

Arkeologisen tietokanta-aineiston luokittelu itseorganisoidulla piirrekartalla

Kai Nikulainen ja Hanna-Maria Pellinen

Arkeologinen luokittelu

Luokittelua voidaan edelleen pitää arkeologian metodisena perustana (Odell 1981, 321). Hyvin usein typologian muodostaminen on kuitenkin perustunut sarjaan intuitiivisia valintoja. Typologiaa on pidetty joukkona työkaluja, joihin ollaan tyytyväisiä niin kauan kun ne toimivat ja antavat halutun lopputuloksen. Tälläkin hetkellä arkeologit käyttävät säännöllisesti satoja eri typologioita. (Adams & Adams 1991:8.)

Arkeologin ensisijainen luokittelukohde on materiaallinen kulttuuri. Luokittelut voivat myös liittyä löytöaineiston perusteella tehtäviin yhteiskunnallisiin tulkintoihin. Tutkijoilla on luonnollisesti erilaisia näemyksiä siitä, kuinka pitkälti luokittelut perustuvat eri kulttuureissa samanlaisille kriteereille ja miltä osin ne ovat kulttuurisidonnaisia.

Perinteinen luokittelu on yleensä ollut intuitiivista (Lavento 2001:16). Luokittelujen pohjana olevia ominaisuuksia on voitu luokitella (laatueroasteikko, järjestysasteikko, välimatka-asteikko, suhdeasteikko ja absoluuttinen asteikko), mutta niiden valinta luokittelun perustaksi on pohjautunut intuition. Valitut attributit kuitenkin määrittelevät typologian. Samat ominaisuudet voivat tuottaa erilaisen luokittelutuloksen silloin kun luokittelun sääntöjä muutetaan. Käsitteet/luokat eivät ole sisällöltään tarkka-

rajaisia. Myöskään mitään yhtenäistä ja muuttumatonta attributtijoukkoa, joka erottaisi käsitteen/luokan muista käsitteistä/luokista ei ole. (Andrefsky 1998:62-64.)

Myös typologioita on luokiteltu, mutta niiden valinta on perustunut parhaimmillaan tietoiseen valintaan siitä, mitä sosiaalista todellisuutta on haluttu kuvata (Pihlman 1990:22-26). Jossakin määrin luokittelun kelpoisuus täytyykin suhteuttaa tutkimusongelmaan. Valintoja joudutaan tekemään myös teoriaan perustuvassa luokittelussa.

Intuitiivinen luokittelu ei sinänsä välttämättä ole huonompi kuin muut menetelmät; onhan sillä melko vahvat perusteet havaintopsykologiassa. Ihmisellä on mitä ilmeisimmin kyky erottaa toisistaan olennainen ja epäolennainen. Niin objektiiviselta kuin pelkkiin mittauksiin perustuva luokittelu voikin vaikuttaa, jättää se ainakin nominaalimuuttujat kuten värin ja kivilajin kokonaan pois. Tällainen menettely on tietysti ymmärrettävää siinä tapauksessa, että kyseisissä ominaisuuksissa ei esiinny hajontaa.

Tilastolliset menetelmät puoltavat paikkaansa yhtenä tapana tutkia aineiston välisiä suhteita. Tässä ei puututa niihin sen enempää. Ne ovat perustaltaan erilaisia ja toimivat eri periaatteilla kuin hermoverkko-ohjelmat. Koodaamisen ongelmat sen sijaan ovat samat kaikessa numeerisessa aineistossa. Mika Lavento (2000:45) on tiivistänyt ne kolmeen pääongelmaan: 1) koodaus heijastaa

vain osaa olemassa olevasta tiedosta, 2) havaitut ominaisuudet eivät ole aina verrattavissa keskenään ja 3) numeerinen luokittelu voi luoda illuusion objektiivisuudesta, prosessiinhan liittyy samalla väistämättä kuitenkin subjektiivista valintaa.

Subjektiivisuuden ongelma on olemassa myös nyt esiteltävällä ohjelmalla muodostetuissa luokitteluissa. Toivomme kuitenkin, että rajoittamalla valintaa aineiston muodostamiseen, attribuuttien valintaan ja luokkien tulkintaan, voimme tällä ArcSOMiksi nimetyllä ohjelmalla eliminoida pois yhden subjektiivisen valintavaiheen luokitteluprosessista, helpottaa laajojen aineistojen käsittelyä ja mahdollistaa uudentyyppisten luokittelujen tuottamisen.

Itseorganisoituvat piirrekartat

Yleistä hermoverkoista

Itseorganisoituva piirrekartta on eräs ns. hermoverkkojen alatyyppejä. Hermoverkolla (neuroverkolla) tarkoitetaan tietokoneohjelmaa, joka pyrkii jäljittelemään hermoston ja erityisesti aivojen hermosolujen toimintaa. Hermoverkot soveltuvat hyvin ongelmiin, joita on perinteisesti pidetty vaikeina ratkaista tietokoneella, mutta jotka sujuvat ihmiseltä lähes automaattisesti. Tällaisia ongelmia ovat esimerkiksi hahmontunnistus ja puheen ymmärtäminen.

Toisin kuin perinteisissä ohjelmissa, hermoverkon toimintalogiikka ei ole ennalta määriteltä, vaan hermoverkko oppii sille annettavien esimerkki-vastaus -parien perusteella. Opetus tapahtuu antamalla verkolle esimerkki ja vertaamalla haluttua (ja oikeaa) vastausta verkon antamaan tulokseen. Tätä eroa käytetään verkon sisäisen rakenteen muokkaamiseen siten, että seuraavalla kerralla tulos on lähempänä haluttua tulosta. Opetusesimerkkejä voi olla satoja tai tuhansia ja jokaista joudutaan mahdollisesti opettamaan verkolle tuhansia kertoja, joten

opetusvaihe voi kestää hyvinkin kauan.

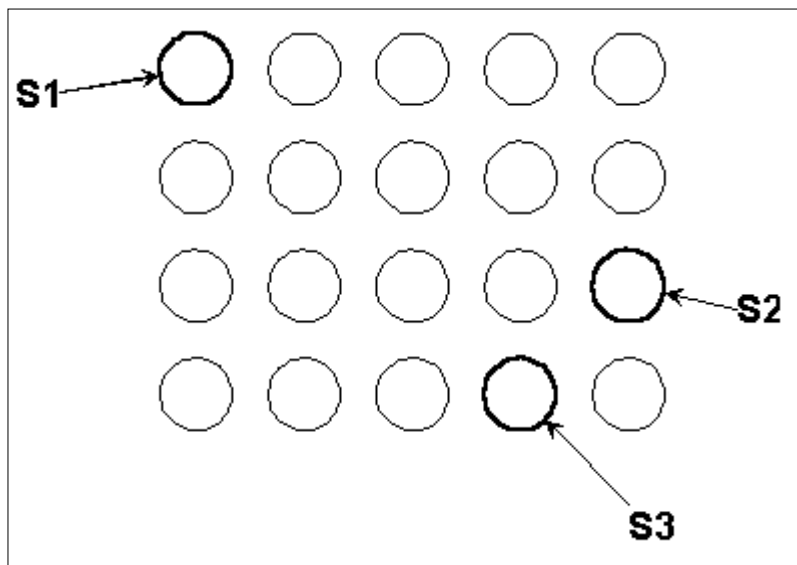
Yrityksiä soveltaa opetettavia hermoverkko-ohjelmia on tehty vaihtelevalla menestyksellä myös arkeologisen tutkimuksen piirissä. Uusin julkaisu lienee Bellin & Jantzin tutkimus, jossa he yrittävät opettaa hermoverkko-ohjelmaa tunnistamaan luurangon osia. Tulokset eivät kaikilta osin vaikuta lupaavilta. Parhaimmillaan ohjelman tuottamat ennusteet ovat tilastollisesti merkitseviä, mutta eivät kovin selkeästi. Tekijät itse kuitenkin käyttävät tuloksia päätelmien tekemiseen. (Bell & Janz 2002:205-211.)

Lisää tietoa hermoverkoista löytyy mm. internet-osoitteesta <http://ftp.sas.com/pub/neural/FAQ.html>

Itseorganisoituva piirrekartta

Itseorganisoituva piirrekartta (self organizing map, SOM) eroaa muista hermoverkoista siten, että opetuksessa ei ole tiedossa tiettyä esimerkkiä vastaavaa oikeaa vastausta, vaan piirrekartan on opetuksen aikana pyrittävä valvomatta löytämään opetusaineistosta samankaltaisuuksia. Piirrekarttojen keksijä on suomalainen professori, akateemikko Teuvo Kohonen, joka julkaisi ajatuksensa SOM:in periaatteista vuonna 1982 ilmestyneessä artikkelissa "Self-Organized Formation of Topologically Correct Feature Maps". Tässä artikkelissa käsitelty asiat löytyvät Kohosen kirjasta "Self-Organizing Maps" (2001). Tietääksemme SOM-ohjelmia ei ole aikaisemmin yritetty soveltaa arkeologisen aineiston käsittelyyn.

Piirrekartta koostuu joukosta solmuja, jotka yleensä muodostavat tason. Kartalle annetaan syötteitä, joista jokainen aiheuttaa yhden solmun aktivoitumisen. Karttaa pyritään opettamaan siten, että toisiaan muistuttavien syötteiden aktivoimat solut ovat lähellä toisiaan ja erilaisten syötteiden aktivoimat solut ovat kaukana toisistaan. (Kohonen 2001:106-109.) Esimerkkikuvassa 1 syötteiden S2 ja S3 pitäisi olla hyvin



Kuva 1. Piirrekartta.

samankaltaisia ja molempien pitäisi erota melkoisesti syötteestä S1.

Toiminnaltaan piirrekartan on tarkoitus muistuttaa aivojen pintakerrosta, jossa samankaltaiset ärsykkeet aiheuttavat reaktion läheisillä alueilla. Tietueet voidaan tulkita pisteiksi moniulotteisessa avaruudessa. Ulotuvuuksia on yhtä paljon kuin muuttujia. Käytännössä piirrekartta on muunnos tästä moniulotteisesta avaruudesta kaksiulotteiseksi tasoksi tai yksiulotteiseksi janaksi. Yleensä kuvauksessa käytetään kaksiulotteista tasoa, koska siinä muodossa informaatiota säilyy enemmän, mutta sen lukeminen on kuitenkin helpompaa kuin esim. kolmiulotteisen avaruuden. (Kohonen 2001:99-105.)

Muuttujien tulkinnasta

Piirrekartta tulkitsee kaikkien käytettyjen muuttujien olevan suhdeasteikolla. Tämä voi aiheuttaa ongelmia käytettäessä laatuero- eli nominaaliasteikon muuttujia, joita arkeologisessa aineistossa esiintyy. Usein tietokannoissa nominaaliasteikon muuttuja on koodattu kokonaisluvuiksi, esimerkiksi 1 = punainen, 2 = harmaa, 3 = vihreä jne. Koska

piirrekartta tulkitsee kaikki muuttujat suhdeasteikolla, tämän seurauksena punainen ja vihreä sijoitetaan kartalle kauemmas toisistaan kuin punainen ja harmaa, vaikka pelkästään värien erolla ei olisikaan tällaista merkitystä. Tämän vuoksi ohjelman syötteessä on käytettävä luokkien nimiä, jotta ohjelma osaa koodata ne tavalla, joka ei anna arvoille ylimääräistä merkitystä. Ohjelma muuntaa luokkien nimet koodeiksi, esim. punainen = 1, 0, 0, harmaa = 0, 1, 0 jne. Samalla yhdestä muuttujasta tulee niin monta muuttujaa kuin on luokkiakin, tässä tapauksessa kolme, ja luokkien väliset erot muuttuvat yhtä suuriksi.

Koska ohjelman toiminta perustuu suurelta osin pisteiden geometrisen välimatkan laskemiseen, suuren arvoalueen omaavien muuttujien merkitys saattaa korostua (välillä 0-100 vaihtelevan muuttujan merkitys välimatkaan on suurempi kuin välillä 0-2 vaihtelevan). Tämä ei ole haluttu vaikutus, koska piirrekartta luokittelee aineiston periaatteessa polyteettisesti. Vaikutusten eroa voidaan pienentää siten, että muunnetaan muuttujien arvoalue välille 0-1 (Kohonen 2001:115). Normalisoinnilla on kuitenkin lievä kaksiarvoisia (0 ja 1, kyllä ja ei ym.)

luokkia korostava vaikutus, koska dikotomisilla muuttujilla mahdollisten arvojen (0 ja 1) välimatka on aina suurin mahdollinen.

ArcSOM

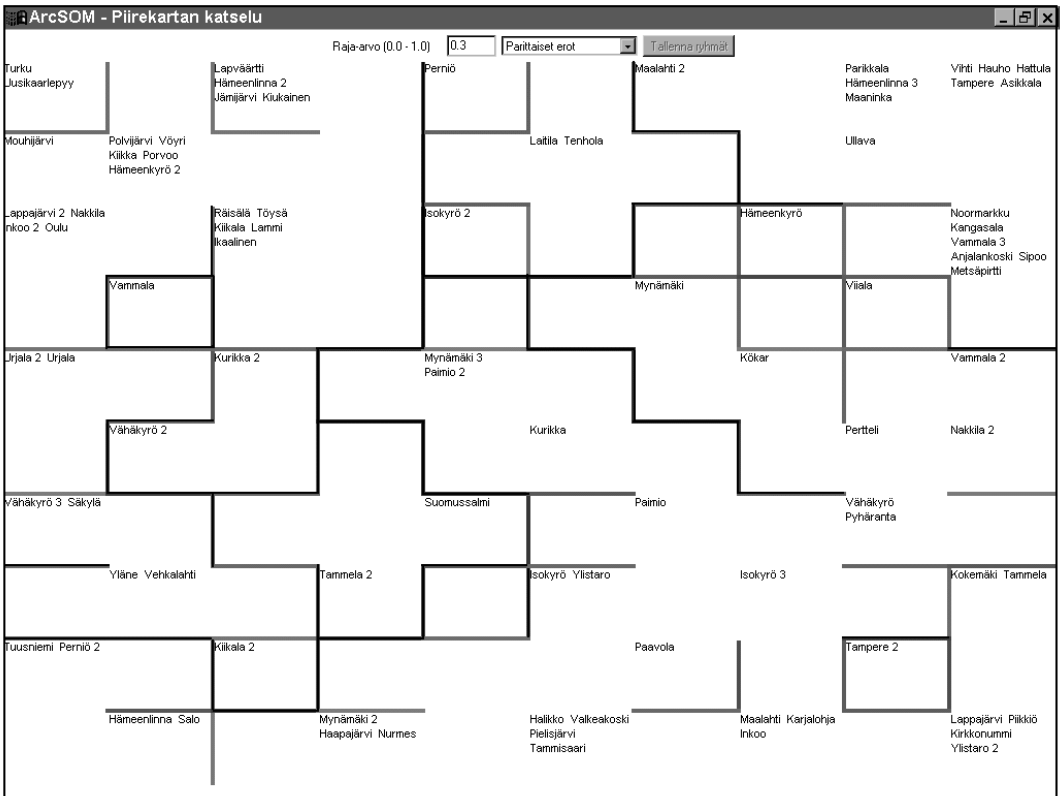
Ohjelma on kirjoitettu Java-kielellä ja se toimii kaikissa koneissa, joihin on asennettu Java versio 1.1:tä tukeva virtuaalikone (= ohjelma, joka suorittaa Java-ohjelmaa). Windows-koneisiin sellainen on yleensä asennettu Internet Explorerin ohessa. Jos koneesta ei kuitenkaan Java-virtuaalikonetta löydy, sellaisen voi ladata ilmaiseksi osoitteesta <http://java.sun.com>.

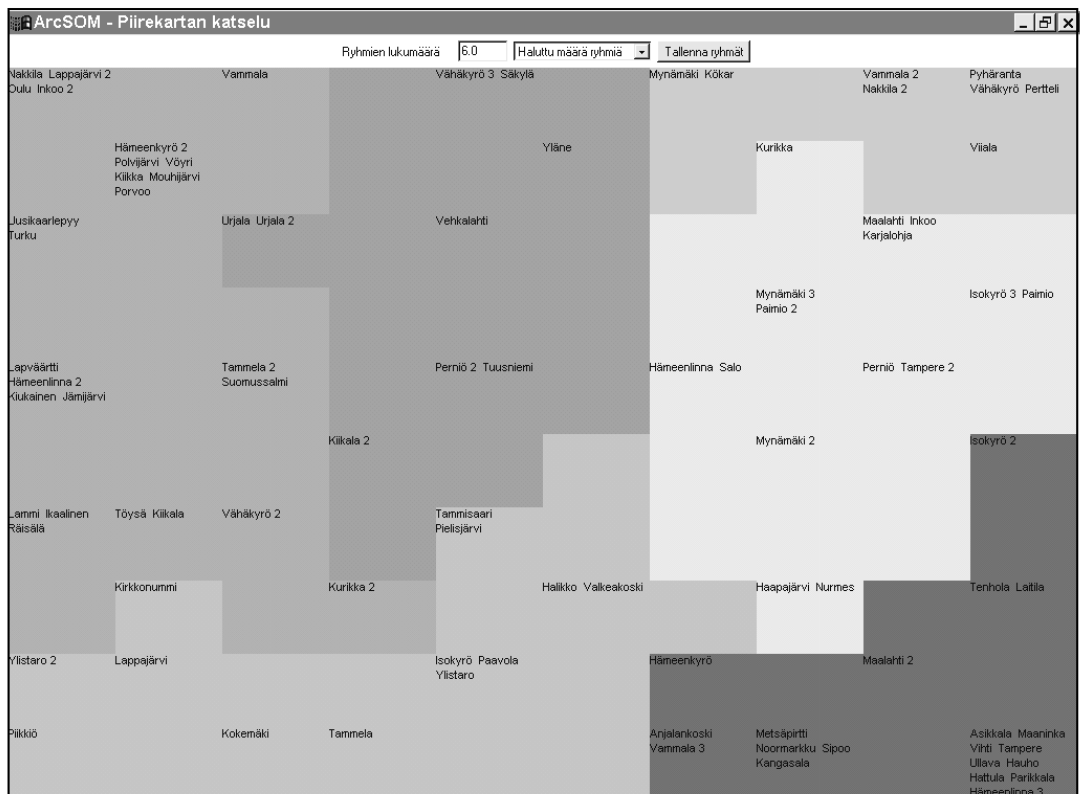
Syötteeksi ohjelmalle annetaan tekstitiedostoja, joissa jokainen rivi tulkitaan omaksi tietueeksi. Kunkin rivin ensimmäinen

alkio on tietueen tunniste, loput muuttujien arvoja. Rivin alkiot erotellaan toisistaan pilkuilla ja jokaisella rivillä on oltava yhtä monta alkioita. Jos rivi alkaa puolipisteellä, se tulkitaan kommenttiriviksi ja jätetään huomiotta. Luokkaeromuuttujien mahdolliset arvot on esiteltävä tiedoston alussa. Esittelyrivillä on sana luokka, luokan nimi ja arvot lueteltuna pilkuilla sulkujen sisällä, esim:
luokka värit (punainen, vihreä, sininen)

ArcSOM koostuu kahdesta eri modulista; toisella luodaan ja talletetaan piirrekarttoja ja toisella tutkitaan talletettuja karttoja ja voidaan tallettaa erilaisia ryhmittelyjä. Luontimodulissa määrätään piirrekartan koko; soluja tulisi olla vähintään yhtä monta kuin

Kuva 2. Piirrekartta, joka kuvaa solujen välisiä eroja tuluskiviaineistossa.





Kuva 3. Samankaltaisina alueina esitetty piirrekartta, jossa tuluskivet on jaettu kuudeksi ryhmäksi.

esimerkkejäkin, mielellään jonkin verran enemmän. Lisäksi voidaan määrätä muuttujat normalisoitaviksi. Tarkastelumodulissa ryhmittelyjen apuna käytetään suhteellisen eron raja-arvoa, jossa 0 tarkoittaa samanlaisia esimerkkejä ja 1 tarkoittaa aineistosta löytyvää suurinta mahdollista eroa kahden opetusesimerkin välillä. Ryhmittely voidaan tehdä kolmella eri tavalla:

1. Parittaiset erot piirtää solujen väliin sitä tummemman viivan, mitä suurempi ero solujen välillä on. Jos solujen ero on pienempi kuin annettu raja-arvo, ei viivaa piirretä lainkaan (kuva 2).

2. Samankaltaiset alueet luo ryhmiä, joiden sisällä solujen välinen ero on korkeintaan annettu raja-arvo. Koska ryhmät voidaan yleensä muodostaa usealla eri tavalla, syöttämällä sama arvo useamman kerran peräkkäin voidaan saada hieman erilaisia ryhmittelyjä. Ohjelma pyrkii

vähentämään ryhmien lukumäärää ja muistaa kullekin raja-arvolle aiemmin saadun vähi-ten ryhmiä antavan luokittelun (kuva 3).

3. Haluttu määrä ryhmiä yhdistää toistuvasti kaksi toisiaan lähimpänä olevaa vierekkäistä ryhmää kunnes ryhmiä on haluttu määrä.

Ohjelman luomaa ryhmittelyä on pidettävä vain viitteellisenä tuloksinna tukena, ryhmien tarkkoja rajoja merkittävää on opetusmerkkien sijoittuminen kartalla suhteessa toisiinsa. Vaikka ihminen haluaa usein määritellä luokille selvät rajat, ovat rajat kuitenkin aina muuttuvia.

Piirrekartan luominen voi kestää jonkin aikaa, koska suoritettavien laskutoimitusten lukumäärä on varsin

suuri. Esimerkiksi 16*16 -soluisen piirrekartan luomiseksi 8-muuttujaisesta aineistosta ohjelma laskee noin 600 miljoonaa laskutoimitusta.

Aineisto ja testaus

Arkeologisessa aineistossa on tiettyjä erityispiirteitä, jotka oli otettava huomioon ohjelmaa tehtäessä. Kuten edellä on kuvattu, ohjelman suorittamista varten muuttujat piti siis yhdenmukaistaa siten, että ohjelma antaa samat painoarvot erilaisille attribuuteille. Periaatteessa ohjelmaan voisi lisätä ominaisuuden, jolla tietyn attribuutin painoarvoa lisättäisiin. Tällä tavalla olisi mahdollista havainnoida yksittäisen attribuutin vaikutusta luokitteluun.

Koetestauksissa havaittiin, että aina ei ole järkevää esittää mittalukuja vaan informatiivisempaa on muuttaa ne suoraan mittasuhteista kertoviksi luvuiksi, esim. esineen pituuden suhde leveyteen, jolloin esineen muoto tulee paremmin esille. Luokkien nimeämisessä ei missään tapauksessa kannata käyttää monimutkaisia kirjain- tai numerokoodeja, kuten KM-numeroita, koska tuloksen hahmottaminen ja analysoiminen piirrekartalta tulee tuolloin todella vaikeaksi.

Testauksessa on toistaiseksi käytetty yhtä esineaineistoa, soikeita tuluskiviä, koska sellainen oli valmiina käytettävissä tietokantamuodossa. Tuluskiviaineisto perustuu Museoviraston ja eräiden muiden museoiden esineluetteloiden kuvauksiin. Aineiston rajausta on tehty kuvausten tarkkuuden perusteella ja vain tietyt kriteerit omaavat kuvaukset on otettu mukaan analyysiin. (Enemmän tuluskivistä ks. Pellinen 1999).

Aineistossa on mukana 91 tuluskiveä. Mukaan valittujen muuttujien määrää rajoitti huomattavasti työskentely sekundaarilähteiden kautta. Aikoinaan löytöluetteloista kopioidut kuvaukset olivat osin tarkkoja, mutta mitallisia muuttujia oli pääsääntöisesti vain kolmesta tekijästä; tuluskiven pituudesta,

paksuimman kohdan paksuudesta ja leveimmän kohdan leveydestä. Mukaan otettiin näiden mittojen tai niiden suhdelukujen lisäksi kärjen muoto, koska se ei näyttänyt olevan suorassa yhteydessä esineen leveyteen. Se määriteltiin piirretyn kuvan perusteella (tylppä, pyöreä, soikea tai terävä). Muita muuttujia ovat sivu-urien olemassaolo (1 tai 0), laakaurat (E = ei ole, T = toisella pinnalla, M = molemmilla pinnoilla) ja poikkileikkaus (tasapaksu, koverankupera, puolikupera, kupera).

Mitallisia muuttujia olisi voinut ehkä ottaa mukaan enemmänkin, mm. paksuuden ja leveyden osalta kolme eri mittausta. Pois on jätetty myös väri ja kivilaji. Edellinen sen vuoksi, että todellista vertailua ei voinut tehdä sanallisen kuvauksen perusteella, joka ei perustunut mihinkään sovittuihin värikoodeihin, jälkimmäinen osin tietojen virheellisyyden vuoksi ja osin siksi, että tuluskivien valmistusaineena on käytetty lähes pelkästään kvartsiittia, eikä kivilaji muuttujana sen vuoksi toisi mukaan juuri ollenkaan variaatiota.

Luokittelun tavoitteena ei siis ole luoda uutta tuluskivitypologiaa, vaan kehittää ja testata ohjelmaa.

Analyyysi

Ohjelma suoritettiin tuluskiviaineistolla kuudella eri raja-arvolla (0.2, 0.3, 0.4, 0.5, 0.6 ja 0.65), joista jokainen raja-arvo testattiin useaan kertaan. Ajoja tehtiin myös määrittämällä tietty määrä luokkia, 6 kpl, eli määrä johon intuitiivisessa luokittelussa olin aikoinani päätenyt. Näin vertailu oli helpompaa.

Lyhyesti voidaan tässä todeta tuloksista sen verran, että tiettyjen luokkien sijainti ja koko vaikuttivat melko stabiileilta. Toinen silmiinpistävä seikka oli joidenkin tuluskivien erottuminen yksittäisinä tai pienissä ryhmissä esiintyvänä tapauksina toistuvasti eri ajoissa. Tällaisia artefakteja voi varmasti pitää

poikkeavina suhteessa muuhun aineistoon. Vaikutti myös siltä, että ryhmissä oli yleensä yksi muuttuja, joka oli lähes kaikilla ryhmän jäsenillä sama. Muissa muuttujissa oli enemmän variaatiota. Tämä yhtenäinen muuttuja saattoi olla eri attribuutti eri ajoissa. Tämä voi johtua toisaalta koko aineiston homogeenisyydestä, toisaalta ehkä siitä, että kun ohjelma piirtää koordinaatistoon aina jonkin "tason", voi tämä sijaita eri ajoissa ei kohdalla.

Selvää on, että luokat ja erityisesti luokkien rajat poikkeavat intuitiivisella menetelmällä tehdystä luokittelusta. Intuitiivinen menetelmä perustuu epäolennaisten ja olennaisten muuttujien erottamiseen. Tietokone-ohjelma ei arvota muuttujia ja se käyttää erotteluun matemaattista kaavaa eli laskee pisteiden etäisyyksiä. Periaatteessa ohjelman voisi opettaa myös luokittelemaan halutulla tavalla. Vanhaan piirrekartalla luokiteltuun aineistoon voi lisätä myöhemmin uuden aineiston "pälle" ja katsoa miten se sijoittuu suhteessa aikaisempaan tulokseen. Ohjelman hyvänä puolena voi pitää sen mahdollisuutta käsitellä hyvinkin suuria aineistoja ja laatia niistä pieniäkin ryhmiä - tällaisen luokittelun hallinta manuaalisesti olisi erittäin työlästä.

Nyt esitelty ohjelma on vasta kehitysvaiheessa. Se on kuitenkin julkaistu vapaaseen levitykseen osoitteessa <http://www.sci.fi/~kajun/arsom/>. Tästä osoitteesta löytyy myös joitakin helppotajuisia valmiita aineistoja ohjelman testausta varten sekä analyysissa käytetty tuluskiviaineisto museonumeroineen. Allekirjoittaneet ottavat mielellään palautetta vastaan alla oleviin sähköpostiosoitteisiin.

Kai Nikulainen
kai.nikulainen@sci.fi
Informaatiotekniikan laitos
20014 Turun yliopisto

Hanna-Maria Pellinen
hanna-maria.pellinen@archaeologist.com
Arkeologia
20014 Turun yliopisto

Tekstissä esiintyvien termien selityksiä:

Tietokantamuoto = yksi luokittelukohde (objekti) tietoineen

Tietue = järjestetty muuttujakokoelma

Laatueroasteikko = kaksijakoinen on/ei, +/- luokittelu

Järjestysasteikko = kuvaa muuttujien välisiä suhteita, $A > B$, $B > C$ jne.

Välimatka-asteikko = numeraalinen, ei kiinteää 0-kohtaa, esim. vuosiluvut

Suhdeasteikko = myös kiinteä 0-kohta

Absoluuttinen asteikko = mittayksikkö ja kiinteä 0-kohta

Monoteettinen luokittelu = yksi attribuutti tunnistetaan kerrallaan

Polyteettinen luokittelu = mikään yksittäinen attribuutti ei ole tärkeämpi kuin muut missään vaiheessa

Ohjelman kohta "haluttu määrä ryhmiä" kuuluu ns. kasautuvaan (agglomerative) luokitteluun = tunnistetaan jokainen yksilö itsenäan tyyppiin kuuluvaksi, näitä yksilöitä yhdistetään samanlaisten kanssa ja muodostetaan yhä suurempia ryhmiä. Tämän vastakohta on jakautuva (divisive) luokittelu, jossa tunnistetaan ensin jokin ryhmä ja tehdään sen sisällä vähittäistä jakoa.

Lähteet

- Adams, W.Y. & Adams, E.W. 1991: *Archaeological Typology and Practical Reality*. A dialectical approach to artifact classification and sorting. Cambridge University Press.
- Andrefsky, W. 1998: *Lithics*. Macroscopic approaches to analysis. Cambridge University Press.
- Bell, S. & Jantz, R. 2002: Neural Network Classification of Skeletal Remains. *Archaeological Informatics: Pushing the Envelope. Computer Applications and Quantitative Methods in Archaeology 2001. BAR International Series 1016*.
- Kohonen, T. 1982: Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*. s. 59-69.

- Kohonen, T. 2001: Self-Organizing Maps. Third edition. *Springer Series in Information Sciences*, 30.
- Lavento, M. 2001: Textile Ceramics in Finland and on the Carelian Isthmus. Nine Variations and Fugue on a Theme of C.F. Meinander. *Suomen Muinaismuistoyhdistyksen Aikakauskirja* 109.
- Pellinen, H.-M. 1999: Soikeiden tuluskivien konteksti. *Muinaistutkija* 1/1999.
- Pihlman, S. 1991: Kansainvaellus- ja varhaismerovinkiajan aseet Suomessa. Typologia, kronologia ja aseet ryhmästrategioissa. *Iskos* 10.