Laskennallinen yhteiskuntatiede politiikan tutkimuksessa

Matti Nelimarkka

Helsinki 12.8.2014

LuK –tutkielma

HELSINGIN YLIOPISTO
Tietojenkasittelytieteen laitos

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkasittelytieteen laitos

Tekijä — Författare — Author

Matti Nelimarkka

Työn nimi — Arbetets titel — Title

Laskennallinen yhteiskuntatiede politiikan tutkimuksessa

Oppiaine — Läroämne — Subject

Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
LuK –tutkielma	12.8.2014	27 sivua ja 2 liitesivua

Tiivistelmä — Referat — Abstract

Tutkielmassa esitellään laskennallisen yhteiskuntatieteen (*computational social science*) käsitettä ja esitellään laskennallisten mentelmien käyttöä politiikan tutkimuksen kirjallisuudessa. Lisäksi valittuja menetelmiä esitellään tarkemmin, jonka jälkeen keskustellaan laskennallisten yhteiskuntatieteiden haasteista tulevaisuudessa.

Kirjallisuuskatsauksen perusteella esitellään politiikan tutkimuksen piirissä käytetyn laskennallinen menetelmä on erilaiset simulaatiomallit, erityisesti agenttipohjaiset simulaatiot nousevat esille käytettynä menetelmänä. Toinen tunnistettu alue laskennallisille menetelmille on teksin analysointi, missä sentimenttianalyysi ja teemojen laskennallinen luokitus nousevat esille.

Systemaattisen kirjallisuuskatsauksen ulkopuolelta nostan esille vielä menetelmiä säännönmukaisuuksien ja ryhmien löytämiseen: assosiaatiosäännöt sekä klusterointialgoritmit. Näitä ohjaamattoman oppimisen menetelmiä ei ole (vielä) sovellettu politiikan tutkimuksen merkittävemmissä lehdissä.

Työn päätteeksi nostan esille laskennallisen yhteiskuntatieteen kaksi haastetta: toisaalta kysymykset liittyen validiteetiin ja reliabiliteettiin ja edelleen haasteet liittyen poikkitieteelliseen työskentelyn. Ensimmäiseen ongelmaan kirjallisuus suosittaa ratkaisuksi laskennallisen tuloksen verifioimista, esimerkiksi vertailemalla tuloksia koeasetelmiin, perinteisin menetelmin saatuihin tuloksiin sekä olemassa oleviin vastaaviin ilmiöihin. Jälkimmäinen ongelma kiteytyy siihen, että tietojenkäsittelytieteilijän ja yhteiskuntatieteilijän tavoitteet ovat erilaiset, eräs ratkaisu tähän voisi olla soveltaa käyttäjäkeskeistä suunnittelua osana laskennallisten mallien kehitystä.

Avainsanat — Nyckelord — Keywords

laskennallinen yhteiskuntatiede, agenttipohjainen mallinnus, ohjattu koneoppiminen, ohjaamaton koneoppiminen

Sisältö

1 Johdanto

Laskennallisten menetelmien käyttö tutkimusmetelmänä muiden tieteenalojen osana on merkittävä sovellusalue tietojenkäsittelytieteille. Esimerkiksi bioinformatiikka sekä laskennallinen biologia keskittyvät aineiston analyysiin tarvittavien algoritmien ja tiedonlouhinnan kehitykseen. Vastaavasti laskennallisessa kielitieteessä pyritään löytämään laajoista tekstimassoista *(korpuksista)* säännönmukaisuuksia sekä kehittää kuluttajille hyödyllisiä sovelluksia automaattisen kääntämisen piirissä. Opetusalalla tiedonlouhinta mahdollistaa kehittyneempien vuorovaikutteisten sovellusten kehittämisen: tiedonlouhinnan avulla mallinnetaan opiskelijan oppimista. Myös yhteiskuntatieteiden piirissä on herännyt mielenkiintoa soveltaa tietojenkäsittelytieteen menetelmiä ja osaamista tutkimuksen osana, esimerkiksi aineiston käsittelyssä, säännönmukaisuuksien etsimisessä ja yksilön sekä ryhmien toiminnan mallintamisessa.

Viimeaikoina tutkimuskentällä – varsinkin tietojenkäsittelytieteilijöiden osalta – on noussut esille suurien tietomassojen käsittely laskennallisesti *(big data analysis)*. Erityisesti viimeaikaisen innostuksen taustalla on niin Internetin kautta (Adamic & Glance, 2005; Notess, 2002) kuin elinympäristöstä kerättyjen aineistojen (Eagle & Pentland, 2006; Oulasvirta et al., 2012) kerääminen sekä tarjoaminen muille tutkijoille. Esimerkiksi sosiologian ja sosiaalipsykologian tutkimuksessa mahdollisuus seurata ihmisten välistä viestintää, sijaintia ja muita tietoja *(reality mining)* mahdollistaa esimerkiksi ystävyyssuhteiden tarkastelun uusilla tavoilla (esimerkiksi Karikoski & Nelimarkka, 2011; Nelimarkka & Karikoski, 2012). Kuitenkin suurien tietomassojen perusteella tehtävää tutkimusta kohtaan on nostettu esille myös kritiikkiä. Esimerkiksi Boyd & Crawford (2012) huomauttavat, että suuret tietomassojen laatu ja hyödyllisyys riippuvat aineistossa käytössä olevista mittareista sekä ympäristön ja taustalla olevien ilmiöiden ymmärtämisestä.

Suurten tietomassojen laskennallinen käsittely on vain osa laskennallisten menetelmien mahdollisista käyttösovelluksista yhteiskuntatieteessä: laskennallisia menetelmiä voidaan käyttää myös perinteisen aineiston käsittelyn tukena ja apuna, käyttäen menetelmiä pienten aineistojen analyysin apuna (Ahonen, 2014). Laskennallisten menetelmien mielenkiintoisuutta voidaan perustella kahdella eri syyllä

1. laskennalliset menetelmät mahdollistavatuusien menetelmien käytön ja mahdollistavat näin *uudenlaisten kysymysten* kysymisen
2. laskennalliset menetelmät *tehostavat* aineiston käsittelyä ja näin esimerkiksi vähentävät tarvetta manuaaliselle työlle

Erityisesti mielenkiintoisuutta nyt lisää se, että tietotekniikka on jokapäiväistynyt ja laskennallisten menetelmien soveltaminen on yleistynyt muiden tieteiden parissa. Laskennallisten menetelmien soveltaminen on myös nousemassa oleva trendi yhteiskuntatieteissä, osittain työtä tehdään tietojenkäsittelytieteen

puolelta – mutta enevemissä määrin myös yhteiskuntatieteilijät soveltavat laskennallisia menetelmiä. Tarpeen onkin pyrkiä esittelemään laskennallisia menetelmiä ymmärrettävästi myös suomen kielellä.

Kirjallisuuden rajaamiseksi tässä työssä yhteiskuntatieteiden osalta keskitytään politiikan tutkimukseen. Ennen tarkempaa paneutumista laskennallisuuteen ja tietojenkäsittelytieteeseen on tarpeen esittää politiikan tutkimuksen oppihistoriaa ja nykyisin vakiintuneita menetelmiä lyhyesti, jolloin laskennalliset menetelmät voidaan sijoittaa osaksi politiikan tutkimuksen traditiota. Tämän jälkeen esitellään olemassa olevia määritelmiä laskennalliselle yhteiskuntatieteelle ja esitellään kirjallisuuskatsauksen tuloksia laskennallisten menetelmien käytöstä yhteiskuntatieteissä. Kolme menetelmäryhmää esitellään tarkemmin niin sovellutusten kuin laskennan pohjalta: simulaatiot, tekstianalyysi ja säännönmukaisuuksien etsiminen. Kirjallisuuskatsauksen perusteella keskustellaan laskennallisen yhteiskuntatieteen haasteista.

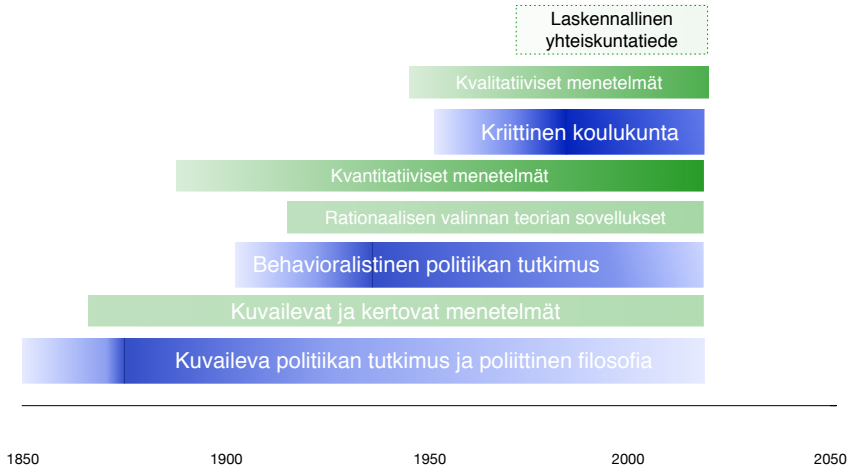
1.1 Politiikan tutkimuksen oppihistoriaa

Politiikan tutkimuksen ensimmäiset teemat keskittyivät historialliseen sekä normatiiviseen tutkimukseen, esimerkiksi aatehistorian sekä käsitteiden määrittelyyn. Politiikan tutkimuksen toinen vaihe 1900–luvun alussa kehitti lähestymistapoja empiirisemmäksi ja tutkimusmenetelmissä pyrittiin lähestymään luonnontieteitä (Berndtson, 2009). Edelleen, menetelmien lisäksi tutkimuskysymykset muuttuivat tutkimaan ihmisten toimintaa ja käyttäytymistä, jonka takia tästä komannesta vaiheesta käytetään nimeä *behavioralismi*.

Behavioralismi myös esitti oletuksia ihmisten käyttäytymisestä, joka yhdistettynä luonnontieteelliseen toimintatapaan johdatti politiikan tutkimuksen kohti erilaisten mittareiden kehitystä ja muuttujien operationalisointia. Behavioralistinen tutkimus voisi siis soveltaa laskennallisen tieteen menetelmiä: esimerkiksi peliteoria ja sen sovellukset olivat tyypillisiä menetelmiä mallintaa ilmiöitä, ja näiden ajatusmallien siirto laskennalliseen muotoon on mahdollista, ja jopa mielekästä.

Tämä nosti esille vastaliikkeitä, joita yhdessä kuvataan termillä kriittinen koulukunta. Kriittisen koulukunnan mukaan yhteiskunnan tutkimuksessa on tarpeen laajempi näkökulma teemoihin. Samoin kriittinen koulukunta kyseenalaisti behavioralismin objektiivisuuden, argumentoiden, että tutkimuksen operationalisoinnit luovat jo subjektiivisen ulottuvuuden ja tulosten tulkinta on edelleen subjektiivista. Tietenkin, myös kriittisen koulukunnan työskentelyä kuvaa subjektiivisuuden keskeinen luonne tutkimuksen osana – mutta, tämä subjektiivisuus onkin usein osa kaikkea yhteiskuntatiedettä. Subjektiivisuuden kehitys osaksi laskennallisia malleja voi kuitenkin olla haastavaa, vaikkei mahdotonta.

Yksinkertaistetusti politiikan tutkimuksen historia voidaan esittää kolmen vaiheen kautta, kuten kuvassa 1 on tehty. Vaihteet ovat liittyneet myös menetelmäkehitykseen, joka on liitetty osaksi alan oppihistoriaa kuvassa. Modernissa



Kuva 1: Politiikan tutkimuksen oppihistoriaa tiivistetysti

empiirisessä yhteiskuntatieteesä voidaan siis nostaa esille kahden laajemman lähestymistavan soveltaminen tutkimuskysymyksiin. Tutkimuksessa sovelletaan yleensä laadullista (kvalitatiivinen) tai määrällistä (kvantitatiivinen) tutkimusotetta, missä ensimmäisessä pääpaino on enemmän kuvailevassa aineistossa ja sen tulkinnassa, jälkimmäinen taas pyrkii mitattaviin aineistoihin ja selittämiseen matemaattisia menetelmiä hyödyntäen. Sellaisenaan voimakasta menetelmäjakoa on kritisoitu, koska useiden ilmiöiden kohdalla olisi mielekästä soveltaa molempia tutkimusmenetelmiä: laadullisen tutkimuksen perinteinen haaste on tutkimuksen yleistettävyys koko joukkoon, kun taas määrällisen tutkimuksen tapauksessa kausaalisen tekijän keskeinen luonne voi helposti jäädä selvittämättä (Ragin et al., 1996; McGraw, 1996; Silverman, 2000).

Kritiikkini kausaalisuuden puutteesta voi olla yllättävä, joten valaistaan tilannetta esimerkillä: on havaittu, että poliitikot eivät suosi vuorovaikutuksellista ja kaksisuuntaista verkkomedian käyttöä, vaan käyttävät verkkomediaa perinteisen median jatkona (Golbeck et al., 2010). Kuitenkaan, nämä määrällistä lähestymistapaa soveltavat tutkimukset eivät täysin pysty selittämään miksi näin on. Tällöin on tarve laadulliselle tutkimukselle, kuten Stromer-Galley (2000) havannointityölle ehdokkainen parissa, missä hän esittää, että ehdokkaat pelkäävät interaktiivisen viestinnän altistavan heidät kyseenalaistamiselle ja mahdollisesti sotkevan kampanjaviestintää. Näin vakuuttavaan yhteiskuntatieteelliseen argumentaatioon vaaditaan toisaalta määrällistä, koko tutkittavan joukon kattavaa havaintoa sekä laadullista työskentelyä, mikä joko selvittää havaintoja – kuten yllä – tai nostaa esille uusia tutkimuskysymyksiä.

1.2 Laskennallinen yhteiskuntatiede

Määrällisestä lähestymistapaa edustavat tilastollisten menetelmien soveltaminen aineiston käsittelyssä, mutta myös formaalit menetelmät, kuten peliteoria

voidaan laskea osaksi tätä lähestymistapaa. Myös tarkastelun kohteena oleva laskennallinen yhteiskuntatiede (*computational social science*) edustaa määrällistä lähestymistapaa, mikä on ilmeistä kirjallisuudessa esitettyjen laskennallisten yhteiskuntatieteiden määritelmistä:

Cioffi-Revilla (2010) määrittelee laskennallisen yhteiskuntatieteen laajasti seuraavien laskennallisten menetelmien soveltamisena: tiedon uuttamisen menetelmät (*information extraction*), sosiaalisten verkostojen analyysi, paikkatietojärjestelmien käyttämisen ja mallinnuksen sekä simulaation menetelmät. Hän myös arvioi, että tiedon visualisointimenetelmät voivat myöhemmin laajentua käytettäväksi laskennallisen yhteiskuntatieteen menetelmänä. Hän siis näkee, että laskennallista tiedettä voidaan määritellä tiettyjen menetelmien käyttönä.

Lazer et al. (2009) taas esittää laskennallisen yhteiskuntatieteen laajojen aineistomäärien keräämisenä ja käsitttelynä määrällisten menetelmien avulla. He nostavat esille, että laskennallinen yhteiskuntatiede on aineistolähtöistä (*data-driven*), minkä tulkitsen tarkoittavan, että perinteinen hypoteeseihin ja niiden tarkasteluun perustuvat tilastolliset menetelmät eivät olisi laskennallista yhteiskuntatiedettä.

Bankes et al. (2002) kuvaavat laskennallisia epistemologioita (*computational epistemology*), eli käsityksiä hyvästä tieteestä. Toisaalta he ehdottavat, että laskennallinen yhteiskuntatieteen tarkoitus on luoda selittäviä malleja, joiden avulla yhteisöä voidaan tutkia. Toisaalta, he nostavat esille myös mahdollisuuden käyttää näitä malleja ei vain yhteisön tutkimiseen vaan myös laskennallisten koeasetelmien suorittamiseen. Tässä määritelmässä siis tärkeintä on tutkittavan ilmiön mallinnus laskennallisin menetelmin.

Kuten havaitaan, laskennallisten menetelmien käsitteet eivät ole kaikilla tutkijoilla samankaltaisia, johtuen myös erilaista lähestymistavoista laskennallisen yhteiskuntatieteen määritelmään. Kuten Cioffi-Revilla (2010) huomauttaa, laskennallisen tieteen osalta voidaan erottaa laskennallisuus menetelmänä ja laskenta teoreettisena lähtökohtana. Jaottelu ei kuitenkään ole ongelmaton: esimerkiksi nykyisin tilastollisten menetelmien kohdalla käytetään poikkeuksetta laskennallista välineistöä. Toisaalta, sosiaalisen verkostojen analyysin kohdalla taustalla on graafiteoria ja sen sovellukset, vaikkakin niitä suoritetaan laskennallisesti.

1.3 Kirjallisuuskatsaus

Havaitaksemme kuinka erilaisia laskennallisten menetelmien käytetään yhteiskuntatieteissä suoritan kirjallisuuskatsauksen laskennallisten mentelmien käytöstä politiikan tutkimuksessa. Etsin kirjallisuuskatsauksessa termejä “computational social science“, “machine learning“, ‘ìnformation extraction“, “simulation“

	CSS	ML	IE	SIM	DM
APSR	0	0	0	20	0
AJPS	0	0	0	21	0
ARPS	0	0	0	1	0
CPS	0	0	0	2	0
EJPR	0	0	0	2	0
SSCR	12	3	0	46	4

Taulukko 1: Hakuosumat kirjallisuudesta
CSS = Computational Social Science, ML = Machine Learning, IE = Information Extraction, SIM = Simulation, DM = Data Mining

sekä "data-mining" artikkeleista käyttäen Web of Science –tietokantaa. Politiikan tutkimuksessa erittäin arvostetut lehdet ovat American Political Science Review (APSR), American Journal of Political Science (AJPS), Annual Review of Political Science (ARPS), Comparative Political Studies (CPS) sekä European Journal of Political Research (EJPR). Tietokanta sisälsi APSRn numerosta 50 lähtien (vuodesta 1956), AJPSn numerosta 17 (1973), ARPSn sekä CPSn ensimmäisestä numerosta (1998, 1968) ja EJPRn numerosta kolme (1975). Tietokannassa olevat artikkelit olivat julkaistu toukokuuhun 2014 mennessä.

Löydettyjen artikkelien on esitetty taulukossa 1, josta havaitaan ettei laskennallisia menetelmiä käytetä vielä alan merkittävimmissä lehdissä, poikkeuksena simulaatiotutkimukset. Simulaatioon liittyviä 46 artikkelia tarkasteltiin tarkemmin niiden abstraktin pohjalta, jonka perusteella simulaatiotiotutkimukset voidaan jakaa neljään perheeseen: simulaatiot aineiston luomisessa tai tilastollisesessa testauksessa, agenttipohjaiset mallit, Monte Carlo –pohjaiset mallit sekä tapauskohtaiset erityismallit. Kappaleessa 2 esitellään simulaatiomenetelmiin liittyviä tuloksia ja käytettäviä menetelmiä tarkemmin.

Koska käytössä olleet menetelmät olivat varsin vähäiset suhteessa yllä esitettyihin, varsin laajoihin kuvauksiin laskennallisesta yhteiskuntatieteestä, on syytä tarkastella muitakin julkaisuja, erityisesti Social Science Computer Review:tä (SSCR)[1] tietojenkäsittelytieteen ja yhteiskuntatieteen yhteisenä lehtenä. Vaikkei kyseessä ole erityisesti politiikan tutkimuksen lehti, lehden menetelmät ovat sovellettavissa myös politiikan tutkimukseen. Edelleen simulaatiomallien määrä suhteissa muihin menetelmiin on suurehko, mutta erityisalan lehdessä on edustettuina myös muita laskennallisen yhteiskuntatieteen töitä. Jälleen simulaatioiden käyttö on yleisin käytössä oleva menetelmä, tiedon louhinna teemana alla nostetaan esille erilaisia menetelmiä tekstiaineistojen analysoimiseksi, näitä käsitellään tarkemmin kappaleessa 4.

Lisäksi kappaleessa 5 nostetaan esille ohjaamattoman oppimisen *(unsupervised learning)* menetelmiä, joita on havaittu kirjallisuuskatsauksen ulkopuolisessa kansainvälisessä vertaisarvioidussa kirjallisuudessa kahden vuoden seurannan aikana.

[1] Saatavilla numerosta kahdeksan, vuodesta 1990.

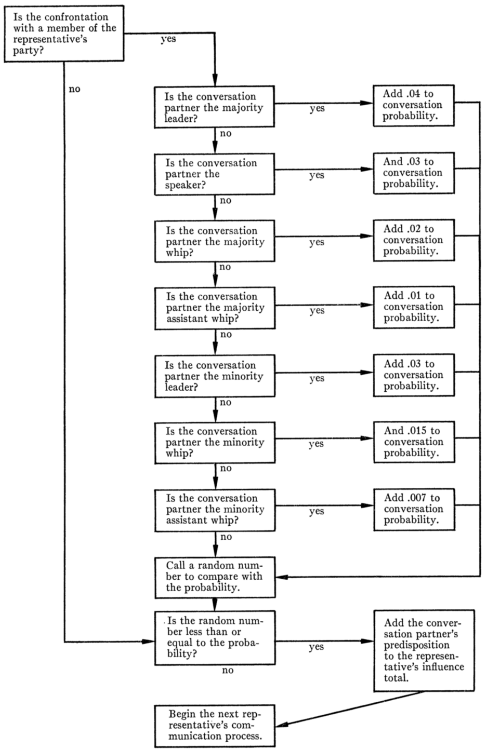
Nostan esille esitettyjä menetelmiä uusina avauksina, joita voitaisiin laskennallisessa yhteiskuntatieteessä soveltaa laajemminkin, vaikkei ne nyt suoritetussa kirjallisuuskatsauksessa nousseet selkeästi esille.

2 Simulaatiot ilmiöiden tarkastelussa

Kirjallisuuskatsauksen pohjalta simulaatiopohjaiset tutkimukset voitiin jakaa neljään ryhmään: simulointi tilastotieteen välineenä (esimerkiksi Clinton et al., 2004; Imai & Tingley, 2012; Mooney, 1996; Slantchev, 2004), toimialakohtaiset mallit ja niiden simulointi (esimerkiksi Shapiro, 1968) sekä agenttipohjaiset mallit (esimerkiksi Orbell et al., 2004; Altaweel et al., 2012; Anderson & Hicks, 2011; Bloomquist, 2006).

Laskennallisen yhteiskuntatieteen kannalta tilastollinen simulointi, esimerkiksi bootsrap-menetelmien käyttö, ei ole mielenkiintoista. Myös toimialakohtaisiin malleihin perustuva simulointi on vanhempaa tutkimusta: esimerkiksi Shapiro (1968) on työssään rakentanut mallin siitä, kuinka Yhdysvaltojen edustajainhuone äänesti vuosina 1963–1964, ja arvio mallin toimintaa suhteissa aitoihin äänestyspäätöksiin. Kuvassa 2 nähdään osa simulaatiomallia. Kyseessä on varsin laaja, mutta yksinkertainen, algoritmi arvioimaan sosiaalisen vuorovaikutuksen merkitystä päätöksentekotilanteessa. Merkittävä ero nykyisiin simulaatiomalleihin on niiden yksinkertaisuus: kullekkin toimijalle luvut ovat samat, eikä satunnaistekijöitä ole mukana mallissa. Tämän yksinkertaisuuden takia nämä mallit eivät ole mielenkiintoisia laskennallisen yhteiskuntatieteen piirissä.

Uudempi simulaatioperinne ottaa paremmin huomioon niin erilaisten toimijoiden kuin satunnaisuuden roolia. Esimerkiksi Orbell et al. (2004) arvioivat yhteistyöhön liittyvien normien kehittymistä useiden vuosisatojen aikana simuloiden normeja ja niiden vaikutusta yhteiskunnassa. Mallin perusteella Orbell et al. argumentoivat, että yhteistyön tapahtumista tukee mahdollisuus arvioida toisen toimijan tilannekuvaa ja motiiveja. Myös Altaweel et al. (2012) käyttävät agenttipohjasta simulointia poliittisten protestien simulointiin ja esittävät mallissaan arvioita erilaisten tekijöiden, kuten rahoituksen ja rangaistusten vaikutusta protestiliikkeiden muodostumisessa. Yllättävää kyllä, kummassakin tapauksessa mallin pohjalta tehtiin varsin vähän ajoja, jolloin mallien satunnaistekijät eivät välttämättä nouse luotettavasti esille. Kuitenkin, muten Villatoro et al. (2013) ovat tehneet, simulaatiota voidaan ajaa usempia kertoja – heidän tapauksessaan 5000 kertaa – tulosten vahvistamiseksi ja simulaation pohjalta muodostettu aineisto on myös luvattu, avoimen tieteen hengessä, tutkijoiden käyttöön. Heidänkin työnsä tarkastelee rankaisun merkitystä normien synnyssä ja ylläpitämisessä. Laajojen yhteiskunnallisten ilmiöiden simuloinnin lisäksi menetelmiä voidaan käyttää niin sairaalan organisaation ja kansanterveystyön ennustamiseen (Pearson et al., 2010) kuin myös veropäätösten vaikutukseen arviointiin (Bloomquist, 2006).



Kuva 2: Toimialakohtaisen mallin simulaation eräs osa (Shapiro, 1968)

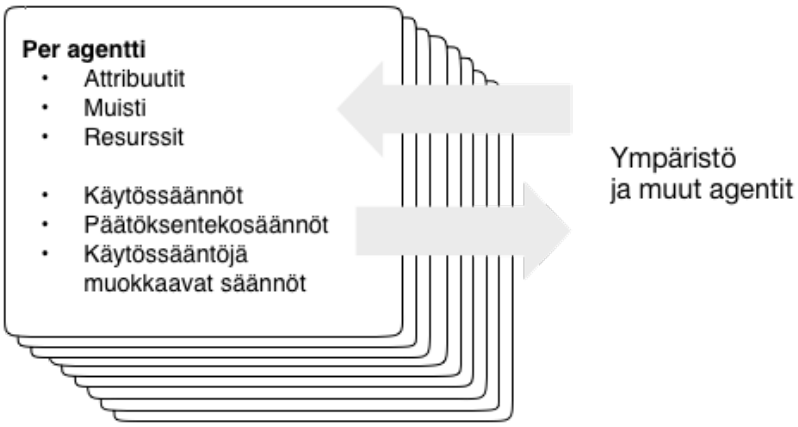
2.1 Agenttipohjainen malli

Agenttipohjaisesissa malleissa (*agent-based models, ABM*) mallinnetaan yhteisön toimintaa yksittäisten toimijoiden kautta. Kukin toimija, eli agentti, on yksilöllinen toimintaheuristiikka, tai prefrenssifunktio $\mathcal{P}$. Tämän preferenssifunktion sekä mahdollisen satunnaisfuntion avulla määritellään agentin tila t seuraavalla ajanhetkellä:

$$t_{i,(t+1)} = \mathcal{P}(t_{i,t}) + \varepsilon.$$

Toistamalla sama simulaatio järjestelmän kaikille agenteille, voidaan määrittää järjestelmän tila ajanhetkellä $t+1$ ajanhetken t perusteella. Luonnollisesti preferenssifuntio $\mathcal{P}$ voi ottaa huomioon myös muiden agenttien toimeinpiteet ja tästä syntyneen järjestelmän tilan tehdessään päätöstä tulevasta tilasta (esimerkiksi Bonabeau, 2002). Mallinnusprosessia siis joudutaan etukäteen määrittelemään merkittäviä muuttujia, joka Gilbert (1993) mukaan helpottaa teorioiden muodostamista ja testaamista: simuloidussa malleissa kun ei voida ottaa huomioon kaikki yhteiskunnallista ilmiötä kuvaavia muuttujia.

Bonabeau (2002) kuvaa tarkemmin agenttipohjaisen mallien hyötyjä. Hänen mukaansa agenttipohjainen mallinnus havaitsee paremmin nousevia ilmiöitä *emergent phenomena*), eli ilmiöitä joita ei voida havaita tarkastellessa vain yksilöitä, vaan yksilöiden välinen vuorovaikutus tai toisten yksilöiden toiminta vaikuttaa yksilön käyttäytymiseen. Esimerkkeinä täläisistä tilanteista on oppiminen ja sopeutuminen (ajallinen korrelaatio), aikaisempi kokemus ja tehdyt



Kuva 3: Agenttipohjainen simulaatio Macal & North (2009) mukaan

päätökset (muisti sekä polkuriippuvuus) sekä epälineaarinen käyttäytyminen, esimerkiksi tietyn kynnysarvon jälkeen tapahtuva toiminta.

Toisena hyötynä Bonabeau (2002) mainitsee agenttipohjaisen mallin helpomman tulkittavuuden. Hän arvioi, että yksilökeskeinen kuvaus toiminnasta on selkeämpi kuin erilaiset tilasiirtymäkuvaukset tai prosessikuvaukset. Tämän takia hän arvioi agenttiohjaisen mallien olevan helpommin asiantuntijoiden tulkittavissa ja arvioitavissa, jolloin mallit ovat selkeämmin validoitavissa. Lisäksi Bonabeau (2002) kritisoi keskisuureiden käyttämistä ilmöiden kuvaamiseksi, koska näissä tilanteissa ääriarvot eivät ole nähtävissä ja keskisuureet tasoittavat vaihtelua.

2.2 Mikrosimulaatiot

Agenttipohjaisten mallien lisäksi on mahdollista käyttää mikrosimulaatiota (*microsimulation model*). Sen toiminta ei merkittävästi eroa agenttipohjaisen simulaation periaatteista, eli malli pyrkii ennustamaan yksilön toimintaa. Kuitenkin taustalla oleva lähestymistapa eroaa. Agenttipohjaisen malli pyrkii luomaan ilmiön parametrien avulla ja tämän jälkeen tarkastelemaan luomiseen tarvittuja parametreja, kun taas mikrosimulaatiossa ilmiöön liittyvät taustamuuttujat ja niiden väliset määritellään etukäteen (Gilbert & Troitzsch, 2005, 58–59). Yksilökeskeisen mallinnustavan takia pidän mikrosimulaatioita ja agenttipohjaisia malleja samankaltaisina agenttipohjaisina simulaatioina, vaikka niiden toiminnan yksityiskohdat eroavatkin toisistaan.

2.3 Tit for tat

Kirjallisuuskatsauksessa ei esiintynyt erästä agenttipohjaisen simulaation varianttia, jossa mallinnetaan erilaisia käyttätysmismalleja ja niiden vaikutusta lopputulokseen. Alan klassikko on Axelrod (1980) simulaatio yhteistoiminnasta,

nimenomaisesti siitä kuinka agenttien kannattaa rankaista toisiaan sääntöjen rikkomisesta. Mallinnus perustui rationaalisen valinnan olettamuksiin yksilöistä oman edun edun tavoittelijoina ja koko asetelman pohjalla onkin peliteorian soveltaminen vuorovaikutuksen analyysiin.

Axelrod (1980) esittelee kilpailua, jossa etsittiin simuloimalla voittavaa strategiaa vangin dilemma-tyyppisessä asetelmassa, jossa yhteistyöstä palkitaan molempia pelaajia, mutta jos pelaaja pettää – ja on ainoa pettäjä – palkinto on suurempi kuin yhteistyössä. Molempien pettäessä kumpikaan pelaaja ei saanut palkintoa. Kilpailun voittanut strategia oli yksinkertaisin: petä jos aiemmalla kierroksella sinua on petetty, muutoin toimi yhteistyön mukaisesti. Tuloksen sovelluksia on ollut erityisesti kansainvälisten suhteiden alalla.

Kirjastot ja kehitysympäristöt

Agenttipohjaisen simulaation toteutukseen on olemassa useita erilaisia valmiita kirjastoja sekä alustoja. Tobias & Hofmann (2004) esittää arviointikriteereitä onnistuneille simulaatioympäristöille: ideaalisti kehitysympäristö pystyy automaattisesti luomaan agentteja tiettyjen todennäköisyysmallien perusteella, yksittäiset agentit voivat olla monimutkaisia ja pystyvät viestimään toisilleen. Lisäksi kehitysympäristön arvioinnissa tulisi heidän mukaansa ottaa huomioon kehittäjien tarpeita, esimerkiksi asennuksen yksinkertaisuus, graafisen käyttöliittymän käyttö ja kattava dokumentaatio ovat hyvän kehitysympäristön tunnusmerkkejä. Lisäksi Tobias & Hofmann (2004) ehdottavat avoimen lähdekoodin lisenssejä positiivisena tekijänä, koska niiden käyttö mahdollistaa mallin muutokset.

Näiden kriteerien pohjalta Tobias & Hofmann (2004) päätyvät suosittelemaan RePast-ympäristöä. RePast sallii useamman kielen käytön simulaation kehityksessä, mikä on mahdollistanut erilaisten välineiden ja kirjastojen käytön kehityksessä North et al. (2006), kuten esimerkiksi Javan käytön. Esimerkki Java-pohjaisesta RePast-mallista on liitteessä A, RePast-ympäristön hyödyn havaitsee valmiista koodista, jotka liittyvät simulaation pohjaan, siinä tehtävään etsintään ja liikkumiseen sekä todennäköisyysjakaumien käsittelyyn. Lisäksi merkitsemällä (*annotate*), voidaan erikseen merkitä kuinka usein kyseinen agentti suorittaa oman preferenssifunktionsa.

3 Ohjattu ja ohjaamaton koneoppiminen

Sekä sentimenttianalyysi että teemojen luokittelu (*topic modeling*) ovat koneoppimiseen perustuvia menetelmiä, mutta niiden toimintaperiaateet ovat erilaisia. Sentimenttianalyysi perustuu ohjattuun oppimiseen *(supervised learning)*, jossa on olemassa syöte-tulos–pareja sisältävää aineistoa. Tätä aineistoa käytetään osana oppimisprosessia. Ohjaamattomassa oppimisessa *(unsupervised learning)* ei tälläistä opetusmahdollisuutta, vaan aineistosta tulee löytää sitä kuvailevat

ominaisuudet erikseen.

Ohjatussa oppimisessa on siis joukko pareja $\{(s_1, t_1), \ldots, (s_n, t_n)\}$, missä s_i kuuluu syötteiden joukkoon $\mathcal{S}$ ja t_i kuuluu tulosjoukkoon $\mathcal{T}$. Opetusvaiheessa ohjattu oppiminen laskee eri syötteiden välisiä yhteyksiä ja päättelee, mitkä tekijät syötteissä vaikuttavat kuhunkin tulokseen. Täsmällisemmin siis optimoimoidaan funktiota $g : \mathcal{S} \mapsto \mathcal{T}$, siten että virheellisten syöte-vastaus–parien määrä pienenee. Ohjatussa koneoppimisessa haasteena onkin ottaa huomioon ylisovituksen ongelma: jos mallin parametrien määrä kasvaa voi se sopia aineistoon erinomaisesti, mutta olla epäsopiva aineiston ulkopuolella olevien syötteiden tulkinnassa. Opetuksen jälkeen ohjattu oppiminen pystyy itsenäisesti arvioimaan mielivaltaista syötettä vastaavan tulosarvon käyttäen opittua funktiota g.

Havainnollistetaan yllä kuvattua konkreettisen yksinkertaisen esimerkin avulla. Olkoon meillä joukko ominaisuuksia ja niitä vastaana dikotominen (binäärinen) arvo, eli pareja $\{(s_1, t_1), \ldots, (s_n, t_n)\}$, missä $s_i \in \mathbb{R}^n$ ja $t_i \in \{0, 1\}$. Eräs menetelemä tähän on naivi Bayes-luokitin, jossa tutkitaan ehdollista riippumatonta todennäköisyyttä $p(0|s_i)$ sekä $p(1|s_i)$. Riippumattomuudella tarkoitetaan, että oletamme, ettei syötteen eri tekijät vaikuta toisiin: tällöin todennäköisyys voidaan kuvata tulomuodossa yli jokaisen piirrevektorien alkion ja luokittimena käyttää siis mallia, joka maksimoi todennäköisyyden kuulua ryhmään 0 tai 1, eli

$$\max_{t \in \{0,1\}} p(t) \prod_{j=1}^{n} p(s_{x,j}|t).$$

Kunkin piirrevektorin alkion ehdollinen todennäköisyys $p(s_{x,j}|t)$ opitaan annetusta aineistosta havaintojen perusteella.

Ohjaamattomassa oppimisessa aineistosta pyritään löytämään säännönmukaisuuksia. Mielivaltainen syöte $\mathcal{S}$ voidaan esimerkikis kuvata akseleilla $k_1, k_2, \ldots, k_n$ tai jakaa ryhmiin $K_1, K_2, \ldots, K_n$. Toisin kuin ohjatussa oppimisessa, ei aineistosta kuitenkaan pyritä luomaan yleistettävää funktiota – vaan tulokset esittelevät aina tiettyä aineistoa. Esimerkkinä tästä lähestymistavasta on pääkomponenttianalyysi, jota esittelen lyhyesti seuraavaksi.

Pääkomponenttianalyysin syöte $\mathcal{S}$ on monidimensionaalinen aineisto $X \in \mathbb{R}^{n \times m}$. Käytössä on matriisimuoto selkeyden vuoksi: matriisissa on n kappaletta havaintoja, joissa kussakin on m kuvaavaa muuttujaa. Tehtävänä on löytää aineiston rakennetta kuvaavat akselit $p = (p_1, p_2, \ldots, p_n)$, missä kukin akseli $p_i \in \mathbb{R}^m$ kuvaa aineiston m muuttujaa uudellleen. Täsmällisemmin, kukin akselin p projisoi alkuperäisen rivin uudelle orgononaaliselle koordinaatiostolle. Näiden akselien $p_1, p_2, \ldots, p_n$ avulla voidaan muodostaa myös $n \times m$ kokoinen matriisi, joka siis rakentuu alkuperäisestä matriisista X lineaarikombinaatioiden avulla.

Tarkastellaan täsmällisemmin proessia jolla nämä akselit löydetään aineistosta.Täsmällisemmin prosessissa muutetaan jokainen X rivi $x_{(i)}$ p_in avulla ja muodostaa uusi matriisi X'. Nyt arvojen välillä ei ole korrelaatiota, eli ne ovat ortogonaaliset. Ensimmäinen komponentti p_1 valitaan siten, että tällä saadaan

selitettyä mahdollisimman suuren osan aineistosta ja sen vaihtelusta, eli

$$\max_{||p||=1} \sum_{i}^{n} (x_i \cdot p_1)^2.$$

Myöhemmät komponentit valitaan poistamalla jo löydetyt komponentit matriisista ja jälleen etsimällä tästä uudesta matriisista suurin selittävä komponentti kuten edellä.

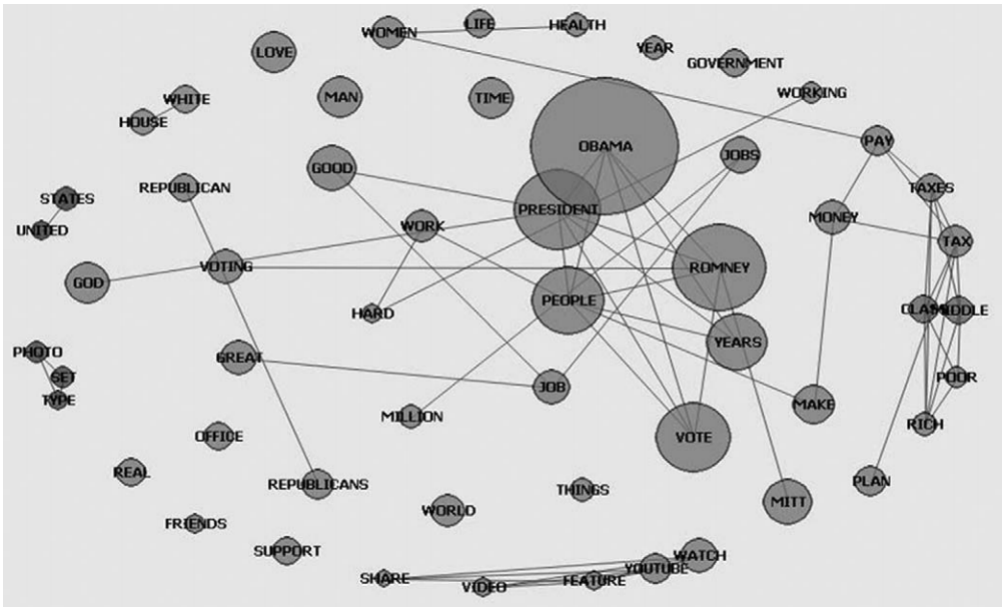
Pääkomponenttianalyysin mielenkiinto on, että on mahdollista valita pienempi joukko akseleita ($p' = (p_1, p_2, \ldots, p_{n'})$, missä $n' < n$) kuvaamaan aineistoa. Tehtäväksi tuleekin valita n' siten, että neliöllinen virhe täydellisen kuvauksen ja osittaisen kuvauksen välillä on hyväksyttävissä, eli $||X - X'||^2 <$ hyväksymiskynnys. Vaihtoehtoisesti n' voidaan valita siten, että komponenttien selitysaste ylittää tietyn kynnyksen tai uuden komponentin lisäämisen vaikutus selitysasteen parantumiseen on vähäinen. Tätä varten kaupallisissa ohjelmissa, kuten IBM SPSS Statistics, esitetään pääkomponenttianalyysin ohella myös selitysasteen muutosta kuvaavia käyriä.

Yhteiskuntatieteen haasteena on löytää ryhmittelevän ohjaamattoman koneoppimisen tuloksille mielekäs merkitys. Esimerkiksi klusterianalyysissä tämä tehdään tarkastelemalla mitä arvoja klustereissa on, pääkomponenttianalyysissä katsotaan erilaisten muuttujien merkitystä kunkin komponentin arvoon ja tällä tavoin muodostetaan komponenteista mielekkäitä akselia. Esimerkiksi, ehdokkaiden vaalikonevastauksissa klusteroinnin avulla saatettaisiin pyrkiä löytämään puolueita, kun taas pääkomponenttianalyysillä voitaisiin arvioida kansanedustajaehdokkaiden sijoittumista ideologisilla akseleilla.

4 Tekstin analysointi

Numeerisen aineiston lisäksi yhteiskuntatieteissä työskennellään usein tekstiaineiston kanssa, kuten haastatteluiden, viranomaistekstien sekä nykyisin myös sosiaalisen median tuotoksien kanssa. Perinteisesti aineiston analyysi on suoritettu lukemalla tekstejä ja tekemällä tästä päätelmiä, kuitenkin tämä on varsin raskasta ja onkin toivottavissa, että tietokoneistetusti tekstin luokittelua voitaisiin systematisoida – Grimmer & Stewart (2013) korostavatkin, että laskennalliset menetelmät toimivat tässä kohtaa ihmisen apuna, mutta tarkoitus ei ole korvata ihmistä analysoinnissa. Tekstin analysointiin käytettäviä menetelmiä on useita, niin ohjatun kuin ohjaamattoman, koneoppimisen tekniikoita (esimerkiksi Grimmer & Stewart, 2013), kirjallisuuskatsauksen perusteella tekstiainestoa on analysoitu sentimenttianalyysillä ja aiheiden tunnistuksella.

Groshek & Al-Rawi (2013) arvioivat Yhdysvaltojen vuoden 2012 presidentin vaalien aikaista viestintää sosiaalisessa mediassa. Aineistona heillä oli yli 1.4 miljoonaa viestiä, osa Facebookissa ja osa Twitterissä – mikä kertoo tarpeesta soveltaa laskennallista menetelmää analyysissä. Analyysi suoritettiin laskemalla sanojen lukumääriä ja termejä, joihin sanat olivat yhteydessä. Erityisesti

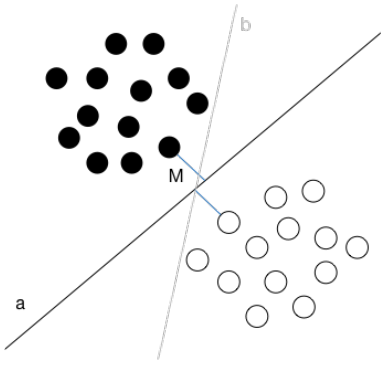


Kuva 4: Barack Obaman sosiaalisen median sivuiden perusteella luotu kuva sanoista (Groshek & Al-Rawi, 2013)

kirjoittajat tarkastelivat, kuinka ehdokkaan vastustajasta puhuttiin ehdokkaan Facebook- ja Twitter-sivuilla ja havaitsivat, ettei vastustajasta puhuttu merkittävästi negatiivisemmin kuin ehdokkaasta. Laskennallisesti kyseessä ei ole erityisen mielenkiintoinen ongelma: kuten kuvasta 4 nähdään, analyysi perustuu sanojen ilmenemiseen toistensa kanssa ja tämän pohjalta laskettuihin todennäköisyyksiin.

Laskennallisesti mielenkiintoisempaa on tekstin sisällön merkitysten arviointi, yksinkertaisimmillaan tekstin positiivisuuden ja negatiivisuuden arviointi. Jatkaen poliitikoiden sosiaalisen median tutkimusta, Park et al. (2011) arvioivat poliitikkojen verkkoprofiilissa esiintyviä positiivisia je negatiivisia sanoja, ja havaitsevat että oppositiopuolueiden sivuilla oli enemmän positiivisia viestejä hallituspuolueisiin verrattuna. Myöskin Tumasjan et al. (2011) ovat kiinnostuneet poliitikkoihin liittyvistä viesteistä Twitterissä, he sentimenttianalyysin perusteella päättelevät, että puolueen kannattajat puhuvat kilpailevista poliitikoista negatiivisemmin kuin omasta ehdokkaastaan, mutta positiivisissa viesteissä ei nähdä selvää eroa.

Lisäksi laskennallisesti on mahdollista erottaa aiheita ja teemoja – mikä vastaa ihmistyötä vastaavassa laadullisen tutkimuksen prosessissa. Levy & Franklin (2013) ovat analysoineet lainsäädäntöön liittyneitä kommentteja ja laskennallisesti eroittaneet aineistosta teemoja, joita kommenteissa käsiteltiin. He analysoivat tarkemmin, ketkä ovat lähettäneet kommentteja ja havaitsevat eroja lobbausorganisaatioiden ja yksittäisten kansalaisten teemojen välillä. Tällä perusteella he arvioivat lainsäädäntötyön läpinäkyvyyttä ja lobbaustahojen merkittävyyttä lainsäädäntötyössä.



Kuva 5: Esimerkki tukivektorikoneen toiminnasta

4.1 Sentimenttianalyysi

Esitellyissä artikkeleissa käytössä oli yksinkertainen, sanoihin ja niiden lukumäärään perustuva sentimenttianalyysi, jossa sanoja luokitellaan jollain asteikolla ja tätä kautta muodostetaan sanakirja, jota voidaan käyttää analysoimisessa (Pennebaker et al., 2001; Thelwall et al., 2010)[2]. Analyysiä voidaan täydentää käyttämällä laajemmin kielellisiä ominaisuuksia, kuten huutomerkejä sekä perättäisille sanoja, sekä niiden vaikutusta yksittäisten termien määrään. On myös mahdollista koneellisesti testata erilaisia yhdistelmiä ominaisuuksia, ja verrata näin saatuja arvoja ihmisen suorittamaan luokitukseen lauseiden osalta, pyrkien optimoimaan eri ominaisuuksien painoarvoja siten, että virhe ihmisen tekemän luokituksen välillä on mahdollisimman pieni (Thelwall et al., 2010).

Sentimenttianalyysi perustuu siis tekstin piirteisiin *(features)* tarkasteluun, laskennallisesti mielenkiintoisempi – tosin, Thelwall et al. (2010)[3] aineiston kohdalla vähemmän tarkka – menetelmä on tukivektorikoneiden *(support vector machine)* käyttö luokittamiseen. Tukivektorikoneita voidaan käyttää myös muuhun tekstuaalisen aineiston käsittelyyn (esimerkiksi Weber et al., 2012).

Kyseessä on ohjatun oppimisen menetelmä, eli tukivektorikoneet perustuvat piirrevektoreiden ja (opetettujen) arvojen pareihin. Olkoon siis aineistona $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ missä $x_i \in \mathbb{R}^n$ on piirrevektori ja $y_i \in \mathbb{Z}$ on opetettu arvo. Kyseessä on lineaarinen luokitin, joka siis jakaa aineisto tasoilla, siten että arvot y ovat mahdollisimman hyvin tasojen muodostamissa ryhmissä, mikä intuitiivisesti voidaan käsittää siten että tasojen väliin jää mahdollisimman paljon tilaa.

Yksinkertaisuuden vuoksi oletamme tässä esityksessä, että $y_i \in \{-1, 1\}$. Tukivek-

[2]Sentimenttianalyysiä voidaan myös tehdä myös ohjaamattomasti, tavoite on aina löytää tekstiaineistosta piirteitä joiden perusteella voidaan ennustaa kielen käyttäytymistä. Käytettyjen menetelmien joukko on laaja, esimerkiksi Bayesilaisia luokittimia ja tukivektorikoneita on vertailtu (Pang et al., 2002; Pang & Lee, 2005; Thelwall et al., 2010). Koska sentimenttianalyysiä tutkitaan aktiivisesti, tässä työssä esittelen yksinkertaisen toteutuksen sentimenttianalyysistä.

[3]Thelwall et al. (2010) on keskittynyt sosiaalisen median viestien analysointiin MySpace–palvelussa, ja samaa menetelmää on sovellettu myös löydetyssä kirjallisuudessa.

torikoneen tavoite on siis jakaa tämä joukko kahteen eri ryhmään yhdellä tasolla. Edelleen, tämän tason halutaan olevan sellainen, että alkioiden ja tason välinen etäisyys – marginaali M – on mahdollisimman suuri, eli yksinkertaistetusti että löydetty taso jakaa aineiston mahdollisimman selvästi kahteen eri ryhmään. Tason määriytyy muutujien $\beta \in \mathbb{R}^n$ sekä $\beta_0 \in \mathbb{R}$ avulla. Tehtävänä on optimoida βn ja β_0 arvoja, siten että M maksimoituu. Kuva 5 esittelee ideaa tarkemmin: kahden joukon välissä oleva suora a jakaa joukot siten, että marginaali M on mahdollisimman iso. Vertaa sitä esimerkiksi vaihtoehtoiseen suoraan b, mikä sekin jakaa joukot kahteen ryhmään, mutta osa alkioista on erittäin lähellä tätä jakosuoraa.

Täsmällisesti siis, tavoite on määrittää β ja β_0 luokittimessa $\text{sign}(x\beta + \beta_0)$. Edelleen, koska halutaan että ryhmien väliin jäävä marginaali M on mahdollisimman iso. On siis tarkoitus löytää $\max M$ kuitenkin niin, että $y_i(x_i \cdot \beta + \beta_0) \geq M - \varepsilon$. ε on lisätty yhtälöön sallimaan osan alkioista olemaan marginaalin sisällä, yksinkertaisuuden vuoksi tässä esityksessä $\varepsilon = 0$.

Yllä kuvattu optimointiongelma voidaan myös esittää muodossa

$$\min \frac{1}{2}||\beta||^2, \text{ kuitenkin huomioiden } y_i(\beta \cdot x_i + \beta_0) \geq 1.$$

Ongelmaan voidaan soveltaa neliöllistä optimointia, jolla ratkaistaan pienin mahdollinen βn arvo. Jos hyväksymme, että $\varepsilon \neq 0$, ongelmaan sovelletaan Lagrangen menetelmää, edelleen jos jakoa ei suoriteta lineaariluokittimella on tarpeen siirtyä käyttämään kernelifunktiota (Vapnik & Kotz, 1982; Cortes & Vapnik, 1995; Hastie et al., 2009).

4.2 Teemojen luokittelu

Teemojen luokittelussa syötteenä on joukko teksidokumenteja $D_1, D_2, \ldots, D_n$ ja tavoitteena on eritellä mistä teemoista $t_1, t_2, \ldots, t_m$ kyseiset dokumentit käsittelivät. Kyseessä on ohjaalamttoman koneoppimisen muoto, eli aiheita ei tarvitse etukäteen määritellä – ja mikä tärkeintä, ei ole tarpeen yhdistää dokumentteja ja aiheita opetusaineiston luomiseksi. Ratkaisu perustuu kunkin tekstidokumentin sanojen ja niiden todennäköisyyksien käyttöön osana mallinnusprosessia.

Menetelmänä teemojen luokittelussa voidaan käyttä latenttien Dirchlet-allokaatioihin *(latent Dirchlet allocation, LDA)* perustuva todennäköisyyslaskenta. Ideana on siis, että aineistolla on piilossa olevia teemoja, joita voidaan havainnoida välillisesti sanojen esiintymisen perusteella dokumenteissa (Blei, 2012; Blei et al., 2003).

Työkalut

Agenttipohjaisten simulaatioiden kohdalla korostettiin dokumentaatiota, avoimuutta ja muokattavuutta eri ympäristöjen arvioinnissa (Tobias & Hofmann,

2004). Sentimenttianalyysissä samanlaiset kriteerit ovat mielekkäitä: kuitenkin, koneoppimista soveltavissa lähestymistavoissa mielekästä on myös arvioida koneopitun aineiston laatua ja luotettavuutta. Esimerkiksi Thelwall et al. (2010) kehittämä SentiStrength tarjoaa valmiina käytettäväksi englannin kielelle soveltuvan sanakirjan, mutta mahdollistaa myös omien painotusten kehittämisen ja optimoinnin.

Sekä tukivektorikoneisiin että teemojen luokitteluun löytyy useita avoimen lähdekoodin toteutuksia. Koska kyseessä on olemassa olevien menetelmien laskennalliset toteutukset, on mielekästä tarjota niistä Hornik et al. (2006) esittelevät neljän erilaisen tukivektorikoneita käsittelevää kirjastoa, ja päätyvät suosittelemaan suorituskykynsä puolesta `kernlab`-pakettia sekä `e1071`-paketissa olevaa, libsvm-kirjastoon perustuvaa toteutusta. Myös aiheiden tunnistukseen on olemassa R-paketti `topicmodels`.

5 Säännönmukaisuuksien ja ryhmien löytäminen

Viimeisenä laskennallisten menetelmien ryhmänä esittelen ohjaamattoman oppimisen menetelmiä, jotka pyrkivät löytämään säännönmukaisuuksia numeerisesta aineistosta. Esimerkiksi politiikan tutkimuksen perinteinen teema on selittää eroja valtioiden välillä, esimerkiksi demokratian tilan arviointia taustamuuttujien perusteella. Jurek & Scime (2013) jatkavat tätä perinteistä linjaa selittämällä valtioiden demokaattisuutta Freedom House –aineiston ja taustamuuttujien, kuten uskonnon ja demokratian keston perusteella. Menetelmänä he käyttävät työssään assosiaatiosääntöjen hakua laskennallisesti, jonka perusteella he löysivät yhteensä 210 erilaista sääntöä. Lisäksi he nostavat esille, että koneoppimisen soveltaminen mahdollistaa selittävien tekijöiden muutoksien valitsemisen vapaammin kuin perinteisten tilastollisten mentelmien soveltaminen, jossa tarkasteltavien sääntöjen määrä on yleisesti merkittävästi pienempi.

Muita ohjaamattoman oppimisen menetelmiä on muuttujien dimensioiden vähentäminen (*dimensionality reduction*), pääkomponenttianalyysi sekä klusterointianalyysi Hastie et al. (2009, 485–586). Dimensioiden vähentäminen esimerkiksi pääkomponenttianalyysin avulla on yleisesti tiedossa oleva menetelmä jo nyt (esimerkiksi Metsämuuronen, 2006, 615–651), mutta klusterointimenetelmistä k-means–mentetelmä on toistaiseksi tuntemattomampi politiikan tutkimuksen piirissä.

5.1 Assosiaatiosäännöt

Assosisaatiosäännöt perustuvat aineiston ominaisuuksien $x_1, x_2, \ldots, x_n$ perustuvat, että voidaan löytää yhteyksiä ominaisuuksien välillä. Yleisesti ominaisuudet ovat binäärisiä, eli $x_i \in \{0, 1\}$. Tavoitteena on löytää sääntöjä muuttujien väli-

sistä suhteista: esimerkiksi

$$A \Rightarrow B, \text{ jossa } A, B \in \mathcal{P}(\{x_1, x_2, \ldots, x_n\}) \text{ ja } A \cap B = \emptyset.^4.$$

Eli, A ja B koostuvat ominaisuuksien joukosta siten, ettei niillä ole yhteisiä ominaisuuksia. Sääntöjen määrä kasvaa 2^n-nopeudella, jolloin haaste on löytää säännöistä mielenkiintoiset, kattavat ja luotettavat säännöt. Jotta tämä ongelma olisi ratkaistavissa, kustakin säännöstä lasketaan erikseen sen selitysaste, luottamus kyseisen säännön yleistettävyyteen ja säännön nosto (*lift*), jolla arvioidaan löydetyn säännön merkittävyyttä (Hastie et al., 2009, 485–586).

Ratkaistaan ongelma sääntöjen löytämisestä kahdessa vaiheessa: ensiksi muodostetaan kattavat säännöt ja näistä valikoidaan milenekiintoiset ja luotettavat myöhemmin. Aineistona sääntöjen löytämiseksi on ryhmä rivejä $r_1, r_2, \ldots, r_n$, joista kukin sisältää arvon kullekkin ominaisuudelle, eli kukin rivi on muotoa $[x_1, x_2, \ldots, x_n]$. Lisäksi käytössä on kaksi valittua kynnysarvoa: ε kuvaamaan vähimmäismäärää havaintoja kattavuudelle ja δ kuvaamaan vähimmäistä tukea luotettavuudelle ($\varepsilon, \delta \in \mathbb{R}$).

Eräs algoritmi kattavien muodostamiseksi on Agrawal & Srikant (1994) esittelemä Apriori. Kuten algoritmi 1 näyttää, Apriori laskee eri sääntökombinaatioiden esiitymisen aineistosta ja riippuen vaaditusta tuesta joko hyväksyy sääntökombinaation osaksi kattavia sääntöjä tai hylkää sen. Uudet sääntöehdokkaat luodaan yksinkertaisesti käymällä aiemmat n alkion kokoisten sääntöjen joukko läpi ja lisäämällä siihen mahdollinen $n+1$ pituisen sääntöjen joukko, kunnes $n+1$ pituisia mahdollisia sääntöjä ei enää ole. Algoritmin jälkeen kerättynä on siis kattavat joukot, joista on tarpeen valita edelleen luotettavat ja mielenkiintoiset säännöt.

Luotettavuuden päättämistä varten määritellään säännön X tuki todennäköisyytenä, että aineistosta löytyy ryhmä X, missä $X \in \mathcal{P}(\{x_1, x_2, \ldots, x_n\})$. Luottamus sääntöön määritellään tuen avulla, ja se normalisoi säännön esiintyvyyden suhteessa mahdollisiin esiintymisiin: luottamus$(A \Rightarrow B) = \frac{\text{tuki}(A \cup B)}{\text{tuki}(A)}$, jälleen $A, B \in \mathcal{P}(\{x_1, x_2, \ldots, x_n\})$. Intuitiivisesti siis, jos aineistossa on paljon Ata, niin tällöin luotettavan säännön muodostamiseksi tarvitaan vahva evidenssi parista (A, B), vastaavasti jos aineistossa on vähän Ata, niin parin (A, B) esiintyminen voi olla vähäisempää.

Toteutuksena jälkimmäinen vaihe on varsin yksinkertainen: jokaiselle kattavan joukon ryhmälle lasketaan luottamus ja valitun kynnyksen δ perusteella joukoista hylätään epäluotettavaksi todetut säännöt. Nyt jäljellä on vain kattavat ja luotettavat säännöt.

[4] $\mathcal{P}(X)$ viittaa joukon X potenssijoukkoon, eli joukkoon jossa on kaikki joukon X osajoukot alkioina.

Algoritmi 1 Apriori

Merkitään K_n on n alkion pituisten kattavien joukkojen joukko.
Merkitään E_n on n alkion pituisten mahdollisten kattavien joukkojen joukko, eli joukko ehdokkaita.
Olkoon $\mathcal{S}$ kaikki aineistossa esiintyvät siirtymät, eli kaikki aineiston rivit $r_1, r_2, \ldots, r_n$

$K_1 \leftarrow$ { kaikki yhden alkion kokoiset kattavat joukot }
$i \leftarrow 2$
while $K_{i-1} \neq \emptyset$ **do**
  $E_i \leftarrow$ { i pituiset joukot siten, että joukon alkioiden $i-1$ pituiset osajoukot ovat kattavia K_{i-1}ssä }
  for all $s \in \mathcal{S}$ **do**
    for all $e \in \{e \mid e \in E_i \wedge e \subseteq \mathcal{S}\}$ **do**
      e.maara $\leftarrow +1$
    end for
  end for
  $K_i \leftarrow \{e \mid e \in E_i \wedge e.\text{maara} \geq \varepsilon\}$
  $i \leftarrow +1$
end while
  return $\bigcup_i K_i$

5.2 Klusterointi

Dimensioiden vähentämisellä viitataan prosessiin, jossa ilmiötä kuvaavien muuttujien määrää vähennetään etsimällä muuttujaryhmiä, joiden kautta aineiston vaihtelua selitetään mahdollisimman hyvin. Yhteiskuntatietelijöiden yleisesti tuntema menetelmä tähän on pääkomponenttianalyysi (*principal component analysysi, PCA*), jonka kautta aineisto jaetaan selittäviin faktoreihin. Laskennalliset menetelmät mahdollistavat myös aineiston samankaltaisten alkioiden esittämisen ryhminä, eli klustereina. Ongelmana siis muotoillaan seuraavasti: joukko alkioita $\{x_1, x_2, \ldots, x_n\}$ halutaan jakaa $k$hon ryhmään, siten että nämä ryhmät edustavat aineiston piirteitä.

Lloyd (1982) esittämä ratkaisu ongelmaan on k-means-algoritmi: se löytää aina lokaalisti parhaan ratkaisun ongelmaan (*local minimum*), mutta ei takaa että tämä ratkaisu on yleisesti paras (*global minimum*). Globaalisti parhaan ratkaisun löytäminen on kuitenkin laskennallisesti vaativa, jolloin sovelluksissa tyydytään heuristiseen lokaalisti parhaaseen ratkaisuun.

Nimensä mukaisesti k-means löytämään keskiarvon kullekkin klusterille ja valitsemaan klustereiden sijainnit siten, että havaintojen sijoittaminen näille klusteripisteille minimoi neliöllisen virheen (Algoritmi 2). Yksinkertaisuuden vuoksi oletamme, että $x_i \in \mathbb{R}^n$, jolloin vektorien välisiä etäisyyksiä voidaan laskea selkeästi normien avulla; k-means menetelmä on toki yleistettävissä mielivaltaiselle

etäisyysfunktiolle $\mathcal{D}(x_i, x_j)$. Täsmällisemmin siis minimoidaan $\sum_i^n ||x_i - k_{x_i}||^2$, missä $k_{x_i} \in \mathbb{R}^n$ on alkiolle x_i määrätty klusteri. Kyseessä on iteratiivinen algoritmi, jota toistetaan kunnes klustereiden sijainnit eivät vaihdu, ensimmäiset klusteripaikat arvotaan satunnaisesti.

Vastaavasti spektriklusterointi perustuu samankaltaisten ominaisarvojen ryhmittämistä, jolloin klusterit korostavat samankaltaisten ominaisuuksien joukkoja paremmin kuin keskiarvoon perustuva k-means klusterointi (Hastie et al., 2009, 485–586).

Algoritmi 2 k-means

Jaetaan aineisto alustavasti klusteriryhmiin $k_1, k_2, \ldots k_k$ ja etsitään jokaiselle alkiolle x_i alkiota lähinnä oleva klusteri
while klusterien paikat eivät merkittävästi muutu **do**
lasketaan jokaiselle klusterille k uusi paikka k' laskemalla keskiarvo klusteriin kuuluvien alkioiden $x_1, x_2, \ldots x_m$ paikoista, eli $k' = \frac{1}{m} \sum_{i=1}^{m} x_i$
uudelleenmääritellään alkoiden kuuluminen klusteriryhmiin tarkastamalla, mitä klusteria k'_i lähinnä alkio x on, eli $\min_{i=1}^{k} ||x - k'_i||^2$
end while

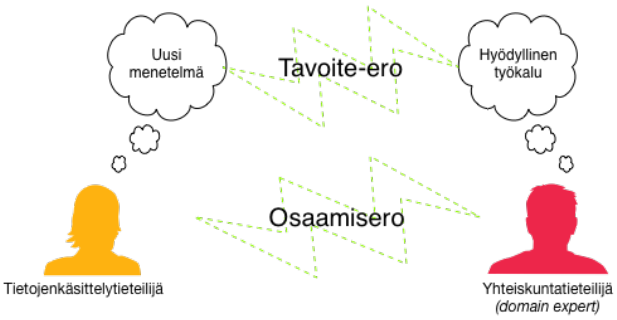
6 Keskustelu

6.1 Validiteetti haasteena

Niin määrällisessä kuin laadullisessa yhteiskuntatieteessä tutkimuksen laatua arvioidaan validiteetin ja reliabiliteetin kannalta. Ensimmäinen viittaa ulkoiseen laatuun, eli siihen tutkitaanko ilmiötä oikein ja jälkimmäinen sisäiseen laatuun, eli onko tutkimus tehty uskottavasti.

Sentimenttianalyysi, teemojen luokittelu, assosiaatiosääntöjen oppiminen sekä klusterointi ovat aineiston käsittelyyn tarkoitettuja algoritmeja, jolloin uskottavuuden kannalta keskeisempää on aineiston laatu: onko aineiston mahdollinen esikäsittely suoritettu oikein ja onko aineisto kerätty oikein. Iteratiivissa algoritmeissa – kuten klusteroinnissa – on tarpeen myös tarkastella tulosten pysyvyyttä eri aloitusarvoilla, eli arvioida saadun tuloksen herkkyyttä (esimerkiksi Pardos & Heffernan, 2010). Käsitellessään tutkimuksen laadukkuutta, Grimmer & Stewart (2013) huomioivat, että laskennallisten menetelmien käytön tulisi vastata ihmisten tekemän analyysin tasoa, eli taustateorioiden, käsin tehdyn aineiston käsittelyn tai tilastollisten arvioiden tulee luoda luottamus menetelmien uskottavuudesta ja sopivuudesta ongelmaan.

Kirjallisuuden pohjalta laskennallisen yhteiskuntatieteen käytetyn menetelmä oli simulointi, joka on mielestäni uskottavuuden kannalta myös haastavin. Toisin kuin yllä mainituissa menetelmissä, simuloinnissa menetelmä ja käytettyjen parametrien määrä voi vaikuttaa tulokseen. Bloomquist (2006) on työssään ar-



Kuva 6: Menetelmäkehityksen ero soveltajaan (mukaillen van Wijk, 2006, 7)

vioinut kolmea julkaistua agenttipohjaista mallia, havaiten ettei tuloksia näissä malleissa ole verrattu ilmiöstä kerättyyn aitoon aineistoon. Lisäksi hän kritisoi sitä, ettei malleilla ole suoritettu useampia ajoja, jolloin mallien tuloksiin liittyy tiettyä epävarmuutta. Samoin Edmonds & Hales (2005) arvioivat erilaisia simulointimalleja, nostaen esille huolensa toistettavuudesta sekä ymmärryksen mallin rajoituksista ja soveltuvuudesta eri teemoihin.

Kritiikki on mielestäni osuvaa ja osoittaakin suurimman huoleni simulaatiotutkimukseen: kuinka voidaan varmistaa, että esitetyt simulaatiot aidosti vastaavat varsin monimutkaisia yhteisöjen käyttäytymissääntöjä? Viimeaikaisessa simulaatiotutkimuksessa näihin kritiikkeihin on pyritty vastaamaan, esimerkiksi simulaatiomalli voidaan rakentaa ilmiötä tutkivien koeasetelmien valossa (Villatoro et al., 2013) tai mallien tuloksia voidaan verrata olemassa oleviin ilmiöihin aktiivisesti ja käyttää näitä mallin rakennuksen tukena (Pearson et al., 2010).

Simulaatiomallit heikkouksistaan huolimatta tarjoavat kiinnostavia mahdollisuuksia mallin luomiseen vuorovaikutuksessa laskennallisten sekä yhteiskuntatieteilijöiden ja muiden toimijoiden välillä. Milne et al. (2014) kuvaavat prosessia, jossa simulaatiomallia ja sen tuloksia arvioitiin yhdessä virkamiesten ja tutkijoiden välillä. Heidän mukaansa mahdollisuus muuttaa mallia lisäsi yhteisymmärrystä ilmiöstä ja helpotti tiedon siirtämistä tutkijoilta soveltajille. Toisaalta, kuten Saunders-Newton & Scott (2001) huomauttavat, eri henkilöt suhtautuvat laskennallisiin menetelmiin myös erilaisin odotuksin. Tällöin luottamuksen rakentaminen ja laskennallisen menetelmän läpinäkyvyys ja selkeys ovat merkittäviä arvoja osana laskennallista yhteiskuntatiedettä, kuten perinteisessä yhteiskuntatieteellisessä tutkimuksessa.

6.2 Työskentely monitieteellisesti

Tutkimuksen uskottavuuden kohdalla nostin esille jo haasteen yhteistyöstä laskennallisen ja yhteiskuntatieteen välillä. van Wijk (2006) käsittelee monitieteellistä yhteistyötä visualisoinnissa. Oman kokemuksensa pohjalta hän huomauttaa, että alan asiantuntijalla (*domain expert*) ja visualisaation tutkijalla on usein erilaiset tavoitteet visualisaation kehittämisessä: kun visulisaation tutkija ensisijaisesti

kehittää uusia visualisaatiomenetelmiä, niin alan asiantuntijan tavoitteena on hyödyllisen työkalun kehittäminen – mikä voidaan saavuttaa myös perinteisillä menetelmillä. Kuvassa 6 esitän saman erottelun sovellettuna laskennalliseen yhteiskuntatieteeseen, alan asiantuntijan ja tietojenkäsittelytietelijöiden välisenä tarkasteluna.

van Wijk (2006) jatkaa, että tavoite-eron takia yhteistyön muoto on joku seuraavista: tietojenkäsittelytieteen asiantuntija voi hänen mukaansa toimia työvälinekehittäjänä tai jatkaa tietojenkäsittelyn menetelmien kehittämistä. Tietojenkäsittelytieteen asiantuntija voi myös soveltaa käyttäjäkeskeisiä menetelmiä kehitystyössään, jolloin laskennallisia menetelmiä kehitetään yhteistyössä alan asiantuntijan kanssa. Toisaalta, tietojenkäsittelyttieteen asiantuntija voi myös itse tutustua alaan tai kehittää visualisaatiotekniikoita aiheisiin, joista hän on kiinnostunut. van Wijk kutsuu viimeistä yhteistyön muotoa mielenkiintovetoiseksi kehitystavaksi (*curiosity driven*), ja nostaa esille muodon haasteen: sellaisenaan tällä toimintatavalla ei voida ratkaista aihealueen haastavimpia ongelmia.

Laskennallisen yhteiskuntatieteen osalta merkittävä kysymys onkin, kuinka yhteistyötä tietojenkäsittelytieteen ja yhteiskuntatieteen välillä voitaisiin parantaa, jolloin aineiston analysoimiseksi olisi käytössä tehokkaat ja uskottavat teknologiat. Eräs keino olisi soveltaa käyttäjäkeskeisen suunnittelun *(user-centered design, UCD)* periaatteita yhteistyöprojekteissa: soveltajan tarpeita ja tutkimustraditioita pyritään ymmärtämään laaja-alaisesti ja tukemaan tutkimustradition mukaista toimintaa. Tällöin kuitenkin haasteena on nimenomaisesti ottaa huomioon myös tietojenkäsittelytieteen tutkimuksen kannalta mielekkäät kysymykset osana tutkimusta.

Vaihtoehtoisesti laskennallinen yhteiskuntatiede voidaan tulkita menetelmäkehityksen kannalta vähemmän oleelliseksi alaksi: laskennalliset menetelmät kehittyvät tietojenkäsittelytieteen tutkijoiden joukossa ja joskus osa näistä menetelmistä siirtyvät käyttöön soveltajien – kuten yhteiskuntatietelijöiden – joukkoon.

6.3 Paradigman muutos?

Mikä merkitys laskennallisella yhteiskuntatieteellä on? Watts (2011, 265) argumentoi, että

> [o]n kuitenkin välttämätöntä soveltaa kaikkia näitä [sekä laskennallisia että kuvailevia] lähetymistapoja samanaikaisesti, pyrkien saavuttamaan johtopäätöksiä siitä, kuinka ihmiset käyttäytyvät ja kuinka maailma toimii – sekä ylhäältä että alhaalta, käyttäen hyödyksi kaikkia menetelmiä jotka ovat käytettävissä. (*oma suomennus*)

Laskennallisille menetelmille yhteiskuntatieteissä on siis tarvetta, koska niiden avulla on mahdollista esittää ratkaisuja uusiin kysymyksiin. Kuitenkin, samaan

aikaan tulee huomioida yhteiskuntatieteiden oppihistoria monimenetelmäisenä ja -paradigmaisena: eri lähestymistapojen käyttö samojen ongelmien käsittelyyn on perinteinen menetelmä yhteiskuntatieteiden osalta.

Kirjallisuuskatsauksen perusteella laskennallisen menetelmät ovat toistaiseksi selkeästi vähemmistössä julkaistujen artikkelien joukossa. Kuitenkin laskennalliset menetelmät tarjoavat kiinnostavia mahdollisuuksia aineistojen, niin määrällisten kuin laadullisten, tarkasteluun ja esikäsittelyyn tutkimusta tukien. En usko laskennallisten menetelmien muuttavan merkittävästi yhteiskuntatieteen tapaa toimia, vaan kehittävän eteenpäin jo nyt vakiintuneita empiirisen tutkimuksen menetelmiä.

7 Johtopäätökset

Työssä käsiteltiin laskennallisia menetelmiä yhteiskuntatieteiden toiminnassa, ja tätä kautta muodostunutta laskennallisen yhteiskuntatieteen toimikenttää. Työn ensimmäinen kontribuutio käsittelee toimintakentän määritelmää: laskennalliselle yhteiskuntatieteelle ei löydy yksikäsitteistä määritelmää ja alan konfferensseissa esitellään esimerkiksi taideprojekteja. Onkin mielekästä puhua laskennallisen yhteiskuntatieteen jatkumosta, jossa laskennallisten menetelmien monimutkaisuus vaihtelee.

Kirjallisuuskatsauksen perusteella suurimmaksi laskennallisuuden sovellusalue on simulaatiotutkimuksessa, erityisesti agenttipohjaisessa simulaaatiossa. Ohjatut ja ohjaamattomat koneoppimisen menetelmät tarjoavat mielenkiintoisia mahdollisuuksia, mutta niiden käyttö ei ole vielä vakiintunut vertailuissa journaaleissa tarkemmin. Yhteiskuntatieteellisissä julkaisuissa on kuitenkin sovellettu esimerkiksi aiheiden tunnistamista, sentimenttianalyysiä ja assosiaatiosääntöjä koneoppimisen menetelminä.

Kirjallisuuskatsauksen perusteella laskennallisella yhteiskuntatieteellä on kaksi merkittävää haastetta: tutkimuksen luotettavuus ja poikkitieteellinen yhteistyö. Tutkimuksen luotettavuuteen ei ole yksikäsitteistä ohjetta, vaan käytössä on esimerkiksi koeasetelmien kautta saavutettu luottamus malleihin (Villatoro et al., 2013), aktiivinen vertailu olemassa oleviin tilastotietoihin (Pearson et al., 2010) sekä saman tutkimusprosessin toistaminen ainakin osittain perinteisin menetelmin (Edmonds & Hales, 2005). Myös ryhmätyöskentelyn rooli osana tutkimuksentekoprosessia on nostettu esille validiteettia parantavana tekijänä (Milne et al., 2014), mikä myös voi olla ratkaisu poikkitieteellisen työskentelyn haasteisiin.

Kiitokset

Kiitän työn ohjaajaa, Antti Ukkosta, keskusteluista ja kyseenalaistavista kysymyksistä. Lisäksi kiitän Airi Lampista sekä Eric Malmia työn aikana käydyistä keskusteluista.

Lähteet

L. A. Adamic & N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 36–43, 2005.

R. Agrawal & R. Srikant. Fast Algorithms for Mining Association Rules. *Proceeding VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.

P. Ahonen. Before big data: a small data political study, 2014.

M. Altaweel, D. Sallach, & C. Macal. Mobilizing for change: Simulating political movements in armed conflicts. *Social Science Computer Review*, page 0894439312451106, 2012.

R. E. Anderson & C. Hicks. Highlights of contemporary microsimulation. *Social Science Computer Review*, 29(1):3–8, 2011.

R. Axelrod. Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution*, 24(1):3–25, 1980.

S. Bankes, R. Lempert, & S. Popper. Making computational social science effective. *Social Science Computer Review*, 20(4):377–388, 2002.

E. Berndtson. 'schools of political science' and the formation of a discipline. 2009.

D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77, Apr. 2012.

D. M. Blei, A. Y. Ng, & M. I. Jordan. Latent dirichlet allocation. *The Journal of machine Learning research*, 3:993–1022, 2003.

K. M. Bloomquist. A comparison of agent-based models of income tax evasion. *Social Science Computer Review*, 24(4):411–425, 2006.

E. Bonabeau. Agent-based modeling: methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, 99 Suppl 3(90003):7280–7, May 2002.

D. Boyd & K. Crawford. Critical Questions for Big Data. *Information, Communication & Society*, 15(5):662–679, June 2012.

C. Cioffi-Revilla. Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):259–271, 2010.

J. Clinton, S. Jackman, & D. Rivers. The statistical analysis of roll call data. *American Political Science Review*, 98(02):355–370, 2004.

C. Cortes & V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

N. Eagle & A. S. Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Computing*, 10(4):255–268, 2006.

B. Edmonds & D. Hales. Computational simulation as theoretical experiment. *Journal of Mathematical Sociology*, 29(3):209–232, 2005.

G. N. Gilbert & K. G. Troitzsch. *Simulation for the Social Scientist*. Open University Press, 2005.

N. Gilbert. Computer simulation of social processes. *Social Research Update*, (6), 1993.

J. Golbeck, J. M. Grimes, & A. Rogers. Twitter use by the U.S. Congress. *Journal of the American Society for Information Science and Technology*, 61 (8):1612—-1621, 2010.

J. Grimmer & B. M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297, 2013.

J. Groshek & A. Al-Rawi. Public sentiment and critical framing in social media content during the 2012 us presidential campaign. *Social Science Computer Review*, page 0894439313490401, 2013.

T. Hastie, R. Tibshirani, & J. Friedman. *The Elements of Statistical Learning*, volume 1 of *Springer Series in Statistics*. Springer New York, New York, NY, 2009.

K. Hornik, D. Meyer, & A. Karatzoglou. Support vector machines in r. *Journal of statistical software*, 15(9):1–28, 2006.

K. Imai & D. Tingley. A statistical method for empirical testing of competing theories. *American Journal of Political Science*, 56(1):218–236, 2012.

S. J. Jurek & A. Scime. Achieving Democratic Leadership: A Data-Mined Prescription. *Social Science Quarterly*, 00(00):n/a–n/a, Apr. 2013.

J. Karikoski & M. Nelimarkka. Measuring social relations with multiple datasets. *International Journal of Social Computing and Cyber-Physical Systems*, 1(1): 98–113, 2011.

D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, & M. Van Alstyne. Life in the network: the coming age of computational social science. *Science*, 323(5915):721–723, 2009.

K. E. Levy & M. Franklin. Driving regulation: Using topic models to examine political contention in the us trucking industry. *Social Science Computer Review*, page 0894439313506847, 2013.

S. Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, Mar 1982.

C. M. Macal & M. J. North. Agent-based modeling and simulation. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 86–98. IEEE, Dec. 2009.

K. M. McGraw. Political Methodology: Research Design and Experimental Methods. In R. E. Goodin & H.-D. Klingemann, editors, *A New Handbook of Political Science*, pages 769–786. Oxford University Press, Oxford, 1996.

J. Metsämuuronen. *Tutkimuksen tekemisen perusteet ihmistieteissä*. International Methelp Ky, 2006.

B. J. Milne, R. Lay-Yee, J. McLay, M. Tobias, P. Tuohy, A. Armstrong, R. Lynn, J. Pearson, O. Mannion, & P. Davis. A collaborative approach to bridging the research-policy gap through the development of policy advice software. *Evidence & Policy: A Journal of Research, Debate and Practice*, 10(1):127–136, Jan. 2014.

C. Z. Mooney. Bootstrap statistical inference: Examples and evaluations for political science. *American Journal of Political Science*, pages 570–602, 1996.

M. Nelimarkka & J. Karikoski. Categorizing and measuring social ties. In *RC33 Eighth International Conference on Social Science Methodology*, 2012.

M. J. North, N. T. Collier, & J. R. Vos. Experiences Creating Three Implementations of the Repast Agent Modeling Toolkit. 16(1):1–25, 2006.

G. R. Notess. The wayback machine: The web's archive. *Online*, 26(2):59–61, 2002.

J. Orbell, T. Morikawa, J. Hartwig, J. Hanley, & N. Allen. "machiavellian" intelligence as a basis for the evolution of cooperative dispositions. *American Political Science Review*, 98(01):1–15, 2004.

A. Oulasvirta, A. Pihlajamaa, J. Perkiö, D. Ray, T. Vähäkangas, T. Hasu, N. Vainio, & P. Myllymäki. Long-term effects of ubiquitous surveillance in the home. In *Proceedings of The 14th International Conference on Ubiquitous Computing*, 2012.

B. Pang & L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115–124, 2005.

B. Pang, L. Lee, & S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

Z. A. Pardos & N. T. Heffernan. Navigating the parameter space of bayesian knowledge tracing models: Visualizations of the convergence of the expectation maximization algorithm. *Educational Data Mining*, 2010:161–170, 2010.

S. J. Park, Y. S. Lim, S. Sams, S. M. Nam, & H. W. Park. Networked politics on cyworld: The text and sentiment of korean political profiles. *Social Science Computer Review*, 29(3):288–299, 2011.

J. Pearson, R. Lay-Yee, P. Davis, D. O'Sullivan, M. von Randow, N. Kerse, & S. Pradhan. Primary Care in an Aging Society: Building and Testing a Microsimulation Model for Policy Purposes. *Social Science Computer Review*, 29(1):21–36, May 2010.

J. W. Pennebaker, M. E. Francis, & R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

C. C. Ragin, D. Berg-Schlosser, & G. de Meur. Political Methodology: Qualitative Methods. In R. E. Goodin & H.-D. Klingemann, editors, *A New Handbook of Political Science*, pages 749–768. Oxford University Press, Oxford, 1996.

D. Saunders-Newton & H. Scott. "But the Computer Said!": Credible Uses of Computational Modeling in Public Sector Decision Making. *Social Science Computer Review*, 19:47–65, 2001.

M. J. Shapiro. The House and the Federal Role : A Computer Simulation of Roll-Call. *The American Political Science Review*, 62(2):494–517, 1968.

D. Silverman. *Doing Qualitative Research. A practical handbook*. SAGE Publications, London, 2000.

B. L. Slantchev. How initiators end their wars: The duration of warfare and the terms of peace. *American Journal of Political Science*, 48(4):813–829, 2004.

J. Stromer-Galley. On-line interaction and why candidates avoid it. *Journal of Communication*, 50(4):111–132, Dec. 2000.

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, & A. Kappas. Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology*, 61:2544–2558, 2010.

R. Tobias & C. Hofmann. Evaluation of free Java-libraries for social-scientific agent based simulation. *Journal of Artificial Societies and Social Simulation*, 7(1), 2004.

A. Tumasjan, T. O. Sprenger, P. G. Sandner, & I. M. Welpe. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4):0894439310386557, 2011.

J. van Wijk. Bridging the Gaps. *Computer Graphics and Applications, IEEE*, 26(6):6–9, 2006.

V. N. Vapnik & S. Kotz. *Estimation of dependences based on empirical data*, volume 40. Springer-Verlag New York, 1982.

D. Villatoro, G. Andrighetto, J. Brandts, L. G. Nardin, J. Sabater-Mir, & R. Conte. The Norm-Signaling Effects of Group Punishment: Combining Agent-Based Simulation and Laboratory Experiments. *Social Science Computer Review*, 32(3):334–353, Dec. 2013.

D. J. Watts. *Everything is obvious. How common sense fails.* Atlantinc Books, Lontoo, 2011.

I. Weber, A. Ukkonen, & A. Gionis. Answers, not links. In *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, pages 613–622, New York, New York, USA, 2012. ACM Press.

A RePast-esimerkki

```
1  package zombies;
2
3  import java.util.*;
4
5  import repast.simphony.query.space.grid.*;
6  import repast.simphony.random.*;
7  import repast.simphony.space.*;
8  import repast.simphony.space.continuous.*;
9  import repast.simphony.space.grid.*;
10 import repast.simphony.util.*;
11
12 public class Human {
13
14  private ContinuousSpace<Object> space;
15  private Grid<Object> grid;
16  private int energy, startingEnergy;
17
18  public Human(ContinuousSpace<Object> space, Grid<Object>
        grid, int energy) {
19   this.space = space;
20   this.grid = grid;
21   this.energy = startingEnergy = energy;
22  }
23
24  @Watch ( watcheeClassName = "zombies.Zombie",
       watcheeFieldNames = "moved", query = "␣within_moore␣1
       ", whenToTrigger = WatcherTriggerSchedule . IMMEDIATE
       )
25  public void run() {
26   GridPoint pt = grid.getLocation(this);
27
28   GridCellNgh<Zombie> nghCreator = new GridCellNgh<Zombie
       >(grid, pt,
29     Zombie.class, 1, 1);
30   List<GridCell<Zombie>> gridCells = nghCreator.
       getNeighborhood(true);
31   SimUtilities.shuffle(gridCells, RandomHelper.getUniform
       ());
32
33   GridPoint pointWithLeastZombies = null;
34   int minCount = Integer.MAX_VALUE;
35   for (GridCell<Zombie> cell : gridCells) {
36    if (cell.size() < minCount) {
```

```
37        pointWithLeastZombies = cell.getPoint();
38        minCount = cell.size();
39       }
40     }
41
42     if (energy > 0) {
43      moveTowards(pointWithLeastZombies);
44     } else {
45      energy = startingEnergy;
46     }
47    }
48
49    private void moveTowards(GridPoint pt) {
50     if (!pt.equals(grid.getLocation(this))) {
51      NdPoint myPoint = space.getLocation(this);
52      NdPoint otherPoint = new NdPoint(pt.getX(), pt.getY())
          ;
53      double angle = SpatialMath.calcAngleFor2DMovement(
          space, myPoint,
54        otherPoint);
55      space.moveByVector(this, 2, angle, 0);
56      myPoint = space.getLocation(this);
57      grid.moveTo(this, (int) myPoint.getX(), (int) myPoint.
          getY());
58      energy--;
59     }
60    }
61
62   }
```

RePast-dokumentaation pohjalta luotu esimerkki.