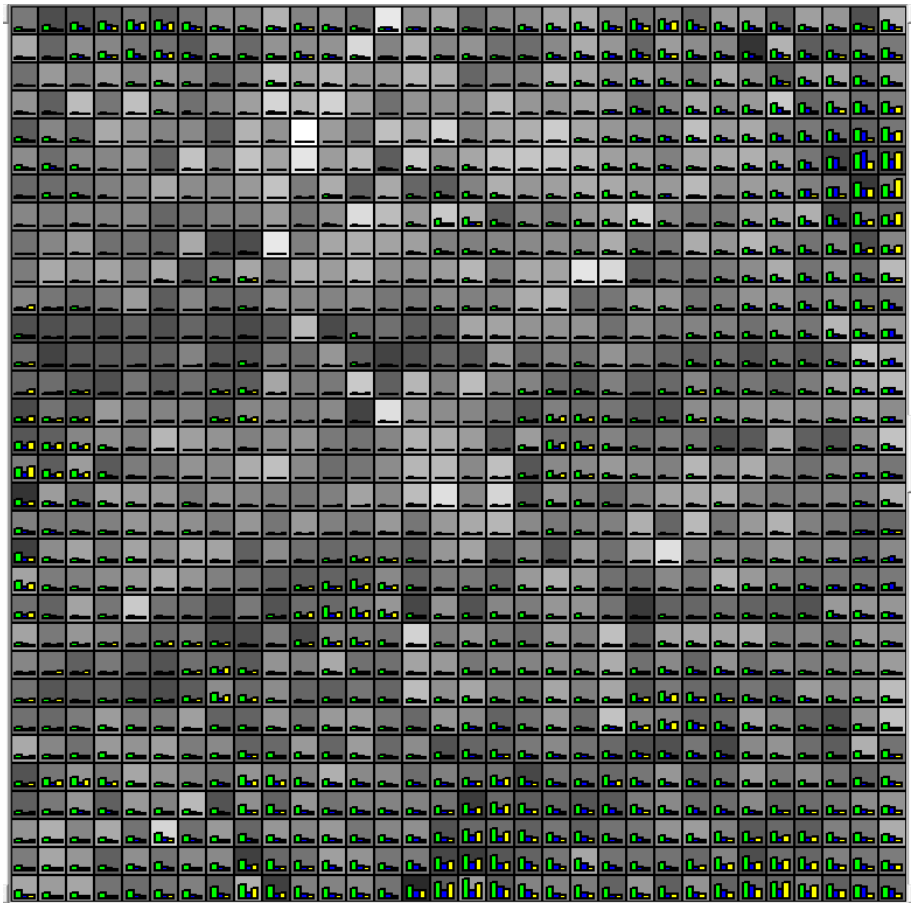
Työssä käsiteltävänä oli ilmanlaatumittausdata (toolodata). Data sisälsi mittaustulokset vuoden ajalta seitsemästä eri muuttujasta, joista neljä kuvaa erilaisten ilmansaasteiden konsentraatioita (NO_2, O_3, CO, PM10) ja kolme säätä (lämpötila, kosteus, tuulennopeus). Datan analysointi suoritettiin Visual Data -ohjelmistolla.

Aluksi datalle suoritettiin esikäsittelyä (preprocessing) vertailemalla erilaisia esikäsittelymetodeja: tasoittaminen (equalizing), normalisointi (normalization) sekä varianssiin perustuva tasoittaminen (variance-based equalizing). Näistä parhaan tuloksen antava menetelmä valittiin jatkokäsittelyä varten. Seuraavaksi muodostettiin datalle SOM-kartta ja tarkasteltiin erilaisten kuvaajien antamaa informaatiota datasta (pylväsdiagrammit, viivadiagrammit, U-matriisi).

SOM-kartan perusteella muodostettiin myös Sammon-kartta, jonka tuottamaa tietoa sopivista klustereista verrattiin SOM-kartan klustereihin. Klusterit muodostettiin karttojen avulla. Klustereille piirrettiin histogrammeja, joiden avulla klustereiden tiedot säästä, ajankohdasta ja ilmansaasteista koottiin taulukkoon ja näiden perusteella klustereille pyrittiin löytämään selitykset ilmanlaadun suhteen.

3 Tulokset

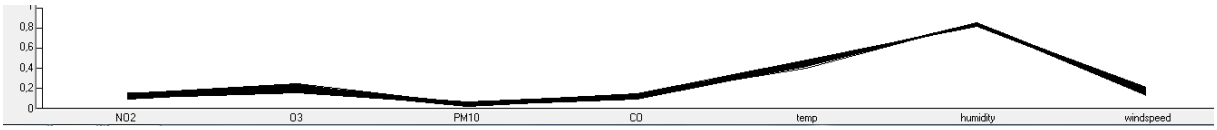
Eri esikäsittelymenetelmistä yhdistelmä equalizing-normalisation vaikutti antavan parhaan tuloksen, joten se valittiin käytettäväksi. Kuvassa 1 on esitetty esikäsitelty data, jossa pylväsdiagrammit kuvaavat saastemuuttujia (NO_2, CO, PM10).



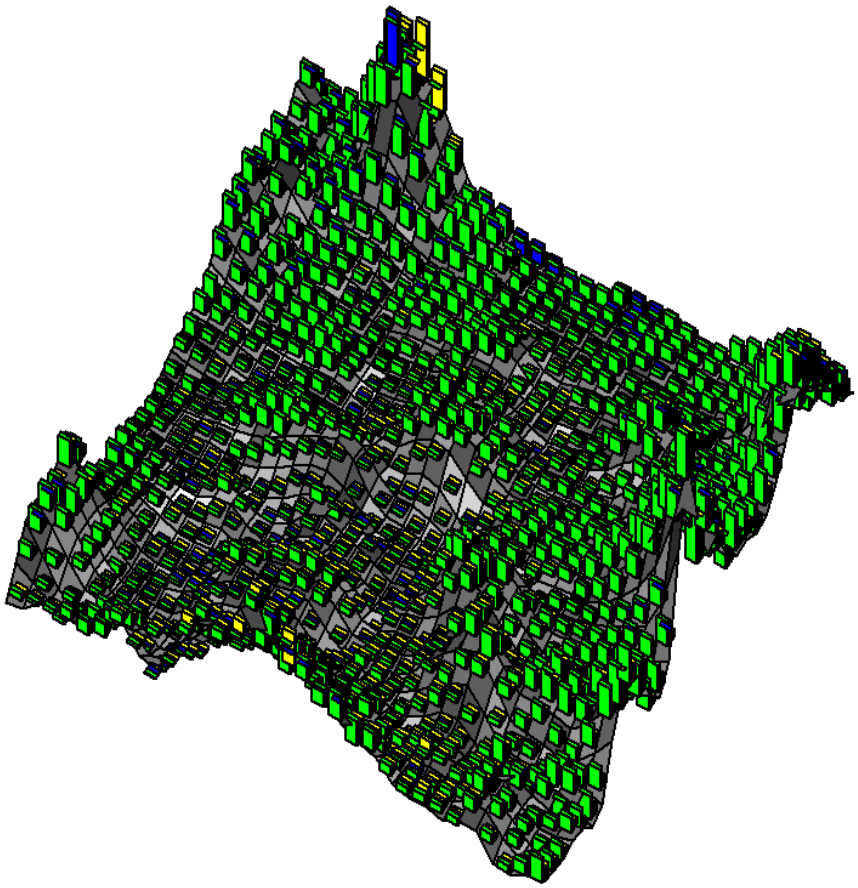
Kuva 1: Esikäsitelty data. Pylväsdiagrammit kuvaavat ilmansaasteiden konsentraatioita: vihreä = NO_2, sininen = CO, keltainen = PM10.

Pylväsdiagrammien avulla voidaan havaita datan jaottumista tietynlaisiin joukkoihin, klustereihin. Solujen viivadiagrammeja vertaamalla voitiin havaita, että hajontaa eri soluissa eri muuttujien suhteen on eri määrä. Esimerkki yhtä solua kuvaavasta viivadiagrammista on esitetty kuvassa 2. Kuvan 2 viivadiagrammissa vaihtelu on varsin pientä.

Datan U-matriisi on esitetty kuvassa 3. U-matriisissa voidaan havaita korkeampia alueita sekä matalampia laaksoja, joita voidaan käyttää hyväksi klustereiden muodostamisessa.

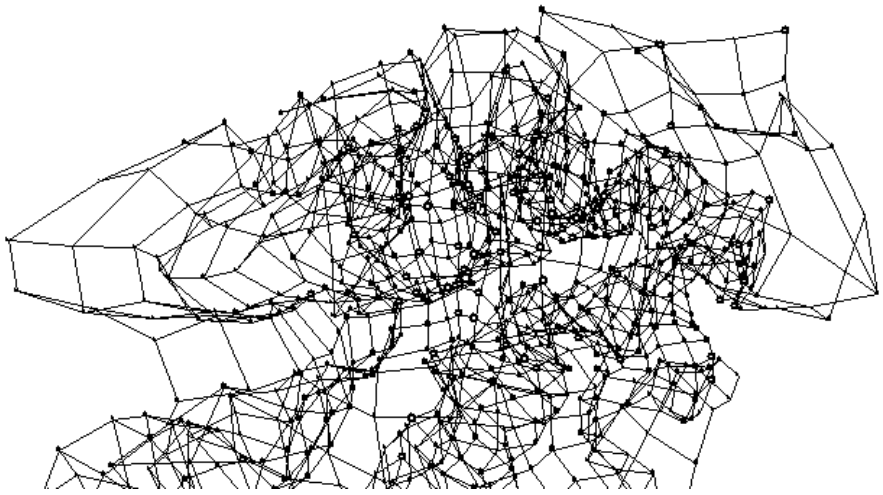


Kuva 2: Viivadiagrammi yhdestä solusta.

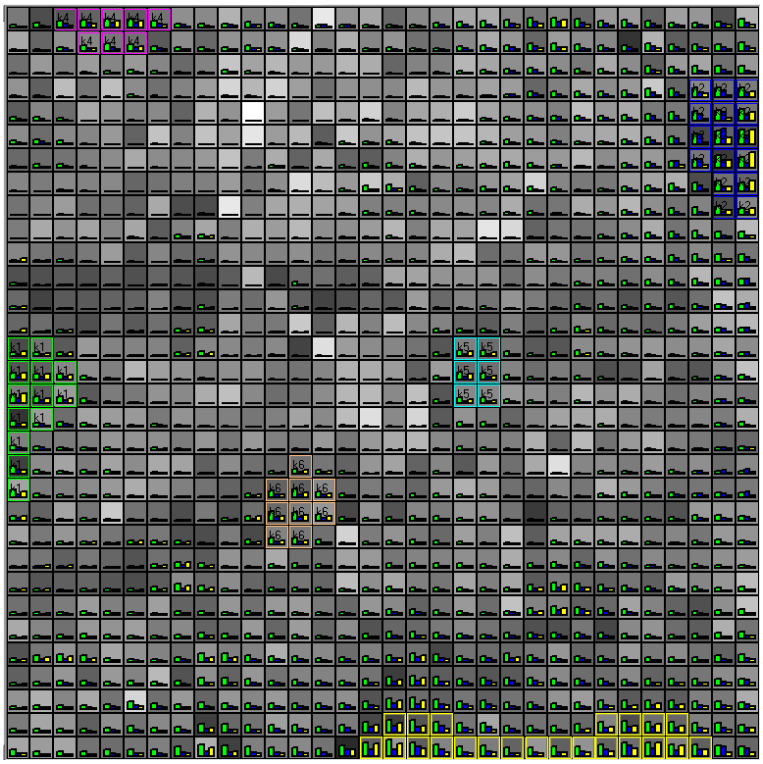


Kuva 3: U-matriisi.

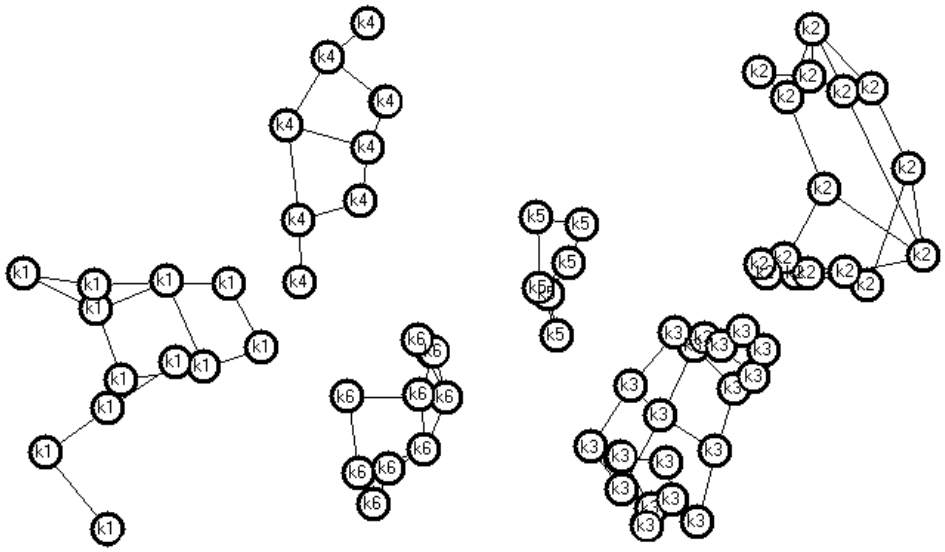
SOM-kartasta muodostettu Sammon-kartta on esitetty kuvassa 4. Kartassa on paljon pisteitä, joten klusterit eivät kovin hyvin erotu toisistaan. Klusterit valittiin SOM-kartasta kuvan 5 mukaisesti. Kun Sammon-kartta piirrettiin uudestaan käyttämällä vain klustereihin kuuluvia pisteitä, kartasta tuli selkeämpi ja klusterit voitiin siitäkin erottaa (Kuva 6).



Kuva 4: Sammon-kartta. Ympyrän sädee kuvaa datapisteiden määrää.

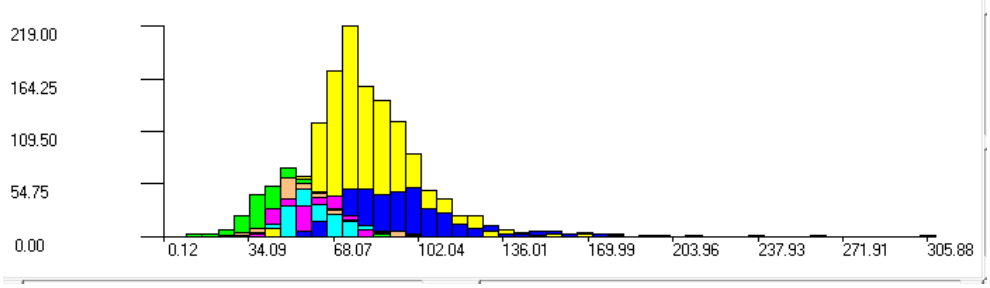


Kuva 5: Valitut klusterit. Vihreä = k1, tummansininen = k2, keltainen = k3, vaaleanpunainen = k4, vaaleansininen = k5, oranssi = k6 (vrt. Sammon-kartta, Kuva 6).

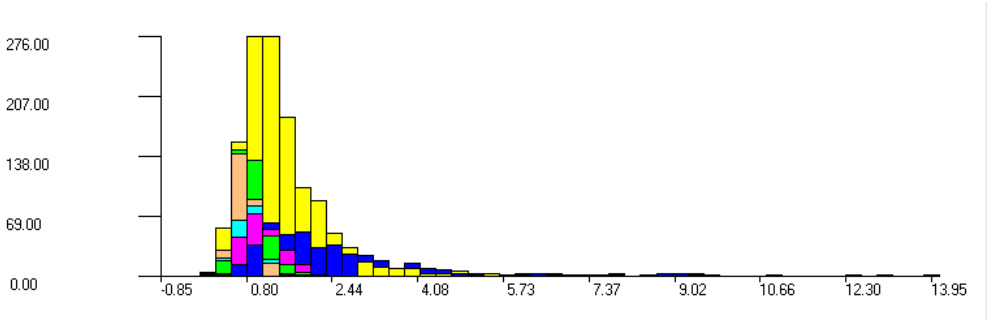


Kuva 6: Sammon-kartta klustereista (kartasta on selkeyden vuoksi poistettu klustereihin kuulumattomat pisteet, eikä ympyröiden sädettä ole skaalattu kuten kuvassa 4).

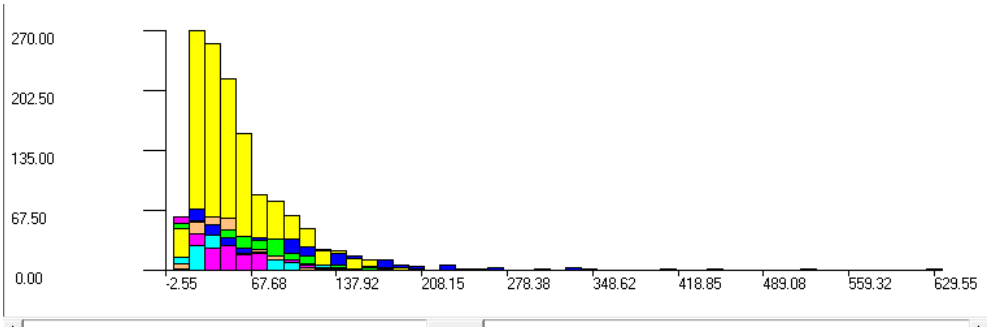
Klustereista muodostetut histogrammit eri ilmansaasteiden osalta on esitetty kuvissa 7 - 9. Histogrammien avulla kootut klustereiden tiedot säästä, ajankohdasta ja ilmansaasteista on esitetty taulukossa 1. Taulukossa 1 on myös tietojen perusteella löydetyt selitykset ilmanlaadulle eri klustereissa.



Kuva 7: Klusterien histogrammi NO_2-konsentraation suhteen.



Kuva 8: Klusterien histogrammi CO-konsentraation suhteen.



Kuva 9: Klusterien histogrammi PM10-konsentraation suhteen.

Taulukko 1: Klusterien ajankohta-, sää- ja ilmanlaatutiedot sekä klustereiden ilmanlaadulle löydetyt selitykset.

	Vuoden-aika	Kello	Tuuli (m/s)	Lämpö-tila	Ilman-saasteet	Selitys
k1	kevät	11-13	5-7	0-5°C	korkea PM10	Katupöly
k2	talvi	4-7	1-2	-15°C	korkeat NO_2, CO ja PM10	Kylmällä talvi-ilmalla esiintyvä inversio-ilmiö estää ilmansaasteiden sekoittumista
k3	kevät, syksy	15-20	1-2	-3-+15°C	korkeat NO_2, CO ja PM10	Liikenne, heikko tuuli ei sekoita ilmaa
k4	talvi	11-13	5-7	-3°C	matalat	Tuuli sekoittaa ilman
k5	kevät	5-12	1-3	-5°C	nousseet CO, PM10	Aamuliikenne, katupöly, heikko tuuli ei sekoita ilmaa
k6	kevät	9-11	1-3	+3°C	noussut PM10	Katupöly, heikko tuuli ei sekoita ilmaa

4 Johtopäätökset

SOM-kartan avulla ilmanlaatumittausdata saatiin järjestettyä selkeäksi kartaksi. Jotta tähän lopputulokseen päästiin, täytyi data ensin esikäsitellä sopivasti ja valita pylväsdiagrammikuvaajiin sopivat muuttujat. Tämän jälkeen SOM-kartasta oli suhteellisen helppo määrittää sopivat klusterit.

SOM-kartan pohjalta piirretty Sammon-kartta oli suuren pistemäärän takia epäselvä, joten klustereiden valintaan SOM-kartta oli parempi kuin Sammon-kartta. Sammon-kartta olisi ehkä täytynyt määrittää pienemmällä joukolla pisteitä, jolloin siitä olisi voinut tulla havainnollisempi. Klusterit nimittäin näkyivät Sammon-kartassakin varsin selkeästi, kun siitä ensin poistettiin ylimääräiset klustereihin kuulumattomat pisteet. Tällöin Sammon-kartta antoi varsin yhdenmukaista tietoa SOM-karttaan verrattuna klustereiden välisistä etäisyyksistä (Kuvat 5 ja 6), eli klusterit ovat sijoittuneet suhteessa toisiinsa nähden kummassakin kartassa suurinpiirtein samalla tavalla. Joka tapauksessa sopivin menetelmä klusterointiin riippuu analysoitavasta datasta ja sen rakenteesta, joten yleensä kumpaakin menetelmää on syytä tarkastella parhaan tuloksen saavuttamiseksi.

Klustereiden selittämiseksi täytyy tarkemmin tutkia niiden tietoja esimerkiksi histogrammien avulla. Histogrammeista nähtiin, että klusterit olivat melko selkeästi jakautuneet eri muuttujien suhteen. Havaittiin myös, että klustereiden analysoiminen oli helpompaa, jos klustereihin ei ollut valittu liikaa pisteitä, jolloin klustereiden väliset erot näkyivät histogrammeissa selkeämmin.

SOM- ja Sammon-kartan avulla ilmanlaatudata saatiin jaettua kuuteen klusteriin, joille etsittiin ajankohdasta, säästä ja ilmansaasteista kertovat tiedot. Näiden perusteella pystyttiin löytämään yksinkertaiset selitykset klustereiden ilmanlaadulle, kun havaittiin klustereiden kuvaavaan tiettyä vuoden- tai vuorokaudenaikaa tietyillä sääparametreilla (Taulukko 1). SOM- ja Sammon-kartat vaikuttivat siis olevan varsin tehokkaita menetelmiä ainakin tämäntyyppisen datan käsittelyyn, koska selittäviin tuloksiin päästiin jo melko pienellä vaivalla. Toki vielä useamman muuttujan dataa analysoitaessa aikaa kuluu paljon sopivien muuttujien valitsemiseen, mutta kaiken kaikkiaan datasta voidaan SOM- ja Sammon-karttojen avulla näppärästi löytää ominaisuuksia, joita ei välttämättä muuten helposti huomaisi.

Viitteet

[1] Kohonen Teuvo *The Self-Organizing Map* Helsinki University of Technology: Laboratory of Computer and Information Science 2005
http://www.cis.hut.fi/projects/somtoolbox/theory/somalgorithm.shtml, 10.10.2011

[2] Kolehmainen Mikko *Lecture Notes on Basics of Multivariate Methods* Itä-Suomen Yliopisto, Luonnontieteiden ja metsätieteiden tiedekunnan opintojakson Basics of Multivariate Methods luentomoniste 2011

[3] Hollmen Jaakko *U-matrix* TKK 1996
http://users.ics.tkk.fi/jhollmen/dippa/node24.html, 10.10.2011