hyväksymispäivä arvosana

arvostelija

Datan visualisointi pikseliperusteisilla menetelmillä

Ella Peltonen

Helsinki 12.12.2011

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet –Faculty	Laitos – Institution – Department
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkäsittelytieteen laitos

Tekijä – Författare – Author

Ella Peltonen

Työn nimi – Arbetets titel – Title

Datan visualisointi pikseliperusteisilla menetelmillä

Oppiaine – Läroämne – Subject

Tietojenkäsittelytiede

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
	12.12.2011	20 sivua

Tiivistelmä – Referat – Abstract

Tiedon louhinnassa pyritään selvittämään suuren tietomäärän muodostaman joukon ja sen osajoukkojen sisäistä rakennetta. Visuaalisessa tiedon louhinnassa nämä rakenteet pyritään esittämään yhtenä tai useampana kuvana. Tarkoituksena on auttaa ihmisiä hahmottamaan datajoukon sisältöä. Tutkielmassani keskityn pikseliperusteisiin visualisointimenetelmiin (eng. pixel-based tai pixel-oriented visualization techniques). Esittelen sellaisia menetelmiä kuin spiraali-, akseli-, käyrä-, rekursio- ja sektorimenetelmät.

Pikseliperusteisten menetelmien perusajatuksena on kuvata jokainen datajoukon piste yhdelle kuvan pikselille. Jokaiselle datajoukon muuttujalle piirretään oma kuvansa. Datapistettä edustavalla pikselillä on sama sijainti jokaisessa piirrettävässä kuvassa, ja pikselin väri määräytyy muuttujan arvon perusteella. Muuttujakohtaisista kuvista kootaan yksi kuvakokonaisuus, jossa datan eri muuttujia vertaillaan rinnakkain.

Pikseliperusteiset menetelmät sopivat suurten datajoukkojen visualisointiin. Näiden menetelmien avulla voidaan järkevästi visualisoida myös kohtuullisen monia muuttujia sisältäviä datajoukkoja. Pikseliperusteisia menetelmiä käytettäessä on pohdittava kolmea peruskysymystä: Miten määritellään datapistettä edustavan pikselin väri? Miten määritellään pikselin sijainti muuttujakohtaisessa kuvassa? Miten muuttujakohtaiset kuvat järjestetään yhdeksi kokonaisuudeksi?

ACM Computing Classification System (CCS):
A.1 [Introductory and Survey],
H.2.8 [Database Applications: Data mining],
I.6.8 [Types of Simulation: Visual]

Avainsanat – Nyckelord – Keywords

visualisointi, pikseliperusteiset visualisointimenetelmät, kyselyriippumaton, kyselyriippuvainen

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Sisältö

1 Johdanto **1**

2 Pikseliperusteisten visualisointimenetelmien luokittelu **2**

2.1 Spiraali- ja akselimenetelmät....3

2.2 Käyrämenetelmä....6

2.3 Rekursiomenetelmä....8

2.4 Sektorimenetelmä....10

3 Menetelmien soveltaminen **13**

3.1 Datapisteen sijainnin määrittäminen....13

3.2 Datapisteen värin määrittäminen....14

3.3 Kokonaiskuvan esittäminen....16

3.4 Kritiikkiä....16

4 Yhteenveto **17**

Lähteet **19**

1 Johdanto

Keim ja Kriegel [KeK96] luokittelevat erilaisia datan visualisoinnin menetelmiä. Heidän työhönsä nojaa myös Ferreira de Oliveiran ja Levkowitzin uudempi luokittelu [FeL03], jota käytän tässä työssäni. Tutkielmassani keskityn *pikseliperusteisiin visualisointimenetelmiin* (eng. pixel-based tai pixel-oriented visualization techniques). Esittelen seuraavat pikseliperusteiset menetelmät: *käyrämenetelmä* (eng. space filling curves technique), *rekursiomenetelmä* (eng. recursive pattern technique), *spiraali- ja akselimenetelmät* (eng. spiral and axes techniques) sekä *sektorimenetelmä* (eng. circle segment technique).

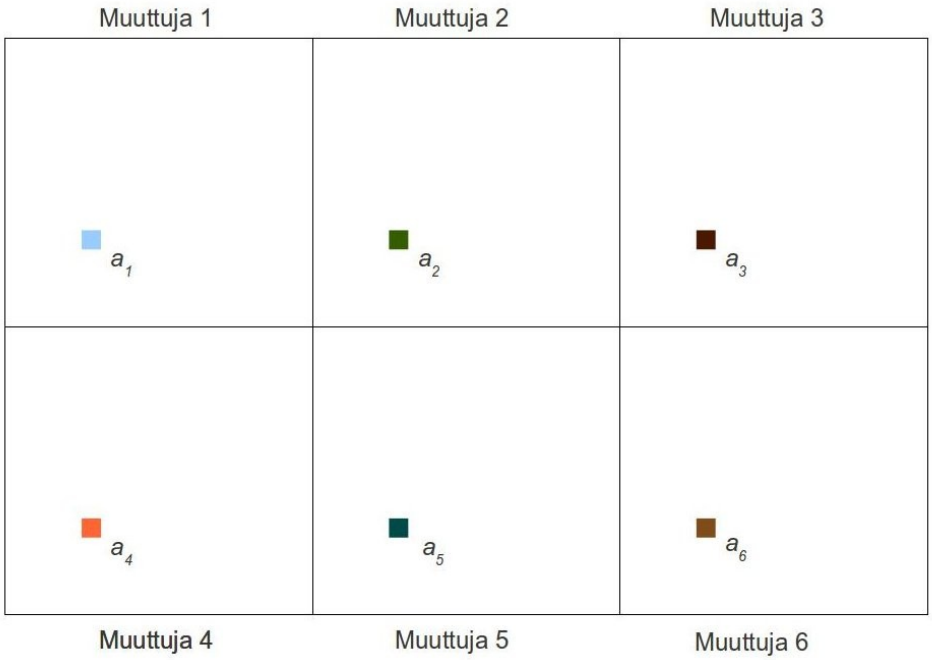
Pikseliperusteisten visualisointimenetelmien perusajatus on seuraavanlainen: Jokainen *datapiste* (eng. data item) kuvataan yhdelle ainoalle pikselille. Jokaiselle datajoukon *muuttujalle* (eng. attribute tai dimension) piirretään oma kuvansa. Jatkossa käytän näistä yhtä muuttujaa esittävistä kuvista nimitystä *muuttujakohtainen kuva* (eng. subwindows). Yksittäisen datapisteen sijainti kussakin muuttujakohtaisessa kuvassa on aina sama, kuten esitän kuvassa 1.

Datapistettä edustavan pikselin väri määräytyy muuttujan arvon perusteella. Jokainen datapiste saa oman pikelinsä, joita ei aseteta limittäin. Sen sijaan toisiaan eniten muistuttavat datapisteet sijoittuvat visualisoinnissa lähelle toisiaan. Datapisteet sijoitetaan yksiulotteiselle käyrälle, jolla peitetään kaksiulotteinen pinta. Käyrän malli valitaan siten, että sen avulla saadaan klusteroitua samanlaiset pikselit mahdollisimman lähekkäin.

Muuttujakohtaiset kuvat kootaan yhdeksi kokonaisuudeksi, jossa eri muuttujia on helppo vertailla rinnakkain. Lopputuloksena on koko datajoukkoa tai sen visualisoitavaksi valittua osajoukkoa esittävä, väritetyistä pikseleistä muodostettu kaksiulotteinen kuva. Visualisoinnin onnistuminen riippuu monesta tekijästä, ennen kaikkea siitä, että käsiteltävälle datalle on valittu sopiva menetelmä. Huono visualisointi voi olla jopa harhaanjohtava.

Keim [Kei00] esittelee pikseliperusteisien visualisointimenetelmien peruskysymyksiä, joiden perusteella voidaan myös valita käytettävä menetelmä: Miten määritellään datapistettä edustavan pikselin väri? Miten määritellään pikselin sijainti muuttujakohtaisessa kuvassa? Ja miten muuttujakohtaiset kuvat järjestetään

yhtenäiseksi, helposti tulkittavaksi kokonaisuudeksi? Eri menetelmät tarjoavat erilaisia vastauksia varsinkin kysymykseen, miten datapisteet järjestetään muuttujakohtaiseen kuvan sisällä. Eroja menetelmien välillä on myös tavassa järjestää muuttujakohtaiset kuvat kokonaisuudeksi.



Kuva 1: Kuvassa on esitetty pikseliperusteisella menetelmällä tuotetun visualisoinnin perusrakenne. Datapisteen $A = (a_1, a_2, a_3, a_4, a_5, a_6)$ jokaiselle kuudelle muuttujalle on laskettu väri perustuen kyseisen muuttujan arvoon. Datapistettä edustavan pikselin sijainti kussakin muuttujakohtaisessa kuvassa on sama.

2 Pikseliperusteisten visualisointimenetelmien luokittelu

Pikseliperusteiset menetelmät jaetaan *kyselyriippuvaisiin* ja *kyselyriippumattomiin menetelmiin* (eng. query dependent techniques ja query independent techniques) sen mukaan, millä tavalla käsiteltävä datajoukko käydään läpi kuvaa piirrettäessä [KeK96, FeL03]. Se, valitaanko käytettäväksi kyselyriippuvainen ja kyselyriippumaton menetelmä, riippuu pitkälti siitä millaista datajoukkoa käsitellään.

Kyselyriippuvaisissa menetelmissä datajoukolle tehdään kysely, jossa määritellään osa datapisteistä arvokkaammiksi kuin toiset [KeK96, FeL03]. Esimerkiksi datapisteiden paremmuus ratkaistaan niiden sisältämien arvojen yleisyyden perusteella, tai kunkin

datajoukon pisteen arvoa verrataan johonkin käyttäjän määrittämään odotusarvoiseen tai ideaaliin arvoon. Parhaiten kyselyä vastaavat datapisteet ovat myös visualisoinnin kannalta kiinnostavimmat. Ne saavat piirrettävässä kuvassa muita pisteitä paremman sijainnin, ja yleensä ne sijoitetaan kuvaan mahdollisimman keskelle. Muut pisteet sijoitetaan kuvaan valitun muotoista käyrää seuraamalla.

Kyselyriippumattomissa menetelmissä datajoukolla on usein jokin luonnollinen järjestys, kuten ajanhetki, jota käytetään visualisoinnin apuna [Kei00, Kei02]. Tällöin datajoukko käydään läpi ja piirretään kuvaksi noudattaen tätä valittua järjestystä. Luonnollisen järjestyksen tilalla voidaan käyttää myös esimerkiksi yhden muuttujan suhdetta datajoukon muihin muuttujiin [KKA95].

Monissa muissa visualisointimenetelmissä toisiaan muistuttavat datapisteet kertyvät keskitetysti samaan visualisointikuvan kohtaan ja muodostavat tälle alueelle erottuvan klusterin. Pikseliperusteisissa menetelmissä jokaiselle datapisteelle piirretään kaikissa tapauksissa oma pikselinsä, joka ei leikkaa tai limity minkään muun datapisteen pikselin kanssa. Toisiaan muistuttavat datapisteet kertyvät lähelle toisiaan datapisteitä yhteen kokoavan käyrän avulla. Keimin mukaan [Kei00, Kei02] klusterit erottuvat erityisen hyvin juuri pikseliperusteisilla menetelmillä. Kun jokainen datapiste kasvattaa klusterin kokoa yhdellä pikselillä, klustereiden suhteellista kokoa on helppo havainnoida.

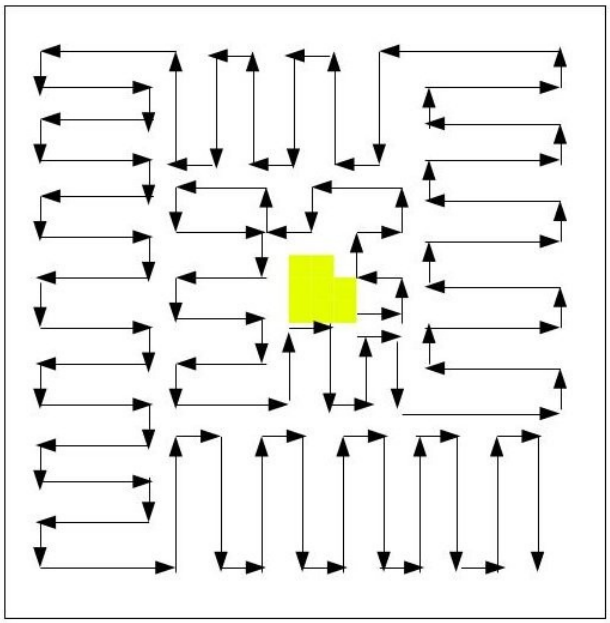
Seuraavaksi esittelen viisi pikseliperusteista menetelmää. Käyttämäni jako perustuu Ferreira de Oliveiran ja Levkowitzin [FeL03] tekemään luokitteluun. Menetelmistä ensin esiteltävät spiraali- ja akselimenetelmät ovat kyselyriippuvaisia menetelmiä, käyrä- ja rekursiomenetelmä puolestaan kyselyriippumattomia menetelmiä. Viimeiseksi esiteltävä sektorimenetelmä tarjoaa ennen kaikkea välineen muuttujakohtaisten kuvien kokoamiselle yhdeksi kuvaksi.

2.1 Spiraali- ja akselimenetelmät

Spiraali- ja akselimenetelmät (eng. spiral and axes techniques) ovat kyselyriippuvaisia menetelmiä. Muun muassa Keimin ja Kriegelin [KeK96] esittelemissä menetelmissä datapisteet järjestetään kuvaan keskiosasta ulospäin kiertyvän käyrän eli spiraalin avulla. Datajoukolle tehdään kysely, jossa parhaat painoarvot saaneet datapisteet sijoitetaan spiraalin alkuun eli kuvan keskelle. Loput pisteet kierretään keskiosan ympärille spiraalin mallin mukaan. Datapisteiden järjestys on siis yksiulotteinen, mutta

spiraalin muodon avulla ne täyttävät kaksiulotteisen pinnan. Toisiaan paljon muistuttavat datapisteet päätyvät lopputuloksessa lähelle toisiaan ja muodostavat klustereita, kun spiraalin muoto ja leveys on valittu oikein. Keim ja Kriegel mainitsevat [KeK96] spiraalimallit Snake, Peano-Hilbert ja Morton, mutta erilaisia malleja on lukuisia.

Keimin ja Kriegelin mukaan [KeK96] sekä spiraaleja että akseleita käytettäessä muuttujakohtaiset kuvat sijoitetaan suurempaan kokonaiskuvaan vierekkäin, esimerkiksi rinnakkain ja allekkain taulukoksi, kuten esitän kuvissa 1 ja 4. Yhden datapisteen paikka kussakin muuttujakohtaisessa kuvassa on sama: pisteen väri kertoo sen osuvuudesta kyseiseen muuttujaan. Keim toteaa [Kei00], että tämän takia värityksen valinta on oleellinen osa visualisoinnin tuottamista.

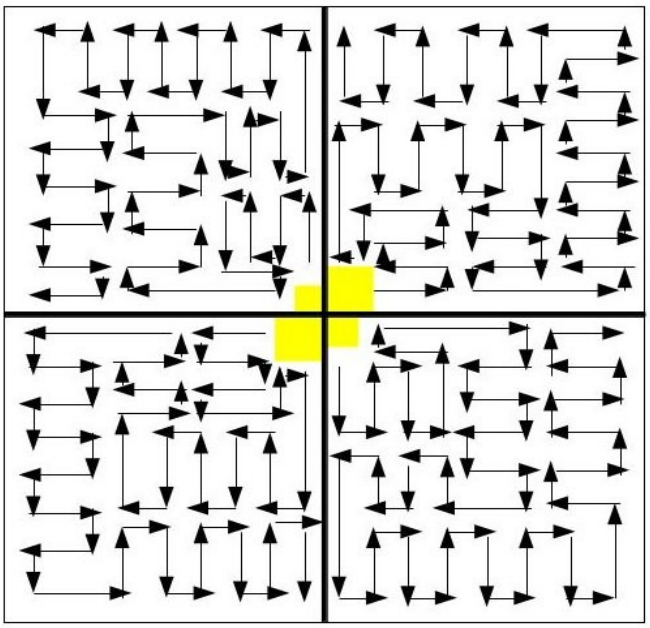


Kuva 2: Snake-spiraali yhdessä muuttujakohtaisessa kuvassa. Keskellä ovat kyselyyn parhaiten vastanneet datapisteet (keltainen väri). Pisteet on sijoitettu kuvaan spiraalin mukaisessa järjestyksessä. Spiraalin leveyttä saa säädettyä nuolten pituutta vaihtamalla.

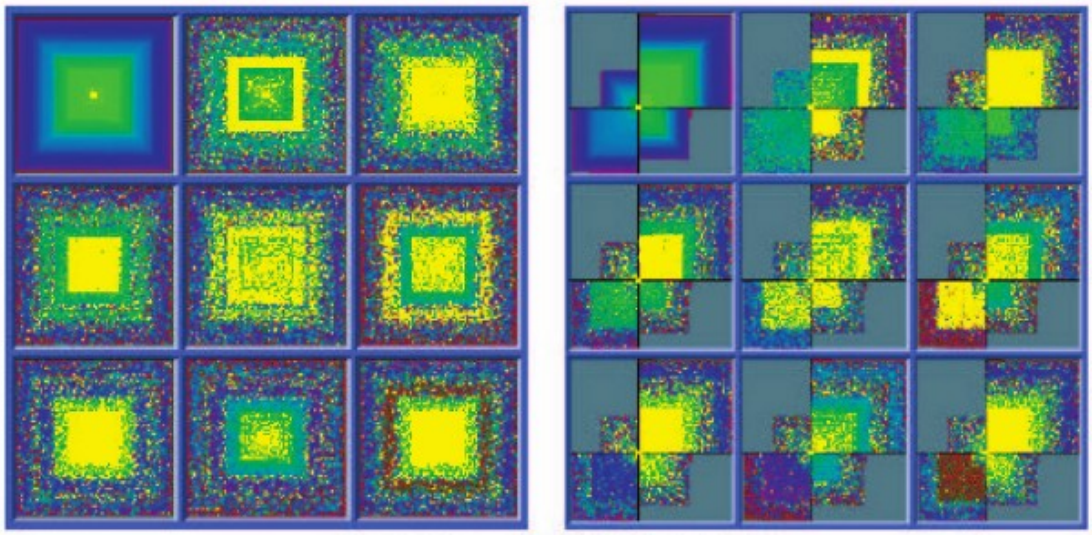
Spiraalien käyttöön liittyy muun muassa seuraavanlaisia ongelmia [Kei00]: Varsinkin pienten klustereiden erottaminen on menetelmällä hankalaa. Piirrettävän kuva-alueen koko täytyy suhteuttaa spiraalin muotoon ja kokoon. Datapisteiden todellinen järjestys saattaa hämärtyä visualisoinnissa. Ratkaisuksi Keim esittää [Kei00] useamman

spiraalimallin käyttöä rinnakkain. Tällöin löytyy todennäköisemmin juuri kyseiselle datalle sopiva spiraali.

Akselimenetelmä pohjautuu spiraalimenetelmälle [KeK96]. Akselimenetelmässä muuttujakohtainen kuva jaetaan kahdella akselilla neljään kenttään. Jokaisen nelikentän osa piirretään yhden spiraalin avulla niin, että jokainen datapiste on mukana vain yhdessä nelikentässä. Akselien eri päät kuvaavat esimerkiksi minimi- ja maksimiarvoja: eniten eroavat muuttujan arvot ovat mahdollisimman kaukana toisistaan. Akselimenetelmän avulla datapisteiden välille saadaan enemmän erottelua kuin spiraalimenetelmässä. Datapisteelle määritellään käyrällä olevan sijainnin lisäksi myös se, mihin nelikentistä se kuuluu.



Kuva 3: Akselimenetelmä Snake-spiraalilla. Kuten kuvassa 2, parhaiten kyselyyn vastanneet datapisteet on saatu kuvan keskelle (keltaiset). Datapisteet sijoitetaan neljän spiraalin avulla nelikenttiin niin, että toisistaan eniten poikkeavat arvot ovat kauimpana toisistaan.



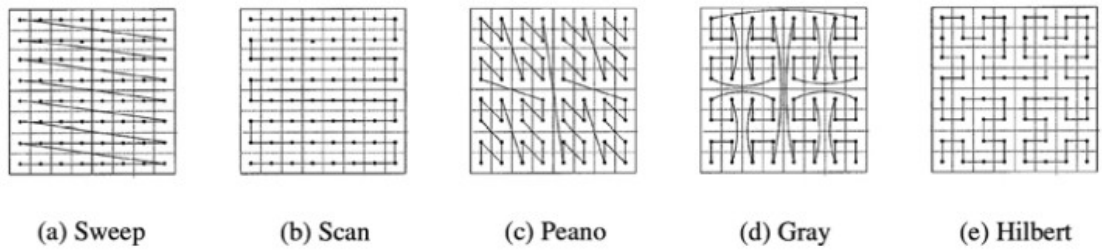
Kuva 4: Ferreira de Oliveiran ja Levkowitzin [FeL03] esimerkki spiraali- ja akselimenetelmien käytöstä. Spiraalimenetelmä vasemmalla, akselimenetelmä oikealla. Vasemmassa yläkulmassa on väriskaalaa havainnoistava verrokkikuva. Muuttujakohtaisia kuvia on kahdeksan.

Ferreira de Oliveira ja Levkowitz [FeL03] näyttävät esimerkin spiraali- ja akselimenetelmistä kuvassa 4. Datajoukko on Keimin tutkimusryhmän synteettistä dataa, eli datajoukko on tuotettu keinotekoisesti visualisoinnin havainnollistamiseksi. Datajoukossa on seitsemän tuhatta datapistettä ja kahdeksan muuttujaa. Datajoukolle on tehty kysely, johon parhaiten vastanneet pisteet ovat saaneet keltaisen arvon. Värikaala on asetettu niin, että seuraavaksi parhaiten kyselyyn vastaavat pisteet ovat vihreitä, sen jälkeen sinisiä, ja huonoiten kyselyyn vastanneet datapisteet ovat punaisia ja mustia. Väriskaalaa on havainnollistettu muuttujakohtaisten kuvien rinnalla vasemmassa yläkulmassa. Muut ruudut ovat muuttujakohtaisia kuvia.

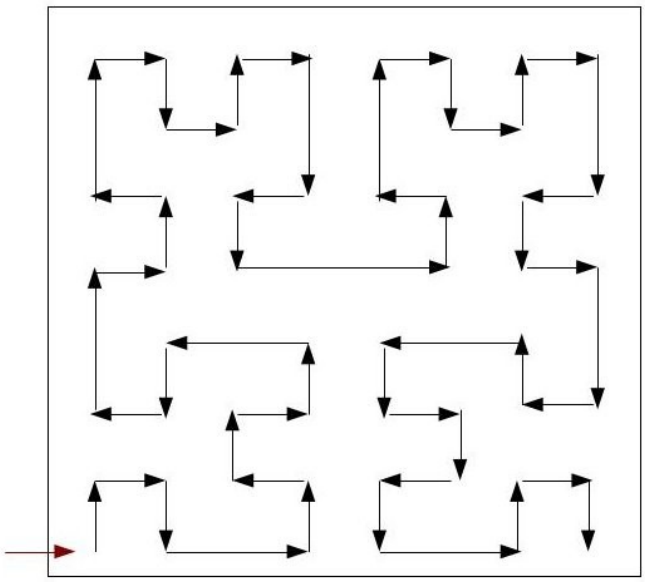
2.2 Käyrämenetelmä

Käyrämenetelmä (eng. space filling curves technique) on kyselyriippumaton visualisointimenetelmä. Sitä käytettäessä datajoukolla on oltava jokin luonnollinen järjestys, kuten ajanhetki, tai vaihtoehtoisesti muu datajoukon ominaisuuksista tuleva järjestämisperiaate. Datapisteet sijoitetaan yksiulotteiselle käyrälle valitun järjestämisperiaatteen mukaisessa järjestyksessä. Käyrällä täytetään tämän jälkeen kaksiulotteisen kuvan pinta. Käyrämenetelmässä alkupiste asetetaan useimmiten johonkin kuvan kulmista, ei kuvan keskelle kuten spiraali- ja akselimenetelmissä.

Erilaisia käyrämalleja on useita. Mokbel, Aref ja Kamel [MAK03] esittelevät kuvassa 5 käyrät Sweep, Scan, Peano, Gray ja Hilbert. Kuvasta näkyy käyrien monipuolisuus, mutta myös se, että suoraviivainen rivi riviltä läpikäyntikin kelpaa käyräksi. Klusteroinnin kannalta on parempi valita polveilevampi ja datapisteitä enemmän kasaava käyrä. Sen sijaan jos kuvassa halutaan säilyttää esimerkiksi vahva yhteys ajanhetkeen, liian monimutkainen käyrä vaikeuttaa kuvan seuraamista.

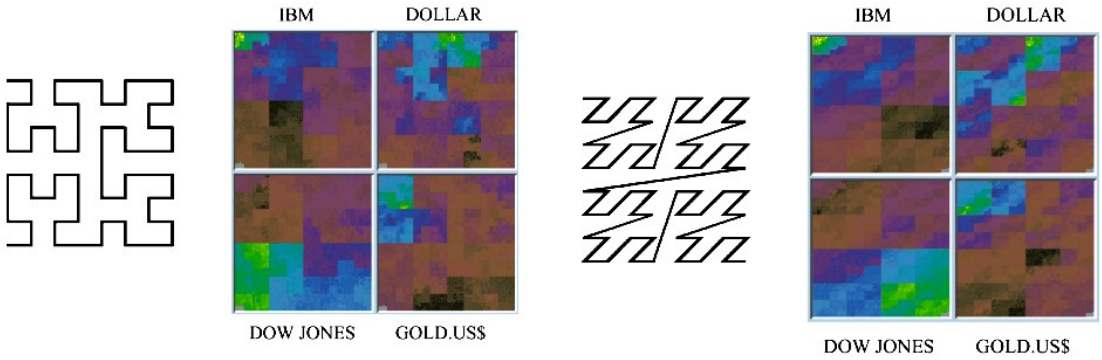


Kuva 5: Mokbel, Aref ja Kamel [MAK03] esittävät erilaisia läpikäyntikäyriä. Keim nimittää käyrää (c) nimellä Morton ja käyrää (e) nimellä Peano-Hilbert.



Kuva 6: Läpikäynti Peano-Hilbert-käyrällä yhden muuttujakohtaisen kuvan sisällä. Punainen nuoli osoittaa lähtöpisteen, mustat nuolet kuvan piirtojärjestyksen. Kuten spiraali- ja akselimenetelmissä, käyrän mallin lisäksi on valittava käyrän leveys, eli tässä kuvassa kunkin nuolen pituus.

Keim [Kei00] antaa esimerkin käyrämenetelmässä talousdatalla kuvassa 7. Datajoukossa on neljä muuttujaa, joista kukin kuvaa yhden kurssin (IBM:n osake, USA:n dollari, Dow Jones -indeksi ja kullan arvo) kehittymistä syyskuusta 1986 helmikuuhun 1995. Kutakin kurssia varten on piirretty oma kuvansa. Käytetty väriskaala on määritelty niin, että sama väri tarkoittaa samaa arvoa kaikissa muuttujakohtaisissa kuvissa. Mitä keltaisempi ja vaaleampi pikseli, sen suurempi on muuttujan arvo. Keimin esimerkissä vasemmanpuoleinen on Peano-Hilbert-käyrä, jonka etenemistä havainnollistan kuvassa 6, ja oikeanpuoleinen Morton-käyrä. Keim toteaa näistä visualisoinneista, että vaikka ne esittävät datajoukon klustereita melko hyvin, käyrän seuraaminen voi olla vaikeaa varsinkin Peano-Hilbert-käyrällä.

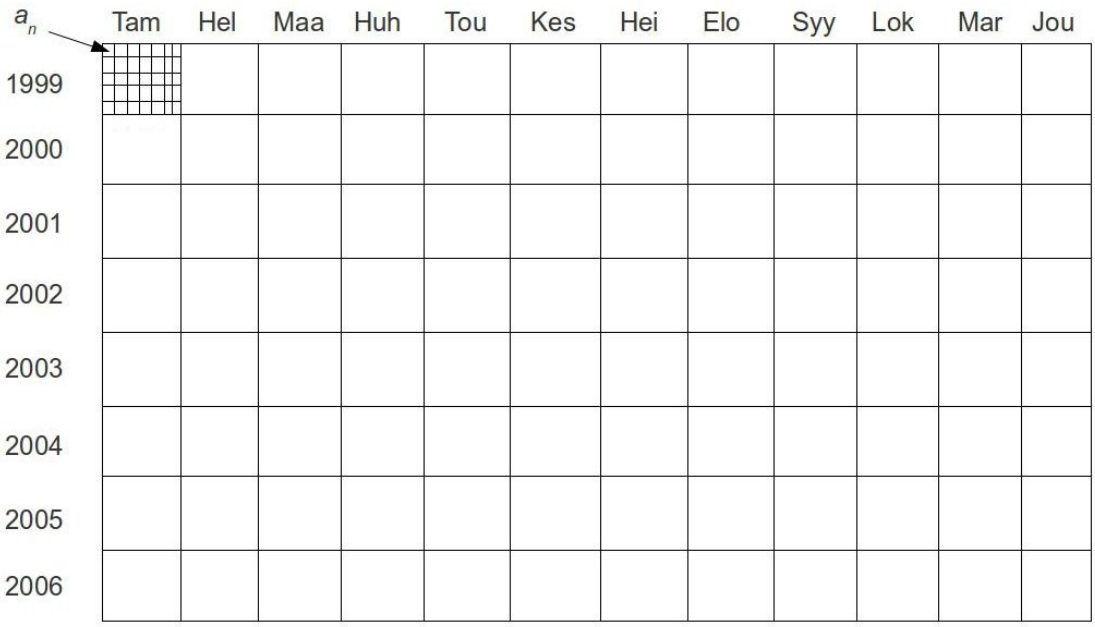


Kuva 7: Keimin [Kei00] esittämiä visualisointeja eri käyrillä. Vasemmalla on käyrä, jota Keim kutsuu nimellä Peano-Hilbert ja Mokbel, Aref ja Kamel [MAK03] nimellä Hilbert. Oikealla Keimin mukaan Morton-käyrä. Käyrän vieressä on esitetty sen avulla tuotettu visualisointi talousdatalla.

2.3 Rekursiomenetelmä

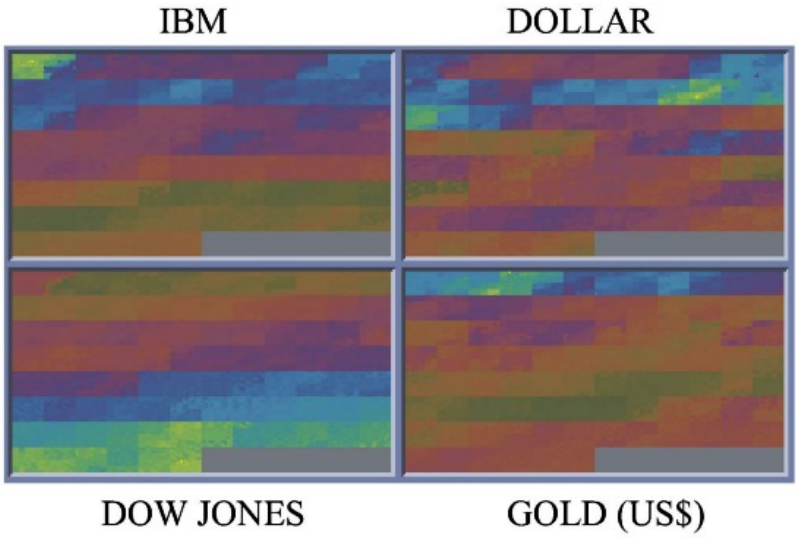
Rekursiomenetelmä (eng. recursive pattern technique) on käyrämenetelmän ohella kyselyriippumaton menetelmä. Keimin [Kei00] sekä Keimin ja Kriegelin [KKA95] mukaan rekursiomenetelmän perusajatuksena on järjestää datapisteet ryhmiin, jotka vastaavat valittua järjestystä, kuten esimerkiksi päiviä, viikkoja ja kuukausia. Datajoukko käydään läpi näiden ryhmien mukaisessa järjestyksessä. Tuloksena on kuva, joka esittää tietyn muuttujan kehittymistä ajan edetessä. Tästä rekursiomaisesta etenemisestä tulee myös menetelmän nimi.

Mahdollisia datajoukon läpikäymisjärjestyksiä on useita. Käytetty läpikäymisjärjestys vaikuttaa menetelmän tuottaman kuvan ulkoasuun. Rekursiosilmukka voi edetä esimerkiksi suoraviivaisesti rivi kerrallaan (eng. line-by-line loop) tai edestakaisin rivien välillä (eng. back-and-forth loop) [KKA95]. Keim, Kriegel ja Ankerst huomauttavat [KKA95], että varsinkin länsimaalaisten on helpoin lukea kuvia rivi kerrallaan kuten tekstiä: vasemmalta oikealle ja ylhäältä alas.



Kuva 8: Rekursiomenetelmällä tehty paikan määrittely datapisteelle $A = (a_1, \ldots, a_n)$, jonka muuttuja a_1 on päivämäärä 1.1.1999. Muuttujien $a_2 - a_n$ muuttujakohtaisissa kuvissa datapisteen sijainti on sama. Datajoukon muiden pisteiden sijainti määritellään samaan tapaan.

Keim [Kei00] antaa rekursiomenetelmästä esimerkin kuvassa 9 käyttäen samaa neljän muuttujan talousdataa kuin käyrämenetelmän yhteydessä kuvassa 7. Lukusuunta tämän esimerkin kullekin muuttujakohtaiselle kuvalle on vasemmasta ylänurkasta alkaen rivi kerrallaan alas ja vasemmalta oikealle kuten lukisi tekstiä. Värin perusteella nähdään kunkin muuttujan arvo kullakin ajan hetkellä: keltaisempi väri kertoo korkeammasta arvosta. Keim toteaa, että tälle talousdatalle rekursiomenetelmä tuottaa selkeämmän visualisoinnin kuin käyrämenetelmä, sillä datan tulkinta liittyy tiiviisti ajanhetkeen. Rekursiomenetelmässä kuvaa katsovien ihmisten on helpompi seurata ajan etenemistä eli datajoukon järjestystä kuvassa.



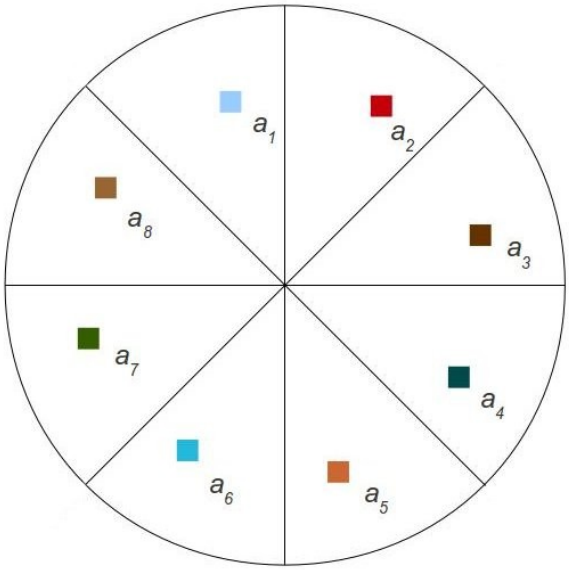
Kuva 9: Keimin [Kei00] esimerkki rekursiomenetelmästä talousdatalla. Data on samaa kuin käyrämenetelmän yhteydessä kuvassa 7.

2.4 Sektorimenetelmä

Spiraali-, akseli-, käyrä- ja rekursiomenetelmissä muuttujakohtaiset kuvat järjestetään useimmiten rinnakkain ja allekkain taulukoksi. Ferreira de Olieveiran ja Levkowitzin esittelemä [FeL03] s*ektorimenetelmä* (eng. circle segments technique) tarjoaa erilaisen muodon datan eri muuttujien kuvaamiselle. Sektorimenetelmässä lopullinen kuva on ympyrä, ja jokainen muuttujakohtainen kuva on yksi ympyrän keskenään samankokoisista sektoreista. Sektorimenetelmän perusajatus on, että se mahdollistaa eri muuttujien helpon vertailu keskenään [FeL03, Kei00, WLL08]. Havainnollistan sektorimenetelmän rakennetta kuvassa 10.

Sektoreita piirretään sama määrä kuin käsiteltäviä muuttujia. Keim [Kei00] sekä Ferreira de Oliveira ja Levkowitz [FeL03] kuvaavat sektorin värittämistä seuraavasta: Värittäminen aloitetaan ympyrän keskeltä edeten kohti ympyrän reunoja. Apuna käytetään ympyrän keskipisteeseen nähden kohtisuorassa olevia piirtoviivoja ja sektorin puolittavaa pystyviivaa. Värittäminen etenee käyränä, joka kulkee edestakaisin sektorin molempien reunojen välillä. Kuva muodostuu kerros kerrokselta. Datapisteiden keskinäinen järjestys sektorin sisällä riippuu käytettävästä datasta. Jos datajoukolla on jokin luonnollinen järjestys, voidaan datapisteet värittää suoraan sen määräämässä

järjestyksessä. Jos datajoukolla ei ole mitään luonnollista järjestystä, datapisteiden järjestys voidaan ratkaista datapisteitä arvottavan kyselyn perusteella.

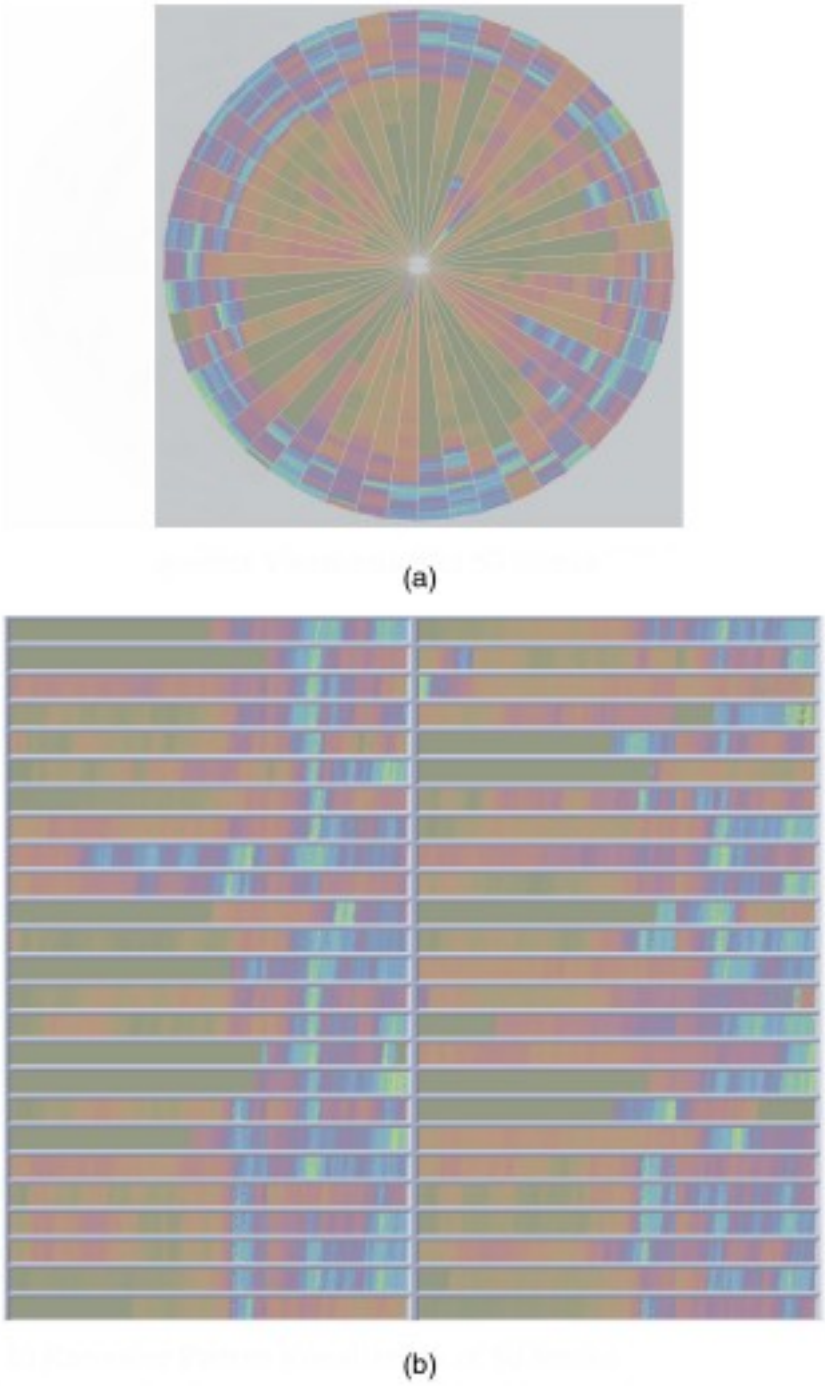


Kuva 10: Datapiste $A = (a_1, \ldots, a_8)$ esitettynä sektorimenetelmällä.
Jokainen muuttujakohtainen kuva on oma ympyrän sektorinsa.
Jokaisen datapisteen sijainti kullakin sektorilla on sama.

Keim vertaa [Kei00, Kei02] sektorimenetelmää ja rekursiomenetelmää toisiinsa. Keim toteaa menetelmien muistuttavan toisiaan, mutta sektorimenetelmän tarjoavan paremmat välineet eri muuttujien keskinäisten riippuvuuksien ja vastaavuuksien löytämiseen. Sektorimenetelmässä kaikki saman datapisteen pikselit muuttujakohtaisten kuvien sisällä ovat lähellä toisiaan. Rekursiomenetelmässä taas muuttujakohtaiset kuvat esitetään taulukkomaisesti allekkain ja päällekkäin, jolloin joidenkin muuttujien kuvat ovat väistämättä taulukon eri äärilaidoilla.

Keimin, Kriegelin ja Ankerstin [KKA95, Kei00] esimerkissä kuvassa 11 on allekkain ensin sektorimenetelmällä toteutettu visualisointi ja sen alla rekursiomenetelmällä tehty visualisointi samasta datajoukosta. Data on FAZ-indeksi kahdenkymmenen vuoden ajalta tammikuusta 1974 huhtikuuhun 1995. Vaaleat datapisteet ovat korkeampia arvoja ja tummemmat datapisteet puolestaan matalampia. Eri muuttujia tällä datajoukolla on viisikymmentä. Menetelmien vertailusta nähdään, miten sektorimenetelmällä on onnistuttu tiivistämään kokonaiskuvaa ja tuomaan saman datapisteen eri muuttujat

lähemmäs toisiaan. Toisaalta esimerkiksi ympyrän keskipistettä lähellä olevia pisteitä voi olla hankala verrata toisiinsa.



Kuva 11: Sama talousdata visualisoituna sektorimenetelmällä (a) ja rekursiomenetelmällä (b). Dataa esittelevät Keim, Kriegel ja Ankerst [KKA], kuva on värillisestä Keimin artikkelista [Kei00].

Keim nostaa [Kei00] sektorimenetelmän ongelmaksi sektoreiden järjestämisen järkevällä tavalla, eritoten jos muuttujia on paljon. Lisäksi ympyrän on oltava riittävän suuri. Nämä samat ongelmat liittyvät myös menetelmiin, joissa muuttujakohtaiset kuvat järjestetään taulukoksi: solujen järjestys on päätettävä ja monisoluinen taulukko voi olla hankala tulkita. Näiden esimerkkien kautta myös huomataan, että pikseliperusteiset menetelmät sopivat vain kohtuulliselle määrälle muuttujia. Liian monen muuttujan vertailu rinnakkain on ihmissilmille hankalaa.

3 Menetelmien soveltaminen

Pikseliperusteisia visualisointimenetelmiä käytettäessä datajoukon jokaiselle datapisteelle on määritettävä sijainti ja väri. Sijainnin määrittelystä huolehtii käytettävä visualisointimenetelmä. Väri sen sijaan saadaan kyseisen datapisteen muuttujan arvosta. Värin määrittelyä varten on valittava käytettävä väriskaala ja huolehdittava datajoukon arvojen skaalaamisesta valitulle väriskaalalle. Suuret erot joidenkin datapisteiden välillä voivat vaikuttaa negatiivisesti visualisointiin, varsinkin jos datajoukossa kiinnostavaa on yleinen käyttäytyminen, eivät yksittäiset suuret tai pienet arvot. Datajoukon muuttujien arvoja täytyy tällöin ennen visualisointia jotenkin esikäsitellä esimerkiksi logaritmien avulla.

Pikseliperusteisten menetelmien perusajatuksena on vertailla jokaista muuttujaa rinnakkain. Kokonaiskuvan esittämistä varten on päätettävä muuttujakohtaisten kuvien muoto ja järjestys kokonaiskuvassa. Jos muuttujia on hyvin paljon, muuttujakohtaisten kuvien keskinäinen järjestys korostuu entisestään. Visualisoinnin päämääränä on tuottaa kuvia, joita ihmissilmät pystyvät tulkitsemaan helposti ja nopeasti [Kei02, Hea96].

3.1 Datapisteen sijainnin määrittäminen

Suurin osa edellisessä luvussa esitellystä menetelmistä paneutuu datapisteen sijainnin määrittämiseen. Spiraalimenetelmä, akselimenetelmä ja käyrämenetelmä käyttävät valitun muotoista ja levyistä yksiulotteista käyrää, jonka avulla datajoukon pisteistä muodostetaan kaksiulotteinen kuva. Käyrä kasaa samanlaiset datapisteet lähekkäin, mikä tekee esimerkiksi klustereiden erottamisesta helppoa. Rekursiomenetelmässä datapisteiden läpikäynti etenee usein muita menetelmiä vieläkin suoraviivaisemmin, jolloin kuvaa on mahdollista lukea rivi kerrallaan kuten tekstiä.

Datapisteen sijainnin määrittämisessä on keskeistä, onko kyse kyselyriippumattomasta vai kyselyriippuvaisesta menetelmästä. Edellisessä luvussa esitellyistä menetelmistä käyrä- ja rekursiomenetelmät ovat kyselyriippumattomia, spiraali- ja akselimenetelmät kyselyriippuvaisia. Sektorimenetelmässä datapisteiden järjestys voidaan määritellä sekä kyselyllä että luonnollisen järjestyksen avulla.

Tarkastellaan seuraavaksi datapisteiden joukkoa $A = \{A_1, A_2, \ldots, A_n\}$, jossa n on joukon alkioiden lukumäärä. Joukon jokainen alkio A_i muodostuu k:sta muuttujasta eli A_i voidaan esittää vektorina $A_i = (a_1, a_2, \ldots, a_k)$. Merkitään muuttujakohtaisen kuvan leveyttä kokonaisluvulla w ja korkeutta kokonaisluvulla h, jolloin muuttujakohtaisen kuvan koko on $(w \times h)$. Keimin [Kei00] määritelmän mukaan kyselyriippumattomissa menetelmissä tulisi löytää datapisteiden joukolle A bijektiivinen kuvaus $f : \{1 \ldots n\} \rightarrow \{1 \ldots w\} \times \{1 \ldots h\}$. Kuvaus määritellään siten, että se minimoi summan S

$$S = \sum_{i=1}^{n} \sum_{j=1}^{n} \left| d\left(f(i), f(j)\right) - d\left((0,0), \left(w \cdot \sqrt{\frac{|i-j|}{n}}, h \cdot \sqrt{\frac{|i-j|}{n}}\right)\right) \right| ,$$

missä i ja j ovat joukon A alkioita edustavien pikseleiden indeksejä, ja $d(f(i), f(j))$ on datapisteiden A_i ja A_j välinen etäisyys kuvassa. Keimin [Kei00] määritelmän ajatuksena on sovittaa monimuuttujaisen datajoukon pisteet optimaalisimmalla tavalla yksiulotteiseen järjestykseen käyrälle. Käyrä täyttää kuvan pinnan, joten lopputuloksena on kaksiulotteinen visualisointi.

Kyselyriippuvaisten menetelmien tapauksessa Keim lisää minimoitavaan summaan S termin T

$$T = \sum_{i=0}^{n} \left| d\left(f(i), \left(\frac{w}{2}, \frac{h}{2}\right)\right) - d\left((0,0), \left(\frac{w}{2} \cdot \sqrt{\frac{i}{n}}, \frac{h}{2} \cdot \sqrt{\frac{i}{n}}\right)\right) \right| ,$$

missä $d(f(i), (w/2, h/2))$ on pikselin A_i etäisyys kuvan keskipisteeseen. Termin T ansiosta parhaiten kyselyä vastaavat datapisteet saadaan lähimmäksi muuttujakohtaisen kuvan keskipistettä.

3.2 Datapisteen värin määrittäminen

Pikseliperusteisissa menetelmissä datapistettä edustaa kuvassa aina yksi pikseli. Pikselin väri perustuu datapisteen muuttujan arvoon kulloisessakin muuttujakohtaisessa kuvassa.

Sekä Healey [Hea96] että Wang et al. [WGM08] korostavat onnistuneen värien valinnan tärkeyttä. Visualisointi tehdään ennen kaikkea ihmisten avuksi, ja suuri osa ihmisistä on tottunut hahmottamaan maailmaa värien kautta.

Värit täytyy tietokoneen näyttöä varten esittää RGB-formaatissa. Tämä tarkoittaa, että väri ilmaistaan kolmella muuttujalla red, green ja blue, joiden numeroarvot ovat väliltä 0 – 255. Visualisointia tuotettaessa värejä voidaan käsitellä myös muussa muodossa, kuten *HSV-väriympyränä* (eng. HSV color wheel tai color space), jolloin värin laskeminen datapisteen muuttujan perusteella on helpompaa [WGM08]. Valmiit värit täytyy jälkeenpäin renderöidä RGB-formaattiin tietokoneen näyttöä varten.

Healey toteaa [Hea96], että tehokas värien käyttö visualisoinnissa riippuu kolmesta tekijästä. Ensimmäinen on *värien etäisyys* (eng. colour distance) toisiinsa nähden, jossa euklidisella etäisyydellä ilmaistaan värien eroavaisuus. Toinen on värien *lineaarinen erottelu* (eng. linear separation) ja kolmas *värien ryhmittely* (eng. colour category), jotka molemmat pyrkivät siihen, ettei visualisoinnissa käytetä liian samanlaisia värejä eri datapisteille.

Healey pyrkii määrittelemään värit niin, että ihmissilmät voivat havaita eri pikselien väliset erot värin perusteella nopeasti ja luotettavasti. Wuerger, Karatzas ja Meyer [WKM05] lisäävät, että väriskaalan asettamisessa on huomioitava eri ihmisten fysiologiset rajoitteet, joiden takia värit nähdään hiukan eri tavalla. Ainakin tyypillisimpiä ongelmatilanteita, kuten punaisen ja vihreän värin sekoittumista, on osattava välttää. Healeyn kolmen periaatteen perusteella asetetaan *väriskaala* (eng. colour scale), jossa esitetään kaikki kyseisessä visualisoinnissa käytettävät värit oikeissa suhteissa toisiinsa nähden.

Kunkin datapisteen väri määritetään skaalaamalla käsiteltävän muuttujan arvo asetetulle väriskaalalle. Wang ja muut esittävät [WGM08] mallin, jolla väri saadaan määritettyä kolmiulotteisesta väriskaalasta datapisteen muuttujan arvon perusteella. Ensin lasketaan värin *sävy* (eng. hue) eli mistä väristä on kyse: punaisesta, sinisestä, vihreästä ja niin edelleen. Toiseksi lasketaan värin *värikylläisyys* (eng. saturation tai vividness), joka kertoo kuinka voimakas väri on eli kuinka paljon siinä on mukana harmaata. Kolmanneksi lasketaan värin *kirkkaus* (eng. lightness) eli kuinka paljon värissä on mukana valkoista.

3.3 Kokonaiskuvan esittäminen

Pikseliperusteisten menetelmien keskeinen ajatus on, että jokainen datajoukon muuttuja esitetään omassa kuvassaan. Näitä muuttujakohtaisia kuvia vertaillaan rinnakkain suurempana kokonaisuutena. Kokonaisuuden rakentaminen oikein helpottaa vertailua. Suurimmalla osalla menetelmistä muuttujakohtaiset kuvat kootaan taulukkoon. Tähän vaihtoehdon tarjoaa sektorimenetelmä, jossa jokainen muuttujakohtainen kuva on yksi ympyrän sektori.

Keim [Kei00] antaa määritelmän muuttujakohtaisten kuvien esittämiselle. Olkoon A_i datapiste n-alkioisesta joukosta $A = \{A_1, A_2, \ldots, A_n\}$. Merkitään muuttujien lukumäärää kokonaisluvulla k, jolloin datapiste $A_i = (a_1, a_2, \ldots, a_k)$. Määritelmän mukaan on minimoitava summa U

$$U = \frac{1}{n} \sum_{l=1}^{n} \left(\frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{k} d\left(f\left(a_i\right), f\left(a_j\right)\right) \right) ,$$

missä $d(f(a_i), f(a_j))$ on datapisteen A_i kahden muuttujan a_i ja a_j välinen erotus. Summaa U minimoimalla pyritään tilanteeseen, jossa saman datapisteen eri muuttujien arvot ovat mahdollisimman lähellä toisiaan siitäkin huolimatta, että jokaiselle muuttujalle on piirretty oma muuttujakohtainen kuvansa. Näin muuttujien vertailu rinnakkain on mielekästä. Sekä taulukon solujen että sektoreiden esittämisjärjestys on huomioitava, varsinkin jos muuttujia on hyvin monta. Muuttujakohtaisten kuvien järjestämiseen on kehitetty erilaisia algoritmeja [Kei00, WLL08].

3.4 Kritiikkiä

Kaikilla visualisointimenetelmillä on hyvät ja huonot puolensa. Ferreira de Oliveiran ja Levkowitzin [FeL03] sekä Keimin ja Kriegelin [KeK96] mukaan pikseliperusteiset menetelmät sopivat suurille datajoukoille. Jokainen datapiste tarvitsee oman pikselinsä, joten näytön resoluutio asettaa rajoitteen sille, kuinka monta pikseliä kuvaan voidaan piirtää. Pikselit eivät sekoitu toisiinsa, sillä samankaltaiset datapisteet sijoitetaan lähekkäin, ei limittäin tai päällekkäin. Tämän ansiosta pikseliperusteisilla menetelmillä on helppo erottaa erityisesti datajoukossa olevia klustereita.

Ferreira de Oliveira ja Levkowitz [FeL03] toteavat pikseliperusteisten menetelmien sopivan myös kohtuullisen monen muuttujan datajoukoille. Jokaiselle muuttujalle

piirretään omat kuvansa, joten eri muuttujat erottaa helposti toisistaan. Muuttujien määrä on kuitenkin pidettävä kohtuullisena: liian monta muuttujakohtaista kuvaa tekee kokonaiskuvan tulkinnasta vaikeaa.

Käytettävä visualisointimenetelmä on aina valittava kulloisenkin datajoukon mukaan. Visualisoinnista on paljon apua ihmisille datajoukon rakenteen ymmärtämisessä, mutta huonosta visualisoinnista on vain haittaa. Väärin valittu visualisointimenetelmä voi tuottaa sekavia ja epäinformatiivisia, jopa virheellisiä kuvia.

4 Yhteenveto

Pikseliperusteisten menetelmien keskeinen ajatus on esittää monimuuttujaisen datajoukon jokainen muuttuja omassa kuvassaan ja vertailla niitä rinnakkain. Muuttujan arvo määrittää kutakin datapistettä edustavan pikselin värin. Datapisteet piirretään kuvaan sijoittamalla ne ensin yksiulotteiseen järjestykseen käyrälle: valitun mallista käyrää läpikäymällä täytetään kuvan kaksiulotteinen pinta. Käyrän tarkoituksena on kerätä toisiaan muistuttavia pisteitä lähekkäin klustereiksi. Jotta muuttujakohtaisten kuvien vertailu olisi mielekästä, jokaisessa muuttujakohtaisessa kuvassa datapistettä edustavan pikselin sijainti on sama.

Pikseliperusteisten menetelmiin liittyy kolme peruskysymystä, joihin eri menetelmät tarjoavat erilaisia vastauksia: Miten pikselit sijoitetaan muuttujakohtaisen kuvan sisään? Miten pikselin väri määritellään? Ja miten muuttujakohtaiset kuvat esitetään yhtenä selkeänä kokonaisuutena? Tässä tutkielmassa olen esitellyt spiraali-, akseli-, käyrä-, rekursio- ja sektorimenetelmät. Spiraali- ja akselimenetelmät sopivat tilanteeseen, jossa datapisteiden järjestys ratkaistaan datajoukolle tehtävällä kyselyllä, esimerkiksi verrataan datapisteitä johonkin oletusarvoon. Käyrä- ja rekursiomenetelmiä käytetään tilanteissa, joissa datajoukolla on olemassa jokin luontainen järjestys, esimerkiksi ajanhetki, jonka perusteella datapisteet voidaan järjestää. Sektorimenetelmä mahdollistaa muuttujakohtaisten kuvien vertailun lähellä toisiaan.

Jokaisessa menetelmässä käytettävien värien valintaan on kiinnitettävä huomiota. Värit tulee valita niin, että kuva on helposti luettavissa ja kuvasta nähdään datajoukon ominaisuuksia, kuten klustereita ja yksittäisten datapiteiden tai datapistejoukkojen eroja. Pikseleiden värittämistä varten lasketaan väriskaala, jolle jokaisen datapisteen

muuttujan arvo skaalataan. Käytettävä väriskaala on sama kaikille muuttujille, mikä mahdollistaa muuttujakohtaisten kuvien vertailun.

Spiraali-, akseli-, käyrä- ja rekursiomenetelmissä muuttujakohtaiset kuvat sijoitetaan taulukkomaisesti rinnakkain yhdeksi kokonaisuudeksi. Sektorimenetelmä tarjoaa vaihtoehtoisen mallin kokonaiskuvalle: siinä kukin muuttujakohtainen kuva esitetään ympyrän sektorina. Muuttujakohtaisten kuvien järjestys kokonaiskuvassa on tärkeää muuttujien onnistuneen vertailun kannalta.

Pikseliperusteiset menetelmät sopivat suurten datajoukkojen visualisointiin. Jokaiselle datapisteelle piirretään oma pikselinsä, joten datapisteet eivät pääse sekoittumaan tai limittymään toisiinsa. Pikseliperusteiset menetelmät sopivat myös kohtuullisen monen muuttujan datalle, sillä jokainen muuttuja saa visualisoinnissa oman kuvansa, eivätkä eri muuttujat sekoitu toisiinsa. Liian monen muuttujan vertailu voi olla hankalaa, ja muuttujakohtaiset kuvat on tarvittaessa järjestettävä. Visualisoinnissa käytettävän näytön resoluution tulee olla riittävän suuri.

Pikseliperusteisilla menetelmillä on mahdollista tuottaa epäselviä ja vaikeasti luettavia visualisointeja. Visualisointimenetelmä on valittava niin, että se sopii juuri kulloinkin käsiteltävälle datajoukolle. Oikean menetelmän selvittämiseksi täytyy ehkä kokeilla useampaa eri menetelmää. Tärkeintä on, että visualisointi auttaa ihmisiä ymmärtämään datajoukon rakennetta.

Lähteet

FeL03 M. C. Ferreira de Oliveira ja H. Levkowitz, From visual data exploration to visual data mining, *IEEE Trans. Visualization and Computer Graphics,* vol. 9, nro. 3, s. 378-394, 2003

Hae96 C. G. Haeley, Choosing effective colours for data visualization, IEEE Visualization: Proceedings of the 7th conference on Visualization '96, s. 263-270, 1996

Kei00 D. A. Keim, Designing pixel-oriented visualization techniques: theory and applications, *IEEE Trans. Visualization and Computer Graphics*, vol. 6, nro. 1, s. 59-78, 2000

Kei02 D. A. Keim, Information visualization and visual data mining, *IEEE Trans. Visualization and Computer Graphics*, vol. 8, nro. 1, 2002

KeK96 D. A. Keim ja H.-P. Kriegel, Visualization techniques for mining large databases: a comparison, *IEEE Trans. Knowledge and Data Eng.,* vol. 8, nro. 6, s. 923-936, 1996

KKA95 D.A. Keim, H.-P. Kriegel ja M. Ankerst, Recursive pattern: a technique for visualizing very large amounts of data, Proc. Visualization '95, s. 279-286, 1995

MAK03 M. F. Mokbel, W. G. Aref ja I. Kamel, Analysis of multi-dimensional space-filling curves, *GeoInformatica*, vol. 7 nro. 3, s. 179-209, 2003

SSK07 J. Schneidewind, M. Sips ja D. A. Keim, An automated approach for the optimization of pixel-based visualizations, *Information visualization*, nro. 6, s. 75-88, 2007

WKM05 S. M. Wuerger, D. Karatzas ja G. F. Meyer, A display calibration technique based on invariant human colour mechanism, Proceedings - APGV 2005: 2nd Symposium on Applied Perception in Graphics and Visualization, s. 171, 2005

WGM08 L. Wang, J. Giesen, K. T. McDonnell, P. Zolliker ja K. Mueller, Color design for illustrative visualization, *IEEE Trans. Visualization and Computer Graphics*, vol. 14, nro. 6, s. 1739-1746, 2008

WLL08 S. L. Wang, C. C. Loy, C. P. Lim, W. K. Lai ja K. S.Tan, Use of circle-segments as a data visualization technique for feature selection in pattern classification, *Lecture Notes in Computer Science*, vol. 4984/2008, s. 625-634, 2008