hyväksymispäivä arvosana

arvostelija

Spatiaalisen tiedon louhinta ja sen käyttö rikosten tutkimisessa

Anu Tanninen

Helsinki 16. huhtikuuta 2003
Spatiaalisen tiedon louhinta -seminaari
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

1 Johdanto

Tiedon louhinnassa pyritään löytämään annetusta tiedosta säännönmukaisuutta ja toistuvia hahmoja. Lisäksi etsitään sellaisia hahmoja, jotka esiintyvät usein yhdessä. Tällaisia yhteyksiä haetaan *assosiaatiosääntöjen* avulla. Kullakin säännöllä on *frekvenssi*, joka ilmaisee, kuinka usein säännön osapuolet esiintyvät yhdessä [EsL01]. Esimerkiksi sääntö $A \Rightarrow B$ *(c %)* ilmaisee, että *c %* annetusta datasta täyttää ehdon *B,* jos se täyttää ehdon A. Asettamalla säännöille minimifrekvenssit saadaan sääntöjoukkoa karsittua pienemmäksi [MaT02].

Frekvenssien lisäksi assosiaatiosäännölle $A \Rightarrow B$ *(c %)* lasketaan *konfidenssi*, joka on ehdollinen todennäköisyys sille, että ehto B esiintyy jos A esiintyy [EsL01, MaT02]. Myös konfidenssille voidaan määrätä sellainen minimi, joka rajoittaa tarkasteltavien sääntöjen määrää riittävästi.

Spatiaalisen tiedon louhinnassa tarkastellaan maantieteellisiä hahmoja ja ympäristöjä sekä etsitään ei-satunnaisia tapahtumia [Ans99]. Spatiaalinen tietokanta sisältää objekteja, jotka on kuvailtu spatiaalisen paikan perusteella. Koska spatiaalisen tiedon louhinnassa mahdolliset hahmot ovat monimutkaisia, on se tavallista tiedon louhintaa hankalampaa [EKS01]. Spatiaalisen tiedon louhinnassa voidaan käyttää esimerkiksi *spatiaalista klusterointia*, jossa tietokannan objektit jaetaan aliryhmiin niiden ominaisuuksien perusteella [EKS01, EsL01, GrM01]. Lisäksi voidaan tarkastella objektien välisiä spatiaalisia suhteita niiden *naapuruussuhteiden* ja näistä suhteista saatavien *naapuruuskarttojen* avulla [EKS01].

Spatiaalisen tiedon louhintaa voidaan hyödyntää rikosten tutkimisessa ja ennaltaehkäisemisessä. Työ ei kuitenkaan ole helppoa, koska rikoksista talletettava tieto on liian yksityiskohtaista [EsL01]. Jotta rikoksia voitaisiin ennaltaehkäistä, on niiden yleisyyteen vaikuttavien tekijöiden tunnistaminen tärkeää. Esimerkiksi tutkittavan alueen maantieteellinen sijainti saattaa olla yksi osatekijä rikosten yleisyydessä. Rikosten tutkimisessa voidaan käyttää avuksi esimerkiksi *rikoskarttoja*, joiden avulla pystytään identifioimaan rikoksia visuaalisesti [Car99]. Monimutkaisten rikosten ymmärtämiseen voidaan käyttää *kuumia pisteitä* [Lev02, EsL01], joiden paikallistamiseen käytetään esimerkiksi *GIS-työkalua* (Geographic Information Systems) [GAA99, GrM01]. *CrimeStat* on eräs spatiaalinen tilastollinen ohjelma, jolla voidaan analysoida paikallistettuja rikostapauksia [Lev02]. Kyseinen ohjelma on suunniteltu rikosten

kartoituskeskukselle.

Luvussa kaksi ja sen aliluvuissa tarkastellaan spatiaalisen tiedon louhintaa yleisesti. Kolmas luku käsittelee rikosten kartoitusta ja niiden ennaltaehkäisyä. Neljännessä luvussa on esitelty muutama spatiaalisen tiedon louhinnan menetelmä ja niiden mahdollinen käyttö kriminologiassa. Viimeinen luku on yhteenveto käsitellyistä asioista.

2 Spatiaalisen tiedon louhinta

Tiedon louhinnassa etsitään annetusta datasta sellaisia ymmärrettävissä olevia hahmoja, jotka ovat mahdollisesti hyödyllisiä. Spatiaalinen tietokanta sisältää spatiaalisen paikan perusteella kuvailtuja objekteja, kuten esimerkiksi vuorijonoja tai alueellista eristyneisyyttä [EKS01]. Koska hahmot, joita spatiaalisen tiedon louhinnassa etsitään ovat monimutkaisempia kuin tavallisen tietokannan tieto, on spatiaalisen tiedon louhinta vaikeampaa.

Jotta saataisiin mahdollisimman paljon hyödyllistä tietoa objekteista, on tutkittavien hahmojen *naapurit* otettava huomioon spatiaalisen tiedon louhinnan algoritmeissa. Tarkasteltavien naapureiden attribuuteilla saattaa olla merkittävää vaikutusta itse tarkasteltavaan objektiin. Esimerkiksi jonkin tarkasteltavan pikkukylän suureen rikollismäärään saattaa yhtenä osatekijänä olla läheisyydessä oleva slummialue. Tällöin myös naapuri, eli slummialue, on otettava lähempään tarkasteluun. Spatiaalisen tiedon louhinnassa tutkitaankin muun muassa objektien välisiä naapuruuskarttoja ja naapuruussuhteita [EKS01].

Objektit tai hahmot voidaan jakaa pienempiin aliryhmiin joidenkin niiden ominaisuuksien perusteella. Tällaisessa spatiaalisessa klusteroinnissa voidaan sitten tutkia näitä ryhmiä erikseen ja tehdä niiden perusteella arvioita ja analyyseja.

Seuraavissa aliluvuissa tarkastellaan spatiaalisia naapuruussuhteita ja -karttoja sekä spatiaalista klusterointia. Etenkin spatiaalista klusterointia käytetään rikosten tutkimisessa ja analysoinnissa.

2.1 Spatiaaliset naapurussuhteet ja naapuruuskartat

Topologinen, etäisyys ja *suunta* ovat kolme perustyyppiä spatiaalisissa naapuruussuhteissa [EKS01]. Ensimmäinen näistä perustuu kahden toisiinsa jollakin tavalla liittyvän objektin rajoihin, sisuksiin ja komplementteihin sekä niiden välisiin erilaisiin suhteisiin. Olkoot esimerkiksi alueet *A* ja *B* kaksi tarkasteltavaa objektia. Tällöin niiden välinen suhde voi olla muun muassa se, että ne ovat erillisiä, ne leikkaavat toisensa, ovat samat tai alue *A* peittää alueen *B*.

Etäisyyssuhteet puolestaan vertailevat kahden tarkasteltavan objektin välistä etäisyyttä. Tällöin käytetään vertailuoperaattoreina $<$, $>$ tai $=$. Suuntasuhteet perustuvat objektien sijaintiin toisiinsa verrattuina. Suuntina ovat yleiset ilmansuunnat eli pohjoinen, itä, etelä, länsi, koillinen, kaakko, lounas ja luode.

Jos naapuri määritellään jollakin edellä mainitulla naapuruussuhteella ja *DB* on spatiaalisia objekteja sisältävä tietokanta, voidaan naapuruuskartat määritellä seuraavasti [EKS01]. Naapuruuskartta $G = (N, E)$ on kartta, missä $N = DB$ on *solmujoukko* ja $E \subseteq N \times N$ on kartan *kaaret*. Kaari $e = (n_1, n_2)$ on olemassa kartassa, jos on olemassa naapuruus (n_1, n_2). *K:n* pituinen *naapuruuspolku* on määritelty solmusegvenssinä $[n_1, n_2, ..., n_k]$, missä naapuruus (n_i, n_{i+1}) on voimassa kaikilla solmuilla n_i, kun $1 \leq i \leq k$. Naapuruuskarttojen pohjalta voidaan tehdä spatiaalisen tiedon louhintaa.

Koska spatiaaliset hahmot ovat usein jonkin objektin tai sen naapuruussuhteiden vaikutuksen seurauksia, on tärkeää tarkastella oleellisia polkuja. Jos poluille ei anneta mitään rajoittavaa attribuuttia, saattaa tarkasteltavien polkujen määrä kasvaa suureksi. Attribuutti, jonka tehtävänä on suodattaa vähemmän merkitykselliset tekijät pois, voi olla joko spatiaalinen tai ei-spatiaalinen. Näitä vähemmän merkityksellisiä tekijöitä voi olla esimerkiksi lähimmän naapurin liian pitkä välimatka tarkastelun kannalta.

2.2 Spatiaalinen klusterointi

Spatiaalisessa klusteroinnissa annetun tietokannan objektit ryhmitellään merkityksellisiin aliryhmiin, joita sitten tarkastellaan. Saman ryhmän sisällä kullakin objektilla on jokin sama piirre, kuten esimerkiksi sijainti suurten maanteiden läheisyydessä. Lisäksi aliryhmien väliset erot ovat mahdollisimman erilaiset [EKS01]. Edellisen aliryhmäesimerkin toinen aliryhmä voisi sisältää ne alueet, jotka ovat kaukana kaikista suurista teistä.

Spatiaalista klusterointia käytetään muun muassa etsimään suurista spatiaalisista tietokannoista hahmoja, jotka ovat keskittyneet spatiaalisesti esimerkiksi jonkin rautatien varteen [EsL01]. Kun jokin spatiaalinen kokonaisuus löydetään, on se merkki kenties mielenkiintoisesta alueesta. Tällainen alue tarvitsee lisää tutkimusta, jotta löydettäisiin mahdolliset korrelaatiot.

Koska suuret, spatiaaliset tietokannat sisältävät paljon erityyppistä tietoa, on spatiaalinen klusterointi käytännöllinen menetelmä tiedon ryhmittelemiseksi [GrM01]. Klusterianalyysejä tehtäessä voidaan käyttää erilaisia tilastollisia ohjelmia kuten SAS ja CrimeStat [Lev02]. Jälkimmäinen ohjelma on suunniteltu rikosten kartoitustutkimuskeskukselle ja sillä voidaan analysoida paikallistettuja rikostapauksia.

Hierarkisessa klusteroinnissa aliryhmien muodostaminen aloitetaan sijoittamalla kukin objekti omaan ryhmäänsä [GrM01]. Siis ryhmiä on yhtä monta kuin on objektejakin. Tämän jälkeen yhdistellään eniten samoja ominaisuuksia sisältäviä ryhmiä yhteen jonkin määrätyn kriteerin pohjalta. Yhdistämiskriteerinä voi olla esimerkiksi kahden naapurin tai pisteen etäisyys, jota verrataan kaikkien tutkittavien pisteiden etäisyyksien keskiarvoon. Jos kyseinen kriteeri täyttyy, eli etäisyys alittuu, yhdistetään tarkastellut ryhmät yhdeksi klusteriksi. Yhdistämistä jatketaan, kunnes on päästy sopivalle tarkkuustasolle. Hierarkisessa klusteroinnissa on mahdollista tutkia rikosten keskittymiä pienillä alueilla ja linkkejä muodostettujen klustereiden välillä. Ongelmana on kuitenkin, että hierarkinen klusterointi tuottaa usein ennemmin lokaalin kuin globaalin optimin. Tämä johtuu siitä, että klusterointi voidaan tuottaa useammalla eri kriteerillä. Myös kriteerien rajojen määrittäminen on hankalaa. Esimerkiksi, jos rikoksia ei lajitella niiden vakavuuksien perusteella, saattaa jokin alue vaikuttaa suotuisammalta rikoksien mahdollisuuksille kuin jokin toinen. Tämä voisi tapahtua silloin, jos jokainen ryöstö ja murha asetetaan samantasoisiksi rikoksiksi. Jos nyt tutkittaisiin poliisimäärän alueellisen lisäämisen tarvetta, voisi resurssit mennä väärään paikkaan.

Toinen tapa muodostaa aliryhmät on *ositus* [GrM01], missä objektit pyritään jakamaan ennalta määriteltyyn määrään ryhmiä. Ositus siis poikkeaa hierarkisesta klusteroinnista muun muassa siinä, että aliryhmien määrä on valmiiksi tiedossa. Ongelmana on kuitenkin määritellä sopiva määrä klustereita annetulle joukolle rikostapauksia. Tähän ongelmaan on kehitelty useita erilaisia ratkaisuja, joita tässä esitelmässä ei kuitenkaan tarkastella.

Tutkittaessa rikoksia ja niiden alueellista esiintymistä, tarkastellaan *kuumia pisteitä*, jotka ovat spatiaalinen ilmiö. Näiden kuumien pisteiden etsimisessä on ymmärrettävä, mikä tekee juuri tietyistä klustereista merkittävämpiä kuumien pisteiden analyysissä. Esimerkiksi ne aliryhmät, joissa on suurempi määrä rikostapauksia kertovat selvästi suuremmasta rikollisesta toiminnasta. Tosin ne eivät välttämättä kerro, kuinka vakavista rikoksista on kyse.

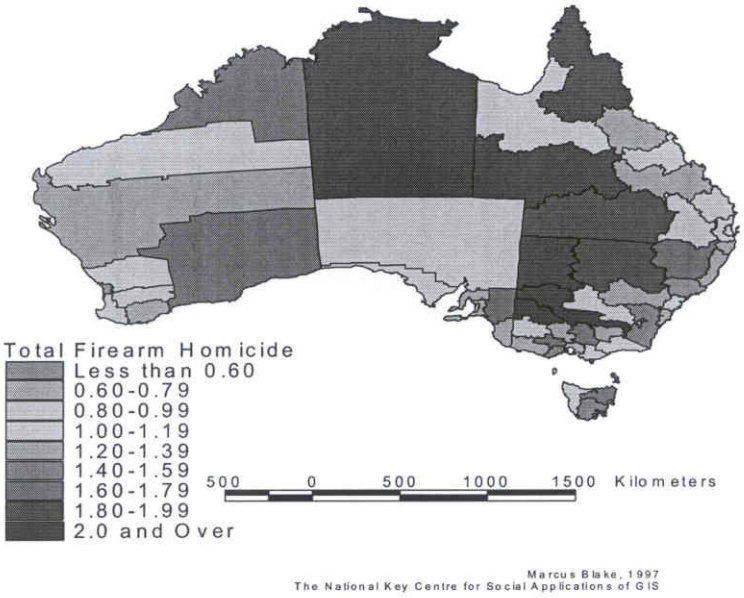
3 Rikollisuuden kartoitus ja ennaltaehkäisy

Rikosten tutkiminen ja ennaltaehkäiseminen on haastava tehtävä. Tärkeää onkin löytää ja pyrkiä tunnistamaan rikosten yleisyyteen vaikuttavia tekijöitä, jotta uusia rikoksia voitaisiin ehkäistä esimerkiksi poliisiresursseja lisäämällä. Tekijöiden tai hahmojen tunnistaminen puhtaasta, käsittelemättömästä tiedosta on kuitenkin vaikeaa, koska rikoksista talletettava tieto on liian yksityiskohtaista ja saattaa olla keskenään erityyppistä [EsL01].

Analysoitaessa rikosaktiviteettia ja sen keskittyneisyyttä, on tärkeää tutkia rikostapausten pohjalta tehtyjä klustereita. Esimerkiksi *tietoköyhissä ympäristöissä* ei ole mahdollista ottaa huomioon naapureiden vaikutusta tutkittavasta objektista kerättävään informaatioon, vaan on tarkasteltava vain jotakin tiettyä klusteria. Tietoköyhistä ympäristöistä ei siis saada paljoakaan taustatietoa tutkimuksen avuksi. *Tietorikkaissa ympäristöissä* puolestaan kerätään kaikki mahdollinen taustatieto valtavaan spatiaaliseen tietokantaan. Tätä tietokantaa käytetään sitten spatiaaliseen klusterointiin ja klustereiden välisten yhteyksien tutkimiseen. Toisin kuin tietoköyhässä ympäristössä, saadaan tietorikkaissa ympäristöissä tietoa myös naapureilta.

Rikoksia tutkittaessa on esimerkiksi erilaisista ohjelmista hyötyä, jolloin saadaan visualisoitua rikosten alueellinen keskittyminen paremmin. Kuvassa 1 on esimerkki rikoskartasta, jossa on kuvattu Australiassa tapahtuneet aseista johtuneet kuolemat. Kartan perusteella voidaan nopeasti ja selkeällä tavalla huomata ne alueet, joilla aseelliset rikosmäärät ovat suuria.

GIS (Geographic Information Systems) on eräs työkalu, joka on antanut mahdollisuuden kartoittaa rikoksia. Kartoituksen avulla rikollisuutta voidaan yrittää kontrolloida ja ennaltaehkäistä tehokkaammin [Car99]. Rikosten kartoituksen tarkoituksena on kuvata informaatiota maantieteellisten alueiden, rikosten ja erilaisten riskitekijöiden välisistä



Kuva 1 [Car99]. Rikoskartta Australiassa tapahtuneista aseista johtuvista kuolemista

suhteista. Esimerkiksi voitaisiin tutkia nuorten miesten tekemien aseellisten ryöstöjen määrää ja keskittyneisyyttä Australiassa. Tällöin tutkittaisiin alueita, joilla on mahdollisia yhteyksiä nuorten miesten ja aseellisten ryöstöjen välillä. Tarkastellut alueet voitaisiin kartoittaa kuvan 1 tapaiseen karttaan, josta olisi helpompi tehdä joitakin päätelmiä.

Carlos Carcach [Car99] tutki rikollisuutta Australiassa ja huomasi, että erityyppiset alueelliset hahmot vaikuttavat eri tavalla naisten ja miesten aseista johtuvien kuolleisuuksien määrään. Esimerkiksi miesten kuolleisuus oli suurempaa usein maantieteellisesti eristyneillä alueilla. Myös sellaisten alueiden, joissa on paljon työttömyyttä, sosiaalista eristyneisyyttä tai huonot kulkuyhteydet muualle, vaikutusta aseista johtuvien kuolemien yleisyyteen tutkittiin. Carcachin mukaan sellaiset alueet, joista on huonot kulkuyhteydet muualle, eivät ole niin riskialttiita aseista johtuviin kuolemantapauksiin kuin hyvien kulkuyhteyksien päässä olevat alueet. Yhtenä mahdollisuutena pyrkiä ennaltaehkäisemään rikollisuutta, on tutkia ihmisten iän merkitystä rikoksiin ja tehdä tämän pohjalta alueellinen ikäkartoitus. Tällöin voitaisiin mahdollisesti pyrkiä vähentämään tiettyjen rikoksten lukumäärää, esimerkiksi nuorten väkivaltarikoksia.

4 Spatiaalisen tiedon louhinnan menetelmiä kriminologiassa

Kriminologiassa voidaan käyttää jollakin tasolla samoja periaatteita hahmojen tunnistamisessa kuin tavallisenkin tai spatiaalisen tiedon louhinnassa. Haasteena on kuitenkin se, että rikoksista talletettava tieto on liian monimuotoista ja yksityiskohtaista. Talletettu tieto on muutettava ennen tutkimista yhdenmukaisempaan muotoon. Seuraavissa aliluvuissa tarkastellaan tiedon louhinnan assosiaatiosääntöjen etsimistä sekä kriminologiassa tarkasteltavien kuumien pisteiden analysointia.

4.1 Assosiaatiosääntöjen louhinta

Assosiaatiosääntöjen louhintaa käytetään suurten tietokantojen korrelaatioiden löytämiseksi. Assosiaatiosääntö on ilmaus $A \Rightarrow B$ *(c %)*, missä *A* on edeltävä tapahtuma ja *B* on seuraus. Säännön avulla kuvataan, kuinka suuri osuus, *c %,* tiedosta, joka täyttää ehdon *A* täyttää myös ehdon *B*. Esimerkiksi sääntö *is_a(x, house)* $\wedge$ *near_by(x, beach)* $\Rightarrow$ *is_expensive(x)(95 %)* ilmaisee, että 95 prosenttia taloista, jotka ovat rannalla*,* ovat myös kalliita [EsL01]. Luvussa 3 mainittiin muutamia riskitekijöitä aseellisiin rikoksiin. Näihin perustuen voisi assosiaatiosääntö olla esimerkiksi muotoa *on(x, mies)* $\wedge$ *asuu(x, eristyneisyys)* $\Rightarrow$ *aseesta_johtuva_kuolema(x)(95 %)*.

Oletetaan, että tarkastellaan 0/1-relaatiota, johon tieto on talletettu. Kullakin säännöllä $A \Rightarrow B$ *(c %)* on frekvenssi, joka kertoo kuinka usein *A* ja *B* esiintyvät samalla rivillä annetussa datassa. Lisäksi jokaisella säännöllä on konfidenssi, joka on ehdollinen todennäköisyys sille, että ehto *B* esiintyy, jos *A* esiintyy. Frekvenssi saadaan laskettua, kun jaetaan niiden rivien lukumäärä, joilla *A* ja *B* esiintyvät, kaikkien rivien lukumäärällä. Konfidenssi saadaan puolestaan, kun jaetaan niiden rivien lukumäärä, joilla sekä *A* että *B* esiintyvät niiden rivien lukumäärällä, joilla esiintyy *B* [MaT02].

Koska assosiaatiosääntöjen lukumäärä saattaa kasvaa valtavaksi, on sääntöjen lukumäärää voitava karsia. Määräämällä sopivat minimifrekvenssi ja minimikonfidenssi saadaan sääntöjen joukosta eroteltua ne, jotka saattavat olla mielenkiintoisia tai merkittäviä esimerkiksi rikoksia tutkittaessa. Kun säännöille, joita halutaan tarkastella, on määritelty minimifrekvenssi ja minimikonfidenssi, voidaan nämä ehdot täyttävät säännöt etsiä. Louhintaan voidaan käyttää esimerkiksi *Apriori-algoritmia* [MaT02], joka etsii kriteerit täyttävät säännöt annetusta relaatiosta.

Taulukko 1 on pieni esimerkki 0/1-relaatiosta, jonka pohjalta voidaan etsiä sääntöjä. Esimerkiksi ehdot *murha* ja *mies* esiintyvät kahdesti samalla rivillä. Tällöin säännön *murha* $\Rightarrow$ *mies* frekvenssi on 2/4 ja konfidenssi on 2/3. Jos ennalta olisi määritelty minifrekvenssiksi 3/4 ja minimikonfidenssiksi 3/4, ei kyseistä sääntöä olisi valittu suurempaa tarkastelua vaativien sääntöjen joukkoon.

Koska jotkut alueelliset ominaisuudet antavat suotuisammat mahdollisuudet rikollisuudelle, syntyy rikosten keskittyneisyyttä. Spatiaalinen assosiaatiosääntöjen louhinta pyrkii löytämään mahdolliset vaikutteet tai jotkut spatiaaliseen klusteroitumiseen vaikuttavat tekijät. Mahdollisia selityksiä rikollisuuden yleisyyteen spatiaalisissa yhteyksissä ovat esimerkiksi alueen fyysinen olemus ja erilaisten palveluiden läheisyys [EsL01]. Esimerkiksi, jos aseelliset ryöstöt ovat yleisiä pikkukylissä, joissa on mahdollista hankkia ase, voidaan samantyyppisille alueille lisätä poliisiresursseja. Spatiaalista klusterointia ja assosiaatiosääntöjen louhintaa käytetään kuumien pisteiden paikallistamisessa.

4.2 Kuumat pisteet

Kuumien pisteiden analyysiä käytetään varsinkin monimutkaisten rikosten ymmärtämiseen ja niiden analysoiminen onkin tärkeää rikostutkijoille, kun määritellään onnettomuuksien keskittyneisyyttä. Löydettäessä kuumia pisteitä, voidaan poliiseja ja kunnallisia resursseja lisätä tarpeen mukaan näille alueille [Lev02]. Kuumat pisteet ovat siis spatiaalinen ilmiö.

Yksi tapa löytää kuumia pisteitä on käyttää klusterianalyysiä [GrM01]. Esimerkiksi CrimeStat- ja SAS-ohjelmat voivat etsiä tehokkaasti alueita, joille rikokset ovat keskittyneet. Kuitenkin yhtenä ongelmana on määritellä nämä kuumat pisteet

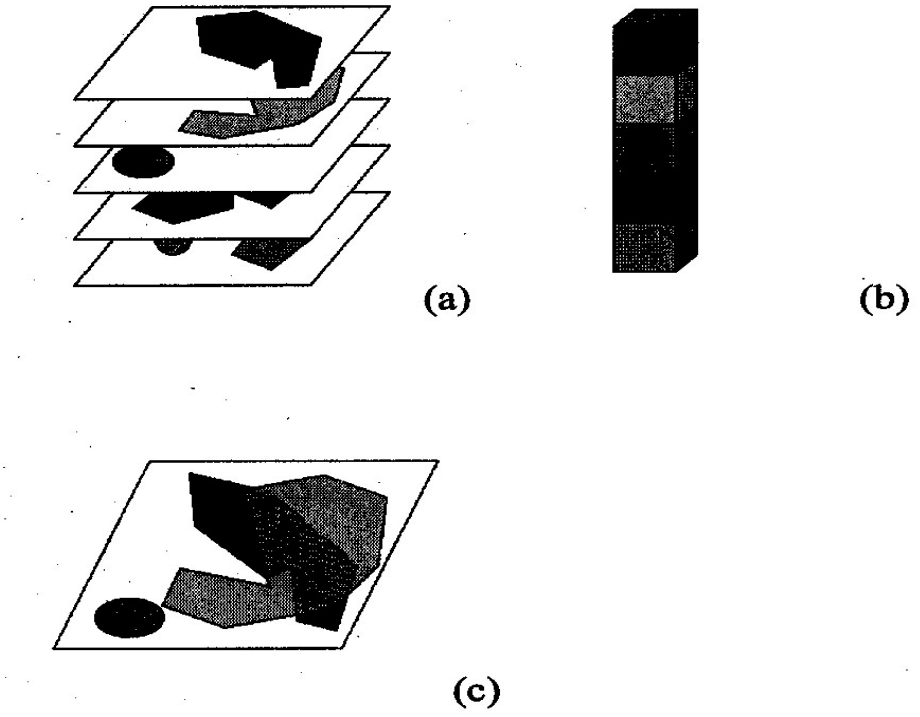
	Aseellinen ryöstö	Murha	Eristynyt alue	Suuri kaupunki	Mies	Nainen
Tapaus 1	1	1	0	1	1	0
Tapaus 2	1	0	1	0	1	0
Tapaus 3	0	1	0	1	0	1
Tapaus 4	0	1	0	1	1	0

Taulukko 1. Esimerkki 0/1-relaatiosta rikostapauksissa

klusteroinnin avulla, koska välttämättä ei löydetä sopivaa aliryhmien määrää. Myös GIS-työkalua hyödynnetään rikosten spatiaalisen jakauman määrittelemiseen. Hyötynä tässä työkalussa on se, että se pystyy yhdistämään spatiaalista tietoa muuhun dataan.

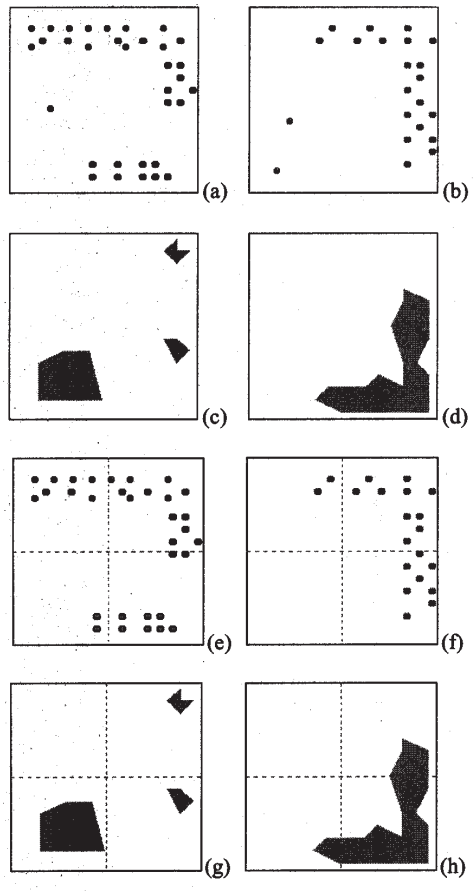
Vladimir Estivill-Castro ja Ickjai Lee [EsL01] määrittelivät kuumat pisteet klusterianalyysin avulla ja yrittivät löytää mahdollisia rikoksiin vaikuttavia tekijöitä assosiaatiosääntöjen louhinnalla. He ehdottivat vertikaalia ja horisontaalia lähestymistapaa yhteyksien löytämiseen.

Tarkastellaan ensin vertikaalia tapausta. Kuvassa 2 on esimerkkinä viisi maantieteellistä tasoa. Nämä tasot on saatu etsimällä spatiaaliset klusterit joltakin tutkittavalta alueelta. Jos edellä mainitulta alueelta valitaan jokin piste, jota halutaan tutkia, saa se viisi eri attribuuttiarvoa, yhden kutakin tasoa kohden. Kuvassa 2 oleva pystysuora palkki esittää valittua pistettä ja sen kunkin tason arvoa. Arvo on 0, jos tasolla ei ole klusteria valitun pisteen kohdalla, muutoin arvo on 1. Tasot jaetaan siis tiettyyn määrään soluja tai pisteitä, joiden avulla pyritään etsimään kiinnostavia assosiaatioita eri tasojen välillä. Esimerkiksi



Kuva 2 [EsL01]. a) Vertikaalisen lähestymistavan tasot, b) Palkki, johon on kuvattu tasojen pisteen arvot (0 tai 1). c) Horisontaalisen lähestymistavan tarkasteltava taso

assosiaatiosääntö *kerros(1)* $\wedge$ *kerros(2)* $\Rightarrow$ *kerros(4) (70 %)* ilmaisee, että 70 prosenttia muodostetuista palkeista (kuva 2) sisältää tasolla neljä attribuuttiarvon 1, jos saman palkin tasoilla yksi ja kaksi on arvo 1. Kun tutkittava alue ja sen tasot on jaettu säännöllisiin soluihin, muodostetaan $m \times n$ kokoinen taulukko. Taulukkoon asetetaan binääriarvot *{0, 1}* sen mukaan, onko kyseinen attribuuttiarvo tosi vai epätosi. Kuvassa 3 on esimerkki siitä, kuinka alueet voidaan jakaa soluihin. Ongelmana on valita sopiva rakeisuustaso, jolla alueita tarkastellaan. Taulukossa 2 on kuvan 3 solujaottelun pohjalta saatu taulukko. Muodostetun 0/1-relaation pohjalta voidaan etsiä assosiaatiosääntöjä, jotka ylittävät annetut minimifrekvenssit ja minimikonfidenssit. Koska tarkastelun kohteena on 0/1-relaatio, on assosiaatiosääntöjen louhintaan käytettävissä perinteisiä tekniikoita [EsL01], kuten Apriori-algoritmi.



Kuva 3 [EsL01]. Tutkittavan alueen jako tasakokoisiin soluihin

	layer(1)	*layer*(2)	*layer*(3)	*layer*(4)
loc(1)	1	1	0	0
loc(2)	1	1	1	1
loc(3)	1	0	1	1
loc(4)	1	1	1	1

Taulukko 2 [EsL01]. Kuvan 3 pohjalta muodostettu taulukko attribuuttien arvoista

Horisontaalinen lähestymistapa kuumien pisteiden analyysissä eroaa vertikaalista siinä, että useamman maantieteellisen tason sijaan tarkastellaan yhtä tasoa. Voidaan siis ajatella, että kuvan 2 viisi tasoa on laitettu yhdelle tasolle, josta sitten etsitään assosiaatioita klustereiden leikkauskohdista. Toinen eroavaisuus vertikaaliseen tapaan on siinä, että aluetta ei jaeta säännöllisiin soluihin vaan sääntöjä etsitään muodostettujen klustereiden avulla. Horisontaalinen lähestymistapa on autonominen ja sen takia se on tehokkaampi valtaville tietokannoille kuin vertikaalinen.

5 Yhteenveto

Spatiaalisen tiedon louhinta etsii spatiaalisesta tietokannasta toistuvia, ymmärrettävissä olevia hahmoja. Spatiaalinen tietokanta sisältää objekteja, jotka on kuvailtu maantieteellisen paikan perusteella. Koska mahdolliset hahmot ovat monimutkaisempia kuin tavallisessa tietokannassa, on spatiaalisen tiedon louhinta hankalampaa. Lisäksi tarkastelussa on otettava huomioon objektin naapurit ja naapuruussuhteet.

Spatiaalisen tiedon louhinnassa ja tiedon hyödyntämisessä rikosten tutkinnassa käytetään usein spatiaalista klusterointia. Tietokannan objektit jaetaan aliryhmiin, joissa kullakin edustajalla on jokin samanlainen piirre, esimerkiksi sijainti jonkin suuren valtatien varrella. Näitä muodostettuja klustereita voidaan analysoida erilaisilla tilastollisilla ohjelmilla, kuten CrimeStat:lla [Lev02]. Saatujen tulosten perusteella pyritään ymmärtämään rikosten yleisyyttä joillakin alueilla.

Spatiaalista klusterointia käytetään myös kuumien pisteiden paikallistamisessa. Kuumien pisteiden analysointi on tärkeää rikostutkijoille, kun tutkitaan onnettomuuksien alueellista keskittyneisyyttä. Tällöin voidaan lisätä poliiseja ja muita kunnallisia resursseja alueille,

joissa on paljon rikollisuutta. Vladimir Estivill-Castro ja Ickjai Lee [EsL01] yrittivät löytää mahdollisia rikoksiin vaikuttavia tekijöitä assosiaatiosääntöjen louhinnalla. He ehdottivat vertikaalia ja horisontaalia lähestymistapaa yhteyksien löytämiseen. Molemmissa tavoissa muodostetaan klustereita, joita sitten tarkastellaan. Vertikaalissa tavassa muodostetaan 0/1-relaatio, johon voidaan käyttää perinteisiä assosiaatiosääntöjen louhintaperiaatteita, esimerkiksi *Apriori-algoritmia* [MaT02].

Spatiaalisen tiedon louhinta on esimerkiksi FBI:lla käytössä rikosten selvittämisessä [EsL01]. Rikostapausten kartoitus ja esittäminen visuaalisesti, esimerkiksi rikoskarttojen avulla, helpottavat alueellisten keskittymien havaitsemista. Karttojen avulla voidaan tehdä jonkinlaisia päätelmiä siitä, millaiset alueet ovat alttiimpia rikoksille ja tarvitsevat suurempaa seurantaa. Spatiaalisen tiedon louhinta ei kuitenkaan ratkaise kaikkia rikoksiin vaikuttavia tekijöitä, mutta se voi auttaa ymmärtämään, millainen ympäristö saattaa olla osatekijänä niiden yleisyyteen.

Lähteet

[Ans99]: Luc Anselin, The Future of Spatial Analysis in the Social Sciences. Geographic Information Sciences 5, 67-76

[Car99]: Carlos Carcach, Spatial Analysis of Crime Data: Firearms Related Homicide in Australia. Paper presented at the 3 rd National Outlook Symposium on Crime in Australia, Mapping the Boundaries of Australia's Criminal Justice System convened by the Australian Institute of Criminology and held in Canberra, 22-23 March 1999

[EKS01]: Martin Ester, Hans-Peter Kriegel, Jörg Sander, Algorithms and Applications for Spatial Data Mining. Published in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis, 2001, 160-187.

[EsL01]: Vladimir Estill-Castro and Ickjai Lee, Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data. 6th International Conference on Geocomputation, 24-26, September, 2001, Brisbane, Australia

[GAA99]: Michael F. Goodchild, Luc Anselin, Richard P. Appelbaum, Barbara Herr Harthorn, Toward Spatially Integrated Social Science.International Regional Science Review 23, 139-159.

[GrM01]: Tony H. Grubesic, Alan T. Murray, Detecting Hot Spots Using Cluster Analysis and GIS. Proceedings from the Fifth Annual International Crime Mapping Research Conference. December 1, 2001, Dallas, TX.

[Lev02]: Ned Levine, CrimeStat A Spatial Statistics Program for the Analysis of Crime Incident Locations (Version 2.0) [Computerfile]. Houston, TX: Ned Levine & Associates/Washington, DC: U.S. Dept. of Justice, National Institute of Justice [producers], 2002. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2002.

[MaT02]: Heikki Mannila, Hannu Toivonen, Knowledge Discovery in Databases: Search for Frequent Patterns. Kurssin Tietämyksen muodostaminen, syksy 2002 kurssimateriaali