Maantieteellisen epämääräisyyden mallintamisesta tietokannoissa

Oulun yliopisto
Luonnontieteellinen tiedekunta
Tietojenkäsittelytieteiden laitos
Kandidaatin tutkielma
Lauri Imppola
20.12.2017

Tiivistelmä

Maantieteellisten tietokantojen ja GIS:in kehittyessä ja uusien kehityssuuntien avautuessa erääksi merkittäväksi tekijäksi on noussut epämääräisyyden hallinta. Maantieteellisen todellisuuden eräs perustavanlaatuinen ominaisuus on epämääräisyys, joka ilmenee useilla eri tavoilla maantieteellistä todellisuutta mallinnettaessa maantieteellisissä tietokannoissa ja GIS:issä. Kyetäkseen mallintamaan todellisuutta tarkasti ja oikeellisesti, GIS-sovellutusten on otettava tämä epämääräisyys huomioon. Epämääräisyyden mallintamisella ja hallinnalla on merkitystä etenkin näihin järjestelmiin pohjautuvan päätöksenteon kannalta.

Tämän kandidaatin tutkielman tavoitteena oli kartoittaa tutkimuskenttää, joka tarkastelee maantieteellistä epämääräisyyttä, sekä pyrkiä löytämään niitä ratkaisuja epämääräisyyden mallintamiseen, joita tuo tutkimuskenttä on tuottanut. Tutkimusmenetelmänä oli kirjallisuuskatsaus.

Tutkimuksen tuloksena löydettiin joukko erilaisia menetelmiä vaihtelevin sovellutusaluein ja lähtökohdin. Menetelmät jakautuivat karkeasti neljään pääjoukkoon: sumeiden joukkojen teoriaan pohjaaviin, karkeiden joukkojen teoriaan pohjaaviin, laajennettuihin malleihin pohjaaviin sekä todennäköisyysmalleihin pohjaaviin. Tutkimuksen perusteella tutkimuskenttä ei ole yksimielinen siitä, mikä menetelmä on tutkimuksen tulevaisuuden kannalta olennainen, ja kysymys vaikutti olevan vielä hyvin auki. Myöskään mitään laajempaa käyttöä näkeviä sovellutuksia tai standardeja ei tutkimuksen perusteella löydetty.

Avainsanat
Maantieteelliset tietokannat, spatiaaliset tietokannat, GIS, epämääräisyys, epätarkkuus, sumeus

Ohjaaja
Yliopisto-opettaja, Jouni Lappalainen

Sisällysluettelo

1. Johdanto

GIS (*Geographic Information Systems*) ja erilaiset karttasovellukset ovat olleet viime vuosien kasvava ja kehittyvä trendi. Niitä ei käytetä enää pelkästään karttojen luomiseen ja erilaiseen ammattilaistyöhön, kuten kaavoittamiseen, maankäytön suunnitteluun tai kartoittamiseen, vaan erilaisia karttasovelluksia on saatavilla tavallisten kuluttajien käyttöön esimerkiksi autonavigaattoreissa ja älypuhelimissa. Tasaisesta kehityksestä huolimatta eräs GIS:in ja maantieteellisten tietokantojen pysyvä ongelma on liittynyt maantieteellisen epämääräisyyden hallintaan. Epämääräisyys on olennainen maantieteellisten objektien ominaisuus (Goodchild & Zhang, 2002, p. 6). Epämääräisyys on myös hyvin olennainen kysymys maantieteellisten tietokantojen hyödyntämisessä, sillä ne toimivat usein erilaisen päätöksenteon tukena. Mitä vähemmän maantieteelliset tietokannat kykenevät mallintamaan todellisuuden epämääräisyyttä, sitä huonompia välineitä ne ovat päätöksenteon tukena. Maantieteellisiä objekteja käsitellään kuin ne olisivat tarkkarajaisia, vaikka usein todellisuus on toinen. Toistaiseksi GIS-sovellukset ja maantieteelliset tietokannat ovat tyytyneet kuvaamaan maantieteellistä todellisuutta tällaisten selvärajaisten objektien kautta (Pauly & Schneider, 2010). Koska mallinnettava maantieteellinen todellisuus sisältää tätä väistämätöntä epämääräisyyttä, minkä tahansa pyrkimyksen mallintaa sitä tulisi myös kyetä sisällyttämään ja hallitsemaan tätä epämääräisyyden elementtiä (Beaubouef & Petry, 2010).

Epämääräisyyden ongelma itsessään maantieteellisten tietokantojen parissa on jo hyvin aikaisin havaittu ongelma. Goodchild ja Gopal (1989) tiivistivät GIS:in ja maantieteellisten tietokantojen epämääräisyyden ongelman suhteessa tyypilliseen kartanpiirtämiseen kirjansa esipuheessa seuraavasti:

1. GIS:in prosessoinnin täsmällisyys (*precision*, skaala) on käytännössä rajaton.
2. Kaiken spatiaalisen datan on tarkkuus (*accuracy*, datan oikeellisuus) on rajallinen.
3. GIS:in täsmällisyys ylittää datan tarkkuuden.
4. Tyypillisessä kartta-analyysissä täsmällisyys sovitetaan yleensä tarkkuuteen.
5. Mahdollisuus muuttaa skaalaa ja yhdistää dataa eri lähteistä ja skaaloista GIS:issä tarkoittaa, että täsmällisyyttä ei yleensä soviteta tarkkuuteen.
6. Ei ole käytössä riittäviä keinoja kuvailla kompleksisten spatiaalisten objektien tarkkuutta.
7. Epämääräisyyden (*uncertainty*) mitta tulisi olla tavoitteena jokaisessa GIS-tuotteessa.

Tärkeimpiä huomioita tässä ovat kaiken spatiaalisen datan rajallinen tarkkuus, yhtäältä mallinnettavan todellisuuden kompleksisuutena ja toisaalta datan keräämisen ja tallettamisen rajallisuutena, sekä GIS:in kyky muuttaa skaalaa ja yhdistää erilaista dataa useista lähteistä. Täsmällisten, selvärajaisten objektien käsittely on laajalti todettu riittämättömäksi, tarve epämääräisyyttä implementoivalle tietokantajärjestelmälle on todettu tärkeäksi ongelmaksi, ja on havaittu, etteivät nykyiset GIS- ja spatiaalisten tietokantojen ratkaisut ole toistaiseksi pystyneet vakuuttavasti vastaamaan tähän maantieteellisen epämääräisyyden käsittelyn ongelmaan (Beaubouef & Petry, 2010; Carniel, et.al., 2014; Pauly & Schneider, 2008).

2. Tutkimuskysymykset ja tutkimusmenetelmä

Tämä tutkielma käsittelee maantieteellistä epämääräisyyttä ja sen hallintaa GIS:in ja tietokantojen kontekstissa. Se pyrkii hahmottamaan maantieteellisen epämääräisyyden käsitettä ja olemusta, sen ilmenemistä tietokantojen yhteydessä, sekä erilaisia menetelmiä tämän epämääräisyyden hallintaan. Tarkemmin tutkielma yrittää vastata seuraaviin tutkimuskysymyksiin:

Miksi epämääräisyys on olennaista maantieteen kontekstissa?
Miten epämääräisyyden ongelma ilmenee maantieteellisissä tietokannoissa?

Ja olennaisimpana päätutkimuskysymyksenä:

Millä keinoin tätä epämääräisyyttä pyritään hallitsemaan ja mallintamaan?

Pääpaino tutkielmassa on maantieteellisen epämääräisyyden olemuksessa sekä niiden datamallien tutkailussa, jotka pyrkivät tuota epämääräisyyttä hallitsemaan. Tutkielmassa ei paneuduta tietokantojen tarkkaan rakenteeseen, sillä näkökulma on yleispiirteinen ja tutkielma luonteeltaan kokoava sekä vertaileva. Myöskään tietokantakyseilyiden tarkkaan toteutukseen ei tässä tutkielmassa paneuduta. Maantieteellisen epämääräisyyden suhteen pysyttäydytään sen käytännönläheisissä ongelmissa, eikä niinkään poiketa filosofiaan epämääräisyyden olemuksen taustalla. Tutkielman kannalta olennaiset käsitteet määritellään tuonnempana tutkielmassa.

Tutkimusmenetelmänä on kirjallisuuskatsaus. Tutkielman tarkoituksena on kartoittaa ja koota olemassaolevaa teoriaa ja käytäntöä maantieteellisen epämääräisyyden hallintaan tietokannoissa, muodostaa tästä tutkimuksen kentästä kokonaiskuva, tutkailla ja vertailla erilaisia menetelmiä, sekä havaita puutteita ja jatkotutkimuksen kohteita nykyisen tutkimuksen kentällä. Tutkielma kirjoitusprosessi on aloitettu hakemalla lähdekirjallisuutta pääkäsitteistä koskien tietokantoja, GIS:iä ja maantieteellistä epämääräisyyttä. Tämän jälkeen lähdehaku on keskittynyt tieteellisiin artikkeleihin, jotka käsittelevät maantieteellisen epämääräisyyden ongelmaan vastaamisen tutkimuskenttää ja ratkaisuja. Artikkeleiden etsimiseen käytetyt tietokannat ovat tieteellisistä tietokannoista Scopus, Web of Science, IEEE Xplore, sekä ACM Digital Library. Kaksi ensimmäistä on valittu niiden laajuuden vuoksi, kaksi jälkimmäistä tarkemman aihealueen vuoksi. Lisäksi artikkeleita on etsitty Google Scholar:in avulla. Hakuprosessin aikana Scopus ja Google Scholar muodostuivat merkittävimmiksi kanaviksi relevanttien artikkeleiden löytämiseen. Artikkeleita etsittiin maantieteen ja tietokantojen osalta käyttäen esimerkiksi hakusanoja ”geographical database”, ”spatial database”, ”GIS”, ”Geographical Information Systems”, ”DBMS”, ja epämääräisyyden osalta esimerkiksi hakusanoja ”uncertainty”, ”impercision”, ”fuzziness”, ”vagueness”, niin että erilaisia hakusanayhdistelmiä ja -variantteja on kokeiltu laajempien ja oikeellisempien tulosten saavuttamiseksi. Kun yleiskäsitys tutkittavasta aihepiiristä muodostui, hakua voitiin täsmentää koskemaan esimerkiksi tiettyjä maantieteellisen epämääräisyyden hallinnan menetelmiä. Lisäksi sopivia artikkeleita etsittiin jo löydettyjen artikkeleiden lähdeviitteiden avulla. Sopivia artikkeleita on haarukoitu ensin otsikon ja sittemmin tiivistelmien perusteella tarkastellen, vaikuttavatko ne käsittelevän tutkimuskysymyksen kannalta olennaisia asioita. Haussa on vältetty myös liian vanhoja artikkeleita ja keskitytty 2000-luvulle, jotta kuva muodostuisi juuri tutkimuksen

nykytilasta, eikä historiallisesta kehitysprosessista. Poikkeuksia ovat tietenkin teorian etenkin matemaattinen perusta ja muut käsitteet, joiden alkuperä ei välttämättä ole 2000-luvulla.

3. Aiempi tutkimus

Seuraavassa kappaleessa määritellään tutkielman kannalta olennaiset käsitteet sekä esitellään aiheeseen liittyvää aiempaa tutkimusta ja teoriaa. Maantieteellisen epämääräisyyden mallintamisen ja esittämisen ongelma ei luonnollisestikaan ole mikään irrallinen osa-alueensa, vaan se liittyy hyvin vahvasti datan epämääräisyyteen ja rajallisuuteen sinänsä sekä sen esittämiseen rajallisissa tietokannoissa, ja ilmenee näiden ongelmien erityistapauksena omine erityispiirteineen. Näin ollen on ilmeistä, että tarkasteltava ongelma ja sen tutkimuskenttä hyötyy aiemmasta ei vain maantieteelliseen dataan liittyvästä tutkimuksesta, vaan tietokantojen, datan ja sen epämääräisyyden sekä matemaattisen teorian tutkimuskentästä. Maantieteellinen epämääräisyys juuri GIS:in ja tietokantojen kontekstissa sinänsä on verrattain uusi asia. Kysymys nousi laajempaan tutkimuksen huomioon 90-luvulla ja valtaosa sen tutkimuksesta on tehty 2000-luvulla (Stefanakis, Vazirgiannis, & Sellis, 1999). Tutkimus kuitenkin hyödyntää paljon huomattavasti aiemmin tehtyä tietojenkäsittelytieteen ja matematiikan teoriapohjaa ja sovittaa tätä maantieteen ja GIS:in kontekstiin ja erityispiirteisiin.

3.1. Tietokannat

Tietokanta on loogisesti johdonmukainen kokoelma yhteen liittyvää dataa, jossa datalla tarkoitetaan tiedettyjä faktoja, jotka voidaan tallentaa ja joilla on implisiittinen merkitys. Tietokanta mallintaa jotain aspektia todellisesta maailmasta, ja se on tietoisesti rakennettu jotain tarkoitusta ja käyttöyhteyttä varten. (Elmasri & Navathe, 2004, p. 4).Olennaisimmalta määritelmältään tietokanta on siis *rakenne, joka kykenee varastoimaan dataa erilaisista entiteeteistä sekä näiden entiteettien välisistä suhteista* (Pratt & Adamski, 1987, p. 14). Tietokannan hallintajärjestelmällä (Database Management System, DBMS) viitataan niihin ohjelmistoihin, jotka liittyvät tietokannan määrittelyyn, luomiseen ja ylläpitämiseen, ja jotka tarjoavat kontrolloidun pääsyn tietokantaan (Connolly, Begg & Strachan, 1996, p. 16). Toisinaan tietokannan hallintajärjestelmään sisällytetään myös tietokanta itse, jolloin tietokantaa ja sen hallintajärjestelmää käsitellään funktionaalisesti yhtenäisenä entiteettinä (Silberschatz, Korth & Sudarshan, 2006, p. 1).

Modernin tietokantojen hallinnan historia voidaan johtaa aina 1960-luvulle Yhdysvaltain Apollo-projektiin, jonka yhteydessä IBM:ää pyydettiin kehittämään järjestelmä, joka kykenee käsittelemään suurien datamäärien koordinointia. Tämän johdosta syntynyt Generalized Update Access Method (GUAM) siirtyi tuotantoon vuonna 1964. Vuonna 1966 järjestelmää laajennettiin suuremmalle yleisölle nimellä Data Language/I (DL/I), joka oli osa laajempaa kokonaisuutta nimeltä Information Management System (IMS). 60-luvulla saivat alkunsa myös 60-luvun puolivälissä kehitetty Integrated Data Store -järjestelmä (I-D-S), sekä 60-luvun loppupuolella Conference on Data System Languages -ryhmässä kehitetty CODASYL- järjestelmä. (Pratt & Adamski, 1987, p. 25)

Eräs merkittävimpiä virstanpylväitä tietokantojen kehityksessä oli Ted Codd:in kehittämä relaatiotietokantajärjestelmä. Relaatiomalli esiteltiin Codd:in paperissa vuonna 1970 ja ensimmäiset kaupalliset sovellutukset kehitettiin 80-luvun alussa, kuten

Oracle:n DBMS sekä IBM:n MVS-käyttöjärjestelmän SQL/DL . Relaatiotietokanta perustuu matemaattisen relaation konseptille ja relaatiotauluihin. (Elmasri & Navathe, 2004, p.125). Relaatiotietokanta on kokoelma normalisoituja relaatioita (Connolly, Begg & Strachan, 1996, p. 88). Relaatio on kaksiulotteinen taulu, joka järjestyy riveihin ja sarakkeisiin. Jokainen rivi edustaa yhteen liittyvää dataa ja sillä on yksilöllinen nimi. Sarakkeet ovat attribuutteja ja sarakkeen jokainen kohta saa saman attribuutin arvoja. Samaan tapaan attribuuteilla on yksilöllinen nimi, joka kuvaa attribuuttia. Relaatiossa rivien ja sarakkeiden järjestys on merkityksetön. Relaatioiden väliset yhteydet kuvataan omina relaatiotauluinaan. (Pratt & Adamski, 1987, p. 76). Relaatiotietokantamallilla pyrittiin parantamaan datan itsenäisyyttä sekä vastaamaan konsistenssi- ja redundanssiongelmiin etenkin normalisoinnin avulla. Relaatiotietokantamallia hyödynnettiin monessa kehitysprojektissa, joista merkittävin oli kenties IBM:n System R, joka johti esimerkiksi SQL-kielen syntyyn. (Commolly, Begg & Strachan, 1996, p. 83).

3.1.1. Oliotietokannat

Myöhemmin ilmeni useita käyttöyhteyksiä, joista relaatiotietokannat eivät suoriutuneet tarpeeksi tehokkaasti. Tällaisia ovat monimutkaisemmat tietokannat, jotka vaativat monimutkaisten objektien käsittelyä, pitkäaikaisempia transaktioita, uusia datatyyppejä ja mahdollisuuksia määritellä ulkostandardillisia operaatioita. Tällaisia käyttöyhteyksiä ovat esimerkiksi tekninen suunnittelu ja tuotanto, tieteelliset kokeet, telekommunikaatio, multimedia ja tämän tutkielman kannalta oleelliset paikkatietojärjestelmät. Ratkaisijaksi tälle ongelmalle ovat nousseet olio-orientoituneet tietokannat (*object-oriented databases*) tai lyhyemmin oliotietokannat. Toinen merkittävä syy olio-orientoituneille tietokannoille on ollut olio-orientoituneiden ohjelmointikielien, kuten Java:n ja C++:n yleistyminen ohjelmistokehityksessä. Näillä kielillä kehitetyt ohjelmistot voivat olla saumattomammin yhteydessä olio-orientoituneisiin tietokantoihin. (Elmasri & Navathe, 2004, p. 639).

Oliotietokannat perustuvat olion käsitteelle. Olio on yksilöllisesti tunnistettava entiteetti, joka sisältää ne attribuutit, jotka kuvaavat tosielämän objektin tilan ja ne toiminnot, jotka liittyvät tosielämän objektiin (Commolly, Begg & Strachan, 1996, p. 687). Toisin sanoen olio rakentuu kahdesta komponentista: tilasta (arvo) ja käyttäytymisestä (operaatiot). Näin tietokannan suunnittelija kykenee määrittämään sekä monimutkaisten objektien rakenteen että operaatiot, joita niille voi suorittaa, ja näin olioiden käyttö tietokannoissa mahdollistaa monimutkaisempien datarakenteiden ja interaktioiden käsittelyn. Oliotietokannat laajentavat olioiden pysyvyyttä niin, että ne selviävät sovelluksen sulkemisen jälkeenkin ja että muut sovellukset voivat noutaa, käyttää ja jakaa niitä. (Elmasri & Navathe, 2004, p. 640). Muita oliotietokantojen ominaisuuksia ovat esimerkiksi enkapsulaatio eli tiedon piilottaminen olion sisälle, joka kontribuoi datan itsenäisyyteen, sekä periytyvyys eli tapa jolla olion luokasta voidaan johtaa aliluokkia (Commolly, Begg & Strachan, 1996, p. 686, 691).

Relaatiotietokantojen laajan ja vakiintuneen suosion vuoksi oliotietokantojen on ollut vaikea saavuttaa suosiota (Elmasri & Navathe, 2004, p. 640). Tämän lisäksi vastaukseksi mainittuihin relaatiotietokantojen haasteisiin olio-relaatiotietokannat laajensivat perinteisten relaatiotietokantojen datamallia sisällyttämään monimutkaisempia tietotyyppejä ja olio-orientaatiota, samalla kuitenkin säilyttäen

relaatiomallin perusteet (Silberschatz, Korth & Sudarshan, 2006, p. 361).

3.1.2. Spatiaaliset tietokannat ja spatiaalinen data

Spatiaaliset tietokannat ovat tietokantojen erityistyyppi, joka keskittyy objektien tilallisen ulottuvuuden mallintamiseen ja tallentamiseen. Toisin sanoen spatiaaliset tietokannat varastoivat spatiaalista dataa, joka määritellään tuonnempana. Periaatteessa spatiaalisista tietokannoista puhuttaessa voidaan erottaa myös CAD-mallinnus, jota käytetään erilaisten objektien, esimerkiksi rakennusten, laitteiden tai käyttöesineiden suunnitteluun, mutta tässä tutkielmassa keskitytään nimenomaan maantieteellisiin tietokantoihin, jotka kuvaavat maanpinnan ja ilmakehän ilmiöitä, esimerkiksi teitä, jokia, hallinnollisia alueita tai sadepilviä. (Silberschatz, Korth & Sudarshan, 2006, p. 908).

Spatiaalisten tietokantojen erityispiirteet liittyvät ennen kaikkea spatiaalisiin datatyyppeihin sekä maantieteellisille objekteille olennaisiin kyselyihin. Maantieteellisten tietokantojen tulee sisältää malleja, jotka kykenevät tulkitsemaan datan spatiaalisia piirteitä, sekä usein lisäksi erityistä indeksointia ja varastointia (Elmasri & Navathe, 2004, p. 781). Tyypillisiä esimerkkejä spatiaalisista datatyypeistä ovat pisteet, viivat, polygonit, alueet sekä rasteridata (Heywood, Cornelius & Carver, 1998, p. 15). Spatiaalisten datatyyppien tallettamisen lisäksi spatiaalisten tietokantojen on kyettävä suoriutumaan tehokkaasti kyselyistä, jotka koskevat objektien topologisia suhteita, esimerkiksi spatiaalisten objektien etäisyyttä toisistaan tai niiden päällekkäisyyksiä, sekä suunta- ja etäisyyssuhteita (Elmasri & Navathe, 2004, p. 781).

Yksinkertaisimmillaan spatiaalinen tai maantieteellinen data tarkoittaa dataa, jolla on jokin spatiaalinen viite. Spatiaalisen datan voi redusoida ilmaisuksi, että jossain määrätyssä lokaatiossa on jokin asia, jossa tämä asia voi olla esimerkiksi entiteetti, spatiaalinen kategoria, aktiviteetti, tai jonkin muuttujan mittatulos. (Goodchild & Zhang, 2002, p. 2). Spatiaalisen datan erityispiirteitä ovat viittaus maantieteelliseen sijaintiin - useimmiten karttakoordinaatteina - yhteydet muuhun spatiaaliseen dataan, ja dataan liittyvät ei-spatiaaliset yksityiskohdat (Burrough, 1986, viitattu lähteessä Heywood, Cornelius & Carver, 1998, p. 13). Termejä "spatiaalinen data" ja "maantieteellinen data" käytetään usein synonyymeinä niiden hienovaraisista eroista huolimatta (Goodchild & Zhang, 2002, p. 2). Kuten spatiaalisten tietokantojen määrittelyssä rajattiin, tässä tutkielmassa keskitytään nimenomaan maantieteelliseen dataan, eli spatiaaliseen dataan, joka kuvaa maanpinnan ja ilmakehän ilmiöitä, ja esimerkiksi CAD-data rajataan tutkielman ulkopuolelle. Tutkielman pääpaino on kaksiulotteisen datan tarkastelulla ja operaatioilla ja kolmiulotteisen datan käsittely rajataan sen ulkopuolelle. Tässä tutkielmassa on päädytty käyttämään termiä "maantieteellinen" termin "spatiaalinen" sijaan, vaikka lähdeaineistossa sana "spatiaalinen" huomattavasti yleisempi onkin, koska näin halutaan korostaa tutkielman käsittelevän spatiaalisia tietokantoja ja dataa nimenomaan maantieteellisessä yhteydessä; siis spatiaalista dataa, jolla on jokin viite maanpallon pinnalle. Termejä "spatiaalinen tietokanta" ja "spatiaalinen data" käytetään yhä niissä yhteyksissä, joissa puhutaan tietokannoista ja datasta yleisemmin.

Spatiaalisen datan tyypillisimmät datarakenteet ovat vektorimalli ja rasterimalli. Rasterimallissa spatiaalinen data on jaettu samankokoisiin soluihin kaksiulotteiseen

taulukkoon. Yksittäinen solu voi saada arvoja esimerkiksi binäärisesti 1 tai 0 riippuen siitä, esiintyykö jotain ilmiötä rasterin solussa vai ei. Solujen ryhmällä voidaan myös ilmaista jotain yksittäistä entiteettiä, kuten rakennusta tai jokea. Vektorimalli taas perustuu kartesiolaiseen koordinaatistoon, jonka perusteella spatiaaliset elementit sijoitellaan. Vektorimallissa spatiaalisen datan perustyyppinä voidaan pitää pistettä. Piste ilmaisee yhden spatiaalisen sijainnin ja siihen mahdollisesti liittyvää tietoa. (Heywood, Cornelius & Carver, 1998, p. 47). Yksittäisillä pisteillä voidaan maantieteessä ilmaista kohteita, jotka ovat luonteeltaan yksilöllisiä, esimerkiksi vuoren huippu, tai liian pieniä esitettäviksi alueina, joita voivat olla skaalasta riippuen esimerkiksi lyhtypylväät, rakennukset tai kaupungit. Pisteet ovat myös muiden spatiaalisten datatyyppien rakennuspalikoita. Viiva on sarja pisteitä, jotka yhdistyvät toisiinsa suorien janojen tai kaarien avulla. Viivat sopivat esitysmuodoksi luontaisesti lineaarisille entiteeteille, esimerkiksi joille tai teille. Alue tai polygoni taas on viiva, joka sulkeutuu itseensä. Polygoneilla voi esittää esimerkiksi metsäalueita tai hallinnollisia alueita. (Heywood, Cornelius & Carver, 1998, p. 24) Olennainen spatiaalisen datan ominaisuus on sen skaala, joka tulee vastaan etenkin karttojen yhteydessä. Tuolloin skaalalla tarkoitetaan pituuden kartalla suhdetta vastaavaan pituuteen mallinnettavassa todellisuudessa (Heywood, Cornelius & Carver, 1998, p. 22).

3.1.3. GIS

Spatiaalinen data ja spatiaaliset tietokannat liittyvät olennaisesti paikkotietojärjestelmiin tai GIS:iin (*Geographic Information Systems*). GIS:ille ei vaikuta olevan olemassa yleisesti hyväksyttyä ja tarkkaa määritelmää, vaan se on väljä kokoelma erilaisia työkaluja ja metodeja, jotka liittyvät maanpinnan datan käsittelyyn. Burrough (1986) määrittelee GIS:n "työkaluiksi, joilla voidaan kerätä ja tallettaa spatiaalista dataa oikeasta maailmasta ja hakea, muokata sekä esittää sitä halutessa." (viitattu lähteessä Heywood, Cornelius & Carver, 1998, p. 12). Iso-Britannian ympäristöministeriö taas määrittelee GIS:n "järjestelmäksi, jolla voidaan kerätä, tallettaa, tarkistaa, integroida, muokata, analysoida ja esittää dataa, joka viittaa spatiaalisesti maapallolle." (Heywood, Cornelius & Carver, 1998, p. 12). GIS:n oleellisimmat operaatiot liittyvät siis datan syöttämiseen, varastoimiseen, hallintaan, analysointiin ja esittämiseen. GIS:n tulee käsitellä sekä spatiaalista dataa, että siihen liittyvää ei-spatiaalista dataa, esimerkiksi spatiaalisen datan ominaisuustietoja (Heywood, Cornelius & Carver, 1998, p. 14).

Vähimmillään GIS:llä voidaan viitata johonkin ohjelmistoon, joka suorittaa näitä spatiaalisen datan käsittelyoperaatioita. Toisaalta Morris (2001) korostaa spatiaalista tietokantaa kaiken GIS:in ytimenä. Laajemmin GIS voidaan ymmärtää järjestelmäksi, johon kuuluvat niin tietokonelaitteet ja niiden käyttöjärjestelmät, ohjelmisto, spatiaalinen tietokanta ja tietokannan hallintajärjestelmä sekä järjestelmää johonkin tarkoitukseen käyttävät käyttäjät (Heywood, Cornelius & Carver, 1998, p. 13). GIS ei sinänsä tallenna karttoja, vaan ominaisuuksia, jotka voidaan esittää karttoina, mutta myös esimerkiksi käyrinä, taulukoina tai graafeina (Morris, 2001). Tyypillinen tapa esittää spatiaalista dataa GIS:llä on jakaa se käyttötarpeiden mukaan erilaisiin kerroksiin, esimerkiksi maatyyppeihin, vesistöihin, metsäalueisiin ja topografiaan. Eräs GIS:n etu suhteessa perinteisiin karttoihin on juuri tämä kyky käsitellä dataa erillisinä kerroksina (*layer*). Jokainen kerros vastaa jotakin mielenkiinnon kohteena olevaa teemaa ja jakaantuu pisteisiin, janoihin tai alueisiin, joilla on yhteinen attribuuttiarvo (Stefanakis, Vazirgiannis, & Sellis, 1999).

3.2. Maantieteellinen epämääräisyys

Maantieteellinen todellisuus on äärimmäisen monimutkainen. Näin ollen jotta sitä kyetään tallettamaan, käsittelemään ja esittämään, se täytyy abstrahoida ja yksinkertaistaa. Myös digitaalisessa mallintamisessa on rajallinen lukujen koko, rajallinen tallennustila ja rajallinen prosessointikyky. Toisin sanoen todellisuutta ei voida tallettaa ja esittää yhtä täsmällisenä kuin se todellisuudessa ilmenee. (Goodchild & Zhang, 2002, p. 3). Goodchild ja Zhang (2002, p. 4) toteavatkin: "On äärimmäisen epätodennäköistä, että tämä äärimmäisen monimutkainen maantieteellinen kompleksisuus olisi pelkistettävissä malleiksi täydellisellä tarkkuudella." Tämä on perustavanlaatuinen maantieteellisen epämääräisyyden lähde. Tuo itsessään epämääräinen maantieteellinen data voi myös kulkeutua erilaisten transaktioiden ja käsittelijöiden kautta, joilla on omat protokollansa ja tulkintatapansa. Tässä kontekstissa maantieteellisen epämääräisyyden voi määritellä maantieteellisen datan ja käyttäjän tuohon dataan liittämän tarkoituksen eroksi, toisin sanoen eroksi sen suhteen, mitä tietokanta sisältää ja mikä olisi käyttäjän luoman tietokannan sisältö, mikäli se kykenisi äärimmäisen suoraan ja tarkkaan todellisuuden mallintamiseen. (Goodchild & Zhang, 2002, p. 5).

Pauly ja Schneider (2008) käsittelevät maantieteellistä epämääräisyyttä epätäydellisyyden (*imperfection*) käsitteen kautta. Heidän mallissaan datan epätäydellisyys koostuu virheellisestä (*inaccurate*) ja epätarkasta (*impercise*) datasta, jossa ensimmäinen tarkoittaa yksinkertaisesti väärää tai puuttuvaa dataa, esimerkiksi inhimillisen erehdyksen tai laitteistovian aiheuttamana, ja jälkimmäinen taas kuvaa sinänsä oikeaa mutta rajallisten tallentamis- ja esityskeinojen vuoksi tarkkuudeltaan epämääräistä dataa. Juuri tämä epätarkka data on maantieteellisen epämääräisyyden (*spatial vagueness*) ilmentymä tietokantatasolla. Epämääräisyys voidaan taas jakaa epävarmuudesta (*incertainty*) ja sumeudesta (*fuzziness*) johtuvaksi. Epävarmassa datassa itse kohde on tarkasti määriteltävissä, mutta sitä ei esimerkiksi kyetä mittaamaan tarkasti tai ollenkaan. Sumean datan yhteydessä taas puhumme maantieteellisistä olioista, joiden olemukseen itse epämääräisyys kuuluu. Goodchild ja Zhang (2002, p. 7) taasen käyttävät termiä epävarmuus (*uncertainty*) kattoterminä virheille (*error*), satunnaisuudelle (*randomness*) - joka tässä yhteydessä oletetaan ennustamattomuudeksi - ja epämääräisyydelle (*vagueness*).

Kun spatiaalisten tietokantojen yhteydessä puhutaan spatiaalisen datan ja objektien tallentamisesta ja esittämisestä, eli pisteistä, viivoista, alueista tai rasteridatasta, puhutaan käytännössä usein datasta, joka on tarkasti rajattua ja määriteltyä (*crisp*) (Schneider, 2014). Epämääräisyys kuitenkin on maantieteellisten objektien perustavanlaatuinen ominaisuus. (Goodchild & Zhang, 2002, p. 6) Useat maantieteellisten objektien käsitteet itsessään kuvaavat jotakin luontaisen jatkumon osaa, jonka täsmällinen erottaminen ympäristöstään ei ole yksiselitteistä. Klassinen esimerkki tästä ovat erilaiset vaihettumisvyöhykkeet, esimerkiksi metsäalueen alkaminen ja loppuminen; Ei ole yksiselitteistä missä vaiheessa puiden tiheys ja alueen laajuus on sellainen, että voimme käsitellä sitä metsänä, ja mihin vedämme rajan metsän ja ei-metsän, esimerkiksi puistikkoalueen välille (Morris, 2003). Toisin sanoen luonto ei sisällä niin tarkkarajaisia erotuksia osiensa välillä kuin ihminen luontoa mallintaessaan tekee. Esimerkiksi metsää mallinnettaessa tietokantaan harvemmin kirjataan yksittäisten puiden sijainteja karttatasolla, joiden tiheys muodostaisi sen alueen, jota kutsumme metsäksi, vaan spatiaalisessa datassa metsät esitetään usein metsä*alueina*. Lähes kaikki

maantieteelliset objektit ovat reunoiltaan epämääräisiä. Esimerkiksi vuorilla, metsillä tai merillä ei ole tarkkaan määriteltyjä rajoja. Metsän ja suoalueen tai maaseudun ja kaupunkialueen vaihettumisvyöhykkeet ovat luonteeltaan jatkuvia ja epämääräisiä. Tällaisia rajoiltaan epämääräisiä objekteja, jotka eivät salli yksiselitteistä tulkintaa sen suhteen, kuuluuko jokin spatiaalinen alue niihin vai ei, kutsutaan epämääräisiksi objekteiksi (*vague object*). (Kulik, 2001). Spatiaalisten tietotyyppien mukaisesti Pauly ja Schneider (2004) jakavat epämääräiset spatiaaliset objektit epämääräisiin pisteisiin (*vague point object*), epämääräisiin janoihin (*vague line object*) ja epämääräisiin alueisiin (*vague region object*). Perinteinen Boolen logiikka, jossa todellisuuden elementit on jaettavissa erillisiin, tarkasti rajattuihin ja toisensa poissulkeviin sarjoihin, ei kykene esittämään tällaisia epämääräisiä objekteja ja kategorioita. (Goodchild & Zhang, 2002, p. 91). Tämä monimutkaisten spatiaalisten kategorioiden konseptualisointi tarkkarajaisina objekteina on merkittävä potentiaalinen epämääräisyyden lähde (Goodchild & Zhang, 2002, p. 134). Täytyy kuitenkin huomioida, että tässäkin tutkielmassa käsitellään usein juuri näiden maantieteellisten *objektien* esittämistä tietokannoissa; Tällä termillä viitataan väistämättömään ihmisperspektiiviin maantieteellisen todellisuuden kategorisoinnissa, eikä niinkään tähän monimutkaisten spatiaalisten kategorioiden todellisuuteen. Maantieteelliset objektit ovat objekteja niin kuin ne ihmiskontekstissa ymmärretään, huolimatta niiden mallinnustavasta tietokanta- ja esitystasolla, jota juuri tässä tutkielmassa tarkastellaan.

Skaala ja datan yhdistäminen eri lähteistä ovat eräitä olennaisia maantieteellisen epämääräisyyden lähteitä. Käytännössä kaikki spatiaalisen datan lähteet ovat pienempiä kuin todellisuus, jota ne mallintavat. Koska spatiaalinen data on aina yksinkertaistettu mallinnus todellisuudesta, eli todellisuuden kaikkia elementtejä ei voida tallentaa, tuo skaalan määräämä pienempi tarkkuus ja siitä aiheutuva datahäviö aiheuttaa epämääräisyyttä spatiaalisen datan tulkinnassa ja käytössä. Tätä datahäviötä tapahtuu aina mallinnettaessa, puhuttiin sitten spatiaalisen datan sijoittamisesta koordinaatistoon (vektori), jolloin esimerkiksi kartan piste on käytännössä lähellä todellista kohdetta, tai ruudukkoon (rasteri), jolloin yhden ruudun kompleksisuus yksinkertaistetaan yhdeksi arvoksi (Beaubouef, et. al. 2007). Maantieteellisellä datalla on siis aina oma rajallinen skaalansa ja tarkkuutensa, ja eri lähteistä koottuna nämä datan skaalat ja tarkkuudet voivat vaihdella. Datan integraatioprosessissa tätä erilaisista tarkkuuksista syntyvää epämääräisyyttä ei kuitenkaan usein oteta huomioon, vaan dataa käsitellään sellaisenaan skaalattomana ja täsmällisenä, joka voi johtaa virheellisyyksiin lopullisessa kootussa datassa (Abler 1987, viitattu lähteessä Goodchild & Zhang, 2002, p. 5).

Vaikka maantieteelliset tietokannat ovat tarkkoja ja yksiselitteisiä, ihmiskonteksti, jossa ne on luotu voi olla erittäin epämääräinen. Oman ongelmansa datan yhdistämiseen tuovat eri perustein laaditut maantieteelliset luokat. Monimutkaisten ilmiöiden luokituksessa diskreetteihin luokkiin esiintyy väistämätöntä subjektiivisuutta esimerkiksi luokkakuvauksissa ja näytevalinnassa. Luokitusten tekeminen itsessään on olennainen osa maantieteellisen datan epämääräisyyttä, koska siinä toteutuu juurikin tämä maantieteellisen todellisuuden yksinkertaistaminen ja täten datan väistämätön häviäminen. Näiden luokkien määrittely ei ole yksiselitteistä ja tarkkaa, vaan luonteeltaan epämääräistä. (Goodchild & Zhang, 2002, p. 167) Loppujen lopuksi siinä on kyse skaalasta ja tarpeellisuudesta. Esimerkiksi puistoaluetta mallintaessa on mallintajan päätettävä kuuluvatko puiston metsäläikät ympäröivästä ruohomaasta erilliseksi metsän luokaksi vai kuuluvatko ne molemmat samaan puistoluokkaan. (Goodchild & Zhang, 2002, p. 7). Luokitukset kuten korkea tai matala heinikko, tai

kaupungistumisen korkea, keskitasoinen ja matala aste ovat luonteeltaan epämääräisiä (Goodchild & Zhang, 2002, p. 168). Spatiaalista dataa käyttäessä törmätään myös luonnollisen kielen epämääräisyyden ongelmaan, etenkin tietokantakyselyiden kohdalla. Käyttäjä voi esimerkiksi etsiä kohteita, jotka ovat ”lähellä” jotain toista kohdetta, jolloin kohteen läheisyys on hyvin tulkinnanvarainen ja liukuva ja siten epämääräinen käsite.

Eräs maantieteellisen datan luonteenomainen piirre on, että se on myös spatiaalisesti ja temporaalisesti riippuvaista. Täten myös sen epämääräisyys ja virhealttius on spatiaalisesti riippuvaista. Spatiaalisella riippuvuudella tarkoitetaan sitä, että toisiaan spatiaalisesti ja temporaalisesti lähellä olevat objektit muistuttavat toisiaan enemmän kuin toisistaan etäällä olevat objektit (Tobler, 1970). Kaikki spatiaalisesti jakautunut data on jossain määrin spatiaalisesti riippuvaista. Esimerkki spatiaalisesti riippuvaisesta datasta on korkeuskäyrien mukaan mallinnetut korkeuserot, jossa yhden pisteen korkeus on riippuvainen ympäröivien alueiden korkeuksista. Epämääräisyys ei tässä tapauksessa synny pelkästään korkeusdatan mittavirheistä, vaan myös läheisten lokaatioiden virheiden spatiaalisesta kovarianssista. (Goodchild & Zhang, 2002, p. 87). Spatiaalisen datan temporaalisuudella tarkoitetaan sen sidonnaisuutta tiettyyn mittaamisajankohtaan. Koska maantieteellinen todellisuus on paikoin alati muuttuvaa, esimerkiksi eroosioalueiden muuttumiset tai napajäätiköiden sulamiset, sen esittäminen täsmällisenä ja tarkkarajaisena voi olla epäkäytännöllistä.

Tämä tutkielman kannalta ei tehdä tarkkaa rajausta siihen, minkälaisesta epämääräisyydestä ja sen lähteestä puhutaan, koska myöskään eri menetelmät eivät pyri vastaamaan mihinkään tiettyyn maantieteellisen epämääräisyyden tyyppiin, vaan ne käsittelevät epämääräisyyttä sinällään ja menetelmästä riippuen niitä voi soveltaa vastaamaan erilaisiin maantieteellisen epämääräisyyden tyyppeihin tai lähteisiin. Pääsääntöisesti fokus on kuitenkin *maantieteellisten objektien ominaislaatuisessa epämääräisyydessä sekä mittaus- ja tallettamisepätarkkuuden aiheuttamassa epämääräisyydessä kaksiulotteisessa spatiaalisessa datassa, GIS-käyttöympäristön kontekstissa*. Pienemmälle huomiolle tässä yhteydessä jätetään spatiaalisen datan temporaalisuus sekä luonnollisen kielen aiheuttama epämääräisyys, joka liittyy ennen kaikkea kyselyjen toteuttamiseen tietokantarakenteen sijaan.

3.3. Menetelmät

Ratkaisua maantieteellisen epämääräisyyden hallintaan on viime vuosikymmeninä etsitty useista erilaisista menetelmistä. Useimmat näistä keskittyvät etenkin epämääräisiin alueisiin ja niiden operaatioihin. Dilo, De By ja Stein (2007) jakavat nämä menetelmät kahteen kategoriaan: menetelmiin, jotka mallintavat asteittaista vaihettumista, esimerkiksi hyödyntäen sumeita joukkoja (*fuzzy sets*), ja menetelmiin, jotka laajentavat epämääräisten alueiden rajan yksiulotteisesta kaksiulotteiseksi homogeeniseksi alueeksi. Pauly ja Schneider esittelevät tarkemmin neljä tyypillistä menetelmää yleisiksi suunnitteluratkaisuksi epämääräisyyden ongelmaan: sumeisiin joukkoihin (*fuzzy sets*) perustuvat mallit, karkeisiin joukkoihin (*rough sets*) perustuvat mallit, todennäköisyysmallit (*probabilistic models*) sekä tarkkojen objektien mallit (*exact object models*) tai toiselta nimeltään laajennetut mallit (*extended models)* (Schneider, 1999, 2008; Pauly & Schneider, 2004, 2008). Näistä sumeisiin joukkoihin perustuvat mallit, karkeisiin joukkoihin perustuvat mallit ja todennäköisyysmallit kuuluvat Dilo:n, De By:n ja Stein:in esittelemään ensimmäiseen joukkoon, tarkkojen

objektien mallit taas jälkimmäiseen. Sumeisiin joukkoihin perustuvat mallit pohjaavat Zadehin (1965) esittelemään sumean joukon teoriaan, jossa alkion kuuluminen joukkoon saa binäärisen ”0 tai 1” tunnuksen sijaan liukuvan [0,1] välille sijoittuvan todennäköisyyden. Karkeisiin joukkoihin perustuvat mallit pohjaavat Pawlakin (1982) esittelemään karkean joukon teoriaan, joka perustuu joukon ylempään ja alempaan approksimaation. Todennäköisyysmallit keskittyvät mallintamis- ja mittausepävarmuuteen ja pohjaavat todennäköisyysteoriaan, joka perustuu tilastolliseen ennustamiseen jo tiedetyn pohjalta. Takkojen objektien mallit tai laajennetut mallit perustuvat olemassaolevien rakenteiden laajentamiseen niin, että ne kykenevät käsittelemään maantieteellistä epämääräisyyttä. Laajennettujen mallien osalta käsitellään laajennuksia sekä kolmiarvoiseen logiikkaan (*three-valued logic*) että moniarvoiseen logiikkaan (*multi-valued logic*) kuitenkin pysyen tarkkarajaisissa objekteissa, näin erottuen esimerkiksi sumeiden joukkojen malleista. (Schneider, 2008; Pauly & Schneider, 2008).

Tässä tutkielmassa tarkasteluun otetaan sumeisiin joukkoihin perustuvat mallit, karkeisiin joukkoihin perustuvat mallit, ja laajennetut mallit. Tämä sisältää myös ne menetelmät, jotka pyrkivät yhdistelemään elementtejä näistä eri menetelmistä. Rajaus on tehty löydetyn lähdekirjallisuuden laajuuden ja menetelmien relevanssin perusteella. Todennäköisyysmallit jätetään tarkastelusta pois, koska ne eivät vaikuta olevan relevantteja juuri tämän tutkimuskysymyksen kannalta nykytutkimuksen kentällä, vaan vastaavat epämääräisyyden ongelmaan kapea-alaisesti lähinnä ennustettavuuden osalta, eivätkä sellaisenaan vastaa maantieteellisille objekteille ominaiseen epämääräisyyteen.

3.3.1. Sumeiden joukkojen menetelmät

Sumeisiin joukkoihin perustuvat menetelmät ovat selvästi yleisin tapa vastata maantieteellisen epämääräisyyden ongelmaan (Dilo, By & Stein, 2007). Se on ollut suosittu etenkin GIS-yhteisön piirissä mallintamisen näkökulmasta, koska sillä voidaan mallintaa useita erilaisia epämääräisiä maantieteellisiä ilmiöitä ja objekteja (Schneider, 2008). Sumeisiin joukkoihin perustuvat menetelmät ovatkin hyvin luontainen tapa esittää etenkin epämääräisiä rajoja omaavia spatiaalisia objekteja, koska tuo sumeiden joukkojen käsitteellinen sumeus sopii hyvin yhteen maantieteellisten objektien sumeuden kanssa (Burrough, 1996, p. 18; Goodchild & Zhang, 2002, p. 168).

Sumeiden joukkojen menetelmät perustuvat Zadehin vuonna 1965 esittelemään sumeiden joukkojen teoriaan. Sumean joukon teoria on laajennus klassiseen Boolen logiikkaan, jossa alkiot joko kuuluvat tai eivät kuulu määrättyyn joukkoon. Formaalisti määriteltynä tämä kuuluisi seuraavasti:
Olkoon X klassinen objektien joukko (universumi), ja jäsenyys tuon joukon X osajoukossa A voidaan kuvata tunnuslukufunktiolla (*characteristic function*)

$$\chi_A : X \rightarrow \{0,1\}$$

, jossa kaikkiin $x \in X$ pätee:

$$\chi_A(x) = \begin{cases} 1 \text{ jos ja vain jos } x \in A \\ 0 \text{ jos ja vain jos } x \notin A. \end{cases}$$

Tämä funktio voidaan yleistää siten, että jokainen X:n alkio saa osajoukossa A osallisuusarvon (*degree of membership*) 1 mikäli se kuuluu osajoukkoon ja osallisuusarvon 0 mikäli se ei kuulu. Spatiaalisten tietokantojen suhteen tämä tarkoittaa, että jokin koordinaattipiste yksiselitteisesti joko kuuluu tai ei kuulu johonkin alueeseen tai objektiin, esimerkiksi metsäalueeseen tai jokeen. Sumeisiin joukkoihin perustuvat menetelmät taas sallivat alkioille osittaisen kuulumisen yhteen tai useampaan sumeaan osajoukkoon. Olkoon X taas klassinen objektien joukko (universumi). Tuolloin:

$$\mu_{\tilde{A}} : X \rightarrow [0,1]$$

on Ã:n osallisuusfunktio (*membership function*), ja joukko

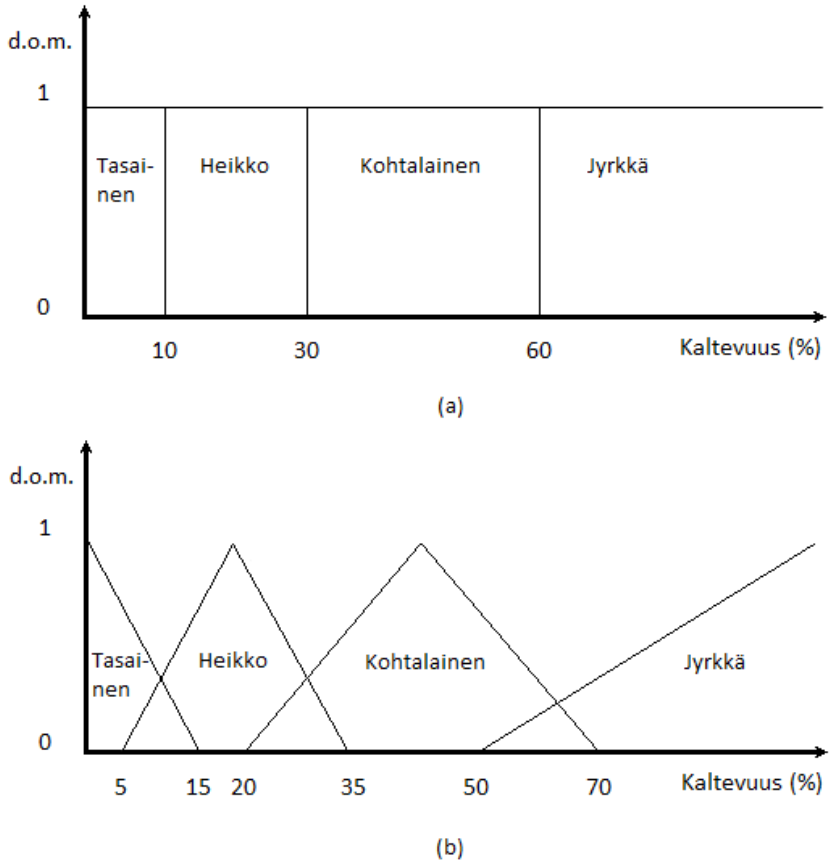
$$\tilde{A} = \{(x,\mu_{\tilde{A}}(x)) \mid x \in X\}$$

on sumea joukko X:ssä. Jokainen X.n alkio saa arvon suhteessa niiden osallisuuteen osajoukossa Ã. Ne $x \in X$ alkiot, jotka klassisessa mielessä eivät kuulu osajoukkoon Ã saavat osallisuusarvon $\mu_{\tilde{A}}(x) = 0$; ne $x \in X$ alkiot, jotka kuuluvat kokonaan osajoukkoon Ã saavat osallisuusarvon $\mu_{\tilde{A}}(x) = 1$. Toisin kuin tarkkarajaisissa joukoissa, sumeissa joukoissa alkiot voivat saada osallisuusarvoja väliltä [0,1]. Spatiaalisten tietokantojen suhteen tämä tarkoittaa, että jokin koordinaattipiste voi kuulua johonkin alueeseen tai objektiin kokonaan (1), olla kuulumatta siihen (0) tai kuulua siihen tietyn asteen verran, (esimerkiksi 0,6). Tämä osallisuusarvon liukuvuus mallintaa maantieteellistä epämääräisyyttä. (Schneider, 1999, 2008; Beaubouef & Petry, 2010). Tyypillisesti sumeiden alueiden mallintamisessa puhutaan juuri osallisuusarvosta, mutta esimerkiksi kyselyiden kontekstissa voidaan puhua myös totuusarvosta (*degree of truth*) (Verstraete et al., 2007).

Kuva 1. Visualisointi epämääräisistä alueista sumeiden joukkojen menetelmällä. Mukailtu lähteestä Schneider (1999).

Sumeiden joukkojen menetelmät hyötyvät mallin intuitiivisesta vastaavuudesta objektien luonnolliseen sumeuteen (tai epämääräisyyteen). Kuvan 1 visualisoinnit voisivat kuvata esimerkiksi ilmansaasteiden leviämistä tuulettomalla (a) ja tuulisella (b) säällä. Tuolloin värin vahvuus kuvaa karttapisteen osallisuusarvoa. Sumeiden joukkojen menetelmät ovat myös hyvin yleispätevä epämääräisyyden mallintamisen muoto ja sitä voidaan soveltaa useisiin erilaisiin käyttökohteisiin. Tietyissä käyttökohteissa ne voivat vähentää otannan tiheyden tarvetta, koska puuttuvat arvot voidaan sumeiden joukkojen mallissa interpoloida (Beaubouef & Petry, 2010). Ne voivat myös vähentää subjektiivisen luokittelun tarvetta, koska malli tukee sumeaa luokkaan kuulumista, jossa

entiteetti, esimerkiksi karttapiste, ei kuulu yksiselitteisesti vain yhteen luokkaan kuten perinteisessä luokittelussa (Kuva 2a), vaan osallisuusarvo on liukuva ja entiteetti voi kuulua yhtaikaa useampaan luokkaan (Kuva 2b). Sumeiden joukkojen menetelmät ovat myös tehokkaita tilanteessa, jossa arvioidaan useamman muuttujan vaikutuksia. (Goodchild & Zhang, 2002, p. 187).



Kuva 2. Perinteisen (a) ja sumean (b) luokkaan kuulumisen vertailu, esimerkkinä kaltevuus. Mukailtu lähteestä Stefanakis, Vazirgiannis, & Sellis, (1999).

Sumeiden joukkojen menetelmien haittapuolet liittyvät paljolti menetelmän raskauteen. Siinä missä ne kykenevät mallintamaan maantieteellisen todellisuuden epämääräisyyttä tarkkoja malleja paremmin ja tuottavat tarkempia tuloksia, ne myös raskauttavat datan prosessointia ja monimutkaistavat tulosten tulkintaa. GIS-tietokannan tasojen määrä kasvaa moninkertaisesti, kun jokainen erillinen teema esitetään niin monella tasolla kuin siihen on liitetty osallisuusarvoja (Stefanakis, Vazirgiannis, & Sellis, 1999). Eräs perustavanlaatuinen ongelma liittyy sumeiden luokkien ja sumeuden spatiaalisen toteutumisen määrittelyyn. Tämä toimi kysyy edelleen ihmisarviota ja on väistämättä arvio, eikä sumeiden joukkojen menetelmää näin ollen voida pohjimmiltaan pitää sen enempää varsinaisesti objektiiviseen todellisuuteen pohjaavana kuin perinteisiä tarkkarajaisia menetelmiä (Goodchild & Zhang, 2002, p. 170).

3.3.2. Karkeiden joukkojen menetelmät

Karkeiden joukkojen menetelmät perustuvat Pawlakin vuonna 1982 esittelemään karkeiden joukkojen teoriaan. Teoria perustuu ylempiin ja alempiin

approksimaatioluokkiin (*upper and lower approximations*), jotka ovat molemmat täsmällisiä (*crisp*) joukkoja, sekä erottamattomuusrelaatioihin (*indiscernibility relations*) – toiselta nimeltään ekvivalenssirelaatio (*equivalence relation*) - alkioiden ominaisuuksien välillä. (Pauly & Schneider, 2008). Karkeat joukot sisältävät seuraavat:

- U on universumi, joka ei voi olla tyhjä,
- R on ekvivalenssirelaatio,
- A = (U,R), järjestetty pari, on approksimaatiotila (*approximation space*),
- $[x]_R$ osoittaa ekvivalenssiluokan (*equivalence class*) x:n sisältämälle R:lle jokaiselle U:n alkiolle x,
- alkeisjoukot (*elementary sets*) A:ssa – R:n ekvivalenssiluokat,
- määriteltävät joukot (*definable sets*) A:ssa – mikä tahansa rajallinen A:n alkeisjoukkojen unioni.

Approksimaatiotila syntyy kun erottamattomuusrelaatio R asemoidaan universumiin U. Tämä jakaa U:n alkeisjoukkoihin, joita voi käyttää määrittämään muita joukkoja A:ssa. Karkea joukko X määritellään A:n määriteltävien joukkojen avulla:

X:n alempi approksimaatio A:ssa on joukko

$$RX = \{x \in U \mid [x]_R \subseteq X\};$$

X:n ylempi approksimaatio A:ssa on joukko

$$RX = \{x \in U \mid [x]_R \cap X \neq \varnothing\};$$

Joukkojen approksimaatiot voidaan kuvata myös alueiden avulla. X:n ylemmän ja alemman approksimaation $\overline{R}X$ ja $\underline{R}X$ suhteen X.n R-positiivinen alue on $POS_R(X) = \underline{R}X$, R-negatiivinen alue on $NEGR(X) = U - \overline{R}X$, ja R-raja-alue on $BN_R(X) = \overline{R}X - \underline{R}X$. X:ää kutsutaan R-määritellyksi jos ja vain jos $\overline{R}X = \underline{R}X$. Muussa tapauksessa ylempi ja alempi approksimaatioalue eivät ole samat, ja X on karkea suhteessa R:ään. (Beaubouef, Petry, & Ladner, 2007; Beaubouef, & Petry, 2009).

Esimerkki karkeiden joukkojen käytöstä spatiaalisen datan yhteydessä:

Olkoon U = {torni, virtaus, puro, joki, metsä, metsämaa, laidunmaa, niitty} ja olkoon ekvivalenssirelaatio R:
R* = {[torni], [virtaus, puro, joki], [metsä, metsämaa], [laidunmaa, niitty]}.

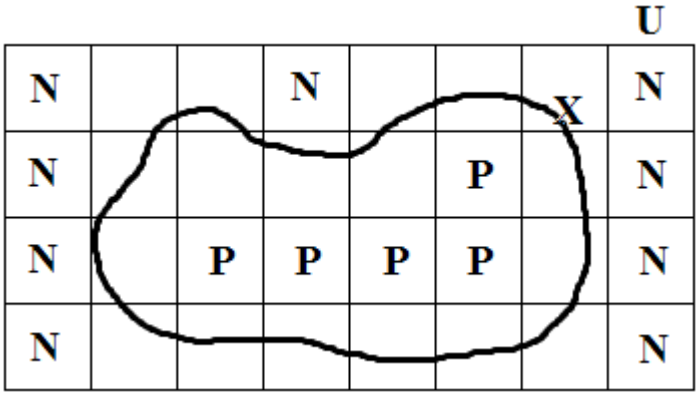
Määrätty joukko X = {torni, virtaus, puro, joki, metsä, laidunmaa} määritellään sitten sen ylemmän ja alemman approksimaation suhteen:

$\overline{R}X$ = {torni, virtaus, puro, joki, metsä, metsämaa, laidunmaa, niitty}, ja
$\underline{R}X$ = {torni, virtaus, puro, joki}.

Alempi approksimaatio sisältää ne ekvivalenssiluokat, jotka sisältyvät kokonaan joukkoon X. Ylempi approksimaatio sisältää alemman approksimaation lisäksi ne luokat, jotka kuuluvat vain osittaisesti joukkoon X. Karkea joukko A on U:n osajoukkojen ryhmä samoilla ylemmillä ja alemmilla approksimaatioilla. Esimerkissä karkea joukko on siis:

{{torni, virtaus, puro, joki, metsä, laidunmaa}
{torni, virtaus, puro, joki, metsä, niitty}
{torni, virtaus, puro, joki, metsämaa, laidunmaa}
{torni, virtaus, puro, joki, metsämaa, niitty}}

(Beaubouef, Petry, & Ladner, 2007; Beaubouef, & Petry, 2010; Petry, & Elmore, 2015).



Kuva 3. Esimerkki karkeasta joukosta X. Mukailtu lähteestä Beaubouef, Petry, & Ladner (2007).

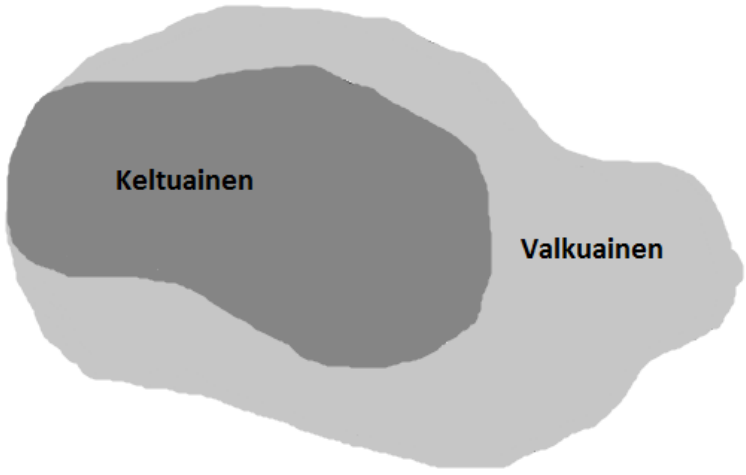
Ekvivalenssirelaatiot mahdollistavat entiteettien ryhmittelyn perustuen jonkinlaiseen samanlaisuuden käsitteeseen, joka taas mahdollistaa tarkastelualueen rakeisuuden pienentämisen tai kasvattamisen. Mahdollisten tai epämääräisten alueiden huomioiminen tietokantakyseilyissä toteutetaan käsittelemällä alempaa approksimaatioaluetta tarkkoja tai varmoja tuloksia vastaavana ja ylemmän approksimaation reuna-aluetta epämääräisenä tai mahdollisena. (Beaubouef, Petry, & Ladner, 2007). Joukon alempi approksimaatio on kaikkien niiden alkeisjoukkojen unioni, jotka kuuluvat siihen, ja joukon ylempi approksimaatio taas on kaikkien niiden alkeisjoukkojen unioni, joilla on ei-tyhjä leikkauspiste sen kanssa. Toisin sanoen joukon alempi approksimaatio sisältää kaikki ne alkiot, jotka varmasti kuuluvat joukkoon, ja ylempi approksimaatio taas ne alkiot, jotka mahdollisesti kuuluvat joukkoon. (Goodchild & Zhang, 2002, p. 181). Kuvassa 3 joukon X alemman approksimaation alkiot on merkitty kirjaimella P ja joukon ulkopuolelle jäävät alkiot kirjaimella N.

Sumeiden joukkojen menetelmien ohella karkeiden joukkojen menetelmät ovat toinen yleinen ja kiinnostusta herättänyt tapa vastata maantieteellisen epämääräisyyden ongelmaan (Beaubouef, Petry, & Ladner, 2007). Karkeiden joukkojen menetelmät käyttökohteet vaikuttaisivat painottuvan ensisijaisesti rasteridataan, jossa sijainnit ja ominaisuudet on sijoitettu ruudukkoon, sekä kuvadataan, toisin kuin useat vektoridataan keskittyvät menetelmät (Pauly, A. & Schneider, M., 2007). Etenkin spatiaalisen datan rasteroinnissa eli sijoittamisessa ruudukkoon karkeiden joukkojen menetelmistä on hyötyä, koska se säilyttää informaation epämääräisyyden siinä tilanteessa, jossa objektit jakaantuvat ruudukoille osittain tai jossa yksi ruudukko sisällyttää useamman objektin. Näin spatiaalinen alue partitioidaan ekvivalenssiluokkiin ja jokainen rasterin ruudukko kuuluu johonkin ekvivalenttiluokkaan. Ruudukon resoluution muuttaminen tarkoittaa tuolloin partitioinnin rakeisuuden muuttamista, joka johtaa vähempiin mutta isompiin luokkiin sekä kasvattaa tai pienentää epämääräisyyden määrää. Näin karkeiden

joukkojen menetelmistä on hyötyä myös muiden menetelmien kannalta epämääräisyyden määrän määrittämisessä. (Beaubouef, Petry, & Ladner, 2007; Beaubouef, & Petry, 2010). Pyrkiessään häivyttämään subjektiivisuutta osallisuusfunktioiden määrittelyssä, karkeiden joukkojen menetelmät toimivat eräänlaisena kompromissina subjektiivisuuden ja objektiivisuuden välillä epämääräisten kategorioiden käsittelyssä. Karkeiden joukkojen menetelmät perustuvat periaatteessa ennemminkin tarkkojen kuin sumeiden joukkojen käsittelyyn. Näin ollen niitä ei tule pitää sumeiden joukkojen erityiskäsittelynä, vaan omanlaisena tekniikkana potentiaaliseen epämääräisyyden hallintaan. (Goodchild & Zhang, 2002, p. 183).

3.3.3. Laajennetut mallit

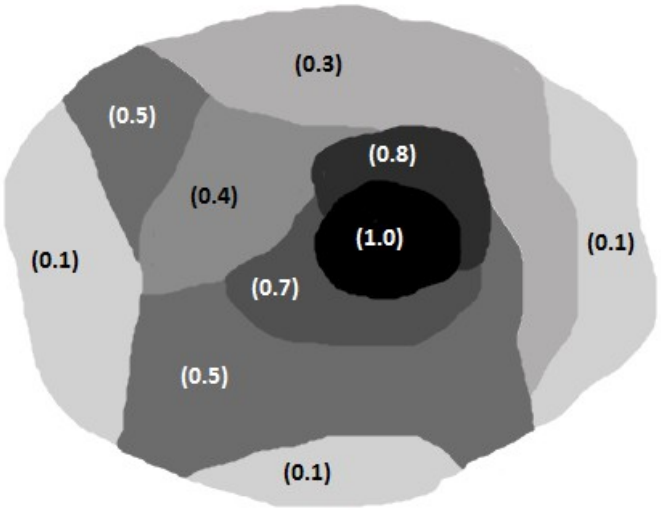
Yksinkertaisuudessaan suosittu tapa vastata maantieteelliseen epämääräisyyteen, etenkin epämääräisten rajojen ongelmaan ovat niin sanonut tarkkojen objektien mallit (*exact object models*) tai laajennetut mallit (*extended models*). Nämä mallit perustuvat maantieteellisen sumeuden implementointiin perinteisten tarkkojen objektien avulla sen sijaan, että epämääräisyyttä mallinnettaisiin koko järjestelmän tasolla. Jo olemassa olevia tarkkarajaisia määritelmiä, tekniikoita, datarakenteita ja algoritmeja laajennetaan mallintamaan datan epämääräisyyttä. Tällaisia menetelmiä on esitetty useita, mutta käytännössä ne lähes kaikki perustuvat objektin tarkkojen rajojen laajentamiseen raja-alueeksi, tämän raja-alueen käsittelyyn omana tarkkarajaisena objektinaan, sekä Boolen logiikan laajentamiseen kolmiarvoiseksi logiikaksi. Näissä malleissa objekti esitetään useimmiten kahtena tarkkarajaisena alueena: ydinalueena, joka varmasti kuuluu objektiin; ja raja-alueena, joka epävarmasti tai epämääräisesti kuuluu objektiin. Näin binäärinen Boolen logiikka (kuuluu, ei kuulu objektiin) on laajennettu kolmiarvoiseksi (kuuluu, ei kuulu, ehkä kuuluu) (Pauly & Schneider, 2008; Schneider, 1999; Schneider, 2008).



Kuva 4. Esimerkki keltuais-valkuaismallista. Mukailtu lähteestä Pauly, & Schneider, (2008).

Keltuais-valkuaismalli (*Egg-Yolk Model*) (Cohn & Gotts, 1996) esittää nimensä mukaisesti epämääräisen objektin kahtena tarkkarajaisena alueena, joista yksi alue – valkuainen – sisällyttää toisen – keltuaisen (Kuva 4). Keltuainen vastaa epämääräisen alueen tai objektin osaa, joka on varma; valkuaisen sisä- ja ulkoraja rajaavat objektin

epämääräisen reuna-alueen. Valkuaisen ulkopuoli ei varmasti kuulu objektiin, valkuaisen sisäpuoli kuuluu objektiin epävarmasti tai epämääräisesti, ja keltuaisen sisäpuoli kuuluu objektiin varmasti. Valkuainen ja keltuainen yhdessä muodostavat epämääräisen objektin. Keltuais-valkuaismalli hyödyntää RCC-metodeja (*Regional Connection Calculus*) alueiden topologisten yhteyksien - esimerkiksi erillisyys, päällekkäisyys, sisäkkäisyys – mallintamisessa. Clementinin ja di Felicen (2001) esittelemä leveiden rajojen malli on hyvin samankaltainen keltuais-valkuaismallin kanssa, määritelmien perustuen yleiseen topologiaan. Epämääräinen alue A koostuu kahdesta $\mathbb{R}^2$:n joukosta A_1 ja A_2 niin, että $A_1 \subseteq A_2$. Leveä raja ΔA on näiden joukkojen erotuksen sulkeuma $\Delta A = A_1/A_2$. Jokainen näistä joukoista on tarkka-rajainen, tavallinen suljettu $\mathbb{R}^2$:n joukko yhtenäisellä sisäpuolella. Keltuais-valkuaismallin tapaan sisempi alue A_1 määrittää alueen varman osan ja leveä raja ΔA rajaa epämääräisyyden rajat. Malli kuvaa topologisia relaatioita 9-intersektiomallin mukaan (Clementin & di Felice, 2001). Erwigin ja Schneiderin (1997) malli eroaa keltuais-valkuaismallista sekä Clementinin ja di Felicen mallista siten, että vain epämääräisen objektin varma ydin on alue, kun taas epävarma raja voi olla joko viiva tai alue. Tämä mahdollistaa tarkka-rajaisen alueen käsittelemisen epämääräisen alueen erityistapauksena. (Dilo, De By, & Stein, 2007).



Kuva 5. Esimerkki laakioaluemallista. Mukailtu lähteestä Kanjilal, Liu, & Schneider, (2010).

Kanjilalin, Liun ja Schneiderin (2010) malli, joka perustuu laakioalueisiin (*plateau region*), kuuluu myös laajennettujen mallien piiriin siinä, että se hyödyntää tarkkarajaisia alueita epämääräisyyden mallintamiseen. Mainittuihin nähden se eroaa kuitenkin siinä, ettei se perustu laajennettuihin reuna-alueisiin ja kolmiarvoiseen logiikkaan, vaan laakioalueisiin ja moniarvoiseen logiikkaan. Laakioalue määritellään rajalliseksi kokoelmaksi tarkkarajaisia ala-alueita niin, että jokaisella ala-alueella on oma osallisuusarvonsa, ja näin muodostavat ”laakion”, joka koostuu konseptuaalisesti rajattomasta määrästä pisteitä, joilla on sama osallisuus (Kuva 5). Jokainen laakioalue esitetään rajallisena sarjana pareja, jotka koostuvat tarkkarajaisesta alueesta sekä sille osoitetusta osallisuusarvosta, joka osoittaa osallisuutta epämääräiseen alueeseen. Parejen, ja siten ala-alueiden, määrä riippuu mallinnettavasta epämääräisestä objektista. Mikäli pareja on nolla, kyseessä on tyhjä laakioalue. Mitkä tahansa kaksi ala-aluetta ovat topologisesti joko vierekkäisiä tai erillisiä. Yksi ala-alue voi koostua useasta

erillisestä komponentista, joilla on sama osallisuusarvo. Kaikki osallisuusarvot ovat erisuuruisia ja parit järjestetään niiden osallisuusarvon mukaan. (Kanjilal, Liu, & Schneider, 2010).

Laakioaluemalli hyötyy samoista seikoista kuin muutkin laajennettujen mallien metodit, eli se pystyy hyödyntämään tarkkarajaisten spatiaalisten objektien jo olemassaolevia määritelmiä, tekniikoita, datarakenteita ja algoritmeja, joita ei siten tarvitse kehittää alusta, vain muokata ja laajentaa. Koska kukin laakioalue koostuu tarkkarajaisista alueista, niihin voi soveltaa olemassaolevia ja tunnettuja tarkkarajaisten alueiden topologisia operaatioita. Samaan tapaan laakioaluemalli pystyy nojaamaan vankkaan tarkkarajaisten objektien teoriaan sekä jo pitkälle kehitettyihin sovelletuksiin myös tietokanta- ja kyselytasolla, ja hyödyntämään niitä epämääräisen datan mallintamisessa. (Schneider, 2014).

4. Pohdinta

Aineistoa tarkastelemalla välittyy hyvin vahvasti kysymyksen kahtiajakautuneisuus: teorian vahvuus ja käytännön heikkous. Maantieteellisen epämääräisyyden mallintamisen kysymys hyötyy sen laajasta teoriapohjasta. Sen taustalla vaikuttaa laaja traditio maantieteen filosofiaa lähtien aina todellisuuden hahmottamisesta mallintamisen tarkkuuteen, matemaattista teoriaa, ja tietoteknistä kehitystyötä. Ongelma on hyvin selvästi tunnistettu sekä määritelty ja tarve sen ratkaisulle tunnistettu GIS:in kehittämisen ja tulevaisuuden kannalta. Esimerkiksi Goodchild ja Zhang (2002) tarkastelevat hyvin yksityiskohtaisesti epämääräisyyden olemusta tarkemmissa maantieteellisissä konteksteissa. Pauly ja Schneider (2004, 2007, 2008, 2010) sekä Kulik (2001) ovat tehneet laajamittaista työtä ongelman parissa kehittämällä epämääräisten objektien käsitettä ja niiden topologisia suhteita sekä spatiaalisia operaatioita sekä yleisluontoisemmin että yhteydessä tässä tutkielmassa käsiteltyihin menetelmiin. Beaubouef ja Petry (2010) sekä Verstraete et.al. (2007) käsittelevät myös *triangular irregular networks*:in (TIN), joka perustuu numeraalisen datan esittämiseen spatiaalisen tason jakamisella toisistaan erillisiin kolmioihin, käyttämistä sumeiden metodien yhteydessä. TIN on yleinen ja kehittynyt digitaalinen maastonmallinnusmenetelmä vektoridatalle (Heywood, Cornelius & Carver, 1998, p. 57).

Maantieteellisen epämääräisyyden pääfokuksena vaikuttavat olevan erityisesti epämääräiset objektit, tarkemmin epämääräiset alueet, etenkin laidoiltaan epämääräiset alueet, ja esimerkiksi erilaiset vaihettumisvyöhykkeet. Tämä selittää sumeiden joukkojen menetelmien suosiota GIS:in yhteydessä, sillä se muistuttaa luonnollisesti tällaista maantieteellistä epämääräisyyttä (Burrough, 1996, p. 18; Goodchild & Zhang, 2002, p. 168). Sumeiden joukkojen menetelmillä on GIS:in piirissä yleisenä trendinä asema sinä viitekehyksenä, jonka piiristä vastausta odotetaan (Pauly, & Schneider, 2007). Tästä huolimatta tämä lähestymistapa ei ole toistaiseksi tutkimuksen perusteella saavuttanut sellaista asemaa ja sellaista käytännön toteutusmuotoa, että sitä voisi pitää *de facto* vastauksena maantieteellisen epämääräisyyden ongelmaan. Suosiosta ja vahvasta teoreettisesta pohjasta huolimatta sumeiden joukkojen menetelmille ei vielä toistaiseksi ole tiedossa käytännön implementaatiota, saati standardisoitunutta käytäntöä (Kanjilal, Liu, & Schneider, 2010; Schneider, 2014). Käytännössä sumeiden joukkojen teorian soveltaminen spatiaalisessa ympäristössä on osoittautunut ongelmalliseksi (Pauly, & Schneider, 2007).

Myöskään muut menetelmät eivät tiedettävästi ole toteuttaneet vielä käytännön sovellutuksia (Pauly, & Schneider, 2008). Laajennetut mallit sinänsä pystyisivät hyötymään nykyisistä tekniikoista, datarakenteista ja algoritmeista, mutta etenkin kolmiarvoisen logiikan tapauksessa herää kysymys, ovatko tällaiset menetelmät riittäviä vastaamaan maantieteellisen epämääräisyyden ongelmaan. Ne ovat kaikista yksinkertaisin taso ongelmaan vastaamiseen siinä, että ne laajentavat yksiselitteisen binäärilogiikan sisältämään myös tuon epämääräisyyden komponentin. Mutta olennainen kysymys on, riittääkö tuo yksitasoinen epämääräisyys, joka ei kykene esittämään epämääräisyyden tasoa tai osallisuuden voimakkuutta, vastaamaan esimerkiksi päätöksenteon tarpeisiin. Siinä, missä kolmiarvoisia malleja laajennetaan moniarvoisen logiikan piiriin, esimerkiksi laakioaluemallissa (Kanjilal, Liu, &

Schneider, 2010), ne alkavat kärsiä samoista ongelmista kuin sumeiden joukkojen menetelmät, esimerkiksi objektien raskaudesta, datan ja karttatasojen määrän kasvamisesta, sekä epämääräisyyden määrittelemisen ongelmasta. Dilo et.al. (2006) pyrkivät hyödyntämään kehittelemässään GRASS-järjestelmässä jo olemassaolevia GIS-rakenteita, esimerkiksi epämääräisen osallisuusarvon tallentamista kolmiulotteisen GIS-järjestelmän korkeuskoordinaattiin.

Eräs teorian ja käytännön välinen kuilu liittyy siihen, että spatiaalisen datan mallintaminen nojaa pääosin euklidisen avaruuden ja geometrian teoriaan ja siten äärettömään tarkkaan aritmetiikkaan, joka taas on ristiriidassa äärellisen tarkkuuden tietokonejärjestelmiin ja äärellisen kapasiteetin tietokantamallinnukseen (Schneider, 2003). Kanjilal, Liu ja Schneider (2010) ovat toisaalta myös kritisoineet sumeiden joukkojen menetelmän riittämättömyyttä objektien sisäisen epämääräisyyden, ei pelkästään rajan epämääräisyyden mallintamisessa. Karkeiden joukkojen menetelmät toisaalta vastaavat epämääräisyyteen rasteridataympäristössä, jossa objektien esitys perustuu niiden ilmenemiseen rasteriruudukolla. Rasteridatalla on tyypillisempää esittää skaalasta riippuen liukuvia muuttujia kuten asukastiheyttä tai korkeuseroja, siinä missä objektien ja siten epämääräisten objektien esittämisen kysymys liittyy olennaisesti vektoridataan (Petry, & Elmore, 2015). Rasterimuotoinen data on kuitenkin hyvin olennainen osa GIS-sovelluksia ottaen huomioon sen käyttökelpoisuuden kaukokartoituksen ja kuvadatan konvertoinnissa (Goodchild & Zhang, 2002, p. 15). Schneider (2003) ehdottaa euklidiseen avaruuteen perustuvan mallin sijaan tietynlaiseen ruudukkojakoon (*grid partition*) sekä vektori- ja rasteridatan käsittelyn yhdistämiseen perustuvaa mallia ja epämääräisyyden implementointia siinä.

Aineistoa tarkastelemalla tulee vaikutelma fokuksen puutteesta. Sumeiden joukkojen menetelmät nousevat esiin ensisijaisena ja suosituimpana vaihtoehtona, mutta yhtäältä tunnustetaan sen ongelmallisuus käytännön sovelletuksissa. Samalla muut menetelmät elävät vahvoina vaihtoehtoina. Epämääräisyyden hallinnan varsinaisista ääneenlausutuista tavoitteista ja siten keinoista ei vaikuta olevan yhdenmukaista ajatusta. Paitsi ettei tutkimuskenttä ole kyennyt luomaan mitään käytännön toimintaan implementoitavia sovelletuksia ja standardeja, ei se ole myöskään päässyt selvyyteen siitä, mikä tuon ongelman ratkaisumuoto tulisi olla. Täytyy korostaa, että useat esitetyt mallit ovat kuitenkin väitetysti sellaisenaan implementoitavia joko nykyisissä GIS-järjestelmissä tai niiden päälle rakennetuissa moduuleissa. Ratkaisuista sinänsä ei siis ole pulaa. Samaiset mallit eivät kuitenkaan käsittele tuota käytännön toteutuksen puolta, esimerkiksi tietokantarakenteita, mallien suoriutumista spatiaalisista operaatioista tai ongelmia tehokkuuden ja tietokantakyselyiden suhteen, taikka sitä, miten mallit vastaavat niihin käytännön ongelmiin, joita epämääräisyys GIS-yhteisössä aiheuttaa. Käytännön sovellutusten niukkuus tulee ilmi myös aineiston luonteessa ja fokuksessa. Mallit ovat luonteeltaan hyvin teoreettisia ja pidättäytyvät aiheen matemaattisessa käsittelyssä. Useimmissa artikkeleissa keskitettyyn erilaisten menetelmien periaatteelliseen soveltamiseen, mutta varsinaisen tietokannan rakenteeseen ja sitä kautta kyselyiden muotoon ei juurikaan puututa. Kuitenkin jotain kehitystyötä myös tällä saralla on saavutettu. Beaubouef ja Petry (2005) hahmottelevat artikkelissaan oliotietokantojen ja luokkaperiytymisen hyödyntämistä karkeiden joukkojen menetelmissä. Morris (2003) on kehittänyt olio-orientoituneen tietokannan prototyypin, joka hyödyntää sumeiden joukkojen menetelmää spatiaalisen läheisyyden ja useiden muuttujien laskennassa. Sözer ja Yazici (2004) ovat kehittäneet sumeaa spatiaalista indeksointijärjestelmää, jonka on tarkoitus nopeuttaa sumeita spatiaalisia kyselyitä.

Tietokantakyseilyiden suhteen keskitytään usein tämän tutkielman ulkopuolelle jätetyn luonnollisen kielen implementointiin sumeuden yhteydessä.

Vahvin tutkimuksen nykytilan perusteella ilmenevä asia on tuo maantieteellisen epämääräisyyden ongelman hyötyminen sen laajasta teoriapohjasta. Tutkimus aiheen ympärillä on jatkunut systemaattisena jo pitkään ja teoriaa on saatu hiottua. Monet teorian osat ovat jo sellaisenaan implementoitavassa muodossa. Käytännön toteutus kaipaa kuitenkin monien eri osa-alueiden kehitystyötä ja näiden osa-alueiden saattamista yhteen optimaaliseen ja GIS-yhteisön tarpeita vastaavaan toteutukseen. Etenkin sumeiden joukkojen metodien osalta kehitystyössä on tehtävää käytännön toteutuksissa, standardien luomisessa, datarakenteissa, kyselyiden toteuttamisessa ja tehokkuuden parantamisessa. Tehokkuus on periniteisesti ollut sumeiden tietokantajärjestelmien ongelma (Morris, 2003). Eräs tutkimushaara on erilaisten menetelmien yhdistäminen hyvien ja huonojen puolien optimoinniksi. Ratkaisut keskittyvät järjestelmällisesti kaksiulotteisen ja temporaalisesti staattisen datan käsittelyyn, joten kolmiulotteisuuden ja temporaalisuuden mukaan ottaminen epämääräisyyteen ovat toinen tulevaisuuden tutkimuksen ala. Lisäksi vektori- ja rasteridatan käsittely epämääräisyyden suhteen vaatii joko kummallekin erillisiä menetelmiä tai vaihtoehtoisesti yhdistävää menetelmää.

5. Yhteenveto

Tämä tutkielma pyrkii hahmottamaan maantieteellisen epämääräisyyden ilmenemistä ja hallintaa maantieteellisissä tietokannoissa ja GIS-käyttöympäristössä. Sen tavoitteena on kartoittaa, miten maantieteellinen epämääräisyys ymmärretään spatiaalisten tietokantojen kontekstissa, mitä tuo epämääräisyys aiheuttaa, ja millä menetelmillä tuota epämääräisyyttä pyritään mallintamaan ja esittämään spatiaalisten tietokantojen ja GIS:in piirissä; mikä on tämän kysymyksen parissa toimivan tutkimuksen nykyinen tilanne, mitä puutteita ja minkälaisia tutkimusavauksia on havaittavissa tulevaisuuden tutkimuksen kannalta. Tutkimusmenetelmänä on kirjallisuuskatsaus. Tarkemmin tutkielma pyrkii vastaamaan seuraaviin tutkimuskysymyksiin:

Miksi epämääräisyys on olennaista maantieteen kontekstissa?

Maantieteellinen todellisuus sisältää väistämätöntä epämääräisyyttä (Goodchild & Zhang, 2002, p. 6). Jotta tuota todellisuutta kyettäisiin kuvaamaan ja siten hyödyntämään, täytyy tuo epämääräisyys ottaa huomioon. Tällä on merkitystä etenkin päätöksenteossa, joka pohjautuu siihen dataan ja esitysmalliin, joka tuosta todellisuudesta on käytettävissä, ja siten myös tuon päätöksenteon oikeellisuus perustuu datan ja esitysmallin oikeellisuuteen. Pohjimmiltaan kysymys palautuu mallintamistarkkuuden ja ihmisperspektiivin ongelmiin. Maantieteellisen todellisuuden kartoittamisessa ja mallintamisessa on käytössä varsin rajallinen kapasiteetti ensinäkin datan keräämisen ja toisaalta sen prosessoinnin ja esittämisen tarkkuuden suhteen. Ihmisperspektiivi taipuu käsittelemään maantieteellistä todellisuutta erilaisten objektien kautta, kuten esimerkiksi ”vuori”, ”metsä”, ”sadealue” tai ”kaupungin keskusta”. Nämä objektit ovat kuitenkin ihmiskategorisointeja vailla yksiselitteistä viitettä tuossa maantieteellisessä todellisuudessa ja sellaisina luonteeltaan hyvin epämääräisiä.

Miten epämääräisyyden ongelma ilmenee maantieteellisissä tietokannoissa?

Perinteisesti maantieteelliset tietokannat ja GIS ovat perustuneet binääriseen logiikkaan, jossa esimerkiksi jokin karttapiste yksiselitteisesti joko kuuluu tai ei kuulu johonkin objektiin. Näin GIS-tietokanta rakentuu erilaisista karttatasoista, joihin erilaiset yksittäiset objektit kuuluvat, ja joita voidaan käsittely- ja esitysvaiheissa valikoida. Tällainen perinteisissä GIS-tietokannoissa käytetty binäärinen logiikka ei sellaisenaan tue objektien epämääräisyyttä, jossa karttapisteen kuuluminen objektiin ei ole yksiselitteistä, vaan epämääräistä, sumeaa, osittaista tai epävarmaa. (Goodchild & Zhang, 2002, p. 91)

Millä keinoin tätä epämääräisyyttä pyritään hallitsemaan ja mallintamaan?

Vastausta epämääräisyyden hallintaan maantieteellisissä tietokannoissa on haettu useista erilaisista menetelmistä. Tyypillisimmät aineistossa ilmenevät menetelmät ovat sumeisiin joukkoihin perustuvat menetelmät, karkeisiin joukkoihin perustuvat menetelmät, todennäköisyysmallit ja laajennetut mallit. Aineiston perusteella selvästi yleisimmät ja lupaavimmat menetelmät perustuvat matemaattiseen sumeiden joukkojen teoriaan, jossa karttapisteille annetaan binäärisen logiikan sijaan osallisuusarvo kuhunkin objektiin välillä [0,1]. Toinen hyvin yleinen menetelmien joukko on

matemaattiseen karkeiden joukkojen teoriaan perustuvat menetelmät, joka perustuu ylempiin ja alempiin approksimaatioluokkiin. Laajennetut mallit pyrkivät hyödyntämään nykyisiä GIS-järjestemien rakenteita ja tarkkoja alueita, sekä moniarvoista logiikkaa epämääräisyyden mallintamisessa. Tämän tutkielman käsittelyn ulkopuolelle jätetyt todennäköisyysmenetelmät perustuvat epämääräisyyden hallinnalle sen ennustettavuuden perusteella. (Schneider, 1999, 2008; Pauly & Schneider, 2004, 2008).

Tutkielmassa paneudutaan näihin menetelmiin ja niiden erityispiirteisiin, sekä niiden asemaan yleisemmässä pyrkimyksessä vastata maantieteellisen epämääräisyyden ongelmaan. Pääasiallisena löydöksenä tutkimusaineistosta käy ilmi teorian vahvuus niin maantieteellisen epämääräisyyden kysymyksen suhteen yleisemmin kuin eri menetelmien osalta. Tutkimuskentän vajaavaisuudet taasen liittyvät teorian soveltamiseen käytäntöön, yleisesti hyväksyttyjen ja todettujen menetelmien sekä standardien kehittämiseen.

Tämän tutkielman kirjallisuuskatsaus toimii verrattain yleispiirteisenä näkökulmana aiheeseen. Tämä ei ole systemaattinen kirjallisuuskatsaus, vaan suppeampi, intuitiivinen kirjallisuuskatsaus, joka pyrkii muodostamaan hyvän yleiskuvan aiheen tutkimuksen nykykentästä ja mahdollisista tulevaisuuden suuntauksista. Katsantoon on valittu olennaisimmat ja merkittävimmät menetelmät ja tutkailtu niiden erityispiirteitä sekä hyviä ja huonoja puolia. Lisätutkimuksessa olisi mahdollista paneutua syvemmin paitsi erilaisiin marginaalisempiin vaihtoehtoihin ja näkökulmiin tutkimuskysymyksen osalta, myös tässä tutkielmassa käsiteltyihin menetelmiin, niiden erityispiirteisiin esimerkiksi filosofisten lähtökohtien ja objektiivisuuteen suhtautumisen suhteen, spatiaalisen algebran, topologisten suhteiden ja operaatioiden suhteen, tietokantarakenteen suhteen esimerkiksi luokkarakenteina, tai epämääräisyyden huomioimisen suhteen tietokantakyseilyissä.

Lähteet:

Abler, R. F. (1987). The national science foundation national center for geographic information and analysis. *International Journal of Geographical Information System*, *1*(4), 303-326.

Beaubouef, T., & Petry, F. E. (2005). Representation of spatial data in an OODB using roughand fuzzy set modeling. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, *9*(5), 364-373.

Beaubouef, T., & Petry, F. E. (2009). Uncertainty modeling for database design using intuitionistic and rough set theory. *Journal of Intelligent & Fuzzy Systems*, *20*(3), 105-117.

Beaubouef, T., & Petry, F. E. (2010). Fuzzy and Rough Set Approaches for Uncertainty in Spatial Data. *Methods for Handling Imperfect Spatial Information*, 103-129.

Beaubouef, T., Petry, F. E., & Ladner, R. (2007). Spatial data methods and vague regions: A rough set approach. *Applied Soft Computing*, *7*(1), 425-440.

Burrough, P. A. (1986). Principles of geographical information systems for land resources assessment.

Burrough, P. A. (1996). Natural objects with indeterminate boundaries. *Geographic objects with indeterminate boundaries*, *2*, 3-28.

Carniel, A. C., Schneider, M., Ciferri, R. R., & de Aguiar Ciferri, C. D. (2014). Modeling fuzzy topological predicates for fuzzy regions. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 529-532). ACM.

Clementini, E., & Di Felice, P. (2001). A spatial model for complex objects with a broad boundary supporting queries on uncertain data. *Data & Knowledge Engineering*, *37*(3), 285-305.

Cohn, A. G., & Gotts, N. M. (1996). The ‘egg-yolk’representation of regions with indeterminate boundaries. *Geographic objects with indeterminate boundaries*, *2*, 171-187.

Connolly, T. M., Begg, C. E., & Strachan, A. (1996). *Database systems: a practical approach to design, implementation, and management*. Pearson Education.

Dilo, A., Bos, P., Kraipeerapun, P., & Rolf, A. (2006). Storage and manipulation of vague spatial objects using existing GIS functionality. In *Flexible Databases Supporting Imprecision and Uncertainty* (pp. 293-321). Springer, Berlin, Heidelberg.

Dilo, A., De By, R. A., & Stein, A. (2007). A system of types and operators for handling vague spatial objects. *International Journal of Geographical Information Science*, *21*(4), 397-426.

Elmasri, R. & Navathe, S. (2004) Fundamentals of Database Systems. Pearson Education.

Erwig, M., & Schneider, M. (1997, July). Vague regions. In *International Symposium on Spatial Databases* (pp. 298-320). Springer, Berlin, Heidelberg.

Goodchild, M. & Gopal, S. (1989) Accuracy of Spatial Databases. Taylor & Francis.

Goodchild, M. & Zhang, J. (2002) Uncertainty in Geographical Information. Taylor & Francis.

Heywood, I., Cornelius, S. & Carver, S. (1998) An Introduction to Geographical Information Systems. Pearson Education.

Kanjilal, V., Liu, H., & Schneider, M. (2010). Plateau regions: An implementation concept for fuzzy regions in spatial databases and GIS. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 624-633). Springer, Berlin, Heidelberg.

Kulik, L. (2001). A geometric theory of vague boundaries based on supervaluation. In *International Conference on Spatial Information Theory* (pp. 44-59). Springer Berlin Heidelberg.

Morris, A. (2001). Why spatial databases need fuzziness. In *IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th* (Vol. 4, pp. 2446-2451). IEEE.

Morris, A. (2003). A framework for modeling uncertainty in spatial databases. *Transactions in GIS*, *7*(1), 83-101.

Pauly, A. & Schneider, M. (2004) Vague spatial data types, set operations, and predicates. Lecture Notes in Computer Science, Volume 3255, 379-392.

Pauly, A. & Schneider, M. (2007). Rosa: An algebra for rough spatial objects in databases. *Rough Sets and Knowledge Technology*, 411-418.

Pauly, A., & Schneider, M. (2008). Spatial vagueness and imprecision in databases. In *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 875-879). ACM.

Pauly, A., & Schneider, M. (2010). VASA: An algebra for vague spatial data in databases. *Information Systems*, *35*(1), 111-138.

Pawlak, Z. (1982). Rough sets. *International Journal of Parallel Programming*, *11*(5), 341-356.

Petry, F., & Elmore, P. (2015). Geospatial Uncertainty Representation: Fuzzy and Rough Set Approaches. In *Fifty Years of Fuzzy Logic and its Applications* (pp. 483-497). Springer International Publishing.

Pratt, P. & Adamski, J. (1987) Database Systems: Management and Design. Boyd & Fraser.

Schneider, M. (1999). Uncertainty management for spatial datain databases: Fuzzy spatial data types. In *International Symposium on Spatial Databases* (pp. 330-351). Springer, Berlin, Heidelberg.

Schneider, M. (2003). Design and implementation of finite resolution crisp and fuzzy spatial objects. *Data & Knowledge Engineering*, *44*(1), 81-108.

Schneider, M. (2008). Fuzzy Spatial Data Types for Spatial Uncertainty Management in Databases. *Handbook of research on fuzzy information processing in databases*, *2*, 490-515.

Schneider, M. (2014). Spatial Plateau Algebra for implementing fuzzy spatial objects in databases and GIS: Spatial plateau data types and operations. *Applied Soft Computing*, *16*, 148-170.

Silberschatz, A., Korth, H. & Sudarshan, S. (2006) Database System Concepts. McGraw-Hill.

Stefanakis, E., Vazirgiannis, M., & Sellis, T. (1999). Incorporating fuzzy set methodologies in a DBMS repository for the application domain of GIS. *International Journal of Geographical Information Science*, *13*(7), 657-675.

Sözer, A., & Yazici, A. (2004). Index structures for flexible querying in fuzzy spatial databases. In *Fuzzy Systems, 2004. Proceedings. 2004 IEEE International Conference on* (Vol. 1, pp. 559-564). IEEE.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*(sup1), 234-240.

Verstraete, J., De Tré, G., Hallez, A., & De Caluwe, R. (2007). Using tin-based structures for the modelling of fuzzy gis objects in a database. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *15*(supp01), 1-20.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338-353.