

MAANPUOLUSTUSKORKEAKOULU

**SOTILAALLISEN AKTIIVISUUDEN ARVIOINTI ASiantuntija-ARVIOIHIN
JA DATAAN PERUSTUVIEN BAYES-VERKKOJEN AVULLA**

Pro gradu -tutkielma

Yliluutnantti
Ville Parkkinen

Maisterikurssi 11
Maaliskuu 2023

Kurssi Maisterikurssi 11	Linja Maasotalinja
Tekijä Yliluutnantti Ville Parkkinen	
Tutkielman nimi Sotilaallisen aktiivisuuden arviointi asiantuntija-arvioihin ja dataan perustuvien Bayes-verkkojen avulla	
Oppiaine, johon työ liittyy Sotatekniikka	Säilytyspaikka Maanpuolustuskorkeakoulun kirjasto
Aika Maaliskuu 2023	Tekstisivuja 61, Liitesivuja 25
<p>TIIVISTELMÄ</p> <p>Koneoppimisen hyödyntämisen päätöksenteon tukemisessa on ennakoitu yleistyvän monilla aloilla. Sotilassovelluksissa päätöksenteon tueksi tarvitaan perusteltuja analyyskejä epävarmoissa tilanteissa. Dataa on usein käytössä vain vähän, mutta sen lisäksi voidaan hyödyntää asiantuntija-arvioita. Tässä opinnäytetyössä tutkitaan, miten sotilaallisen aktiivisuuden arviointia voidaan tukea dataan ja asiantuntija-arvioihin perustuvalla Bayes-verkolla. Erityisesti tutkimuksessa selvitetään, miten data ja asiantuntija-arviot yhdistetään ja hyödynnetään järjestysasteikollisten solmujen menetelmään perustuvassa Bayes-verkossa.</p> <p>Tutkimusmenetelminä käytettiin kirjallisuuskatsausta, matemaattista analyysiä sekä mallintamista. Työn päätuloksena esiteltiin uusi Bayes-päätelyä ja järjestysasteikollisten solmujen menetelmää hyödyntävä lähestymistapa asiantuntija-arvioiden ja datan yhdistämiseen. Aikaisemmassa tutkimuksessa järjestysasteikollisten solmujen menetelmän parametrit on määritelty tunnettuina arvoina. Tässä työssä esitellyssä lähestymistavassa nämä parametrit määritellään satunnaismuuttujina. Parametrien määrittely satunnaismuuttujina mahdollistaa asiantuntija-arvioiden tarkentamisen datan perusteella. Järjestysasteikollisten solmujen menetelmää laajennettiin tässä työssä lisäksi kahdella muunnoksella. Ensimmäinen näistä mahdollistaa ehdollisten todennäköisyysjakaumien tarkemman muodostamisen välimatka-asteikollista dataa käytettäessä ja toinen suurilla Bayes-verkon isäsolmujen määrillä. Tässä työssä esiteltyä uutta lähestymistapaa ja molempia järjestysasteikollisten solmujen menetelmän muunnoksia voidaan käyttää moniin sovelluskohteisiin.</p> <p>Tässä työssä esiteltyä Bayes-päätelyä ja järjestysasteikollisten solmujen menetelmää hyödyntävää lähestymistapaa testattiin sotilaallisen aktiivisuuden arviointiin kahdella aineistolla. Lähestymistavan todettiin voivan tuottaa tarkkoja arvioita sotilaallisesta aktiivisuudesta. Lähteiden merkitystä ja luotettavuutta voidaan arvioida estimoitavien parametrien perusteella. Erilaisten asiantuntija-arvioiden yhteensopivuutta datan kanssa voidaan verrata virhe-estimaatin avulla. Lähestymistavan tuottamaa arviota voidaan hyödyntää ohjaamalla ihminen arvioimaan dataa tarkemmin silloin, kun datan tarkempi arviointi vaikuttaa tarpeelliselta. Jatkotutkimuksia tarvitaan lähestymistavan systemaattiseksi arvioimiseksi käyttäjäkokemusten perusteella.</p>	
<p>AVAINSANAT</p> <p>Menetelmä tutkimus, koneoppiminen, tiedusteluanalyysi, Bayes-päätely, Bayes-verkot</p>	

SISÄLLYS

1	JOHDANTO	1
1.1	Aihealueen esittely	1
1.2	Tärkeimmät käsitteet	2
1.3	Aikaisempi tutkimus	3
1.4	Tutkimuskysymykset ja tutkimusmenetelmät	4
1.5	Viitekehys ja rajaus	4
2	KONEOPPIMISEN SOVELTUVUUS TIEDUSTELUN TUEKSI	6
2.1	Hyötyjä ja haasteita	6
2.2	Yhteenveto	8
3	TODENNÄKÖISYYSPÄÄTTELYN MENETELMIÄ	9
3.1	Bayes-päättele	9
3.2	Bayes-verkot	12
3.3	Järjestysasteikollisten solmujen menetelmä	16
3.4	Gibbsin otanta	24
3.5	Yhteenveto	27
4	BAYES-PÄÄTTELYÄ JA JÄRJESTYSASTEIKOLLISTEN SOLMUJEN MENETELMÄÄ HYÖDYNTÄVÄ LÄHESTYMISTAPA	29
4.1	Lähestymistavan yleiskuva	29
4.2	Asiantuntija-arvio	31
4.3	Aineiston kerääminen	38
4.4	Aineiston esikäsittely	39
4.5	Estimointi	45
4.6	Validointi	49
4.7	Arvion käyttäminen	51
4.8	Yhteenveto	52
5	LÄHESTYMISTAVAN KÄYTTÖKELPOISUUDEN TESTAAMINEN	54
6	DISKUSSIO	56
6.1	Työn metodologinen kontribuutio	56
6.2	Työn pragmaattinen kontribuutio	57
6.3	Työn luotettavuuden arviointi	58

6.4	Jatkotutkimusaiheita	59
-----	--------------------------------	----

7	JOHTOPÄÄTÖKSET	60
----------	-----------------------	-----------

LÄHTEET

LIITTEET

Liite 1: Satunnaislukuihin perustuvien muunnosten testaus

Liite 2: Välimatka-asteikollisen muunnoksen testaus

Liite 3: Tarkkuusasteikon testaus

Liite 4: Lähestymistavan käyttökelpoisuuden testaaminen (TL IV)

SOTILAALLISEN AKTIIVISUUDEN ARVIOINTI ASiantuntija- ARVIOIHIN JA DATAAN PERUSTUVIEN BAYES-VERKKOJEN AVULLA

“Teknologinen kehitys erityisesti digitalisaation, tekoälyn, koneautonomian ja sensoriteknologioiden aloilla vaikuttaa puolustuksen kaikkiin toimintaympäristöihin [- -]. Kehityksen yhtenä tavoitteena on tukea päätöksentekoa tarkemmalla ja nopeammin käyttöön saatavalla tiedolla.”
- Valtioneuvoston puolustuselonteko 2021

1 JOHDANTO

1.1 Aihealueen esittely

Sotilastiedustelu seuraa ja arvioi Suomen sotilaallista toimintaympäristöä ennakkovaroituksen antamiseksi, Puolustusvoimien ja valtiojohdon päätöksenteon tukemiseksi sekä Puolustusvoimien lakisääteisten tehtävien suorittamiseksi [1]. Sotilaallisen toimintaympäristön arviointiin kuuluu ulkomaisten asevoimien sotilaallisen aktiivisuuden arviointi. Aktiivisuus on esimerkiksi joukkojen siirtoja, perustamista, harjoittelua ja operointia. Sotilaallisesta aktiivisuudesta voidaan saada dataa kirjallisista lähteistä sekä erilaisilla sensoreilla.

Sotilaallisesta aktiivisuudesta tehtyjä havaintoja voidaan analysoida asiantuntija-arvioiden tai aikaisemmin kerätyn datan perusteella. Aikaisemmin kerätystä datasta voidaan löytää sotilaalliseen toimintaan usein liittyviä säännönmukaisuuksia [2], ja koneoppimista voidaan käyttää näiden säännönmukaisuuksien tunnistamiseen suuristakin datamääristä. Luotettavan datan hankkiminen sotilaallisesta aktiivisuudesta voi kuitenkin olla vaikeaa. Olemassa oleva data voi lisäksi vanhentua, jos arvioitava joukko vaihtaa toimintamallejaan. Data kertoo menneestä, mutta asiantuntija voi ennakoida tulevaa. Asiantuntija voi esimerkiksi arvioida havaittujen säännönmukaisuuksien jatkumisen todennäköisyyttä.

Tarve datan ja asiantuntija-arvioiden samanaikaiselle hyödyntämiselle on tunnistettu monilla aloilla [3–7]. Datan ja asiantuntija-arvioiden yhdistämistä tutkitaan tässä työssä yleisesti, jolloin tästä saatuja tuloksia voidaan hyödyntää alasta riippumatta. Tuloksia sovelletaan tässä työssä sotilaallisen aktiivisuuden arviointiin. Työssä esitellään datan ja asiantuntija-arvioiden yhdistämisen mahdollistava lähestymistapa, joka perustuu Bayes-verkon määrittämiseen asiantuntija-arvioita ja dataa käyttäen. Bayes-verkko [8] on matemaattinen malli satunnaismuuttujien välisille riippuvuuksille. Bayes-verkon määrittäminen pelkän asiantuntija-arvion perusteella voi olla työlästä, mutta vaadittua asiantuntijan työmäärää voidaan vähentää pääättelemällä verkko osittain datasta koneoppimista hyödyntäen.

1.2 Tärkeimmät käsitteet

Muuttuja on mitattavissa oleva ominaisuus. Mittauksen tuloksena havaitaan muuttujan arvo. Arvo voi vaihdella mittauksen kohteesta tai esimerkiksi mittausajasta riippuen. Järjestysasteikollisen muuttujan arvoilla on suuruusjärjestys, mutta niitä ei voida välttämättä esimerkiksi laskea yhteen. Välimatka-asteikollisen muuttujan arvot ovat lukuja, joten niillä voidaan tehdä laskutoimituksia. Välimatka-asteikolliset muuttujat voivat olla jatkuvia tai diskreettejä. Jatkuvat muuttujat voivat saada arvoinaan reaalilukuja. Diskreettien muuttujien arvojoukko voidaan numeroida kokonaisluvuilla. Data ja aineisto muodostuvat joukosta havaintoja ja niitä käytetään joskus synonyymeinä toisilleen. Datalla tarkoitetaan tässä työssä mitä tahansa havaintojen joukkoa ja aineistolla viitataan jotakin tarkoitusta varten koottuun dataan.

Todennäköisyysmalli on matemaattinen kuvaus tarkasteltavaan ilmiöön liittyvästä epävarmuudesta. Todennäköisyysmalli sisältää yhden tai useampia satunnaismuuttujia, jotka kuvaavat mahdollisten arvojen todennäköisyydet todennäköisyysjakaumana. Empiirisellä jakaumalla tarkoitetaan arvojen esiintymistiheyttä satunnaisesta tapahtumasta kerätyssä aineistossa eli otoksessa. Empiiristä jakaumaa voidaan käyttää esimerkiksi approksimointiin. Approksimointi tarkoittaa laskennallisesti vaikean yhtälön likimääräistä, mutta teoreettisesti lähes oikeaksi taattua ratkaisemista. Todennäköisyysmalliin voi kuulua jakaumiin vaikuttavia parametreja. Koneoppimisessa näiden parametrien arvot optimoidaan tarkasteltavaa ilmiötä kuvaaviksi datan perusteella. Optimoinnista voidaan tässä yhteydessä käyttää termejä oppiminen tai estimointi. Tämän optimointiprosessin täsmälliset vaiheet voidaan esittää algoritmina.

Todennäköisyysmallissa voidaan huomioida asiantuntija-arvioita, joiden hankkimista kutsutaan elisitoinniksi. Tätä ei tule sekoittaa elisitaatioon, joka on peitetty tiedonhankintamenetelmä [9].

Kohdemuuttujalla tarkoitetaan tässä työssä päättelyn kohteena olevaa satunnaismuuttujaa. Tämä voi olla esimerkiksi hypoteesi, jolla tarkoitetaan tässä työssä totuusarvolla väitelausetta. Lähestymistavalla tarkoitetaan tässä työssä todennäköisyysmallin hyödyntämiseksi määritettyjä toimenpiteitä.

1.3 Aikaisempi tutkimus

Koneoppimisen hyödyntämisestä sotilaskontekstissa on julkaistu lukuisia tutkimuksia kuluneiden 20 vuoden aikana. Tuore katsaus aiheeseen on tarjolla esimerkiksi artikkelissa [10] sekä Pro gradu -työssä [11]. Tiedustelussa koneoppimista on sovellettu muun muassa automatisoituun tiedonkeruuseen [12], satelliittikuvien prosessointiin [13], sentimenttianalyysiin [14], faktantarkistukseen [15] sekä terroristisen toiminnan ja radikalisoitumisen havaitsemiseen [16]. Tässä työssä koneoppimismenetelmänä käytetään Bayes-päätelyä, jonka käyttämistä Yhdysvaltain keskustiedustelupalvelu (CIA) tutki analyysimenetelmänä jo 1960-luvulla [17]. Myöhemmässä tutkimuksessa [18] on todettu Bayes-päätelyn voivan tuottaa tarkempia tuloksia kuin esimerkiksi kilpailevien hypoteesien menetelmä (Analysis of Competing Hypotheses, ACH [19]). Lisäksi menetelmän käyttöön annetun koulutuksen on todettu tarkentavan Bayes-päätelyn avulla saatuja tuloksia entisestään [20], toisin kuin ACH:n tapauksessa [21]. Bayes-verkkoja [8] on tutkittu 1990-luvulta lähtien. Bayes-verkkoja on käytetty tiedustelun alalla muun muassa ACH-menetelmän täydentämiseen [22] sekä tiedonkeruun priorisointiin [23]. Lisäksi niitä on käytetty muun muassa riskianalyyseissä [24], taloustieteessä [25], lääketieteessä [26] ja psykologiassa [27].

Koneoppimisalgoritmien tukemista asiantuntija-arvioin on tarkasteltu yleisellä tasolla muun muassa artikkelissa [28]. Bayes-verkkojen rakentamista pelkkiin asiantuntija-arvioihin nojaten on tutkittu paljon. Tutkimuksessa on keskitytty erityisesti mallin rakentamiseksi tarvittavan asiantuntijan työmäärän vähentämiseen [29]. Tässä työssä tarkastellaan järjestysasteikollisten solmujen menetelmän [30–34] hyödyntämistä Bayes-verkkojen rakentamisessa. Järjestysasteikollisten solmujen menetelmä on todettu toimivaksi monessa käytännön sovelluksessa [35]. Vaikka tämä menetelmä mahdollistaa asiantuntija-arvioiden hyödyntämisen, se ei tarjoa keinoja niiden myöhemmän tarkentamiseen datan avulla. Asiantuntija-arvioiden ja datan hyödyntäminen samanaikaisesti on mahdollista Bayes-verkoissa hyvin tunnetun multinomi-Dirichlet-mallin avulla sekä lineaaristen gaussisten Bayes-verkkojen tapauksessa. Näistä ensimmäisen huonona puolena on mallin rakentamiseksi vaadittu suuri asiantuntijan työmäärä. Jälkimmäinen menetelmä ei sovellu järjestysasteikollisen datan käsittelyyn [36].

1.4 Tutkimuskysymykset ja tutkimusmenetelmät

Tämän työn päätutkimuskysymyksenä on selvittää, miten sotilaallisen aktiivisuuden arviointia voidaan tukea dataan ja asiantuntija-arvioihin perustuvalla Bayes-verkolla. Alatutkimuskysymykset ovat

1. Miten koneoppimismenetelmien soveltuvuutta tiedusteluanalyysiin on aikaisemmin arvioitu?
2. Miten sotilaallisen aktiivisuuden arviointia voidaan tukea Bayes-päätelyyn perustuvilla menetelmillä?
3. Miten data ja asiantuntija-arviot yhdistetään ja hyödynnetään järjestysasteikollisten solmujen menetelmään perustuvassa Bayes-verkossa?
4. Miten asiantuntija-arvioihin ja dataan perustuvaa Bayes-verkkoa hyödyntävän lähestymistavan käyttökelpoisuus testataan sotilaallisen aktiivisuuden arvioinnissa?

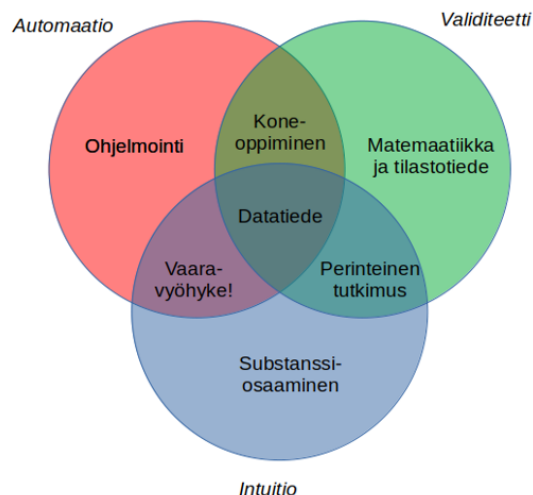
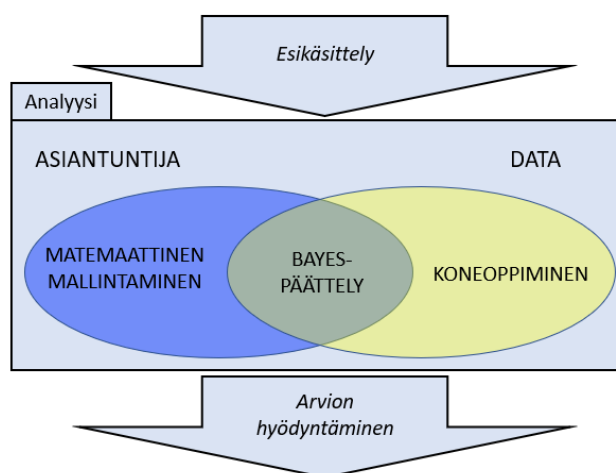
Alatutkimuskysymykseen 1 vastataan luvussa 2 kirjallisuustutkimuksella. Alatutkimuskysymykseen 2 vastataan luvussa 3 kirjallisuustutkimuksella sekä esittelemällä muunnelmia kirjallisuudesta löytyviin menetelmiin. Muunnelmien vertailussa hyödynnetään matemaattista analyysiä sekä algoritmien analyysiä [37]. Alatutkimuskysymykseen 3 vastataan luvussa 4 mallintamalla datan ja asiantuntija-arvioiden hyödyntäminen. Alatutkimuskysymyksen 4 vastataan luvussa 5 mallintamalla sotilaallisen aktiivisuuden arviointi esimerkkitalanteissa.

1.5 Viitekehys ja rajaus

Viitekehys

Tämän työn tutkimuskysymykset liittyvät johtopäätösten tekemiseen epävarmoista ilmiöistä. Painopiste on asiantuntija-arvioon ja dataan perustuvassa analyysissä. Huomiota kiinnitetään kuvan 1 vasemman puolen mukaisesti myös kerätyn datan esikäsittelyyn ja analyysin tuottaman arvion hyödyntämiseen. Nämä ovat myös muun muassa tiedustelu-analyysin vaiheita [1]. Esikäsittelystä datasta käytetään joskus myös termiä informaatio.

Todennäköisyyden käsitettä käytetään tässä työssä epävarmuuden mittana subjektiivisessä merkityksessä. Eri henkilöillä voi olla erisuuruinen arvio saman tapahtuman todennäköisyydestä.



Kuva 1: Tutkimuksen viitekehys. Vasemmalla on esitetty Bayes-päätelyn konteksti tässä tutkimuksessa. Oikealla Drew Conwayn esittämä datatieteen Venn-diagrammi [38] lisäyksillä, jotka esitti Andrew Silver [39].

Todennäköisyyksiä voidaan määrittää toistettavien tapahtumien lisäksi myös harvinaisille tapahtumille. Asiantuntija-arviona osin määriteltävää matemaattista mallia tarkennetaan koneoppimista käyttäen. Koneoppimismenetelmänä käytetään subjektiiviseen todennäköisyyskäsitykseen perustuvaa Bayes-päätelyä. Bayes-päätelyä voidaan käyttää johtopäätösten tekemiseen mallin parametreista, mutta tässä työssä sitä käytetään ensisijaisesti johtopäätöksien tekemiseen kohdemuuttujan arvoista.

Koneoppimisen ja substanssiosaamisesta syntyvien asiantuntija-arvioiden yhdistämistä voidaan kutsua kuvan 1 oikean puolen termeillä datatieteeksi, mutta tässä työssä ei hyödynnetä substanssiosaamista. Tutkimuskysymyksiä ei ole johdettu Puolustusvoimien operatiivisesta toiminnasta tai tarpeesta. Työ perustuu julkisiin lähteisiin ja sen tavoitteena on tiedustelun alalla sovellettavissa olevan uuden lähestymistavan kehittäminen.

Rajaus

Työ rajataan tarkastelemaan Bayes-verkkojen sovelluksia aluvuossa 3.3 esiteltävää järjestysasteikollisten solmujen menetelmää käyttäen. Työssä ei tarkastella Bayes-verkon rakenteen oppimista datasta, kuten ei myöskään dynaamisia Bayes-verkkoja.

Työssä ei hyödynnetä Puolustusvoimien operatiivisia järjestelmiä, dataa tai näiden asiantuntemusta. Työssä ei tarkastella tulosten operatiivista käytettävyyttä. Kehitettävää lähestymistapaa ei arvioida käyttäjäkokemusten perusteella. Asiantuntija-arvioita ei elisitoida substanssiosaajilta, vaan tutkija laatii esimerkeissä tarvittavat asiantuntija-arviot.

2 KONEOPPIMISEN SOVELTUVUUS TIEDUSTELUN TUEKSI

Koneoppimisen soveltamista sotilaskontekstissa käsittelevien tutkimusten määrän merkittävä kasvu 2010-luvulla [10] kertoo kasvaneesta mielenkiinnosta aiheeseen. Julkisesti saataville on kertynyt myös kirjallisuutta koneoppimismenetelmien käytöstä tiedusteluanalyysissä. Tässä luvussa vastataan kirjallisuuskatsauksen avulla alatutkimuskysymykseen 1: “miten koneoppimismenetelmien soveltuvuutta tiedusteluanalyysiin on aikaisemmin arvioitu’.” Luvun johtopäätökset huomioidaan myöhemmissä luvuissa, kun arvioidaan koneoppimismenetelmien soveltuvuutta sotilaallisen aktiivisuuden arviointiin.

Kirjallisuuskatsaus toteutettiin hakemalla “EBSCO military & government collection” -tietokannasta [40] enintään kymmenen vuotta vanhoja kirjoja. Haku kohdistettiin luokkaan “political science / intelligence & espionage”, josta etsittiin hakusanoja “artificial intelligence”, “machine learning” tai “algorithm” sisältäviä lähteitä. Haun tulokset rajattiin teoksen otsikon perusteella tiedusteluanalyysiä käsitteleviin teoksiin. Koneoppimisen soveltamista tiedusteluanalyysiin käsiteltiin sisällysluettelon perusteella kolmessa kirjassa [41–43]. Hakua täydennettiin Google Scholar -internetsivustolla [44] hakusanoilla “military intelligence” ja “machine learning” tehdyllä haulla. Tämä haku tuotti yli 2000 tulosta, joita kaikkia ei tarkastettu. Haun tuloksista valittiin kolme tulosta, joissa käsiteltiin koneoppimisen soveltamista tiedusteluanalyysiin. Nämä olivat vertaisarvioidut artikkelit [28, 45] ja konferenssijulkaisu [46].

2.1 Hyötyjä ja haasteita

Lähteiden käsittelemät hyödyt koneoppimisen soveltamisesta tiedusteluanalyysiin jakoutuivat ajankäytön tehostamiseen sekä analyysin laadun parantamiseen. Tietokoneet ovat hyviä suurten datamäärien ympärivuorokautiseen käsittelyyn, joten koneoppimista on käytetty relevantin tiedon keräämiseen, prosessointiin, visualisointiin ja tiivistämiseen suurista datamääristä [28, 42]. Erityisesti sen on koettu soveltuvan jatkuvaan ja lähes reaaliaikaiseen riskien tunnistamiseen ennakkovaroituksen antamiseksi sekä analyytikon ajan vapauttamiseen tiedon keruusta tiedon analysointiin [42]. Analyysin laadun osalta koneoppimisen etuna on päättelyn toistettavuus. Vaikka tietokonetta ei voi pitää ihmisen tavalla älykkäänä, sen toimintaan eivät vaikuta väsymys, ennakoasenteet, henkilösuhteet, organisaatiokulttuurien erot tai muut inhimilliset tekijät. Koneoppiminen mahdollistaa siten laajan aineiston käsittelyn yhtenäisesti ja virheitä vähentäen. Tämän on koettu voivan mahdollistaa tarkempia ennusteita sekä organisaatioiden sisäistä

ja välistä yhteistyötä [42].

Suurten odotusten ja hyvien kokemusten lisäksi lähteissä esitettiin koneoppimisen soveltamisesta tiedusteluanalyysiin myös kritiikkiä. Tästä erottui neljä haastetta, joista ensimmäinen on datan saatavuus. Monet koneoppimisalgoritmit tarvitsevat paljon dataa toimiakseen, mutta tiedustelusovelluksissa dataa ei välttämättä ole saatavilla [45, 46]. Datan säilytystä rajoittavien käytäntöjen johdosta suuria datamääriä ei välttämättä kerry [41]. Jos käytössä on suuria datamääriä, on tiedettävä, mitä siitä etsitään. Tiedustelusovelluksissa suuri aineisto saattaa sisältää vain pienen määrän relevanttia tietoa, jolloin ihmisen tarve saattaa jopa korostua [28]. Datan määrään liittyviä haasteita voidaan pyrkiä ratkaisemaan parantamalla olemassa olevan datan käytettävyyttä sekä käyttämällä koneoppimismenetelmiä, jotka hyödyntävät dataa mahdollisimman tehokkaasti. Tällaisia menetelmiä voivat olla datan augmentointi eli esimerkiksi datan monistaminen pienin muutoksin [47], siirto-oppiminen eli toisesta datasta opitun mallin hyödyntäminen käsillä olevaan ongelmaan [48] sekä Bayes-päätelyyn perustuvat menetelmät.

Datan määrän lisäksi sen laatu voi olla puutteellista. Vastustaja pyrkii harhauttamaan ja hämäämään, mutta vähäistä dataa ei tämän perusteella voida jättää huomiotta [41, 46]. Lähteiden luotettavuutta ja datan puuttumisen syitä tulee tästä syystä voida arvioida [41]. Kyky käsitellä epävarmoja havaintoja ja ilmaista arvioiden epävarmuuksia voivat siten olla hyödyllisiä ominaisuuksia tiedusteluun sovellettavalle koneoppimismenetelmälle. Merkittävät datan laadun puutteet ovat ongelma mille tahansa analyysille, eikä koneoppiminen ole tästä poikkeus.

Kolmantena haasteena tunnistettiin läpinäkyvyyden toteutuminen. Analyysin taustalla oleva prosessi tulisi kyetä avaamaan jaettavaksi, arvioitavaksi ja hyödynnettäväksi jatkokäyttöön [43]. Monet koneoppimisalgoritmit ovat toiminnaltaan ihmiselle käytännössä avautumattomia “mustia laatikoita”, mikä heikentää luottamusta niiden toimintaan ja hankaloittaa niiden oikeellisuuden tarkistamista [46, 49]. Tämä on merkittävä ongelma, sillä tiedustelusovelluksissa ihmisen tulisi voida ymmärtää ja tulkita päätelyn kulkua [28, 41, 45]. Itse päätelyprosessin tuoma ymmärrys voi joskus olla arvokkaampaa kuin sen tuottama vastaus. Läpinäkyvyyden ongelmaa on pyritty ratkaisemaan kehittämällä selitettävää tekoälyä (engl. explainable AI, XAI) [49, 50]. Tämän tärkeydestä sotilassovelluksissa kertoo osaltaan Yhdysvaltojen asevoimien tutkimusjärjestön vuosina 2017-2021 järjestämä ohjelma selitettävän tekoälyn kehittämiseksi [51].

Lähteiden nostamista haasteista neljäs liittyy mallipohjaisiin kvantitatiivisiin analyysimenetelmiin yleisesti liitettyyn monimutkaisuuteen ja tästä johtuvaan hitauteen [43]. Koneoppimisen soveltaminen vaatii tyypillisesti menetelmän asiantuntijoita ja paljon aikaa, jolloin tehokkaastakin sovelluksesta saatu hyöty voi kokonaisuutena jäädä vähäiseksi [41]. Tärkeänä ratkaisuna

tähän on pidetty uusien koneoppimisprojektien keskittymistä yksittäisen ongelman sijaan kokonaiseen ongelmien luokkaan [45]. Lisäksi koneoppimisen soveltamista voidaan helpottaa ylläpitämällä olemassa oleva data käyttökelpoisena. Esimerkiksi erään vuonna 2020 toteutetun kyselyn [52] mukaan alan asiantuntijat joutuvat käyttämään lähes puolet työajastaan datan käytettävyyden korjauksiin.

2.2 Yhteenveto

Kirjallisuuskatsauksen perusteella koneoppimisen hyödyt tiedusteluanalyysissä liittyvät ajan käytön tehostamiseen sekä analyysin laadun parantamiseen, kun taas haasteet liittyvät datan määrään ja laatuun sekä koneoppimisen läpinäkyvyyteen ja monimutkaisuuteen. Tunnistetuista haasteista läpinäkyvyys on jo avannut uuden suunnan koneoppimisen tutkimukselle. Myös datan määrään ja menetelmien monimutkaisuuteen on etsitty ratkaisuja. Haasteista vaikeimmaksi jää laaduton data, joka on toisaalta ongelma mille tahansa analyysille. Yhdessä lähteessä ei pidetty todennäköisenä tiedusteluanalyysin täyttä automatisointia koneoppimista tai muita tekoälyn menetelmiä käyttäen. Sen sijaan todennäköisenä kehityssuuntana pidettiin analyysin pysymistä kokeneiden ihmisten alana, mutta tapahtuvan yhä enemmän koneoppimisen tukemana [28, 42]. “Fusing algorithms and analysts” -kirjan kirjoittajat ennakoivat ihmisen ja tekoälyn suhteen muotoutuvan kuten tietokoneavusteisessa shakissa [28]. Tässä lajissa parhaiten eivät ole menestyneet yksinpelin mestarit tai parhaita yksinpelaajia voittaneet tietokoneohjelmat, vaan parhaiten tietokonetta apunaan käyttäneet pelaajat.

3 TODENNÄKÖISYYSPÄÄTTELYN MENETELMIÄ

Todennäköisyys on abstrakti epävarmuuden mittari. Sitä voidaan käsitellä todellisesta maailmasta muodostetun mallin perusteella joko intuitiivisesti tai matemaattisesti. Matemaattinen todennäköisyyden käsittely on usein hitaampaa, koska se vaatii täsmällisen mallin määrittelyn. Toisaalta se pakottaa avaamaan päättelyn taustalla olevat oletukset, mahdollistaa mallin arvioinnin ja on edellytys todennäköisyyksiin perustuvan päättelyn automatisoinnille. Matemaattinen todennäköisyyden käsittely ei myöskään kärsi intuitiivisen käsittelyn vinoumista. Päättelyssä käytetyn mallin poikkeamat todellisuudesta aiheuttavat epätarkkuuksia sekä intuitiivisessa että matemaattisessa todennäköisyyden käsittelyssä.

Tässä luvussa vastataan alatutkimuskysymykseen 2: “miten sotilaallisen aktiivisuuden arviointia voidaan tukea Bayes-päätelyyn perustuvilla menetelmillä”. Luvussa esitellään kirjallisuuden perustuen erilaisia Bayes-päätelyn menetelmiä ja arvioidaan esimerkkien avulla näiden käytettävyyttä tiedusteluanalyyseissä. Ensimmäisessä alaluvussa määritellään Bayes-päätely ja tarkastellaan sen tiedusteluanalyyseissä käytettyä sovellusta. Alaluvussa 3.2 tarkastellaan riippuvuussuhteiden määrittelyä Bayes-verkkojen avulla ja alaluvussa 3.3 Bayes-verkkojen rakentamista asiantuntija-arvion perusteella. Alaluvussa 3.4 esitellään menetelmä Bayes-verkkojen parametrien estimointiin. Kaikkia tämän luvun menetelmiä käytetään luvussa 4 esiteltävässä Bayes-päätelyä ja järjestysasteikollisten solmujen menetelmää hyödyntävässä lähestymistavassa.

3.1 Bayes-päätely

Teemme päivittäin johtopäätöksiä epäsuorasti tukeutuen suorien havaintojen sijaan viitteisiin asioiden tilasta. Valitettavasti emme tyypillisesti ole hyviä käsittelemään tällaista epäsuoraa informaatiota. Esimerkiksi Monty Hall -ongelma osoittautuu yksinkertaisuudestaan huolimatta useimmille hankalaksi [53]. Ongelmassa koehenkilö saa valita yhden kolmesta ovesta, joista yhden takana on palkinto. Tämän jälkeen yksi henkilön valitsematta jättämistä vaihtoehdoista näytetään tyhjäksi ja hänelle annetaan mahdollisuus vaihtaa valintaansa toiseen jäljellä olevista ovista. Todennäköisyyden laskusääntöjen perusteella vaihtaminen kaksinkertaistaa voiton mahdollisuuden.

Bayes-päätely on matemaattista todennäköisyyspäättelyä käyttävä menetelmä, joka mahdollistaa epäsuoran informaation käsittelyn analyyttisesti. Sen lähtökohtana on subjektiivisen toden-

näköisyyden käyttö epävarmuuden mittana, ja sen on osoitettu olevan ainoa tämän lähtökohdan kanssa johdonmukainen päättelymenetelmä [54]. Bayes-päätelyssä henkilö ilmaisee näkemyksensä epävarmasta päättelyn kohteesta θ ja päivittää näkemystään järjestelmällisesti saatuaan uutta tietoa. Alustavaa subjektiivista näkemystä kutsutaan prioriksi ja tämä muotoillaan todennäköisyysjakaumana $P(\theta)$. Havaintojen perusteella päivitettyä jakaumaa $P(\theta | D)$ kutsutaan posteriorijakaumaksi. Menetelmän ainoa kaava on ehdollisen todennäköisyyden laskusääntö eli Bayesin kaava

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} = \frac{P(D | \theta)P(\theta)}{\int_{\theta} P(D | \theta)P(\theta)}. \quad (3.1)$$

Kaavan käyttöä varten on määriteltävä uskottavuusfunktio $P(D | \theta)$, joka kertoo havaintojen D todennäköisyyden päättelyn kohteen θ ollessa tiedossa. Esimerkissä 3.1 havainnollistetaan hypoteesin todennäköisyyden päivittäminen datan perusteella Bayes-päätelyä käyttäen.

Esimerkki 3.1. *Hypoteesin H mukaan valtio X valmistelee hyökkäystä valtion Y alueelle seuraavan 90 päivän kuluessa. Vastahypoteesin V mukaan näin ei käy. Ennen uusien havaintojen saamista analyttikko pitää hypoteesin H todennäköisyytenä $P(H) = 2\%$. Vastahypoteesin todennäköisyys on tällöin $P(V) = 98\%$. Ensimmäinen havainto SH on valtion X järjestämä poikkeuksellisen suuri sotaharjoitus. Toinen havainto PR on valtion Y tekemä protesti valtion X toiminnasta. Analyttikko käyttää arvionsa päivittämiseen Bayesin kaavaa ja arvioi tätä varten todennäköisyyden $P(SH | H)$. Tämän mukaan näkisimme vastaavan sotaharjoituksen tilanteessa, jossa hyökkäystä todella valmisteltaisiin. Analyttikko toistaa arvion vastahypoteesille V ja tekee saman toiselle havainnolle PR . Analyttikon arviot ovat:*

$$\begin{aligned} P(SH | H) &= 95\% & P(SH | V) &= 20\% \\ P(PR | H) &= 90\% & P(PR | V) &= 70\%. \end{aligned}$$

Jos tapahtumat SH ja PR oletetaan riippumattomiksi, rationaalisesti todennäköisyyksiä käsittelevän analyttikon tulisi näiden havaintojen jälkeen päivittää hypoteesin H todennäköisyydeksi

$$\begin{aligned} P(H | SH, PR) &= \frac{P(SH, PR | H)P(H)}{P(SH, PR | H)P(H) + P(SH, PR | V)P(V)} \\ &= \frac{P(SH | H)P(PR | H)P(H)}{P(SH | H)P(PR | H)P(H) + P(SH | V)P(PR | V)P(V)} \\ &= \frac{0,95 \cdot 0,9 \cdot 0,02}{0,95 \cdot 0,9 \cdot 0,02 + 0,2 \cdot 0,7 \cdot 0,98} \\ &\approx 11\%. \end{aligned}$$

Arvio hyökkäyksen todennäköisyydestä on havaintojen SH ja PR jälkeen 11 %. Vastahypoteesin V todennäköisyydeksi jää 89 %. Seuraavat havainnot voidaan käsitellä samalla tavalla todennäköisyyksien päivittämistä jatkaen.

Käsitys hypoteesin paikkansa pitämisestä lasketaan esimerkissä 3.1 Bayes-päätelyä käyttäen. Tätä varten esitetään arvio tehtyjen havaintojen todennäköisyyksistä molempien hypoteesien tapauksissa sekä arvio hypoteesin paikkansa pitävyydestä ennen havaintojen saamista. Näiden arvioiden muodostaminen saattaa olla hypoteesin posterioritodennäköisyyden $P(H | D)$ intuitiivista arvioimista helpompaa. Esimerkissä 3.1 tehty Bayes-päätely ei poista tarvetta arvioiden esittämiselle, vaan mahdollistaa vaikean arvion johdonmukaisen muodostamisen helpommin arvioitavien todennäköisyyksien perusteella.

CIA:n 1970-luvulla tuottama käsikirja “Handbook of Bayesian analysis for intelligence” [55] esittelee Bayes-päätelyä hyödyntävän lähestymistavan tiedusteluanalyysin käyttöön. Tämä lähestymistapa vastaa matemaattiselta osuudeltaan esimerkkiä 3.1, sillä erotuksella että se mahdollistaa useampienkin kuin kahden hypoteesin samanaikaisen tarkastelun. Lähestymistapaa kuvataan käsikirjassa rationaaliseksi, johdonmukaiseksi ja objektiiviseksi keinoksi koostaa näennäisesti toisiinsa liittymättömistä faktoista kokonaisarvio. Etuina pidetään arvion koostamista matemaattisesti analyttikon intuition sijaan sekä lopputuloksen luontaista esittämistä numeerisena todennäköisyytenä. Lähestymistavan huomautetaan olevan tarkasti dokumentoitavissa, toistettavissa ja ohjaavan analyttikon harkitsemaan vaihtoehtoisia selityksiä havainnoille. Toisessa CIA:n asiakirjassa [56] samalta vuosikymmeneltä huomautetaan Bayes-päätelyn toimivan myös yksittäisen havainnon diagnostisuuden määrittämiseen. Diagnostisuudella tarkoitetaan havainnon merkittävyyttä todennäköisten hypoteesien erottelussa vähemmän todennäköisistä. Bayes-päätelyn heikkouksina pidettiin inhimillisiä tekijöitä, kuten mahdollisuutta todennäköisyyksien tahalliseen manipulointiin, mahdollisia haasteita todennäköisyyksien käsittelyssä sekä lähestymistavan työläyttä [55].

Mallintamisen näkökulmasta edellä kuvatun lähestymistavan oletus havaintojen riippumattomuudesta toisistaan ja kuluneesta ajasta voi olla heikkous. Hypoteesien todennäköisyyttä päivitetään uusia havaintoja saataessa, mutta edeltävien havaintojen vaikutusta ei arvioida tällöin uudelleen. Vanhojen havaintojen vaikutus pysyy samana, vaikka hypoteesin sisältämä määre, kuten “seuraavan 90 päivän kuluessa”, riippuu tarkasteluhetkestä. CIA:n käsikirja [55] esittää ratkaisuna havaintojen pudottamista pois laskuista jonkin ajan kuluttua. Tämä on vain osittainen ratkaisu, sillä se ei ota kantaa alkuperäisen prioritodennäköisyyden vanhenemiseen. Riippuvuus ajasta voidaan huomioida arvioimalla jokainen havainto uudestaan, kun hypoteesin todennäköi-

syyttä päivitetään. Tämä voi lisätä lähestymistavan kuormittavuutta.

Oletus havaintojen keskinäisestä riippumattomuudesta tuottaa ongelman, jota voidaan kutsua kahdesti laskemiseksi. Esimerkiksi jos yhden sota-aluksen havaitseminen tietyllä alueella todetaan Bayesin kaavalla kohtalaiseksi indikaattoriksi konfliktille, niin kolmen aluksen havaitseminen samalla alueella tulkitaan riippumattomuusoletuksen myötä erittäin vahvaksi indikaattoriksi. Todellisuudessa alukset eivät ole alueella toisistaan riippumatta, vaan yhden havaitseminen voi samalla olla vahva indikaattori useampien alusten havaitsemiselle. Toisen ja kolmannen aluksen havaitseminen ei välttämättä tuota yhtä paljon informaatiota kuin ensimmäinen havainto. Edellä käsitelty CIA:n lähestymistapa [55] ratkaisee ongelman vaatimalla havaintojen erittelyn toisistaan mahdollisimman riippumattomiin osiin ennen varsinaista laskentaa. Esimerkiksi kolme havaintoa eri sota-aluksista samaan aikaan ja samalla alueella tulee käsitellä yhtenä kolmen sota-aluksen havaintona. Ratkaisu ei ole ongelmaton, sillä riippuvuuksien käsittelyä pidetään erityisen hankalana aiheena matemaattisesti suuntautumattomille analyytikoille [56].

Riippumattomuusoletuksesta seuraava kahdesti laskemisen ongelma ei koske ainoastaan Bayes-päätelyä. Siitä kärsii myös esimerkiksi ACH tarjoamatta työkaluja sen ratkaisemiseen [18]. Bayes-päätelyssä riippumattomuusoletusta ei tarvita. Käytännössä tästä luopuminen tarkoittaa kunkin havainnon todennäköisyyden arvioimisessa jonkin hypoteesin voimassa olon olettamisen lisäksi kaikkien edellisten havaintojen huomioimista. Esimerkissä 3.1 todennäköisyyden $P(PR | H)$ sijaan käytetään tällöin todennäköisyyttä $P(PR | SH, H)$, eli arvioidaan todennäköisyys valtion Y tekemälle protestille olettaen, että valtio X on aloittanut hyökkäysvalmistelut ja tämän järjestämä sotaharjoitus on havaittu. Tämä on tapahtumien täyttä riippumattomuutta luontevampi tapa arvioida uusia havaintoja, mutta todennäköisyyksien arviointi kaikkien edeltävien tapahtumien suhteen voi käydä työlääksi, jos saman arvion päivittämistä jatketaan pitkään.

Siinä missä esimerkin 3.1 laskut ovat ratkaistavissa kynällä ja paperilla, Bayes-päätely voi monimutkaisemmissa tilanteissa johtaa jopa analyyttisesti ratkeamattomiin yhtälöihin. Nämä voivat olla approksimoitavissa nykyaikaisten tietokoneiden laskentakapasiteettia hyödyntäen. Yksi tällainen menetelmä esitellään alaluvussa 3.4. Seuraavaksi tarkastellaan Bayes-verkkoja, jotka mahdollistavat riippuvuuksien esittämisen graafisesti.

3.2 Bayes-verkot

Bayes-verkko [8] (engl. Bayesian network, Bayesian belief network tai Bayes net) on monipuolinen ja intuitiivinen todennäköisyysmalli. Se koostuu suunnatusta syklittömästä verkosta

eli DAG:ista (engl. directed asyclic graph) sekä satunnaismuuttujien ehdollisista todennäköisyysjakaumista. DAG koostuu mallin satunnaismuuttujat kuvaavista solmuista sekä näiden väliset suorat riippuvuussuhteet määrittävistä nuolista. Kun solmusta A kulkee nuoli solmuun B , niin solmua A kutsutaan solmun B isäsolmuksi ja solmua B solmun A lapsisolmuksi. Syklittömyydellä tarkoitetaan, että yhdestäkään verkon solmusta A ei ole mahdollista päätyä nuolten suuntaisesti takaisin samaan solmuun A . [36]

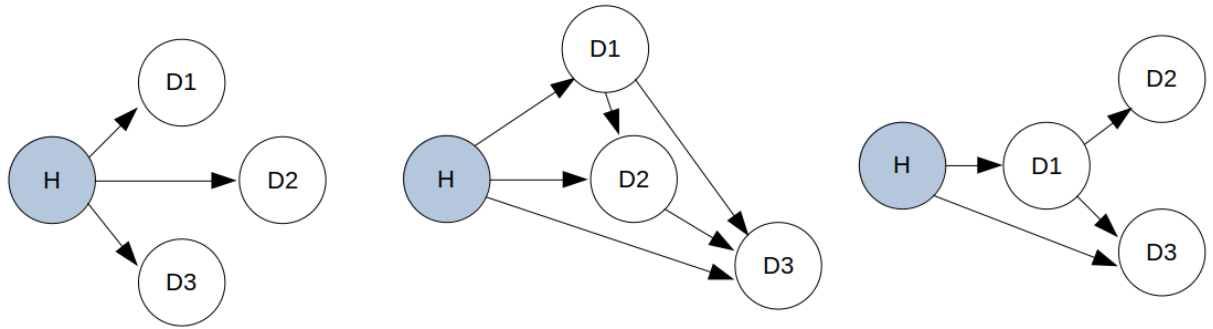
Jokaista solmua vastaavalle satunnaismuuttujalle määritetään ehdollinen todennäköisyysjakauma sen isäsolmujen suhteen. Nuolten määrittämät suorat riippuvuudet merkitsevät isäsolmun tilan tuntemisen antavan tietoa lapsisolmun tilasta. Suorien riippuvuuksien lisäksi kahden satunnaismuuttujan riippuvuus voi olla ehdollista, vaikka niitä vastaavien solmujen välillä ei kulkisikaan nuolta. Kahta satunnaismuuttujaa X ja Y sanotaan ehdollisesti riippumattomiksi satunnaismuuttujien Z suhteen, jos tieto satunnaismuuttujan X arvosta ei anna tietoa satunnaismuuttujan Y arvosta satunnaismuuttujien Z arvojen ollessa tunnetut. Tällöin

$$P(Y | X, Z) = P(Y | Z).$$

Jos X , Y ja Z eivät toteuta tätä ehtoa, satunnaismuuttujien X ja Y sanotaan olevan ehdollisesti riippuvia satunnaismuuttujien Z suhteen. Satunnaismuuttujien välinen ehdollinen tai suora riippuvuus ei välttämättä tarkoita niihin liittyvien tapahtumien olevan kausaalisesti toistensa seurauksia. [36]

Kuvassa 2 on kolme esimerkkiä DAG:ista. Jokainen näistä mallintaa hypoteesin H ja havaintojen $D1$, $D2$ ja $D3$ väliset riippuvuudet eri tavoin. Vasemmanpuoleisessa verkossa kaikki havainnot ovat ehdollisesti riippumattomia toisistaan hypoteesin arvon ollessa tunnettu. Tämä vastaa alaluvussa 3.1 kuvattua CIA:n 1970-luvun lähestymistapaa. Keskimmaisessä täyden riippuvuuden mallissa kaikki satunnaismuuttujat riippuvat suoraan toisistaan. Tämä vastaa alaluvussa 3.1 esitettyä täyden riippuvuuden mallia. Oikeanpuoleisessa verkossa esiintyy erilaisia ehdollisia riippumattomuuksia. Havainnon $D3$ ehdollinen jakauma on määritetty hypoteesin ja havainnon $D1$ suhteen mutta ei havainnon $D2$ suhteen. Havainnon $D2$ ehdollinen jakauma on määritetty vain havainnon $D1$ suhteen. Havainto $D2$ on ehdollisesti riippumaton sekä hypoteesista että havainnosta $D3$, kun havainto $D1$ on tunnettu.

Bayes-verkon ehdollisten todennäköisyyksien määrittämisen työläys riippuu suorien riippuvuuksien määrästä. Kuvan 2 verkoista vasemmanpuoleisimman ehdollisten todennäköisyyksien määrittäminen on yksinkertaisinta, sillä jokaisen lapsisolmun ehdollinen jakauma määritetään



Kuva 2: Kolme esimerkkiä suunnatuista syklittömistä verkoista eli DAG:eista. Jokaisessa verkossa on neljä solmua: hypoteesi H sekä kolme havaintoa D1 - D3.

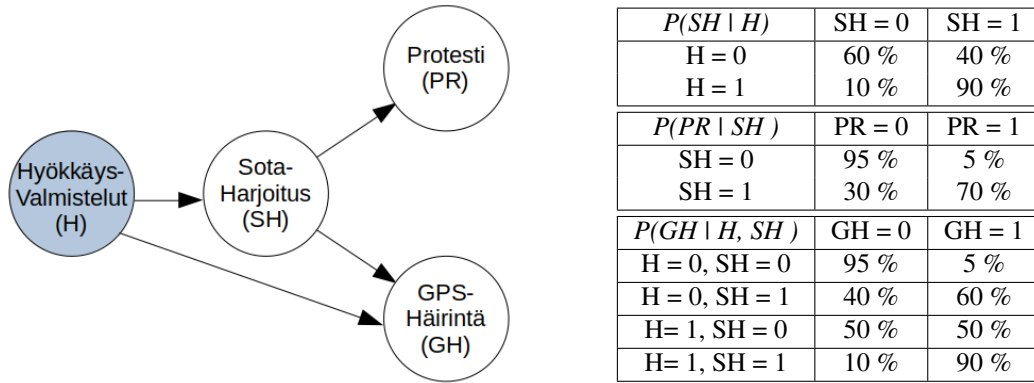
ainoastaan hypoteesin suhteen. Jos kaikki muuttujat ovat joko tosia tai epätosia, ehdollisia todennäköisyysjakaumia tarvitaan kaksi jokaista lapsisolmua kohti, yhteensä $3 \cdot 2 = 6$. Keskimäisessä tapauksessa mallin määrittely vaatii $2 + 4 + 8 = 14$ ja oikeanpuoleisessa $2 + 2 + 4 = 8$ ehdollista jakaumaa. Täyden riippuvuuden malli on työläin, sillä sen työläys kasvaa eksponentiaalisesti muuttujia lisättäessä. Oikeanpuoleista mallia voidaan pitää vasemman ja keskimmäisen verkon edustamien ääripäiden välisenä kompromissina. Suurenkaan mallin ehdollisten todennäköisyyksien määrittämisen työmäärä ei nouse korkeaksi, jos yksittäisen solmun ehdollinen jakauma riippuu korkeintaan muutamasta muusta solmusta. Vain tärkeimmät suorat riippuvuus-suhteet sisältävä malli voi kuvata ilmiön täyden ehdollisen riippumattomuuden mallia tarkemmin, mutta täyden riippuvuuden mallia kevyemmällä työmäärällä.

Bayes-verkko mahdollistaa johtopäätösten tekemisen kaikista mallin satunnaismuuttujista, vaikka niistä vain osan arvo olisi tiedossa. Seuraavassa esimerkissä arvioidaan hypoteesin todennäköisyys yhden havainnon puuttuessa.

Esimerkki 3.2. *Valtio Y väittää protestissaan valtion X käynnistäneen poikkeuksellisen suuren sotaharjoituksen ja valmistelevan hyökkäystä. Valtion X mahdollisesta sotaharjoituksesta ei ole saatavilla luotettavia tietoja. Valtioiden X ja Y raja-alueilla on tehty luotettavia havaintoja GPS-häirinnästä. Analyytikko kuvaa käsityksensä tilanteen riippuvuussuhteista ja ehdollisista todennäköisyyksistä kuvassa 3 esitetyllä Bayes-verkolla. Analyytikon alustavan arvion mukaan valtion X todennäköisyys hyökkäyksen valmistelulle on $P(H) = 5\%$. Lasketaan seuraavaksi GPS-häirinnän GH ja protestin PR havaintojen jälkeinen todennäköisyys hyökkäysvalmisteluille H ilman tietoa sotaharjoituksesta SH.*

Kokonaistodennäköisyys muodostuu ehdollisten riippumattomuuksien avulla

$$\begin{aligned} P(H, SH, PR, GH) \\ = P(H) P(SH | H) P(PR | SH) P(GH | SH, H). \end{aligned}$$



Kuva 3: Esimerkki Bayes-verkosta. Suunnattu sykliton verkko on vasemmalla ja tätä vastaavat ehdolliset todennäköisyydet oikealla. Kaikilla verkon solmuja vastaavilla satunnaismuuttujilla on kaksi tilaa: epätosi ja tosi. Nämä merkitään 0 ja 1. Ehdolliset jakaumat on määritetty taulukon riveillä. Protesti PR ja sotaharjoitus SH riippuvat suoraan vain yhdestä solmusta, joten niille molemmille on määritetty kaksi ehdollista jakaumaa. GPS-häirintä GH riippuu suoraan kahdesta solmusta, joten sille on määritetty neljä ehdollista jakaumaa. Hyökkäysvalmisteluilla H ei ole isäsolmuja eikä siten myöskään ehdollista jakaumaa.

Todennäköisyys hyökkäysvalmisteluille on

$$\begin{aligned}
 &P(H = 1 | PR = 1, GH = 1) \\
 &= \sum_{i=0}^1 P(H = 1, SH = i | PR = 1, GH = 1) \\
 &= \sum_{i=0}^1 P(H = 1, SH = i, PR = 1, GH = 1) / P(PR = 1, GH = 1) \\
 &= \sum_{i=0}^1 P(H = 1, SH = i, PR = 1, GH = 1) / \left(\sum_{j=0}^1 \sum_{k=0}^1 P(H = j, SH = k, PR = 1, GH = 1) \right) \\
 &= \frac{0,05 \cdot 0,1 \cdot 0,05 \cdot 0,5 + 0,05 \cdot 0,9 \cdot 0,7 \cdot 0,9}{0,95 \cdot 0,6 \cdot 0,05 \cdot 0,05 + 0,95 \cdot 0,4 \cdot 0,7 \cdot 0,6 + 0,05 \cdot 0,1 \cdot 0,05 \cdot 0,5 + 0,05 \cdot 0,9 \cdot 0,7 \cdot 0,9} \\
 &\approx 15\%.
 \end{aligned}$$

Havainnot GPS-häirinnästä ja valtion Y protestista nostavat siis analyytikon määrittämän Bayes-verkon perusteella arvion valtion X hyökkäysvalmistelujen todennäköisyydestä viidestä prosentista 15 prosenttiin.

Esimerkki 3.2 osoittaa Bayes-päätelyä voitavan käyttää Bayes-verkkoon perustuvassa mallissa, vaikka dataa puuttuu. Toisaalta tämä yksinkertainen esimerkki johti työläisiin laskuihin. Laskenta tehdään käytännössä tietokoneella, mutta vaadittu laskentamäärä kasvaa nopeasti mallin monimutkaistuessa [57]. Laskennan helpottamiseksi on olemassa algoritmeja, joista yksi esitellään alaluvussa 3.4.

Bayes-verkkojen vahvuutena on puuttuvan datan käsittelyn lisäksi mahdollisuus hyödyntää sekä

empiiristä dataa että subjektiivisia estimaatteja. Mallin rakenne sekä sen ehdolliset todennäköisyydet voidaan määrittää asiantuntija-arvion, datan tai näiden yhdistelmän perusteella. Tähän palataan luvussa 4. Bayes-verkkojen käyttöä voi rajoittaa vaatimus verkon syklittömyydestä. Solmujen X ja Y välillä ei voida mallintaa takaisinkytkentää. Verkko voidaan kuitenkin rakentaa huomioiden ajan kuluminen. Merkitsemällä satunnaismuuttujalla M_i muuttujaa M ajanhetkellä i voidaan tehdä esimerkiksi malli, jossa satunnaismuuttuja X_1 vaikuttaa suoraan satunnaismuuttujaan Y_2 , joka vaikuttaa suoraan satunnaismuuttujaan X_3 . Tämä on syklitön verkko, sillä X_1 ja X_3 ovat eri satunnaismuuttujia.

3.3 Järjestysasteikollisten solmujen menetelmä

Järjestysasteikollisten solmujen menetelmä eli RNM (engl. ranked nodes method) [30, 32] tarjoaa nopean tavan määrittää järjestysasteikollisista solmuista koostuvan Bayes-verkon ehdolliset todennäköisyydet. Kullekin verkon solmulle määritetään jakauma, jonka odotusarvo muodostuu sen isäsolmujen funktiona. Kun isäsolmuja on n kappaletta, tarvitsee menetelmää käytettäessä määrittää enintään $n + 1$ parametrin arvoa, joista ehdolliset todennäköisyydet tämän jälkeen seuraavat. Jos solmulla on esimerkiksi viisi mahdollista tilaa ja kolme isäsolmua, joista edelleen jokaisella on viisi mahdollista tilaa, tarvitaan kuvan 3 mukaisen ehdollisten todennäköisyyksien taulukon muodostamiseen arvio $5^4 - 5^3 = 500$ todennäköisyydestä. Järjestysasteikollisten solmujen menetelmällä ehdolliset todennäköisyydet määritetään tässä tilanteessa valitsemalla menetelmän neljästä funktiosta yksi ja määrittämällä tälle enintään neljä parametria.

Valittu funktio määrittää periaatteen, jolla lapsisolmun todennäköisin arvo μ muodostuu sen isäsolmujen arvoista. Ehdolliset todennäköisyydet lasketaan välille $[0, 1]$ rajatun normaalijakauksen $N(\mu, \sigma^2)$ perusteella. Hajontaparametrin σ suuruus ei riipu isäsolmujen arvoista. Ehdollisten todennäköisyyksien määrittäminen perustuu solmujen mallintamiseen diskreetteinä muuttujina, joiden jokainen tila vastaa välin $[0, 1]$ osaa. Väli jaetaan tasaisesti solmun tilojen kesken. Jos tilat ovat esimerkiksi ”erittäin pieni”, ”pieni”, ”keskisuuri”, ”suuri” ja ”erittäin suuri”, näitä vastaavat välit ovat $[0, 1/5]$, $[1/5, 2/5]$, $[2/5, 3/5]$, $[3/5, 4/5]$ ja $[4/5, 1]$. Tilojen määrä m voi vaihdella verkon solmujen välillä.

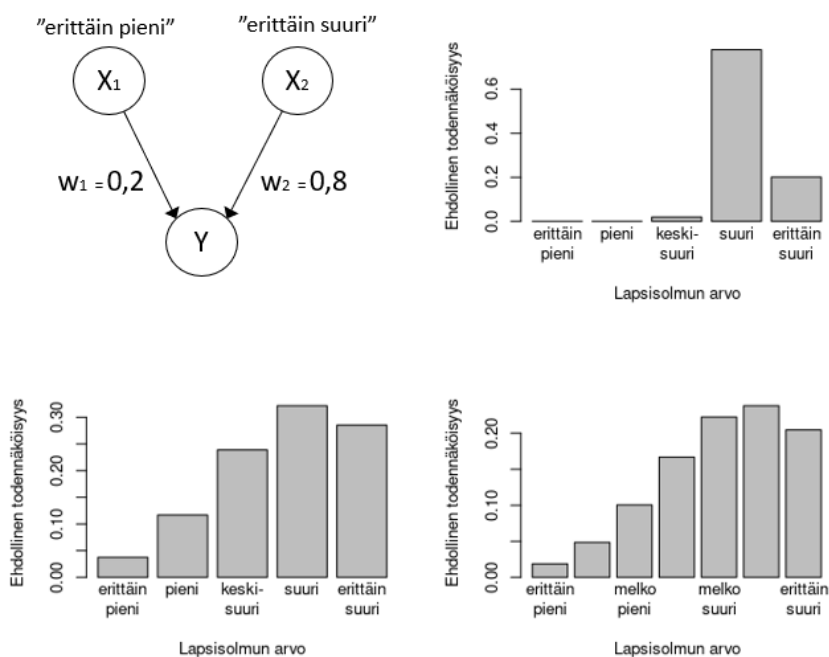
Järjestysasteikollisten solmujen menetelmän neljä RNM-funktiota mahdollistavat erilaisten riippuvuussuhteiden kuvaamisen. Näistä funktioista ensimmäinen on painotettu keskiarvo WMEAN. Lapsisolmun jokaiselle isäsolmulle määritetään tätä käytettäessä painokerroin w_i . Jokaisen painokertoimen on oltava vähintään nolla ja enintään yksi. Kun isäsolmujen lukumäärä on n , pai-

nokertoimien summan on oltava $\sum_{i=1}^n w_i = 1$. WMEAN-funktio on

$$\text{WMEAN}(x, w) = \sum_{i=1}^n w_i x_i, \quad (3.2)$$

missä x_i on yksittäisen isäsolmun tila. Painokertoimien summan muodostaman rajoitteen johdosta yhdelle isäsolmulle ei tarvitse määrittää painokerrointa, sillä tämä on laskettavissa muista painoista.

Ehdollisen jakauman määräytyminen WMEAN-funktiota käytettäen on havainnollistettu kuvassa 4. Satunnaismuuttujan Y ehdollinen jakauma määritetään kuvassa sen isäsolmujen X_1 ja X_2 suhteen. Valitut painot antavat isäsolmulle X_2 suuremman painoarvon, joten satunnaismuuttujan Y ehdollisen jakauman todennäköisin arvo on lähempänä solmun X_2 kuin solmun X_1 arvoa. Suurempi hajontaparametrin arvo tuottaa tasaisemman jakauman. Hajontaparametrin muuttaminen ei kuitenkaan muuta jakauman todennäköisintä arvoa. Lapsisolmun tilojen määrän kasvattaminen muuttaa jakauman todennäköisimmän arvon, koska arvojoukko muuttuu. Lapsisolmun tilojen määrä ei vaikuta jakauman huipukkuuteen.



Kuva 4: Satunnaismuuttujan Y jakauma ehdolla $X_1 = \text{"erittäin pieni"}$ ja $X_2 = \text{"erittäin suuri"}$. Jakauma lasketaan RNM-menetelmän WMEAN-funktiolla. Muuttujan X_1 paino on $w_1 = 0,2$ ja muuttujan X_2 paino on $w_2 = 0,8$. Ylärivillä oikealla esitetty jakauma on laskettu käyttäen pienempää hajontaa σ kuin vasemmalla alhaalla esitetty jakauma. Jakauman todennäköisin arvo pysyy samana. Alarivillä oikealla esitetyn jakauman mahdollisten arvojen määrä m on alarivillä vasemmalla esitettyä jakaumaa suurempi.

Menetelmän toinen funktio on painotettu minimi WMIN. Tämä mahdollistaa isäsolmujen pieniä arvoja painottavien suhteiden mallintamisen. Tällainen voi olla esimerkiksi tehokkuuden

kuvaaminen koulutuksesta ja välineistä riippuvana siten, että koulutus ilman välineitä tai välineet ilman koulutusta ovat merkityksettömiä. Funktion käyttäytymistä säädetään määrittämällä painokerroin jokaiselle isäsolmulle. Jokaisen WMIN-funktion painokertoimen tulee olla vähintään yksi. Funktio on

$$\text{WMIN} = \min_{i=1,\dots,n} \left\{ \frac{w_i x_i + \sum_{j \neq i} x_j}{w_i + n - 1} \right\}. \quad (3.3)$$

Suurilla painojen arvoilla WMIN toimii kuten tavallinen minimifunktio. Toisessa ääripäässä, kaikkien painojen ollessa 1, funktio toimii kuten aritmeettinen keskiarvo.

Menetelmän kolmas funktio WMAX toimii vastaavasti kuin WMIN korvaten minimin maksimilla. Neljäs funktio MIXMINMAX on painotettu keskiarvo isäsolmujen suurimmasta ja pienimmästä arvosta. Tällä funktiolla on kaksi parametria: painokertoimet w_{min} pienimmälle ja w_{max} suurimmalle isäsolmun arvolle. Näiden arvon on oltava vähintään nolla ja niiden summan on oltava yksi, joten jälkimmäinen voidaan määrittää ensimmäisestä $w_{max} = 1 - w_{min}$. Funktio on

$$\text{MIXMINMAX} = w_{min} \min_{i=1,\dots,n} \{x_i\} + w_{max} \max_{i=1,\dots,n} \{x_i\}. \quad (3.4)$$

Ehdollinen jakauma $P(Y | X_1, \dots, X_n)$ muodostetaan seuraavalla tavalla. Jokaisen isäsolmun arvo x_i tulkitaan välinä $[a_i, b_i]$. Tämä väli saadaan jakamalla väli $[0, 1]$ isäsolmujen mahdollisia arvoja vastaaviin osiin. Esimerkiksi viisiportaisen asteikon alin tila tulkitaan välinä $[0, 1/5]$. Ehdolliset todennäköisyydet lasketaan valitun RNM-funktion integraalina näiden välien yli. Tätä voidaan approksimoida seuraavalla menetelmällä [32, 35]:

1. Kunkin isäsolmun X_i tilaa x_i vastaavalta väliltä $[a_i, b_i]$ poimitaan tasaisin välein s pistettä aloittaen pisteestä a_i ja päättäen pisteeseen b_i .
2. Poimituista pisteistä muodostetaan kaikki kombinaatiot $\{z_1, z_2, \dots\}$ siten, että kombinaation $z_k = \{z_k^{(1)}, z_k^{(2)}, \dots\}$ alkio $z_k^{(i)}$ on isäsolmun X_i tilaa vastaavalta väliltä poimittu piste. Kun isäsolmuja on n kappaletta, niin kombinaatioita muodostuu s^n kappaletta.
3. Jokaista kombinaatiota z_k kohti lasketaan arvo $\mu_k = f(z_k, w)$, jossa f on solmulle yllä esitettyistä vaihtoehtoista määritetty funktio painoilla w .

4. Tilaa y vastaavan välin $[a_y, b_y]$ ehdollinen todennäköisyys lasketaan approksimaationa

$$P(Y | X_1, \dots, X_n) \approx \frac{1}{s^n} \sum_{k=1}^{s^n} \int_{a_y}^{b_y} \phi_{[0,1]}(u; \mu_k, \sigma^2) du, \quad (3.5)$$

missä $\phi_{[0,1]}$ on välille $[0, 1]$ rajatun normaalijakauman tiheysfunktio.

RNM-menetelmä vähentää Bayes-verkon ehdollisten todennäköisyyksien määrittämisen työmäärää. Perusmuodossaan se ei kuitenkaan tarjoa keinoa ehdollisten todennäköisyyksien määrittämiseen datan avulla. Luvussa 4 esitetään uusi lähestymistapa, joka mahdollistaa datan hyödyntämisen RNM-menetelmän yhteydessä. Tässä lähestymistavassa approksimaatio (3.5) korvataan datan asteikosta riippuvalla muunnoksella. Järjestysasteikolliselle datalle käytetään satunnaislukuihin perustuvaa muunnosta ja jatkuvalle välimatka-asteikolliselle datalle välimatka-asteikollista muunnosta. Nämä muunnokset määritellään seuraavaksi.

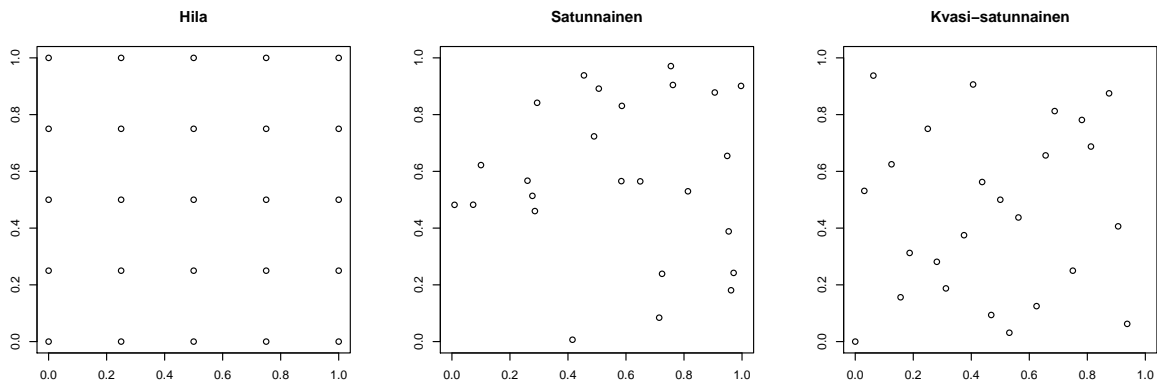
Satunnaislukuihin perustuva muunnos

Perusmuotoisessa RNM-menetelmässä Bayes-verkon solmujen ehdolliset todennäköisyydet lasketaan approksimaationa kaavalla (3.5). Approksimaation tarkkuus riippuu välin $[0, 1]$ jaon tiheyden määrittävästä parametrilla s . Aiemmassa tutkimuksessa [35] osoitettiin kaavan (3.5) tuottamien ehdollisten todennäköisyyksien konvergoivan kohti todellista jakaumaa suurilla parametrin s arvoilla. Toisin sanoen approksimaation keskimääräinen virhe on sitä pienempi mitä suurempaa arvoa s käytetään. Arvon s kasvattaminen pidentää laskenta-aikaa, sillä kaavan (3.5) integraali on laskettava s^n kertaa. Seuraavaksi esitettävillä vaihtoehtoisilla tavoilla valita approksimaation perustana olevat pisteet päästään samaan tulokseen joissakin tilanteissa lievemällä laskennallisella vaatimuudella. Nämä vaihtoehdot ovat satunnaislukujen ja kvasi-satunnaislukujen käyttäminen kuvan 5 tapaan.

Aluksi osoitetaan, että ehdolliset RNM-pistetodennäköisyydet voidaan laskea kaavan (3.5) sijaan Monte Carlo -integraalilla. Aikaisemmin on osoitettu [35], että lapsisolmun todennäköisyys saada arvo väliltä $[a_C, b_C]$, kun isäsolmujen arvot vastaavat välejä $[a_1, b_1], \dots, [a_n, b_n]$, lähestyy suurilla arvoilla s todennäköisyyttä

$$P_{[a_C, b_C]} = P(\chi_C \in [a_C, b_C] | \chi_C \in [0, 1]), \quad (3.6)$$

missä satunnaismuuttuja χ_C riippuu tasajakautuneista satunnaismuuttujista $\chi_1 \sim U(a_1, b_1), \dots,$



Kuva 5: 25 pistettä valittuna yksikköneliön alueelta kolmella vaihtoehoisella tavalla. RNM-menetelmän käyttämässä hilassa pisteet valitaan tasavälein kunkin ulottuvuuden suhteen. Vaihtoehoisesti pisteet voidaan valita satunnaisesti, jolloin tyypillisesti jotkin alueet jäävät muita harvemmiksi. Toisena vaihtoehtona pisteet voidaan valita kvasi-satunnaislukuina, joissa on havaittavissa hieman hilan kaltaista säännöllisyyttä. Erona hilaan, kvasi-satunnaislukuja voidaan varsinaisten satunnaislukujen tapaan generoida yksi kerrallaan, jolloin jokainen uusi piste osuu aiemmin ”tyhjälle” alueelle.

$\chi_n \sim U(a_n, b_n)$, RNM-funktiosta f sekä painoista w seuraavasti:

$$\chi_C = f(\chi_1, \dots, \chi_n, w) + e, \quad e \sim N(0, \sigma^2).$$

Merkitään $\chi = \{\chi_1, \dots, \chi_n\}$. Jakauman kantaja on $\Omega = [a_1, b_1] \times \dots \times [a_n, b_n]$ ja tiheysfunktio on

$$f_\chi(z) = \begin{cases} ((b_1 - a_1) \cdot \dots \cdot (b_n - a_n))^{-1}, & \text{kun } z \in \Omega \\ 0, & \text{muutoin.} \end{cases}$$

Kokonaistodennäköisyyden laskusäännön perusteella todennäköisyys (3.6) voidaan kirjoittaa muotoon

$$P_{[a_C, b_C]} = \int_{\Omega} P(\chi_C \in [a_C, b_C] \mid \chi = z, \chi_C \in [0, 1]) f_\chi(z) dz. \quad (3.7)$$

Integraali (3.7) voidaan approksimoida Monte Carlo -integraalilla

$$\begin{aligned} P_{[a_C, b_C]} &\approx \frac{V}{N} \sum_{i=1}^N P(\chi_C \in [a_C, b_C] \mid \chi = z_i, \chi_C \in [0, 1]) f_\chi(z_i) \\ &= \frac{1}{N} \sum_{i=1}^N \int_{a_C}^{b_C} \phi_{[0,1]}(x, f(z_i, w), \sigma^2) dx, \end{aligned} \quad (3.8)$$

missä $z_1, \dots, z_N \sim \chi$ on N -kokoinen otos edellä määritetystä tasajakaumasta χ ja kantajan Ω tilavuus $V = 1/f_\chi(z)$.

Kaava (3.8) on lähes identtinen kaavan (3.5) kanssa. Ero näiden välillä muodostuu vain approksimaation perustana olevien pisteiden z valintatavasta. Kaavassa (3.5) pisteet valitaan tasavälein ja kaavassa (3.8) satunnaisesti. Tällä voi olla merkitystä approksimaation tarkkuuden kannalta. Integraalin approksimointi tasavälisin pistein tunnetaan puolisuunnikassääntönä [37]. Tämän soveltaminen moniulotteiseen tapaukseen hilana johtaa kaavaan (3.5) sillä erotuksella, että puolisuunnikassääntöä käytettäessä hilan reunoilla ja kulmissa olevat pisteet saavat muita pienemmän painotuksen. Tästä voidaan tehdä kaksi johtopäätöstä. Ensinnäkin kaavaa (3.5) voi olla mahdollista kehittää ottamalla käyttöön puolisuunnikassäännön mukainen painotus. Toiseksi, koska reuna- ja kulmapisteiden määrän suhde muihin pisteisiin vähenee kasvatettaessa pisteiden kokonaismäärää N , kaavan (3.5) approksimaatiovirhe vähenee suurilla pisteiden määrillä samassa suhteessa puolisuunnikassääntöön perustuvan approksimaation kanssa. Tämä virhe on yksiulotteisessa tapauksessa $O(N^{-2})$ [37]. Koska saman tiheyden hilan muodostaminen vaatii eksponentiaalisen määrän pisteitä isäsolmujen määrän n kasvaessa, n -ulotteisessa tapauksessa approksimaatiovirhe on $O(N^{-2/n})$. Kaava (3.8) on Monte Carlo -integraali, joten sen approksimaatiovirhe on isäsolmujen määrästä n riippumatta $O(N^{-1/2})$ [37]. Tämän perusteella kaava (3.8) vaikuttaa keskimäärin tarkemmalta, jos isäsolmuja on enemmän kuin neljä.

Monte Carlo -integrointi voidaan toteuttaa tehokkaammin käyttämällä satunnaislukujen sijaan kvasi-satunnaislukuja. Satunnaislukujen käytössä Monte Carlo -integroinnissa on ongelmana niiden mahdollisesti epätasainen jakautuminen. Kuten kuvasta 5 nähdään, satunnaisluvut voivat täyttää alueen epätasaisesti. Pisteiden epätasainen jakauma vääristää näiden keskiarvona laskettavaa approksimaatiota. Kvasi-satunnaislukujen tarkoitus on rajoittaa tätä ilmiötä korvaamalla satunnaisluvut täysin deterministisellä sekvenssillä, jossa jokainen luku on kaukana edellisistä luvuista. Eräs tunnettu tällainen sekvenssi on Sobol-sekvenssi [58]. Satunnaislukujen korvaaminen tällä sekvenssillä antaa Monte Carlo -integraalille approksimointivirheen $O((\ln N)^n/N)$ [37]. Tämä approksimointivirhe on suurilla pisteiden määrillä N lähes yhtä hyvä kuin $O(N^{-1})$. Kaavan (3.5) virhe vastaa tätä, kun isäsolmuja on kaksi. Jos isäsolmuja on enemmän, Sobol-sekvenssin käyttäminen vaikuttaa tarkastelluista vaihtoehdoista teoreettisesti parhaalta. Tämä teoreettinen tarkastelu perustuu raja-arvoihin, eikä huomioi RNM-menetelmässä approksimoitavien funktioiden erityispiirteitä, jotka saattavat johtaa yleisiä arvoja pienempiin virhetermeihin.

Satunnaislukuihin perustuvia muunnoksia verrattiin kokeella perusmuotoiseen RNM-menetelmään. Kokeen asetelma ja tulokset on esitetty liitteessä 1. Suurta approksimaatiopisteiden määrää käytettäessä parhaat tulokset saatiin teoreettisten tulosten mukaisesti kvasi-satunnaisluvuilla. Heikoimmat tulokset saatiin hilaan perustuvalla perusmuotoisella kaavalla (3.5), mutta käytän-

nössä ero tulosten välillä ei ollut merkittävä. Pienillä approksimaatiopisteiden määrillä kvasi-satunnaisluvut sekä satunnaisluvut tuottivat kahdesta testistä yhdessä merkittävästi perusmuotoista kaavaa paremman tuloksen ja toisessa lähes saman tuloksen.

Välimatka-asteikollinen muunnos

Tiedustelusovelluksissa voidaan kohdata erityyppistä dataa. Esimerkiksi sentimenttianalyysin tulos “kirjoittajan asenne on erittäin myönteinen” on järjestysasteikollista dataa. “Erittäin myönteisten” blogikirjoitusten suhteellinen osuus mediassa on välimatka-asteikollista dataa. Järjestysasteikollisten solmujen menetelmä käsittelee nimensä mukaisesti järjestysasteikollista dataa. Muu kuin järjestysasteikollinen data tulee menetelmän soveltamista varten muuntaa järjestysasteikolliseksi. Esimerkiksi suhteelliset osuudet voidaan muuntaa kaksiarvoiselle järjestysasteikolle: “alle 50 %” ja “vähintään 50 %”.

Muunnettaessa välimatka-asteikollista dataa järjestysasteikolliseksi menetetään informaatiota. Informaation menetystä voidaan vähentää erilaisia muunnoksia kokeilevilla lähestymistavoilla [31, 34] tai lisäämällä järjestysasteikon arvojen määrää. Järjestysasteikollisten solmujen menetelmä voidaan myös muokata käyttämään välimatka-asteikollista dataa. Välimatka-asteikollinen muunnos toteutetaan tässä työssä esittelemällä uusi tapa ehdollisten todennäköisyyksien määrittämiseen. Välimatka-asteikollisen datan käyttäminen on huomioitava lisäksi parametrien estimointiin käytettävän algoritmin valinnassa [59]. Alaluvussa 3.4 esitellään estimointimenetelmä, joka on yhteensopiva sekä välimatka- että järjestysasteikollisen datan kanssa. Seuraavaksi tarkastellaan ehdollisten todennäköisyyksien muodostamista. Lapsisolmuista huomioidaan erikseen niiden kuvaaminen jatkuvina tai diskreetteinä muuttujina.

Solmujen järjestysasteikollisuus huomioidaan RNM-menetelmässä kuvaamalla jokainen asteikon arvo välinä $[a, b] \subset [0, 1]$. Solmun ehdollinen todennäköisyys

$$P_{[a_C, b_C]} = P(\chi_C \in [a_C, b_C] \mid \chi_1 \in [a_1, b_1], \dots, \chi_n \in [a_n, b_n], \chi_C \in [0, 1])$$

integroidaan sekä arvoa C että isäsolmujen χ arvoja vastaavien välien yli. Integraali isäsolmujen arvojen yli ei ratkea analyttisesti, joten se approksimoidaan summana sekä kaavassa (3.5) että tälle tässä työssä vaihtoehtona esitetyssä kaavassa (3.8). Näiden kaavojen muokkaaminen yhteensopiviksi välimatka-asteikollisten arvojen kanssa mahdollistaa ehdollisten jakaumien muodostamisen välimatka-asteikollisesta datasta.

Jos isäsolmujen arvot kuvataan välien sijaan välimatka-asteikollisina lukuina z , jää approksi-

mointia vaativa integraali isäsolmujen arvovälien yli pois. Molemmat kaavat pelkistyvät tällöin muotoon

$$p_{[a_c, b_c]} = \int_{a_c}^{b_c} \phi_{[0,1]}(x, f(z, w), \sigma^2) dx. \quad (3.9)$$

Kun lapsisolmun arvo c on havainto jatkuvasta muuttujasta, sen tiheysfunktio on diskreetin sijaan jatkuva. Lapsisolmulla ei ole integrointia vaativaa arvoväliä, mutta jos isäsolmut ovat järjestysasteikollisia, integroidaan edelleen niiden arvovälien yli. Tiheysfunktioiksi saadaan tällöin

$$p(c) \approx \frac{1}{N} \sum_{i=1}^N \phi_{[0,1]}(c, f(z_i, w), \sigma^2), \quad (3.10)$$

jossa pisteet z_i on generoitava integraalin approksimointia varten isäsolmujen arvoväleiltä jollakin aikaisemmin esitellyistä menetelmistä. Jos sekä lapsisolmun että isäsolmujen arvot ovat välimatka-asteikollisia, jää approksimaatio pois. Tällöin tiheysfunktioiksi saadaan

$$p(c) = \phi_{[0,1]}(c, f(z, w), \sigma^2). \quad (3.11)$$

Välimatka-asteikolliset diskreetit lapsisolmun arvot c voivat olla esimerkiksi lukumääriä. Näiden jakauma tulisi kuvata rajatun normaalijakauman $\phi_{[0,1]}$ sijaan diskreetillä jakaumalla. Sopivia jakaumia voivat olla esimerkiksi binomi- tai Poisson-jakauma. Näissä ei kuitenkaan voida käyttää suoraan samoja parametreja kuin tavanomaisessa RNM-mallissa, eikä tässä työssä tutkita niiden parametrisointia. Diskreetit arvot voidaan kuvata myös havaintoina jatkuvasta jakaumasta, jolloin voidaan käyttää kaavoja (3.10) ja (3.11). Diskreetin datan kuvaaminen jatkuvana ei kuitenkaan ole välttämättä tarkempaa kuin sen muuntaminen järjestysasteikolliseksi.

Asteikon muuntamisen vaikutusta testattiin keinotekoisilla aineistoilla. Kokeen asetelma ja tulokset on esitetty liitteessä 2. Kokeessa todettiin aineistosta laskettujen estimaattien muuttuvan epätarkemmiksi, kun aineiston asteikkoa muunnetaan. Sama tulos saatiin sekä muunnettaessa järjestysasteikollinen aineisto välimatka-asteikolliseksi että muunnettaessa välimatka-asteikollinen aineisto järjestysasteikolliseksi. Kokeen tulos tukee tässä alaluvussa esitellyn välimatka-asteikollisen muunnoksen käyttämistä alkuperäisen aineiston ollessa välimatka-asteikollista.

3.4 Gibbsin otanta

Gibbsin otanta on Markovin ketju Monte Carlo eli MCMC (engl. Markov chain Monte Carlo) -menetelmiin kuuluva menetelmä todennäköisyysjakaumien simulointiin [60, 61]. Sen avulla voidaan simuloida satunnaismuuttujan θ jakaumasta otos $X_1, \dots, X_N \sim \theta$, vaikka tätä jakaumaa ei tunnetaisi. Edellytyksenä on, että satunnaismuuttuja voidaan jakaa komponentteihin $(\theta_1, \dots, \theta_n) = \theta$ ja että kunkin komponentin θ_k täysin ehdollinen jakauma

$$p(\theta_k \mid \theta_0, \dots, \theta_{k-1}, \theta_{k+1}, \dots, \theta_n)$$

tunnetaan. Jakaumasta simuloitua otosta voidaan käyttää approksimointiin. Funktion f odotusarvo satunnaismuuttujan θ jakauman suhteen voidaan approksimoida suurella otoskoolla N [61]. Approksimaatio saadaan keskiarvona

$$E(f(\theta)) \approx \frac{1}{N} \sum_{i=1}^N f(X_i). \quad (3.12)$$

Kaavaa (3.12) käyttäen voidaan approksimoida esimerkiksi satunnaismuuttujan odotusarvo, kun funktioksi f valitaan $f(\theta) = \theta$.

Gibbsin otanta perustuu sopivan Markovin ketjun muodostamiseen halutun jakauman simulointia varten. Markovin ketju on joukko satunnaismuuttujia $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots\}$, joista millä tahansa indeksillä i satunnaismuuttujan $\theta^{(i)}$ jakauma sitä edeltävillä satunnaismuuttujilla ehdollistettuna on sama kuin sen jakauma vain edellisellä satunnaismuuttujalla ehdollistettuna [61]. Toisin sanoen satunnaismuuttujien on kaikilla indekseillä i toteutettava ehto

$$p(\theta^{(i)} \mid \theta^{(i-1)}, \theta^{(i-2)}, \dots, \theta^{(0)}) = p(\theta^{(i)} \mid \theta^{(i-1)}).$$

Markovin ketjulla voidaan mallintaa esimerkiksi pokerinpelaajan pelikassaa. Pelin alussa kassan suuruus on $\theta^{(0)}$, ensimmäisen kierroksen jälkeen $\theta^{(1)}$ ja niin edelleen kunkin pelikassan $\theta^{(i)}$ riippuen vain edellisen kierroksen pelikassasta $\theta^{(i-1)}$. Tiettyjen ehtojen vallitessa Markovin ketjulla on tasapainojakauma, jota satunnaismuuttujien jakaumat lähestyvät suurilla indeksin arvoilla [61]. Tasapainojakauman saavuttamiseen kuluva iteratioiden määrää kutsutaan sisäänpalamisajaksi (engl. burn-in). Tämän pituus on satunnainen, joten tasapainon saavuttaminen on tulkittava esimerkiksi kuvasta [60].

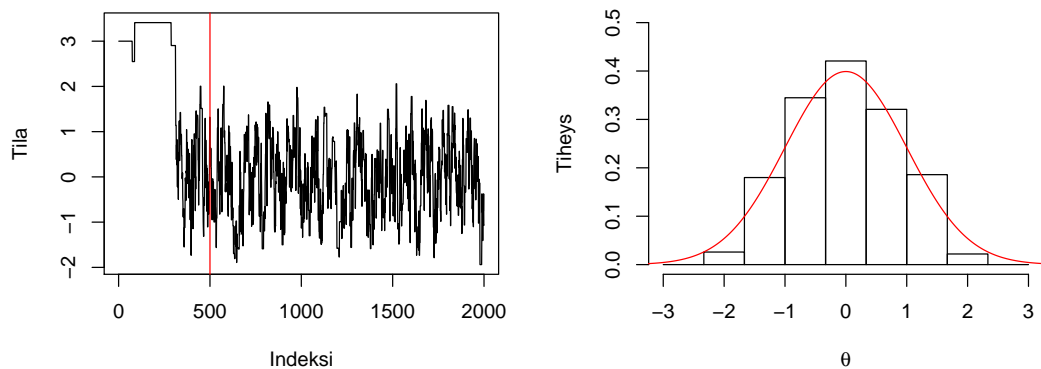
Esimerkissä 3.3 esitellään Markovin ketjun käyttäminen arvojen simulointiin standardinormaa-

lijakaumasta. Samalla havainnollistetaan tasapainojakauman ja sisäänpalamisajan käsitteet.

Esimerkki 3.3. Muodostetaan Markovin ketju seuraavalla tavalla:

1. Aloitetaan ketju tilasta $\theta^{(0)} = 3$.
2. Arvotaan luku x välin $(\theta^{(i)} - 1, \theta^{(i)} + 1)$ tasajakaumasta, missä $\theta^{(i)}$ on ketjun viimeisin tila.
3. Hyväksytään x ketjun uudeksi tilaksi todennäköisyydellä $\min(1, e^{-\frac{1}{2}x^2})$. Jos x hylätään, asetetaan ketjun uudeksi tilaksi sama kuin viimeisin tila. Palataan vaiheeseen 2.

Kuvassa 6 on havainnollistettu näin määritetyn ketjun 2000 ensimmäistä tilaa. Ensimmäiset 200 tilaa poikkeavat selvästi ketjun loppuosasta. Otos valitaan siten, että alun poikkeava osuus jää siitä pois. Tässä tapauksessa otokseksi valitaan ketjun tilat $\{\theta^{(501)}, \theta^{(502)}, \dots, \theta^{(2000)}\}$. Otoksen empiiristä jakaumaa on verrattu kuvassa 6 standardinormaalijakauman tiheysfunktioon.



Kuva 6: Vasemmalla on esitetty esimerkin 3.3 Markovin ketjun 2000 ensimmäistä tilaa. Ketjun voidaan katsoa saavuttaneen tasapainojakaumansa viimeistään 500 iteraation jälkeen. Oikealla sisäänpalamisajan jälkeinen jakauma histogrammina sekä punaisella standardinormaalijakauman tiheysfunktio.

Gibbsin otannassa muodostetaan satunnaismuuttujan θ täysin ehdollisia jakaumia hyödyntäen Markovin ketju, jonka tasapainojakauma on satunnaismuuttujan θ jakauma. Erona esimerkiksi 3.3 Gibbsin otantaa sovelletaan moniulotteisiin jakaumiin. Ketjun viimeisintä tilaa $\theta^{(i)} = (\theta_0^{(i)}, \dots, \theta_n^{(i)})$ ei päivitetä samalla iteraatiolla kokonaan, vaan ainoastaan yksi sen komponenteista saa uuden arvon. Otos valitaan simuloidusta ketjusta samaan tapaan kuin esimerkissä 3.3. Gibbsin otannan Markovin ketju luodaan seuraavalla tavalla:

1. Aloitetaan alustamalla ketju satunnaiseen tilaan $\theta^{(0)}$.

2. Valitaan ketjun viimeisimmästä tilasta $\theta^{(i)}$ satunnaisesti yksi komponentti $\theta_k^{(i)}$.
3. Arvotaan tälle komponentille uusi arvo sen täysin ehdollisesta jakaumasta muilla nykyisen tilan $\theta^{(i)}$ komponenteilla ehdollistettuna. Asetetaan tämä yhdessä komponentissa poikkeava tila ketjun uudeksi tilaksi ja palataan vaiheeseen 2.

Vaihetta 2 voidaan muokata esimerkiksi käymällä kaikki komponentit läpi deterministisesti. Komponenttien päivittämisen järjestyksellä ei ole merkitystä, mutta kukin komponentti tulee päivittää keskimäärin yhtä usein [61]. Esimerkkiin 3.3 verrattuna Gibbsin otannassa tilan muutos hyväksytään aina, mikä mahdollistaa lyhyemmän sisäänpalamisajan. Gibbsin otanta voidaan toteuttaa käyttäen valmiita ohjelmistoja [62–64]. Tällöin käyttäjän on varmistuttava riittävän pitkästä sisäänpalamisajasta ja kyllin suuresta otoskoosta.

Esimerkissä 3.4 tarkastellaan alaluvussa 3.3 esitellyn RNM-menetelmän parametrien päättelystä datasta yksinkertaistetussa tilanteessa.

Esimerkki 3.4. *Solmulla Y on isäsolmut X ja Z . Ehdollinen jakauma kuvataan RNM-menetelmän WMEAN-funktiolla siten, että odotusarvo on $\mu = wx + (1 - w)z$. Satunnaismuuttujan Y ehdollinen jakauma määritetään alaluvussa 3.3 esitetyksi välimatka-asteisen muunnoksen tiheysfunktiksi (3.11). Tämän esimerkin laskujen yksinkertaistamiseksi tiheysfunktion välille $[0, 1]$ rajattu normaalijakauma korvataan tavallisella normaalijakaumalla. Toisin sanoen*

$$(Y \mid X = x, Z = z) \sim N(wx + (1 - w)z, \sigma^2),$$

missä x ja z ovat isäsolmujen arvot ja w sekä σ tuntemattomat parametrit. Kuvassa 7 esitetty aineisto käsittää $N = 5$ havaintoa solmujen arvoista. Priorikäsitystä tuntemattomista parametreista kuvaavat jakaumat

$$w \sim \text{Beta}(1, 1)$$

$$\sigma \sim \text{Gamma}(1, 10^{-3}).$$

Lasketaan Bayes-päätelyllä tuntemattomien parametrien todennäköisimmät arvot aineisto huomioiden.

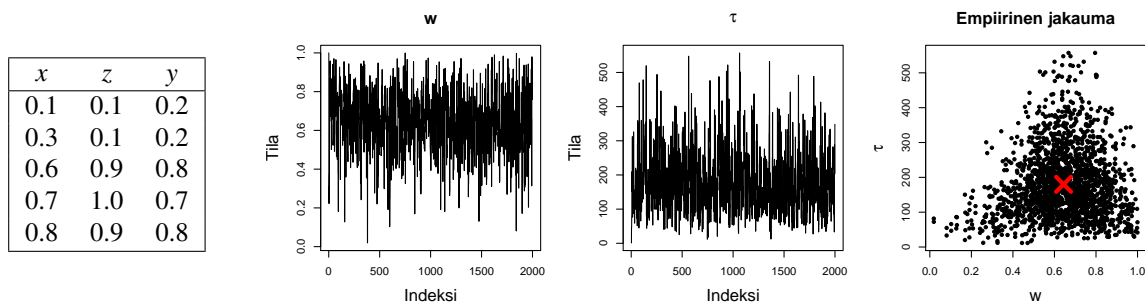
Todennäköisimpien arvojen laskeminen analyttisesti posteriorijakauman $p(w, \sigma \mid x, y, z)$ perusteella on mahdotonta. Parametrien täysin ehdolliset jakaumat voidaan ratkaista, joten postero-

rijakaumasta voidaan simuloida otos Gibbsin otannalla. Täysin ehdolliset jakaumat ovat

$$(w \mid \sigma, x, y, z) \sim N_{[0,1]} \left(\frac{\sum_{i=1}^N (y_i - z_i)(x_i - z_i)}{\sum_{i=1}^N (x_i - z_i)^2}, \frac{\sigma^2}{\sum_{i=1}^N (x_i - z_i)^2} \right)$$

$$(\tau \mid w, x, y, z) \sim \text{Gamma} \left(\frac{N}{2} + 1, 10^{-3} + \frac{1}{2} \sum_{i=1}^N (y_i - z_i - w(x_i - z_i))^2 \right),$$

missä $\tau = 1/\sigma^2$ ja $N_{[0,1]}$ on välille $[0, 1]$ rajoitettu normaali-jakauma. Gibbsin otannalla saadaan kuvan 7 mukainen Markovin ketju. Sisäänpalaminen vaikuttaa tapahtuvan nopeasti. Otokseksi valitaan ketjun tilat indeksistä 101 alkaen. Otoksen empiirinen jakauma on esitetty kuvan 7 oikeassa reunassa. Parametrien odotusarvot voidaan approksimoida käyttäen kaavaa (3.12). Painokertoimelle saadaan estimaatti $E(w) \approx 0,64$. Keskihajonnan estimaatti saadaan tarkkuusparametrin τ funktiona $E(\sigma) = E(1/\sqrt{\tau}) \approx 0,085$.



Kuva 7: Esimerkin 3.4 aineisto, Markovin ketju ja Gibbsin otannalla muodostettu otos. Vasemmalla taulukossa on esimerkissä käytetty aineisto. Keskellä ovat Markovin ketjun 2000 ensimmäistä tilaa komponenteittain. Ketjun voidaan katsoa saavuttaneen tasapainojakaumansa viimeistään 100 iteraation jälkeen. Oikealla on sisäänpalamisajan jälkeinen otos, jonka keskiarvo on merkitty raksilla.

Gibbsin otannan soveltamisen vaatimat täysin ehdolliset jakaumat kyettiin ratkaisemaan esimerkissä 3.4. Vaikka nämä voidaan ratkaista monessa tilanteessa [65, 66], yleistä taetta tästä ei ole. Gibbsin otanta voidaan toteuttaa päivittämällä osa komponenteista täysin ehdollisia jakaumia hyödyntäen ja osa muilla tavoin. Analyttisesti ratkeamatonta täysin ehdollista jakaumaa voidaan simuloida esimerkiksi hylkäyspoiminnalla (engl. rejection sampling) [61, 67].

3.5 Yhteenveto

Tässä luvussa tarkasteltiin tutkimuskysymystä “miten sotilaallisen aktiivisuuden arviointia voidaan tukea Bayes-päätelyyn perustuvilla menetelmillä”. Bayes-päätelyn todettiin mahdollistavan epäsuoran informaation käsittelyn analyttisesti. Se ei poista tarvetta arvioiden esittämiselle, vaan mahdollistaa vaikean arvion johdonmukaisen muodostamisen helpommin arvioitavien

todennäköisyyksien perusteella. Bayes-päätelyä voidaan käyttää sotilaallisen aktiivisuuden arvioinnissa johtopäätösten tekemiseen epäsuorasta informaatiosta. Kun päätelyn kohde kuvataan satunnaismuuttujalla ja sille annetaan priorijakauma, Bayes-päätelyä käyttäen voidaan muodostaa datan huomioiva posteriorijakauma. Tätä periaatetta voidaan käyttää asiantuntija-arvion ja datan yhdistämiseen. Bayes-verkkoja voidaan käyttää kuvaamaan ilmiöön liittyvät epävarmuudet intuitiivisesti helpossa muodossa. Niiden ehdollisten todennäköisyyksien määrittäminen voi kuitenkin olla työlästä. Tätä työn määrää voidaan vähentää järjestysasteikollisten solmujen menetelmällä. Bayes-päätelystä usein seuraavia vaikeita laskuja voidaan approksimoida Gibbsin otannalla.

Alaluvussa 3.3 esiteltiin kaksi uutta muunnosta järjestysasteikollisten solmujen menetelmään. Ensimmäisessä esitellyssä muunnoksessa käytetään ehdollisten todennäköisyyksien laskemiseen RNM-menetelmän tavanomaisen hila-approksimaation (3.5) sijaan Monte Carlo -integraalia. Monte Carlo -integraalin pisteet voidaan valita tavanomaisia satunnaislukuja tai kvasi-satunnaislukuja käyttäen. Tavanomaisten satunnaislukujen käyttäminen vaikuttaa hila-approksimaatiota tarkemmalta silloin, kun isäsolmuja on enemmän kuin neljä. Sobol-sekvenssiin perustuvien kvasi-satunnaislukujen käyttäminen vaikuttaa hila-approksimaatiota tarkemmalta silloin, kun isäsolmuja on enemmän kuin kaksi. Toinen alaluvussa 3.3 esitetty muunnos mahdollistaa järjestysasteikollisten solmujen menetelmän ehdollisten todennäköisyyksien muodostamisen välimatka-asteikollista dataa käytettäessä. Välimatka-asteikollisen datan hyödyntäminen on huomioitava estimointimenetelmän valinnassa.

4 BAYES-PÄÄTTELYÄ JA JÄRJESTYSASTEIKOLLISTEN SOLMUJEN MENETELMÄÄ HYÖDYNTÄVÄ LÄHESTYMISTAPA

Pelkkään dataan perustuva koneoppiminen voi olla ongelmallista, jos dataa on saatavilla vain vähän. Tämän tyyppisiä sovelluskohteita löytyy luvun 2 perusteella esimerkiksi tiedustelun alalta. Ongelma voidaan ratkaista hyödyntämällä datan lisäksi asiantuntija-arvioita. Tässä luvussa vastataan alatutkimuskysymykseen 3: “miten data ja asiantuntija-arviot yhdistetään ja hyödynnetään järjestysasteikollisten solmujen menetelmään perustuvassa Bayes-verkossa”. Luvussa esitellään uusi Bayes-päätelyä ja järjestysasteikollisten solmujen menetelmää hyödyntävä lähestymistapa. Lähestymistavan uutuus on Bayes-päätelyn ja järjestysasteikollisten solmujen menetelmän yhdistämisessä.

Järjestysasteikollisten solmujen menetelmää käytetään tässä luvussa esiteltävässä lähestymistavassa asiantuntijan työmäärän vähentämiseen. Bayes-päätelyä käytetään asiantuntija-arvion ja datan yhdistämiseen. Saatavilla olevaa dataa ei jaeta erillisiin opetus-, validointi- tai testiaineistoihin. Aineiston oletetaan sisältävän joitakin päätelyn kohteena olevan kohdemuuttujan y arvoja. Lähestymistapaa voidaan siten pitää esimerkkinä puoliohjatusta koneoppimisesta [68]. Kohdemuuttujan y tuntemattomat arvot sekä riippuvuuksien mallintamiseen käytettävän Bayes-verkon parametrit estimoidaan samanaikaisesti. Lähestymistapaa voidaan käyttää esimerkiksi sotilaallisen aktiivisuuden arviointiin.

Alaluvussa 4.1 esitellään lähestymistavan vaiheet ja tärkeimmät määritelmät. Lähestymistavan yksityiskohdat esitellään vaiheittain alaluvuissa 4.2 - 4.7. Alaluvussa 4.8 esitetään yhteenveto lähestymistavan ominaisuuksista ja käyttömahdollisuuksista vastauksena alatutkimuskysymykseen 3.

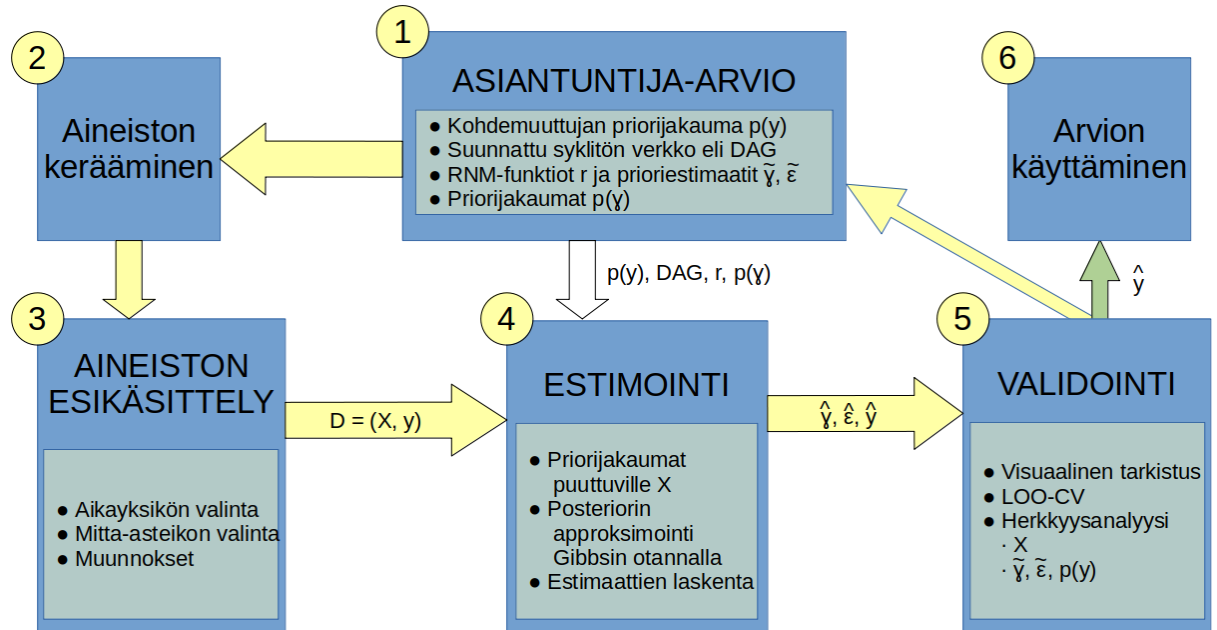
4.1 Lähestymistavan yleiskuva

Tässä luvussa esiteltävää lähestymistapaa käytetään ajan suhteen vaihtelevan kohdemuuttujan y arvioimiseen datan ja asiantuntija-arvion perusteella. Kohdemuuttujan on oltava joko hypoteesi tai välimatka-asteikollinen muuttuja. Sen arvot muodostavat aikasarjan $y = \{y_1, y_2, \dots\}$. Hypoteesin arvo y_i ajanhetkellä i on binäärisesti tosi tai epätosi. Sotilaallista aktiivisuutta seurattaessa hypoteesi voi olla esimerkiksi muotoa “alueella on sotilaallista toimintaa”. Välimatka-asteikollisen muuttujan avulla voidaan tarkastella “kuinka paljon” -tyyppisiä kysymyksiä. Sotilaallista aktiivisuutta arvioitaessa välimatka-asteikollinen muuttuja voi olla esimerkiksi “kuinka

suuri joukko alueella toimii”. Välimatka-asteikollinen muuttuja skaalataan siten, että sen pienin mahdollinen arvo on jokaisella ajanhetkellä nolla ja suurin mahdollinen arvo on 1.

Lähestymistavassa mallinnetaan kohdemuuttujan y sekä selittävien muuttujien X väliset riippuvuudet Bayes-verkolla. Suoraan kohdemuuttujasta riippuvia selittäviä muuttujia kutsutaan indikaattoreiksi ja muita selittäviä muuttujia vaihtoehtoisiksi selityksiksi. Bayes-verkon ehdolliset jakaumat määritetään järjestysasteikollisten solmujen menetelmällä. RNM-parametrit $\gamma = (w, \sigma)$ kuvataan satunnaismuuttujina ja niille määritetään priorijakauma $p(\gamma)$. Priorijakaumat määritetään myös tuntemattomille kohdemuuttujan sekä vaihtoehtoisten selitysten arvoille. Priorijakaumien ja aineiston D perusteella muodostetaan posteriorijakauma Bayes-päätelyä käyttäen. Posteriorijakaumasta lasketaan posterioriestimaatit $(\hat{y}, \hat{\gamma}, \hat{\varepsilon})$. Kohdemuuttujan odotusarvon posterioriestimaatti on $\hat{y} = E(y | D) = (E(y_1 | D), E(y_2 | D), \dots)$. RNM-parametreille lasketaan posterioriestimaatit $\hat{\gamma}$ ja $\hat{\varepsilon}$. Näistä ensimmäinen estimoii odotusarvoa ja jälkimmäinen tarkkuutta. Tarkkuuden prioriestimaatti $\tilde{\varepsilon}$ ja posterioriestimaatti $\hat{\varepsilon}$ määritellään järjestysasteikolla tulkinnan helpottamiseksi.

Tässä luvussa esiteltävä lähestymistapa koostuu kuvan 8 mukaisesti kuudesta vaiheesta. Vaiheessa 1 muodostetaan asiantuntija-arvio. Tähän sisältyvät kohdemuuttujan y priorijakauman $p(y)$ määrittely, suunnatun syklisen verkon määrittely sekä RNM-funktioiden r ja RNM-parametrien prioriestimaattien $\tilde{\gamma}$ ja $\tilde{\varepsilon}$ määrittely. Prioriestimaattien perusteella muodostetaan RNM-parametrien priorijakauma $p(\gamma)$. Aineisto kerätään vaiheessa 2 ja esikäsitellään vaiheessa 3. Esikäsitelyssä valitaan lähestymistavan aikayksikkö. Tämä määrittää aikasarjana (y_1, y_2, \dots) esitettävän kohdemuuttujan yksittäisen arvon y_i kuvaaman aikavälin pituuden. Aineistolle tehdään esikäsitelyssä erilaisia muunnoksia sen sisältämän tiedon käsiteltävyyden parantamiseksi. Vaiheessa 4 estimoidaan satunnaismuuttujat $\theta = (y, \gamma, \varepsilon)$. Tätä varten määritetään ensin priorijakauma aineistosta puuttuville vaihtoehtoisten selitysten arvoille ja muodostetaan otos Bayes-päätelyllä määritettävästä posteriorijakaumasta. Otos muodostetaan Gibbsin otantaa käyttäen ja posterioriestimaatit $\hat{\theta}$ määritetään kaavaa (3.12) käyttäen. Vaiheessa 5 tutkitaan asiantuntija-arvion mielekkyyttä ja tulosten tarkkuutta. Vaihe 5 aloitetaan posterioriestimaattia vastaavan Bayes-verkon visualisoinnilla ja vertaamisella prioriestimaattia vastaavaan Bayes-verkkoon. Tämän jälkeen lasketaan estimaatti tuloksen virheelle yksi-pois -ristiinvalidoinnilla (LOO-CV, englanniksi leave-one-out cross-validation). Herkkyysanalyysillä tutkitaan posterioriestimaatin $\hat{\gamma}$ robustisuutta toistamalla vaiheet 1-4 asiantuntija-arvioita ja dataa muunnellen. Vaiheen 5 perusteella päätetään joko palata vaiheeseen 1 tai siirtyä vaiheeseen 6. Vaiheessa 6 käytetään kohdemuuttujan posterioriestimaattien \hat{y} muodostamaa arviota.



Kuva 8: Bayes-päätelyä ja järjestysasteikollisten solmujen menetelmää hyödyntävän lähestymistavan rakenne. Lähestymistavan käyttäminen alkaa asiantuntija-arvion muodostamisesta ja päättyy arvion käyttämiseen. Lähestymistavan vaiheet on numeroitu.

4.2 Asiantuntija-arvio

Tässä luvussa esiteltävän lähestymistavan ensimmäisessä vaiheessa määritellään asiantuntija-arviona kohdemuuttujan y priorijakauma $p(y)$, suunnattu sykliton verkko, RNM-funktiot r ja RNM-parametrien $\gamma = (w, \sigma)$ priorijakauma $p(\gamma)$. RNM-parametrien priorijakauma muodostetaan erikseen määritettävien prioriestimaattien $\tilde{\gamma} = (\tilde{w}, \tilde{\sigma})$ perusteella. Priorijakaumat kuvaavat asiantuntijan kohdemuuttujaa ja RNM-parametreja koskevan priorikäsityksen. Priorijakaumien määrittäminen mahdollistaa asiantuntija-arvion sekä datan huomioivan posteriorijakauman muodostamisen Bayes-päätelyä hyödyntäen.

RNM-funktioista ja -parametreista tehtävien asiantuntija-arvioiden muodostamista voidaan tukea tarkastelemalla erilaisten vaihtoehtoisten arvioiden tuottamia ehdollisia jakaumia. Ehdollisia jakaumia ei voida muodostaa ennen aineiston mitta-asteikon määrittämistä. Välimatka-asteikollista aineistoa käsiteltäessä ehdolliset jakaumat määritetään jatkuvina ja järjestysasteikollista aineistoa käsiteltäessä diskreetteinä. Aineiston mitta-asteikon valinta tehdään lähestymistavan vaiheessa 3, mutta tämä valinta voi olla ennakoitavissa vaiheessa 1. Ehdollisia jakaumia voidaan tällöin tarkastella esimerkiksi RNM-parametrien prioriestimaattien $\tilde{\gamma}$ funktiona.

Kohdemuuttujan priorijakauma

Kohdemuuttujan y priorijakauma $p(y)$ on subjektiivinen arvio kohdemuuttujan todennäköises-

tä arvosta lähitulevaisuudessa. Arvio antaa Bayes-päätelylle lähtökohdan, jota tullaan päivittämään aineistoon nojaten samalla periaatteella kuin esimerkissä 3.1. Kohdemuuttujan priorijakauma määritetään sen tulevaa arvoa koskevan kysymyksen perusteella. Kysymys esitetään asiantuntijalle ja priorijakauma määritetään vastauksen perusteella. Kysymyksen muotoilu riippuu kohdemuuttujan tyypistä. Hypoteesina määritetyn kohdemuuttujan arvosta kysytään todennäköisyyttä d_H hypoteesin paikkansa pitämiseksi satunnaisena ajanhetkenä lähitulevaisuudessa. Välimatka-asteikollisena määritetyn kohdemuuttujan arvosta kysytään sen keskimääräistä arvoa d_S lähitulevaisuudessa. Molemmissa tapauksissa vastauksen on oltava luku väliltä $[0,1]$.

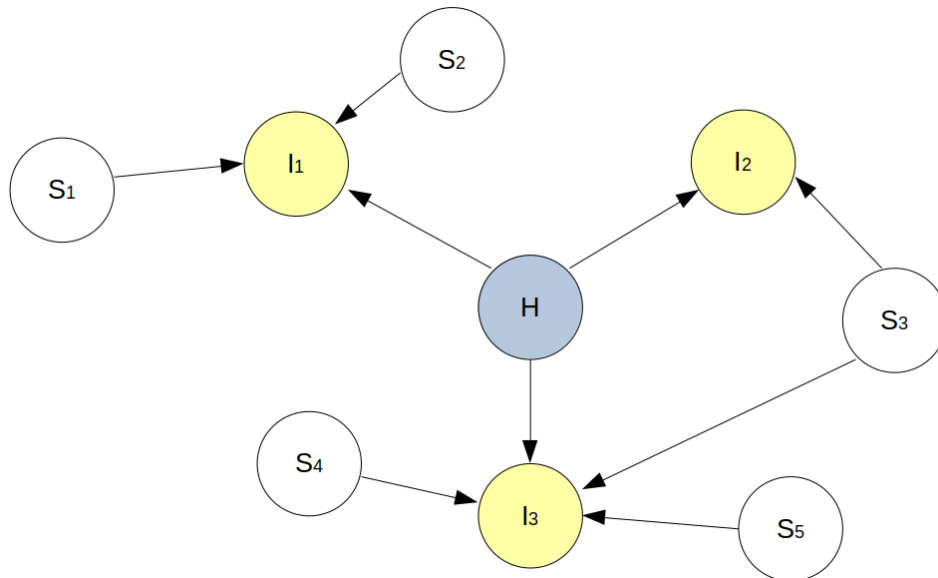
Priorijakauma $p(y)$ määritetään vastauksen d_H tai d_S perusteella. Hypoteesin arvo on binäärinen, joten sen priorijakauma määritetään Bernoulli-jakaumana $y \sim B(d_H)$. Välimatka-asteikollisen muuttujan priorijakaumaa ei voida määrittää Bernoulli-jakaumalla, sillä se voi saada tämän työn lähestymistavassa arvoja väliltä $[0, 1]$. Välimatka-asteikollisen muuttujan priorijakaumana käytetään kohtalaisen hajonnan antavaa, välille $[0, 1]$ rajoitettua normaalijakaumaa $N_{[0,1]}(d_S; 0,3)$. Tämän jakauman hajonnan määrittämistä asiantuntija-arvion perusteella ei tarkastella tässä työssä.

Suunnattu syklitön verkko

Suunnattu syklitön verkko eli DAG määrittää kohdemuuttujan y ja selittävien muuttujien X väliset suorat riippuvuussuhteet. Selittäviä muuttujia ovat tämän työn lähestymistavassa kohdemuuttujasta suoraan riippuvat indikaattorit sekä indikaattoriin suoraan vaikuttavat vaihtoehtoiset selitykset. Asiantuntija-arvion perusteella muodostettava DAG on pohja tämän työn lähestymistavassa määritettävälle Bayes-verkolle.

Suunnatun syklittömän verkon rakentaminen aloitetaan erittelemällä indikaattorit ja vaihtoehtoiset selitykset. Tunnistetut indikaattorit piirretään kuvan 9 tapaan kohdemuuttujan ympärille. Esimerkiksi sotilaallista aktiivisuutta arvioitaessa indikaattorit voivat olla aktiivisuudesta kertovia kirjallisia lähteitä tai sensoreiden havaintoja. Indikaattorien hyödyllisyys arvioidaan myöhemmin lähestymistavan vaiheessa 5, joten tarkasteluun sisällytettävän indikaattorin hyödyllisyydestä ei tarvitse olla varmuutta etukäteen.

Indikaattorien tunnistamisen jälkeen piirretään kunkin niistä ympärille vaihtoehtoisia selityksiä. Esimerkiksi arvioitaessa sotilaallista aktiivisuutta voisi vaihtoehtoisena selityksenä ilmakuviin näkyville ajoneuvojen määrille olla alueella vietettävä juhlapäivä. Ilman vaihtoehtoisia selityksiä kaikki indikaattorin poikkeavat arvot tulkitaan suoriksi viitteiksi kohdemuuttujan poikkeavasta arvosta. Joidenkin indikaattorien osalta tämä voi olla luontevaa, jolloin niille ei määri-



Kuva 9: Kohdemuuttujana olevan hypoteesin H , indikaattorien I_1, I_2 ja I_3 sekä vaihtoehtoisten selitysten S_1, \dots, S_5 riippuvuussuhteet kuvaava suunnattu syklitön verkko.

tetä vaihtoehtoisia selityksiä. Samasta vaihtoehtoisesta selityksestä voidaan olettaa riippuvuus useampaan indikaattoriin kuvan 9 indikaattorin X_3 tapaan.

RNM-funktiot ja RNM-parametrien prioriestimaatit

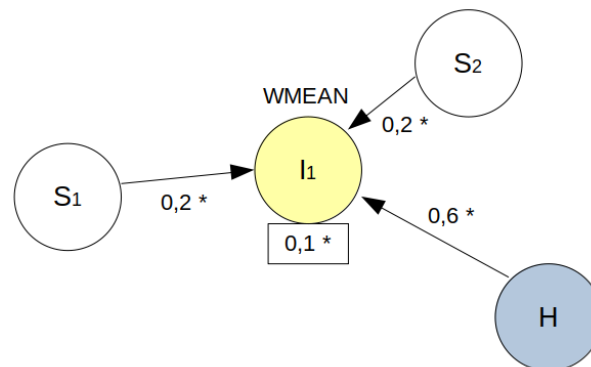
Kohdemuuttujan y sekä selittävien muuttujien X väliset riippuvuudet kuvataan Bayes-verkkona, jonka ehdolliset todennäköisyydet määritetään tämän työn lähestymistavassa järjestysasteikollisten solmujen menetelmällä. Tavanomaisessa RNM-menetelmässä ehdolliset todennäköisyydet määritetään valitsemalla RNM-funktiot r ja RNM-parametrit $\gamma = (w, \sigma)$. Tästä poiketen parametrit γ määritetään tämän työn lähestymistavassa satunnaismuuttujina. Asiantuntija-arvion perusteella määritetään funktioiden r lisäksi satunnaismuuttujien γ priorijakauma $p(\gamma)$. Priorijakauma muodostetaan odotusarvoa kuvaavan prioriestimaatin $\tilde{\gamma} = (\tilde{w}, \tilde{\sigma})$ sekä parametrien γ varianssia kuvaavan prioriestimaatin $\tilde{\epsilon}$ perusteella.

RNM-funktiot ja RNM-parametrien prioriestimaatit määritetään indikaattorikohtaisesti. Indikaattori on tämän työn lähestymistavassa suunnatun syklittömän verkon lapsisolmu, jolla on isäsolmuina kohdemuuttuja sekä mahdollisia vaihtoehtoisia selityksiä. RNM-funktio valitaan asiantuntija-arviona luvussa 3.3 käsitellyistä neljästä funktiosta. WMEAN-funktiolla kuvataan isäsolmujen vaikutus painotettuna keskiarvona. Jos vaikutuksessa korostuu isäsolmujen kulloisenkin suurin tai pienin arvo, on valinta vastaavasti WMAX tai WMEAN. Mikäli riippuvuutta kuvaa parhaiten isäsolmujen kulloisenkin suurimman ja pienimmän arvon yhdistelmä, valinnaksi sopii WMIXMINMAX. Funktion valinnan jälkeen määritetään asiantuntija-arviona sen parametrien odotusarvoa kuvaavat estimaatit $\tilde{\gamma} = (\tilde{w}, \tilde{\sigma})$. Niiden määrittäminen voidaan tehdä sa-

malla tavalla kuin tavanomaisen RNM-menetelmän parametrien määrittäminen. Parametreja voidaan jättää myös kokonaan arvioimatta, jos asiantuntija-arviota ei voida jostakin syystä muodostaa.

RNM-parametrien $\gamma = (w, \sigma)$ hajontaa kuvaava järjestysasteikollinen prioriestimaatti $\tilde{\epsilon}$ perustuu asiantuntija-arvioon prioriestimaattien $\tilde{\gamma}$ tarkkuudesta. Tarkkuus arvioidaan jokaista indikaattoria kohti erikseen indikaattorin RNM-parametreille. Indikaattoriin i liittyville painoille annetaan yksi arvio $\tilde{\epsilon}_{i,w}$. Tämä kuvaa kaikkien indikaattorin painojen vähimmäistarkkuuden. Indikaattorilla i on vain yksi hajontaparametri, joten tätä vastaava arvio $\tilde{\epsilon}_{i,\sigma}$ kuvaa suoraan hajontaparametrin tarkkuuden. Tarkkuusarviot valitaan viisiportaiselta asteikolta: epäselvä (-), suuntaa antava (*), melko tarkka (**), erittäin tarkka (***), täsmällinen (****). Tämän asteikon arvot voidaan ilmaista joko sanallisesti tai suluissa olevaa notaatiota käyttäen. Arvolla “epäselvä” arvioitujen parametrien γ odotusarvon prioriestimaatteja ei huomioida lainkaan myöhemmässä tarkastelussa. Arvioimatta jätettyjen parametrien γ tarkkuudeksi valitaan aina “epäselvä”.

RNM-funktiot ja RNM-parametrien prioriestimaatit voidaan merkitä Bayes-verkon graafiseen esitykseen kuvan 10 tapaan. Kuvassa 10 on arvioitu kuvan 9 indikaattorin I_1 riippuvan isäsolmuistaan WMEAN-funktion mukaisesti. Kohdemuuttujan vaikutus on arvioitu selvästi vaihtoehtoisia selityksiä suuremmaksi, mutta kaikkien parametrien arvioita on pidetty vain suuntaa antavina. Indikaattorin I_1 käsittelyn jälkeen siirryttäisiin kuvan 9 tilanteessa indikaattorien I_2 ja I_3 käsittelyyn. Kun kaikki indikaattorit on käsitelty, muodostetaan prioriestimaattien $\tilde{\gamma}$ ja $\tilde{\epsilon}$ perusteella parametrien γ priorijakauma $p(\gamma)$.



Kuva 10: Yksittäisen indikaattorin I_1 täsmennetyt riippuvuussuhteet. Riippuvuussuhteen luonteen kuvaa RNM-funktio WMEAN. Numerot nuolien vieressä kertovat muuttujia vastaavien painojen w arvot. Indikaattorin alapuolella laatikossa oleva luku on puolestaan hajontaparametrin arvio. Lukujen jäljessä olevat tähdet ovat asiantuntija-arvion tarkkuus.

RNM-parametrien priorijakauma

Seuraavaksi määritellään RNM-parametrien $\gamma = (w, \sigma)$ priorijakauman $p(\gamma) = p(w)p(\sigma)$ muodostaminen tässä luvussa aikaisemmin tehtyjen arvioiden perusteella. Tähän ei liity enää uusien

asiantuntija-arvioiden muodostamista tai valinnaisia toimenpiteitä. Käytettävät jakaumat määräytyvät RNM-funktioiden r perusteella. Jakaumien hyperparametrit määräytyvät prioriestimaattien $\tilde{\gamma} = (\tilde{w}, \tilde{\sigma})$ ja $\tilde{\epsilon}$ perusteella.

Verkon riippuvuussuhteita kuvaavien neljän RNM-funktion painojen w merkitykset eroavat, joten näille määritetään omat priorijakaumansa. WMEAN-funktion (3.2) painot $w = (w_1, \dots, w_k)$ ovat epänegatiiviset siten, että näiden summa on 1. Näitä vaatimuksia vastaa Dirichlet-jakauma

$$w \sim \text{Dir}(\alpha_1, \dots, \alpha_k).$$

WMIN-funktion (3.3) ja WMAX-funktion jokaisen parametrin w_i on oltava vähintään 1, mutta näiden summaa ei rajoiteta. Jokainen parametri $w_i \in (w_1, \dots, w_k)$ voidaan kuvata käyttäen esimerkiksi vasemmalta rajattua normaalijakaumaa

$$w_i \sim N_{(1, \infty)}(\mu, \sigma^2)$$

tai yksinkertaista muunnosta $w'_i = w_i - 1$ ja gamma-jakaumaa

$$w'_i \sim \text{Gamma}(\alpha, \beta).$$

Tässä työssä käytetään jälkimmäistä näistä kahdesta vaihtoehdosta WMIN ja WMAX -jakaumien parametreille, koska gamma-jakauma johtaa rajattua normaalijakaumaa yksinkertaisempiin laskuihin. Myöhemmin palataan vielä erikoistapaukseen, jossa arviota näiden parametrien suuruuksista ei ole, jolloin käytämme vasemmalta rajattua standardinormaalijakaumaa. Tällöin ongelmaa hyperparametrien selvittämisen monimutkaisuudesta ei synny, sillä standardinormaalijakauman hyperparametrit ovat $\mu = 0$ ja $\sigma = 1$.

MIXMINMAX-funktion (3.4) parametrit w_{min} ja w_{max} ovat WMEAN-funktion parametrien tapaan epänegatiivisia ja niiden summa on 1. Jakauman määrittäminen parametrille w_{min} riittää, koska $w_{max} = 1 - w_{min}$. WMEAN-funktion kohdalla käytetty Dirichlet-jakauma pelkistyy yhdellä muuttujalla beta-jakaumaksi

$$w_{min} \sim \text{Beta}(\alpha, \beta).$$

Painojakaumien lisäksi tarvitaan jakauma hajontaparametrille σ , joka määrittää ehdollisen todennäköisyysjakauman huipukkuuden. Hajonta on positiivinen, joten sille voidaan käyttää WMIN-funktion parametrien tavoin gamma-jakaumaa. Toisin kuin WMIN-funktion painot, ha-

jontaparametri ei tarvitse muunnosta. Hajontaparametri kuvataan

$$\sigma \sim \text{Gamma}(\alpha, \beta).$$

RNM-parametrien priorijakaumien hyperparametrit määritetään prioriestimaattien perusteella käyttäen tietoja jakaumien odotusarvoista ja hajonnoista. Indikaattorille i on arvioitu kaksi tarkkuutta: painojen tarkkuus $\tilde{\epsilon}_{i,w}$ ja hajontaparametrin tarkkuus $\tilde{\epsilon}_{i,\sigma}$. Hajontaparametrin järjestysasteikolla arvioitu tarkkuus $\tilde{\epsilon}_{i,\sigma}$ määrittää keskihajonnan s taulukon 1 mukaisesti. Gamma-jakauman odotusarvo on α/β ja varianssi α/β^2 , joten asiantuntijan arvioidessa hajontaparametrin odotusarvoksi $m = \tilde{\sigma}_i$ ja keskihajonnan s saadaan hyperparametrit $\alpha = m^2/s^2$ ja $\beta = m/s^2$.

Taulukko 1: Tarkkuutta kuvaavien termien määrittely keskihajonnan suhteen. "Virhe" vastaa 95 % luottamusväliä normaalijakautuneelle muuttujalle keskihajonnalla s .

Merkintä	Termi	Hajontaparametri σ		WMIN/WMAX painot w		Muut parametrit	
		s	Virhe	s	Virhe	s	Virhe
-	Epäselvä	yli 0,05	yli $\pm 0,1$	yli 0,4	yli $\pm 0,8$	yli 0,1	yli $\pm 0,2$
*	Suuntaa antava	0,05	$\pm 0,1$	0,4	$\pm 0,8$	0,1	$\pm 0,2$
**	Melko tarkka	0,025	$\pm 0,05$	0,2	$\pm 0,4$	0,05	$\pm 0,1$
***	Erittäin tarkka	0,0125	$\pm 0,025$	0,1	$\pm 0,2$	0,025	$\pm 0,05$
****	Täsmällinen	0	0	0	0	0	0

Indikaattorin painojen tarkkuus $\tilde{\epsilon}_{i,w}$ määrittää keskihajonnan s hajontaparametrin tavoin taulukon 1 mukaisesti. RNM-funktiolla voi kuitenkin olla parametrina useita painoja, jolloin s määrittää keskihajonnan ylärajan. Beta-jakauma kuvaa parametrin w_{min} lisäksi epäsuorasti parametrin $w_{max} = 1 - w_{min}$. Tästä syystä näiden parametrien keskihajonnat ovat samat ja niille molemmille käytetään arvoa s . WMIN ja WMAX -funktioiden hajontaparametrit ovat toisistaan riippumattomat, joten myös niiden kohdalla käytetään samaa keskihajontaa s kaikille painoille. Dirichlet-jakauman tapauksessa yksittäisten painojen reunajakaumat riippuvat toisistaan. Jakauman epätarkin parametri w_* on se paino w_i , jonka odotusarvoestimaatti \tilde{w}_i on lähimpänä lukua 0,5. Prioriestimaatti $\tilde{\epsilon}_{i,w}$ määrittää tämän painon keskihajonnan $s_* = s$. Muut funktion hyperparametrit ratkaistaan epätarkimman parametrin keskihajonnan s_* ja saman parametrin odotusarvoestimaatin m_* funktiona. Koonnos hyperparametrit määrittävistä lausekkeista on taulukossa 2.

Keskihajonnan vastaavuus tarkkuutta kuvaaviin termeihin on määritetty normaalijakauman 95 % luottamusvälin suuruuden suhteen parametrin oletettujen arvojoukkojen suuruuksiin. Sama

Taulukko 2: Hyperparametrien muodostaminen valituille jakaumille. Viimeisessä sarakkeessa esitetään hyperparametrien ratkaisu odotusarvon m ja keskihajonnan s funktiona. Dirichlet-jakauman hyperparametrien $(\alpha_1, \dots, \alpha_k)$ määrittäminen tehdään suurimman varianssin s_*^2 ja tätä vastaavan satunnaismuuttujan odotusarvon m_* suhteen.

Parametri	Priorijakauma	Odotusarvo	Varianssi	Hyperparametrit
WMEAN painot	Dirichlet	$\frac{\alpha_i}{\sum_k \alpha_k}$	$\frac{\alpha_i / (\sum_k \alpha_k) (1 - \alpha_i / (\sum_k \alpha_k))}{\sum_k \alpha_k + 1}$	$\alpha_i = m_i \left(\frac{m_*(1-m_*)}{s_*^2} - 1 \right)$
WMIN/WMAX painot	Gamma	α/β	α/β^2	$\alpha = m^2/s^2, \beta = m/s^2$
MIXMINMAX painot	Beta	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\alpha = \frac{m^2(1-m)}{s^2} - m, \beta = \alpha \frac{1-m}{m}$
Hajontaparametrit	Gamma	α/β	α/β^2	$\alpha = m^2/s^2, \beta = m/s^2$

tarkkuutta kuvaava antaa eri RNM-parametreille erisuuruisen hajonnan, kun parametrien arvojoukot ovat erilaiset. Keskihajonnan vastaavuus tarkkuutta kuvaaviin termeihin on määritetty taulukossa 1. WMEAN- ja MIXMINMAX-funktioiden painojen oletusarvojoukko on niiden todellinen arvojoukko $[0, 1]$. Joukon suuruus on 1. Hajontaparametrin sekä WMIN- ja WMAX-funktioiden painojen arvojoukkoja ei kuitenkaan ole rajoitettu ylhäältä. Niiden keskihajonnan suhteutetaan AgenaRisk-ohjelmiston [69] käyttämiin oletusarvojoukkoihin. Tätä ohjelmistoa käytetään yleisesti järjestysasteikollisten solmujen menetelmän soveltamisessa. Hajontaparametrille tämä oletusarvojoukko on $[10^{-4}; 0,5]$ ja WMIN- sekä WMAX-funktioiden parametreille $[1, 5]$. Hajontaparametrin oletusarvojoukon suuruus on lähes 0,5. WMIN- ja WMAX-funktioiden oletusarvojoukon suuruus on 4. Sama tarkkuutta kuvaava termi johtaa siten esimerkiksi WMIN-funktion painojen kohdalla nelinkertaiseen hajontaan WMEAN-funktioon verrattuna. Liitteessä 3 esitetty testi osoittaa priorijakaumien muodostamisen taulukoiden 1 ja 2 mukaisesti mahdollistavan Bayes-päätelyn soveltamisen.

Taulukon 1 mukaan sanallinen arvio “täsmällinen” antaa keskihajonnaksi $s = 0$, jolloin taulukon 2 kaavat eivät ole määriteltyjä. Arvion ilmoittaminen täsmälliseksi tarkoittaa, että ilmiöön ei siltä osin liity epävarmuutta, jolloin havainnot eivät myöskään voi kertoa siitä mitään uutta. Tällöin parametri tulkitaan satunnaismuuttujan sijaan lukuna, eikä sitä päivitetä Bayes-päätelyn keinoin. Tämä voi olla hyödyllinen ominaisuus, kun arvio on varma tai jos datan uskotaan olevan harhaanjohtavaa. Muuksi kuin täsmälliseksi määritetty arvio voi havaintojen lisääntyessä päivittyä “erittäin tarkaksi”, mutta ei koskaan “täsmälliseksi”.

Termin “täsmällinen” lisäksi erityinen merkitys on myös taulukon 1 termillä “epäselvä”, jolle muista taulukon termeistä poiketen ei ole määritetty tarkkaa keskihajontaa. Kyseinen termi on tarkoitettu käytettäväksi tilanteissa, joissa arviota parametrille ei ole saatavilla. Kun priorijakaumaa ei voida määrittää täsmällisen odotusarvon ja hajonnan suhteen, määritetään epäinformatiivinen prioriksi. Tämän tarkoituksena on ilmaista neutraali suhtautuminen parametrin mahdollisiin

arvoihin. Asiantuntija voi käytännössä jättää parametrin arvioimatta, jolloin epäinformatiivinen priorin antaa datalle suuren merkityksen parametrin arvioinnissa. Neutraalius riippuu kontekstista, joten epäinformatiivinen priorin määritetään seuraavaksi jokaiselle parametrille erikseen.

WMEAN-painojen epäinformatiiviseksi prioriksi valitaan $\text{Dir}(1, 1, \dots, 1)$, joka antaa kaikille painoille yhtä suuren prioriarvon. Se sallii posterioriarvioiden nopean muuttumisen prioriarvioon verrattuna datan lisääntyessä. Epäinformatiivinen beta-priorin $\text{Beta}(1, 1)$ saadaan WMEAN-painojen epäinformatiivisesta priorista erikoistapauksena. Gamma-jakauman epäoleellinen priorijakauma $p(\xi) \propto 1$ antaa kaikille parametrin arvoille $\xi > 0$ yhtä suuren prioritodennäköisyyden, mutta painottaa suuria arvoja suhteessa mihin tahansa rajattuun arvojoukkoon. Tätä ei voi pitää neutraalina, sillä paljon käytetty AgenaRisk-ohjelmisto olettaa sekä hajontaparametrin että WMIN- ja WMAX-funktioiden parametrin arvojoukot nolasta alkaviksi ja ylhäältä rajatuiksi. Näiden parametrin pieniä arvoja voidaan siten pitää lähtökohtaisesti todennäköisempinä kuin suuria arvoja. Gamma-jakaumalla kuvatuille parametreille valitaan tästä syystä epäinformatiiviseksi prioriksi vasemmalta rajattu standardinormaalijakauma $N_{(0,\infty)}(0, 1)$. Tämä jakauma sallii kaikki positiiviset arvot, mutta painottaa pieniä arvoja suuriin verrattuna. Valitut epäinformatiiviset priorijakaumat on esitetty kootusti taulukossa 3.

Taulukko 3: Epäinformatiivisten priorijakaumien määrittely parametreittain. WMEAN ja MIXMINMAX -funktioiden painoille epäinformatiivinen priorin saadaan antamalla kaikille niiden hyperparametreille arvoksi 1. WMIN/WMAX -funktioiden painot sekä hajontaparametrit mallinnetaan muutoin niiden kohdalla käytetyn gamma-jakauman sijaan vasemmalta rajatulla standardinormaalijakaumalla.

Parametri	Epäinformatiivinen priorin
WMEAN painot	$\text{Dirichlet}(1, 1, \dots, 1)$
WMIN/WMAX painot	$N_{(0,\infty)}(0, 1)$
MIXMINMAX painot	$\text{Beta}(1, 1)$
Hajontaparametrit	$N_{(0,\infty)}(0, 1)$

4.3 Aineiston kerääminen

Lähestymistavan toisessa vaiheessa kerätään kaikki saatavilla oleva data kohdemuuttujasta y sekä selittävästä muuttujasta X joltakin aikaväliltä. Aineiston kattama aikaväli valitaan siten, että datasta mahdollisesti löytyvät säännönmukaisuudet ovat yleistettävissä koko aikavälillä. Aineistosta ei muodosteta esimerkiksi erillistä opetus- tai testidataa, vaan kaikki saatavilla oleva data käsitellään yhtenä aineistona.

Aineiston tulee sisältää dataa kohdemuuttujasta, mutta kohdemuuttuja voi olla aineiston katta-

malla aikavälillä pääasiassa tuntematon. Jos kohdemuuttujasta ei ole saatavilla dataa aineiston kattamalla aikavälillä, siitä muodostetaan muutamia asiantuntija-arvioita. Nämä asiantuntija-arviot liitetään aineistoon ja niitä käsitellään tässä luvussa esiteltävässä lähestymistavassa kuten dataa.

Tässä luvussa esiteltävällä lähestymistavalla muodostettua arviota voidaan käyttää pohjana uudelle tällä lähestymistavalla muodostettavalle arviolle. Tässä tapauksessa uusi data lisätään edellisellä kerralla käytetyn aineiston jatkoksi. Näin menetellen voidaan hyödyntää esimerkiksi säännöllisesti kertyvää dataa uusien arvioiden tuottamiseen.

4.4 Aineiston esikäsittely

Lähestymistavan kolmannessa vaiheessa alkuperäinen aineisto eli raakadata muokataan estimoinnin mahdollistavaan muotoon. Aineistolle tehdään esikäsittelyssä muunnoksia, joista osa on välttämättömiä ja osa valinnaisia. Esimerkiksi havaintojen jaottelu alaluvussa 4.3 valitun aikayksikön mukaisesti on välttämätön muunnos. Toisaalta valinnaisena muunnoksena voidaan esimerkiksi pyrkiä tuomaan jonkin indikaattorin vaihtelu paremmin esiin vaihtamalla sen mittasteikko. Tässä vaiheessa päätetään, käsitelläänkö aineisto vaiheessa 4 järjestysasteikollisena vai välimatka-asteikollisena. Koko aineisto muunnetaan vaiheen lopuksi tämän valinnan mukaisesti. Tässä alaluvussa tarkastellaan seuraavaksi aineiston jaottelun määrittävää aikayksikön valintaa. Aikayksikön valinnan jälkeen esitellään kaikki aineistolle tehtävät välttämättömät muunnokset sekä eräitä hyödyllisiä valinnaisia muunnoksia. Lopuksi tarkastellaan muunnosten toteuttamisen järjestystä.

Aikayksikön valinta

Jokainen kohdemuuttujan $y = \{y_1, y_2, \dots\}$ arvo y_i estimoidaan ajanhetkestä t_i tietoa antavan datan perusteella. Tätä varten aineisto jaotellaan esikäsittelyvaiheessa aikaväleille t_1, \dots, t_N , missä t_1 sisältää ainakin aineiston ensimmäisen havainnon ja t_N ainakin aineiston viimeisen havainnon. Aikavälien määrittämistä varten valitaan aloitushetki a_0 sekä aikavälin pituuden määrittävä aikayksikkö a . Tämä voi olla esimerkiksi viikko tai kuukausi. Aikavälit muodostuvat tällöin $t_i = [a_0 + (i - 1)a, a_0 + i \cdot a]$. Pidempi aikayksikkö a tuottaa suurempaan aineiston määrään perustuvia arvioita, jotka ovat siten luotettavampia. Toisaalta arvioita saadaan tällöin harvemmin.

Välttämättömät muunnokset

Muunnos välimatka-asteikolle

Järjestysasteikollisella datalla ei voida tehdä laskutoimituksia. Järjestysasteikollinen data on siksi muunnettava välimatka-asteikolliseksi muiden tässä alaluvussa kuvattujen muunnosten mahdollistamiseksi. Järjestysasteikolle määritetään muunnos sen arvojoukosta joukkoon $\mathbb{R} \cap [0, 1]$. Tämä toteutetaan jakamalla väli $[0, 1]$ järjestysasteikkoa vastaaviin yhtä suuriin osiin ja tulkitsemalla jokainen asteikon arvo vastaavan välin keskimmäisenä lukuna. Esimerkiksi jos järjestysasteikolla on kolme arvoa, niin näitä vastaavat välit $[0, 1/3]$, $[1/3, 2/3]$ ja $[2/3, 1]$. Tämän järjestysasteikon arvot kuvautuvat näiden välien keskikohdiksi, jotka ovat $1/6$, $1/2$ ja $5/6$. Muuttuja voidaan jatkokäsittelyn jälkeen muuntaa takaisin järjestysasteikolliseksi. Tähän palataan valinnaisten muunnosten osuudessa. Jos järjestysasteikollista muuttujaa käytetään sellaisenaan, ilman tarvetta muille tässä alaluvussa kuvatuille muunnoksille, voidaan edestakainen asteikon muuttaminen luonnollisesti jättää tekemättä.

Jaottelu aikayksikön mukaan

Kohdemuuttujan $y = \{y_1, y_2, \dots\}$ arvo y_i on kohdemuuttujan arvo ajanhetkellä t_i . Kohdemuuttujan arvoa y_i vastaavat selittävien muuttujien arvot $x_i = \{x_i^{(1)}, \dots, x_i^{(k)}\}$, jotka koostuvat kaikkien indikaattorien ja vaihtoehtoisten selitysten parhaiten ajanhetkeä t_i kuvaavista arvoista. Aineistosta muodostetaan ajanhetkiä t_i parhaiten kuvaavat arvot sekä kohdemuuttujalle että selittäville muuttujille. Ajanhetkeä t_i vastaavat kohdemuuttujan tai selittävien muuttujien arvot merkitään puuttuvaksi, jos niiden muodostamiseen ei ole dataa. Selittävien muuttujien arvoista $x_i = \{x_i^{(1)}, \dots, x_i^{(k)}\}$ ajanhetkellä t_i merkitään puuttuviksi ainoastaan niiden indikaattorien tai vaihtoehtoisten selitysten arvot, joiden muodostamiseen ei ole dataa.

Parhaiten ajanhetkeä t_i kuvaavien arvojen laskentaan voi olla useita keinoja. Esimerkiksi useita lukuja aikayksikössä sisältävästä raakadatasta voidaan laskea aikavälejä vastaavat keskiluvut. Toisaalta raakadatan koostuessa yksittäisten tapahtumien tapahtuma-ajoista jaottelu onnistuu luontevasti laskemalla tapahtumien lukumäärät aikaväleittäin. Puuttuvan raakadatan tulkintaan tulee kiinnittää huomiota. Jos havaintoja ei ole aikavälillä voitu jostakin syystä kerätä, tulee tätä vastaava arvo lähtökohtaisesti merkitä puuttuvaksi. Jos havaintoja on voitu kerätä aikavälillä vain osittain, voidaan harkinnanvaraisesti käyttää puuttuvan arvon sijaan esimerkiksi osittaiselta väliltä laskettua keskiarvoa. Kuvassa 11 on esimerkki aikayksikön mukaan jaottelusta. Tämän kuvan esimerkissä raakadata koostuu tapahtumien päivämääristä. Havaintoja kerännyt sensori ei ollut toimintakunnossa yhdellä aikavälillä. Havaintojen puute on ratkaistu merkitsemällä tämän

Havainnot	Aikaväli	x	x'
1.1.	1.1. - 7.1.	1	0,25
10.1.	8.1. - 14.1.	4	1
11.1.	15.1. - 21.1.	0	0
13.1.	22.1. - 29.1.	NA	NA
14.1.	30.1. - 5.2.	2	0,5
30.1.			
31.1.			

Kuva 11: Esimerkki raakadatan muuntamisesta ensin numeeriseksi ja tämän jälkeen välille $[0, 1]$. Sensorista saatujen havaintojen päivämäärät muodostavat vasemmanpuoleisen taulukon raakadatan. Tästä on laskettu oikeanpuoleiseen taulukkoon viikon mittaisia aikavälejä vastaava muuttuja x sekä tämän muunnos x' . Aineistoon on sisällytetty puuttuvana havaintona viikko, jolloin sensori ei ollut toimintakunnossa. Puuttuvat havainnot on merkitty koodilla NA.

aikavälin havainto puuttuvaksi (NA), sillä arvon 0 käyttäminen olisi harhaanjohtavaa.

Tässä luvussa esiteltävän lähestymistavan oletuksena on, että ainakin osa kohdemuuttujan arvoista tunnetaan. Kohdemuuttujan tunnettujen arvojen indeksit muodostavat joukon

$$T = \{i \mid i \in \mathbb{Z}^+ : y_i \text{ tunnetaan}\}.$$

Arvot y_T eivät ole satunnaismuuttujia eikä niitä estimoida lähestymistavan vaiheessa 4.

Kääntäminen

Kohdemuuttujan, indikaattorien ja vaihtoehtoisten selitysten väliset korrelaatiot oletetaan positiivisiksi. Kohdemuuttujan sekä vaihtoehtoisten selitysten suurien arvojen tulisi siis johtaa suuriin arvoihin indikaattoreille. Kahden muuttujan välinen negatiivinen korrelaatio voidaan kääntää positiiviseksi kääntämällä muuttujista toinen. Kohdemuuttujan y kanssa oletettavasti negatiivisesti korreloivat indikaattorit ja vaihtoehtoiset selitykset x käännetään muunnoksella

$$f(x) = -x. \quad (4.1)$$

Poikkeama keskiluvusta

Joissakin tilanteissa korrelaation suuntaa ei voida olettaa. Tällöin havaintojen normaaliin verrattuna pienet arvot voivat olla merkityksellisiä siinä missä suuret arvot. Korrelaation suunta voidaan jättää huomiotta laskemalla muunnos keskimääräisen arvon suhteen. Tämä saadaan itseisarvona

$$f(x) = |x - \bar{x}|. \quad (4.2)$$

Keskiluku \bar{x} tulee valita vastaamaan muuttujan keskimääräistä arvoa. Aritmeettinen keskiarvo voi olla huono valinta keskiluvuksi, jos muuttuja poikkeaa normaalista systemaattisesti pääasiassa ylös tai alas. Epätasaisten poikkeamien vaikutusta voidaan vähentää jonkin verran käyttämällä keskilukuna mediaania. Epätasaisten poikkeamien tuottaman virheen välttämiseksi muunnosta ei tule käyttää vain poistamaan korrelaatioiden suunnan arvioinnin tarve.

Skaalaus

Lähestymistavan vaiheessa 4 oletetaan kaikki data skaalatuksi välille $[0, 1]$. Tämä toteutetaan muunnoksella

$$f(x) = \frac{x - \min(x)}{\max(x) - \min(x)}. \quad (4.3)$$

Muunnosta on käytetty kuvan 11 esimerkissä.

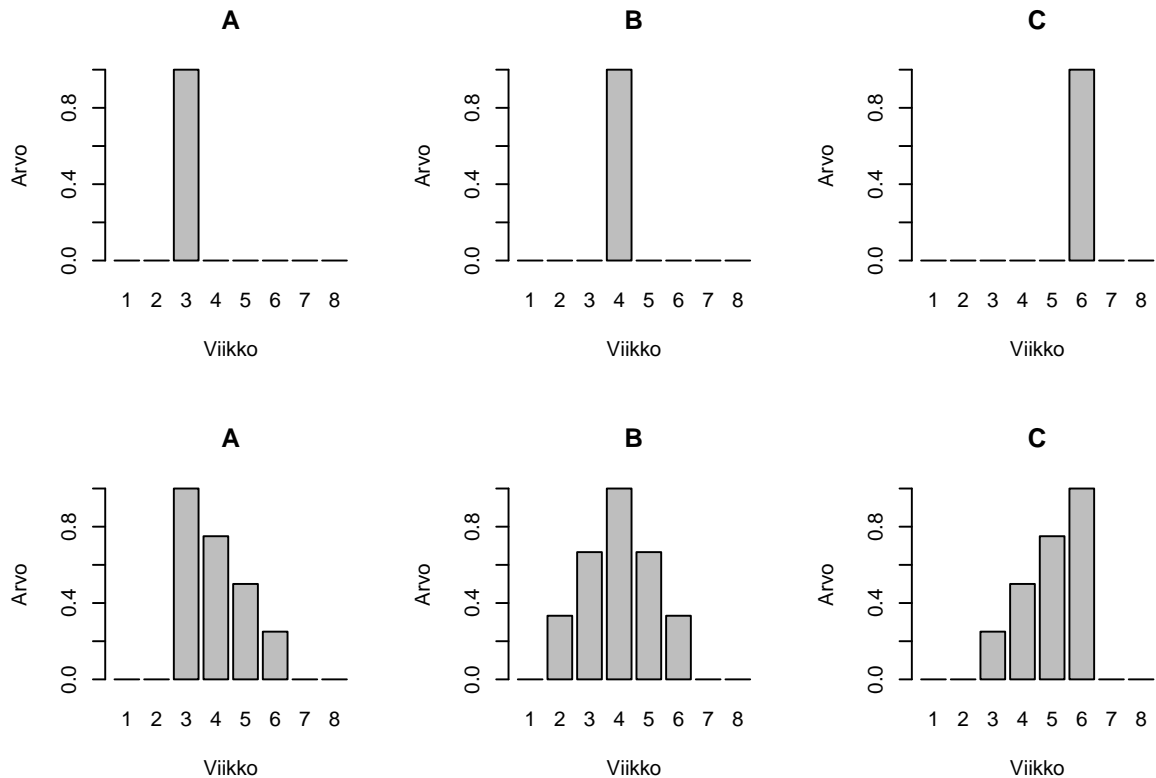
Valinnaiset muunnokset

Painotettu liukuva keskiarvo

Kohdemuuttujaa y arvioidaan erikseen jokaisella aikavälillä t_i ja kaikki nämä arviot ovat toisistaan riippumattomia. Yksittäinen havainto vaikuttaa siis ainoastaan oman aikavälinsä arvioon. Kuitenkin todellisuudessa yksittäinen havainto voi olla syytä huomioida pidemmälläkin aikavälillä. Tämän työn lähestymistavassa havaintojen pidemmälle kantava vaikutus huomioidaan tasoittamalla muuttujia esikäsittelyvaiheessa painotetulla liukuvalla keskiarvolla. Liukuvan keskiarvon pituudeksi valitaan yksittäisen havainnon oletettu vaikutusaika. Havainnon todellinen ajankohta saa painokertoimen 1, josta kerroin laskee lineaarisesti nolnaan. Lähtökohtaisesti liukuva keskiarvo lasketaan symmetrisesti molemmin puolin havaintoa, mutta perustellusta syystä se voidaan laskea toispuoleisesti.

Kuvan 12 esimerkissä on seurattu kohteen viikoittaista aktiivisuutta kolmella indikaattorilla. Indikaattori A viittaa tulevaan, B meneillään olevaan ja C hiljattain päättyneeseen aktiivisuuteen. Jokaisesta indikaattorista on saatu vain yksi havainto. Ihmiselle kokonaisuuden näkeminen havaintojen taustalla voi olla helppoa, mutta jokaiselle viikolle erikseen arviota tekevä tietokone ei pysty yhdistämään kuvan ylärivin mukaisesti eri viikoilla tehtyjä havaintoja. Indikaattorit voivat kuitenkin kertoa viikkoa pidemmästä kokonaisuudesta, joten niitä voidaan tasoittaa painotetulla liukuvalla keskiarvolla. A viittaa tulevaan ja C menneeseen aktiivisuuteen, joten ne tasoitetaan toispuoleisesti. B tasoitetaan symmetrisesti. Tasoitetuista indikaattoreista kuvan alemmalla rivillä nähdään kaikkien niiden saavan nollasta poikkeavia arvoja lähes samanaikaisesti. Tämä

mahdollistaa automaattisen päättelyn eri aikoina saatujen vähäisten havaintojen perusteella.



Kuva 12: Ylärivillä on piirretty kolmen eri indikaattorin arvot viikoittain jaoteltuna. Jokainen indikaattori on havaittu vain kerran ja saa tätä vastaavalla viikolla arvon 1. Jokainen havainto antaa kuitenkin informaatiota noin kuukauden pituiselta ajalta. Alarivillä indikaattorit A ja C on tasoitettu toispuoleisesti ja indikaattori B symmetrisesti käyttäen painotettua liukuvaa keskiarvoa.

Logaritmimuunnos

Esimerkiksi suosiota tai julkisuutta mittaavien muuttujien vaihtelu voi olla eksponentiaalista. Suurin osa havainnoista on tällöin vaihteluvälin alaosassa ja muutama havainto nousee poikkeaviin lukemiin. Eksponentiaalinen vaihtelu voi olla tarpeen tasoittaa vaihtamalla muuttujan asteikko logaritmiseksi. Logaritmiselle asteikolle muunnettuna poikkeavat havainnot jäävät edelleen suurimmiksi, mutta keskimääräiset havainnot erottuvat paremmin pienimmistä havainnoista.

Muunnos järjestysasteikolle

Esikäsittelyvaiheessa tehdään päätös lähestymistavan vaiheessa 4 käytettävän aineiston tyyppistä. Tämän työn lähestymistapa sallii estimoinnin joko järjestysasteikollista tai välimatkaasteikollista aineistoa käyttäen, mutta päätös on tehtävä yhtenevästi koko aineiston suhteen. Aikaisemmin alaluvussa 3.3 esitellyn teorian perusteella estimointi kannattaa toteuttaa aineiston alkuperäisessä muodossa. Jokainen muunnos asteikkojen välillä voi heikentää aineiston alkuperäistä informaation sisältöä. Jos siis suurin osa raakadatasta on järjestysasteikollista, voi olla

hyödyllistä muuntaa loputkin aineistosta sellaiseksi. Muunnos järjestysasteikolle tehdään päätämällä muuttujan arvojen määrä ja jakamalla muuttujan vaihteluväli näin moneen yhtä suureen osaan. Koska esikäsittelyvaiheen muunnokset tehdään välimatka-asteikolla, voi useiden muunnosten jälkeen olla perusteltua jättää aineisto välimatka-asteikolliseksi alkuperäisestä asteikon tyypistä riippumatta.

Muunnosten järjestys

Muunnosten järjestyksellä on merkitystä, kun niitä tehdään useita samalle muuttujalle. Järjestysasteikollisten muuttujien käsittely on aloitettava tekemällä muunnos välimatka-asteikolle, sillä muut muunnokset edellyttävät laskutoimitusten tekoa muuttujien arvoilla. Liukuvien keskiarvojen ottaminen edellyttää aineiston jaottelua aikayksikön mukaan. Lisäksi mahdollinen logaritmuunnos tulee olla tehtynä ennen painotetulla liukuvalla keskiarvolla tehtävää tasoitusta. Poikkeama keskiluvusta lasketaan vasta painotetulla liukuvalla keskiarvolla tehtävän tasoituksen jälkeen, sillä erityisesti vähän havaintoja sisältävässä muuttujassa painotettu liukuva keskiarvo voi vähentää keskiluvun riippuvuutta äärimmäisistä havainnoista. Kääntäminen muuttaa ainoastaan arvojen etumerkin, joten se voidaan tehdä missä tahansa vaiheessa logaritmin ja skaalauksen välissä. Kääntämistä ja poikkeamaa keskiluvusta ei kuitenkaan tehdä samalle muuttujalle, sillä ne liittyvät erilaisiin oletuksiin muuttujan korrelaatiosta muiden muuttujien kanssa. Skaalauksen on oltava välimatka-asteikolliselle muuttujalle tehtävistä toimenpiteistä viimeinen. Muunnos järjestysasteikolle voidaan tehdä tarvittaessa tämän jälkeen.

Muunnokset toteutetaan seuraavassa järjestyksessä:

1. muunnos välimatka-asteikolle,
2. jaottelu aikayksikön mukaan,
3. logaritmi,
4. painotettu liukuva keskiarvo,
5. kääntäminen,
6. poikkeama keskiluvusta,
7. skaalaus,
8. muunnos järjestysasteikolle.

4.5 Estimointi

Lähestymistavan vaiheessa 4 lasketaan kohdemuuttujan y tuntemattomille arvoille sekä RNM-parametreille γ posterioriestimaatit \hat{y} ja $\hat{\gamma}$. Posterioriestimaatit perustuvat priorijakaumiin $p(y)$ ja $p(\gamma)$ sekä aineistoon D . Posterioriestimaatit määritetään odotusarvoina $\hat{y} = E(y | D)$ ja $\hat{\gamma} = E(\gamma | D)$. Posteriorijakauma määritetään Bayes-päätelyllä. Lisäksi RNM-parametrien posterioriestimaatille $\hat{\gamma}$ määritetään posteriorijakauman varianssin perusteella tarkkuutta kuvaava posterioriestimaatti $\hat{\epsilon}$. Posterioriestimaatteja $\hat{\gamma}$ ja $\hat{\epsilon}$ verrataan asiantuntija-arvion perusteella muodostettuihin prioriestimaatteihin $\tilde{\gamma}$ ja $\tilde{\epsilon}$ lähestymistavan vaiheessa 5. Kohdemuuttujan y posterioriestimaatteja \hat{y} tarkastellaan lähestymistavan vaiheessa 6.

Posteriorijakaumasta muodostetaan otos Gibbsin otannalla samanaikaisesti kaikille tuntemattomille muuttujille θ . Näihin kuuluvat RNM-parametrit γ , kohdemuuttujan y tuntemattomat arvot ja selittävien muuttujien X tuntemattomat arvot X_{NA} . Posteriorijakauman muodostamiseen tarvitaan kaikkien tuntemattomien muuttujien priorijakauma $p(\theta)$ sekä havaintojen todennäköisyyden kuvaava uskottavuusfunktio. Priorijakaumat $p(y)$ ja $p(\gamma)$ muodostettiin vaiheessa 1, mutta priorijakauma $p(X_{NA})$ on vielä määritettävä vaiheessa 3 puuttuviksi kirjatuille selittävien muuttujien havainnoille. Priorijakauma kaikille tuntemattomille muuttujille määritetään

$$p(\theta) = p(y) p(\gamma) p(X_{NA}).$$

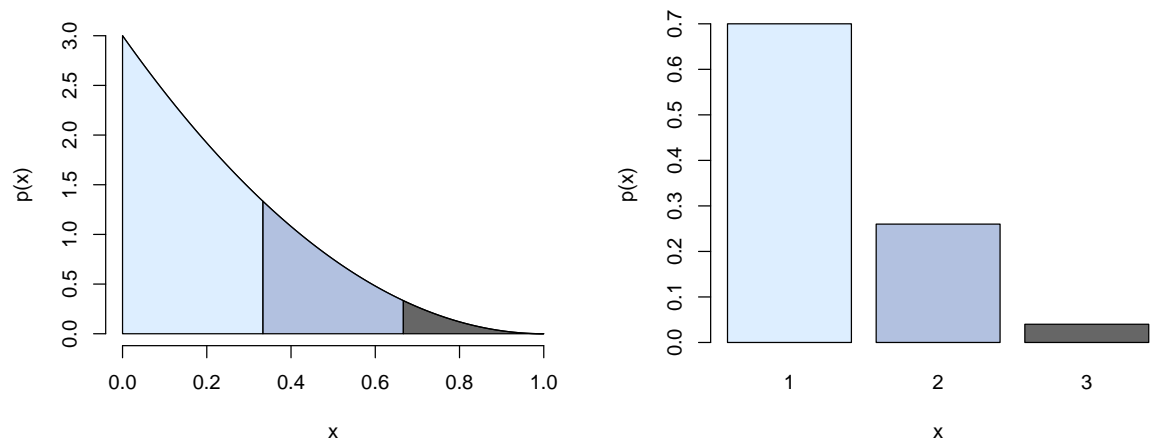
Uskottavuusfunktio $p(X, y_T | \theta)$ havaituille kohdemuuttujan arvoille y_T sekä selittävien muuttujien arvoille X määräytyy vaiheessa 1 asiantuntija-arvion perusteella määritetyn suunnatun syklittömän verkon ja RNM-funktioiden perusteella. Nämä määrittävät arvojen θ funktiona Bayes-verkon, joka mahdollistaa havaintojen kokonaistodennäköisyyden laskemisen.

Puuttuvien havaintojen priorijakauma

Puuttuvat havainnot voivat olla sekä kohdemuuttujan y tuntemattomia arvoja että selittävien muuttujien X puuttuvia arvoja X_{NA} . Kohdemuuttujan arvot ovat tämän työn lähestymistavan varsinainen kiinnostuksen kohde. Selittävien muuttujien puuttuvat arvot X_{NA} ovat ylimääräisiä niin sanottuja kiusaparametreja. Ne estimoidaan kuitenkin samanaikaisesti. Myös kiusaparametreille on määritettävä priorijakauma. Selittävät muuttujat ovat indikaattoreita ja vaihtoehtoisia selityksiä. Indikaattorien ehdolliset jakaumat määritetään kohdemuuttujan ja vaihtoehtoisten selitysten suhteen. Kohdemuuttujan priorijakauma määritettiin vaiheessa 1, joten priorijakauma tarvitaan vielä vaihtoehtoisten selitysten puuttuville arvoille.

Vaihtoehtoisten selitysten puuttuville arvoille määritetään asiantuntija-arvioihin perustumaton priorijakauma. Matalammille arvoille annetaan hieman suurempi todennäköisyys kuin suurille. Kaikki muuttujat on muunnettu esikäsittelyvaiheessa korreloimaan positiivisesti kohdemuuttujan kanssa, joten suurten arvojen painottaminen tai tasaisen priorijakauman käyttäminen voisi useiden havaintojen puuttuessa johtaa korkeaan arvioon kohdemuuttujan arvosta. Priorijakaumana käytetään välimatka-asteikollisille muuttujille jakaumaa $Beta(1, 3)$. Järjestysasteikollisille muuttujille käytetään tästä $Beta$ -jakaumasta diskretisoimalla saatua diskreettiä jakaumaa.

Kuvassa 13 on esimerkki jakauman $Beta(1, 3)$ diskretisoinnista. Kuvan tilanteessa muodostetaan diskreetti pistetodennäköisyysfunktio järjestysasteikolliselle muuttujalle, jolla on kolme arvoa. Jatkuva priorifunktio $Beta(1, 3)$ on jaettu kuvassa väleihin $[0, 1/3]$, $[1/3, 2/3]$ ja $[2/3, 1]$. Näitä vastaavat pistetodennäköisyydet ovat 70 %, 26 % ja 4 %. Pistetodennäköisyydet voidaan muodostaa jakauman $Beta(1, 3)$ perusteella samaan tapaan mille tahansa järjestysasteikollisen muuttujan arvojen määrälle.



Kuva 13: Vasemmalla on esitetty jakauman $Beta(1, 3)$ tiheysfunktio. Tämän tiheysfunktion alapuolelle jäävä pinta-ala on jaettu kolmeen osaan. Osien pinta-alat muodostavat diskreetin jakauman pistetodennäköisyydet. Nämä on esitetty oikealla.

Posteriorijakauman approksimointi Gibbsin otantaa käyttäen

Posteriorijakauman $p(\theta | D)$ analyttinen ratkaiseminen on Bayes-verkkojen tapauksessa haasteellista. Posteriorijakaumasta voidaan muodostaa otos Gibbsin otannalla. Otoksesta voidaan laskea yksittäisten muuttujien posterioriestimaatit kaavalla (3.12).

Gibbsin otanta toteutetaan käytännössä tietokoneohjelmalla. Ohjelmalle annetaan syötteenä todennäköisyysmalli, joka koostuu tässä luvussa määritetyistä priorijakaumista, uskottavuusfunktioista sekä esikäsittelystä aineistosta. Gibbsin otannan toteuttamiseen soveltuvat esimerkiksi

BUGS (Bayesian inference Using Gibbs Sampling) -tyyppiset ohjelmistot [62–64]. Käyttäjän tehtäväksi jää varmistua alaluvussa 3.4 käsitellyn sisäänpalamisajan riittävydestä ja valita estimaattien laskemiseen käytettävän otoksen pituus n .

Esimerkissä 4.1 esitetään todennäköisyysmallin määrittely BUGS-tyyppiseen ohjelmaan tämän luvun lähestymistapaa käyttäen. Todennäköisyysmalli määritellään alaluvun 4.2 mukaisen asiantuntija-arvion perusteella.

Esimerkki 4.1. *Arvioidaan hypoteesia H käyttäen yhtä indikaattoria I . Indikaattorilla on lisäksi vaihtoehtoinen selitys S . Asiantuntija-arvion mukaan hypoteesin H ja vaihtoehtoisen selityksen S vaikutus indikaattoriin I on kuvattavissa painotettuna keskiarvona. Riippuvuussuhdetta kuvaa siis WMEAN-funktio. Asiantuntija on arvioinut painokertoimiksi $w_H = w_S = 0,5$ sekä hajonnaksi $\sigma = 0,2$. Asiantuntija pitää arvioitaan melko tarkkoina. Aineisto D ei sisällä puutteita indikaattorin tai vaihtoehtoisen selityksen osalta, mutta hypoteesin H kaikki arvot eivät ole tiedossa. Hypoteesin tuntemattomat arvot estimoidaan. Asiantuntija-arvion mukaan hypoteesi pitää lähitulevaisuudessa paikkansa 10 % todennäköisyydellä. Aineisto on päätetty analysoida välimatka-asteikollisena.*

Painojen $w = (w_H, w_S)$ priorijakaumaksi saadaan taulukon 2 perusteella $Dir(49,5; 49,5)$. Tämä voidaan ilmaista myös yksinkertaisemmin $w_H \sim Beta(49,5; 49,5)$ ja $w_S = 1 - w_H$. Hajontaparametrin σ jakaumaksi saadaan samaa taulukkoa käyttäen $Gamma(1,9; 8,9)$. Indikaattorin ehdollinen jakauma määritetään välimatka-asteikollista kaavaa (3.11) käyttäen. Siinä esiintyvä funktio f on arvion mukaan WMEAN, eli $f(H, S, w_H) = H \cdot w_H + S \cdot (1 - w_H)$. Hypoteesin priorijakauma on saadun arvion ja alaluvun 4.2 perusteella Bernoulli-jakauma $B(0,1)$. Todennäköisyysmalli annetaan BUGS-ohjelmistolle syötteenä:

```

model{
  wH ~ dbeta(49.5, 49.5)
  sigma ~ dgamma(1.9, 8.9)
  tau <- pow(sigma, -2)
  for (i in 1:N) {
    mean[i] <- wH * H[i] + (1-wH) * S[i]
    H[i] ~ dbern(0.1)
    I[i] ~ dnorm(mean[i], tau) T(0,1)
  }
}

```

Tässä τ on ohjelmiston normaalijakauman parametrina käyttämä varianssin käänteisluku ja

N aineiston aikavälien lukumäärä.

Esimerkissä 4.1 määritettiin todennäköisyysmalli BUGS-tyyppiselle ohjelmistolle. Aineiston syöttämisen jälkeen ohjelmisto laskee posteriorijakaumasta otoksen. Otoksen perusteella voidaan laskea estimaatit Bayes-verkon parametreille sekä hypoteesin tuntemattomille arvoille.

Estimaattien laskenta

Gibbsin otannalla posteriorijakaumasta $p(\theta | D)$ muodostetusta otoksesta (Z_1, \dots, Z_n) lasketaan tuntemattomien muuttujien estimaatit kaavalla (3.12). Satunnaismuuttujan $\xi \in \theta$ odotusarvo saadaan valitsemalla kaavassa (3.12) $f(\theta) = \xi$, jolloin $E(\xi) = E(f(\theta)) \approx \frac{1}{n} \sum_{i=1}^n f(Z_i)$. Tästä saadaan estimaatti \hat{y} tuntemattomille kohdemuuttujan y arvoille valitsemalla $f_y(\theta) = y$, jolloin

$$\hat{y} = E(f_y(\theta)) \approx \frac{1}{n} \sum_{i=1}^n f_y(Z_i).$$

RNM-parametrit $\gamma = (w, \sigma)$ saadaan vastaavasti valitsemalla $f_\gamma(\theta) = \gamma$, jolloin

$$\hat{\gamma} = E(f_\gamma(\theta)) \approx \frac{1}{n} \sum_{i=1}^n f_\gamma(Z_i).$$

RNM-parametrien $\gamma = (w, \sigma)$ tarkkuudelle lasketaan järjestysasteikolliset estimaatit $\hat{\epsilon}$ RNM-parametrien posteriorivarianssin estimaatin \hat{v}^2 perusteella. Satunnaismuuttujan ξ varianssi on määritelmältään odotusarvo $\text{var}(\xi) = E((\xi - E(\xi))^2)$, joten sijoittamalla tähän RNM-parametrin γ_p odotusarvon posterioriestimaatti $\hat{\gamma}_p$ ja valitsemalla $f_{v,p}(\theta) = (f_{\gamma_p}(\theta) - \hat{\gamma}_p)^2$ saadaan kaavalla (3.12) RNM-parametrin γ_p varianssille estimaatti

$$\hat{v}_p^2 = E(f_{v,p}(\theta)) \approx \frac{1}{n} \sum_{i=1}^n (f_{\gamma_p}(Z_i) - \hat{\gamma}_p)^2.$$

RNM-hajontaparametrin σ tarkkuuden posterioriestimaatti $\hat{\epsilon}_\sigma$ määritetään taulukon 1 perusteella. Posterioriestimaatti $\hat{\epsilon}_\sigma$ on pienin taulukon 1 arvo s , jolla $\sqrt{\hat{v}_\sigma^2} < s$. Esimerkiksi jos asiantuntija-arviona määritetyn suunnatun syklittömän verkon lapsisolmun i hajontaparametrin σ_i posteriorivarianssin estimaatti on $0,03^2$, niin tätä vastaava järjestysasteikollinen posterioriestimaatti on $\hat{\epsilon}_{i,\sigma} = \text{”suuntaa antava”}$. Tätä paremman arvon saavuttamiseksi posteriorivarianssin estimaatin tulee taulukon 1 mukaan olla pienempi kuin $0,025^2$. RNM-painojen w tarkkuuden posterioriestimaatti lasketaan muuten samoin, mutta jokaiselle lapsisolmulle i määritetään vain yksi painojen tarkkuusestimaatti $\hat{\epsilon}_{i,w}$. Tämä määritetään solmun i painojen $w^{(i)} = (w_1^{(i)}, \dots, w_k^{(i)})$ suurimman posteriorivarianssin $\max\{\hat{v}_{i,1}^2, \dots, \hat{v}_{i,k}^2\}$ mukaan.

4.6 Validointi

Vaiheessa 4 määritettyjä posterioriestimaatteja \hat{y} tarkastellaan tässä luvussa esiteltävän lähestymistavan vaiheessa 6. Posterioriestimaatit riippuvat aineiston D lisäksi vaiheessa 1 tehdyistä asiantuntija-arvioista, joita voi olla erilaisia. Esimerkiksi jonkin muuttujan sisällyttämisestä tarkasteluun voi olla eriäviä näkemyksiä. Vaiheessa 5 arvioidaan estimaattien tarkkuutta sekä vaiheen 1 asiantuntija-arvioiden merkitystä menetelmillä, jotka eivät vaadi erillistä validointi- tai testidataa.

Validointi aloitetaan vertaamalla visuaalisesti prioriestimaatin $\tilde{\gamma} = (\tilde{w}, \tilde{\sigma})$ mukaista Bayes-verkkoa posterioriestimaatin $\hat{\gamma} = (\hat{w}, \hat{\sigma})$ mukaiseen Bayes-verkkoon. Vertailussa huomioidaan myös parametrien tarkkuuden prioriestimaatit $\tilde{\epsilon}$ ja posterioriestimaatit $\hat{\epsilon}$. Tämän jälkeen lasketaan posterioriestimaateille \hat{y} virhearvio yksi-pois -ristiinvalidoinnilla. Lopuksi tarkastellaan yksittäisiä posterioriestimaatteja y_i herkkyyksanalyysillä. Vaiheeseen 6 siirrytään, jos validointi ei anna aiheutta epäillä posterioriestimaattien käytettävyyttä.

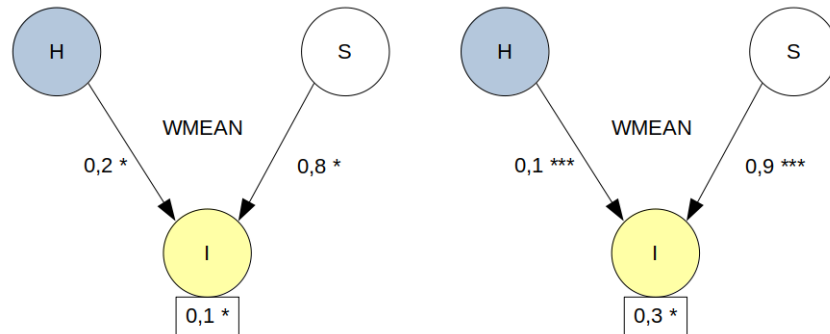
Visuaalinen tarkistus

Vaiheessa 1 määritettiin asiantuntija-arviona suunnattu sykliton verkko, RNM-funktiot r sekä RNM-parametrien $\gamma = (w, \sigma)$ prioriestimaatit $\tilde{\gamma}$. Näiden määrittämää Bayes-verkkoa voidaan verrata posterioriestimaatilla $\hat{\gamma}$ saatavaan verkkoon. Verkkoja tarkastellaan vierekkäin. Asiantuntija-arvioon perustuva verkko täydennetään prioritarkkuuksilla $\tilde{\epsilon}$. Posterioriestimaatteihin perustuva verkko täydennetään vastaavasti posterioritarkkuuksilla $\hat{\epsilon}$. Parametrien muutoksen järkevyyttä arvioidaan. Esimerkiksi prioriarviosta merkittävästi poikkeavat posterioriarviot voivat olla syy tulosten kyseenalaistamiselle. Visuaalisella tarkastelulla voidaan vertailla myös erilaisten asian-tuntija-arvioiden tuottamia posterioriestimaatteja toisiinsa.

Korkea hajontaparametrin posteriorestimaatti viittaa tätä vastaavan indikaattorin vähäiseen merkitykseen. Koska aineisto on skaalattu välille $[0, 1]$, korkeana hajontaparametrin arvona voidaan pitää suurempia arvoja kuin 0,5. Tätä korkeammat posterioriestimaatin arvot hajontaparametrille voivat kertoa virheestä alkuperäisessä arviossa tai aineiston puutteellisuudesta. Tällöin voidaan harkita vaiheeseen 1 palaamista indikaattorin poistamiseksi, indikaattorin merkityksen osoittavan lisäaineiston hankkimista tai alkuperäisen asiantuntija-arvion korjaamista varmemmaksi. Äärimmäisessä tapauksessa asiantuntija voi korjata oman arvionsa viisiportaisen tarkkuusasteikon korkeimmaksi arvoksi eli “täsmälliseksi”, jolloin sitä ei päivitetä aineiston avulla.

Kuvassa 14 on esimerkki vertailusta kolmen muuttujan tapauksessa. Päivitetyn verkon painoil-

le on tässä tapauksessa saatu erittäin tarkka, lähes asiantuntijan alkuperäistä näkemystä vastaava arvio. Indikaattorin hajontaparametri on sen sijaan päivittynyt asiantuntija-arviota korkeammaksi. Tämän arvion tarkkuutta aineiston huomiointi ei ole parantanut.



Kuva 14: Bayes-verkon päivitys visualisoituna. Vasemmalla lähestymistavan vaiheessa 1 asiantuntija-arvion perusteella luotu Bayes-verkko, jossa hypoteesi H ja vaihtoehtoinen selitys S vaikuttavat indikaattoriin I WMEAN-funktion mukaisesti. Funktion hajontaparametri on indikaattorin alla laatikossa. Arvioiden tarkkuudet on ilmoitettu tähdillä alaluvun 4.5 asteikkoa käyttäen. Oikealla sama verkko vaiheen 4 tuloksena saaduilla posterioriestimaateilla.

Yksi-pois -ristiinvalidointi

Lähestymistavan tuloksena saatavat posterioriestimaatit \hat{y} ovat ennusteita kohdemuuttujan tuntemattomille arvoille. Arvioiden onnistumista voidaan tarkastella mittaamalla posterioriestimaattien ja todellisten kohdemuuttujan arvojen välistä eroa. Tähän tarvitaan aineistoa jossa kohdemuuttujan arvot ovat tiedossa mutta jota ei ole käytetty parametrien estimointiin. Tämän työn lähestymistavassa estimointi tehdään koko aineistoon perustuen. Posterioriestimaattien tarkkuutta voidaan tällöin tarkastella yksi-pois -ristiinvalidoinnilla (engl. leave-one-out cross-validation, LOO-CV) [70, 71]. Menetelmällä tutkitaan posterioriestimaatin tarkkuutta poistamalla aineistosta yksi tunnettu kohdemuuttujan arvo, toistamalla estimointi ja vertaamalla tunnettua arvoa näin saatuun uuteen posterioriestimaattiin. Posterioriestimaatin odotetulle virheelle saadaan estimaatti havaittujen virheiden keskiarvona, kun sama toimenpide on toistettu jokaiselle aineiston tunnetulle kohdemuuttujan arvolle.

Virhettä voidaan mitata funktiolla $L(a, b) = |a - b|$. Alkuperäisestä aineistosta kohdemuuttujan i arvo poistamalla saatavaa muokattua aineistoa merkitään D^{-i} . Aineistoa D^{-i} käyttäen vaihe 4 toistamalla saatu posterioriestimaatti kohdemuuttujan arvolle i merkitään \hat{y}_i^{-i} . Yksi-pois -ristiinvalidoinnilla odotetulle virheelle saatava estimaatti on

$$V = \frac{1}{|T|} \sum_{i \in T} L(\hat{y}_i^{-i}, \hat{y}_i) = \frac{1}{|T|} \sum_{i \in T} |\hat{y}_i^{-i} - \hat{y}_i|,$$

missä T on alkuperäisen aineiston tunnettujen kohdemuuttujan arvojen indeksijoukko ja $|T|$

näiden indeksien lukumäärä [70]. Posterioriestimaattien \hat{y} hyödyllisyys tulee aina kyseenalaiseksi, jos virhe-estimaatti on suuri, mutta niitä ei hylätä vain tämän perusteella. Virhe-estimaatin V varianssi on suuri, sillä sen perustana olevat aineistot D^{-i} korreloivat vahvasti [70].

Virhe-estimaattia V voidaan käyttää posterioriestimaattien tarkkuuden arvioinnin lisäksi vaihtoehtoisten asiantuntija-arvioiden vertailemiseen. Vaihtoehtoinen asiantuntija-arvio voi esimerkiksi jättää mallista jonkin indikaattorin pois. Vertailu tehdään toistamalla vaiheet 1-5 vaihtoehtoista asiantuntija-arviota käyttäen. Tämän jälkeen lasketaan virhe-estimaatti vaihtoehtoisen arvion tapauksessa ja verrataan sitä alkuperäiseen virhe-estimaattiin. Pienemmän virhe-estimaatin antava vaihtoehto vaikuttaa yksi-pois -ristiinvalidoinnin perusteella paremmalta.

Herkkyysanalyysi

Tämän työn lähestymistavassa herkkyysanalyysillä tutkitaan kohdemuuttujan posterioriestimaattia \hat{y} . Tarkastelu kohdistetaan kerralla yhteen aikaväliin i ja tätä vastaavaan posterioriestimaattiin \hat{y}_i . Tutkittavaksi valitaan ainakin sellainen indeksi i_1 , jolla \hat{y}_{i_1} on korkea, ja sellainen indeksi i_2 , jolla \hat{y}_{i_2} on matala. Aikavälin i valinnan jälkeen muutetaan asiantuntija-arvioita priorijakaumalle $p(y)$, prioriestimaatteja $\tilde{\gamma}$ ja $\tilde{\epsilon}$ sekä saman aikavälin selittäviä muuttujia X_i . Posterioriestimaatti \hat{y}_i lasketaan jokaiselle yksittäiselle muutokselle vaiheet 1-4 toistamalla. Posterioriestimaatin vahva riippuvuus yksittäisestä tekijästä $p(y)$, $\tilde{\gamma}$, $\tilde{\epsilon}$ tai X voi tapauksesta riippuen olla perusteltua tai merkki tulosten heikosta yleistettävyydestä. Esimerkiksi indikaattorin tai selittävän muuttujan tarpeellisuutta tulee harkita, jos sen arvojen suurikaan vaihtelu ei vaikuta kohdemuuttujan posterioriestimaattiin.

4.7 Arvion käyttäminen

Tässä luvussa esiteltävän lähestymistavan viimeisessä vaiheessa tarkastellaan kohdemuuttujan posterioriestimaattien \hat{y} muodostamaa arviota. Arviota voidaan tarkastella esimerkiksi piirtämällä posterioriestimaatit kuvaajaan halutulla ajanjaksolla. Arviota voidaan käyttää myös automaattisesti esimerkiksi ihmisen huomion herättämiseen. Ihmisen huomio voidaan ohjata tarkastelemaan dataa tarkemmin esimerkiksi jos arvio ylittää ennalta määritetyn arvon. Lähestymistavan käyttäminen voidaan päättää arvion hyödyntämiseen. Lähestymistavan käyttämistä voidaan myös jatkaa keräämällä lisää dataa ja palaamalla vaiheeseen 2.

4.8 Yhteenveto

Tässä luvussa tutkittiin kysymystä “miten asiantuntija-arviot ja kerätty data yhdistetään ja hyödynnetään järjestysasteikollisten solmujen menetelmää käyttäen”. Perusmuotoisessa järjestysasteikollisten solmujen menetelmässä ei hyödynnetä dataa. Dataa voidaan kuitenkin hyödyntää järjestysasteikollisten solmujen menetelmää käyttävässä lähestymistavassa, jossa RNM-parametrit määritellään satunnaismuuttujina. RNM-parametreille määritetään tällöin priorijakauma asiantuntija-arvion perusteella. Datan ja asiantuntija-arvion yhdistävä posteriorijakauma lasketaan Bayes-päätelyä käyttäen. RNM-parametrien priorijakauma määritetään tässä luvussa esitellyssä lähestymistavassa elisitoimalla asiantuntijalta parametrien odotusarvojen lisäksi tarkkuusarvioita. Tämä lisää asiantuntijalta elisitoitavien parametrien määrää kahdella josta Bayes-verkon lapsisolmua kohti. Yhden tai useamman RNM-parametrin posteriorijakauma voidaan muodostaa myös pelkän datan perusteella, jolloin näille parametreille ei elisitoida asiantuntija-arvioita.

Lähestymistapaa voidaan käyttää joko hypoteesin tai välimatka-asteikollisen kohdemuuttujan arviointiin. Sotilaallista aktiivisuutta arvioitaessa hypoteesi voi olla esimerkiksi muotoa “alueella harjoittelee vähintään pataljoonan vahvuinen joukko”. Välimatka-asteikollisella muuttujalla voidaan tarkastella muodossa “kuinka suuri joukko alueella harjoittelee” olevia kysymyksiä. Kohdemuuttujan tyyppin valinnassa huomioidaan analyysin tavoite ja käytettävissä olevat lähteet. Välimatka-asteikollisen kohdemuuttujan tarkastelu ei ole hedelmällistä, jos informaatiota saadaan esimerkiksi sotilaallisen aktiivisuuden arvionnissa vain siitä, onko alueella ylipäätään sotilaallista toimintaa.

Lähestymistapa sallii järjestysasteikollisen ja välimatka-asteikollisen aineiston käytön. Aineiston on kuitenkin oltava tyypiltään yhtenäinen. Tarvittavat muunnokset asteikoiden välillä tehdään esikäsittelyvaiheessa. Asteikon valinta on hyödyllistä tehdä niin, että tarvittavien muunnosten määrä on mahdollisimman pieni. Alun perin järjestysasteikollinenkin aineisto on perusteltua käsitellä välimatka-asteikollisena, jos aineisto vaatii runsaasti esikäsittelyvaiheessa tehtäviä muunnoksia. Tämä johtuu siitä, että aineisto muunnetaan esikäsittelyvaiheessa ensin välimatka-asteikolliseksi muita muunnoksia varten.

Lähestymistapaa voidaan käyttää aikaisemmin tehdyn arvion päivittämiseen tai uusien arvioiden tekemiseen säännöllisesti kertyvän datan perusteella. Vaihe 1 toteutetaan tällöin vain kerran. Vaiheen 6 jälkeen saatu uusi data liitetään alkuperäiseen aineistoon. Vaiheet 2-6 toistetaan päivitettyllä aineistolla. Vaiheessa 5 tehtävä validointi voidaan jättää tekemättä, jos uusi data ei

ole muuttanut posterioriestimaatteja $\hat{\gamma}$ ja $\hat{\epsilon}$ merkittävästi. Näin voi olla esimerkiksi, jos uutta dataa on vähäinen määrä aineiston kokonaisuuteen verrattuna. Lähestymistapaa voidaan käyttää myös ymmärryksen kehittämiseen tarkasteltavasta ilmiöstä. Tällöin voidaan keskittyä esimerkiksi vaihtoehtoisten asiantuntija-arvioiden vertailuun vaiheessa 5 ja jättää vaihe 6 toteuttamatta.

5 LÄHESTYMISTAVAN KÄYTTÖKELPOISUUDEN TESTAAMINEN

Tässä luvussa esitetään julkinen tiivistelmä opinnäytetyön turvallisuusluokitellusta osiosta (TL IV). Turvallisuusluokittelun perusteena on laki viranomaisen toiminnan julkisuudesta (621/1999) 24 § 1. mom. 10k. Turvallisuusluokiteltu osio on opinnäytetyön liitteessä 4.

Julkinen tiivistelmä turvallisuusluokitellun osion sisällöstä

Tässä luvussa vastattiin alatutkimuskysymykseen 4: “miten asiantuntija-arvioihin ja dataan perustuvaa Bayes-verkkoa hyödyntävän lähestymistavan käyttökelpoisuus testataan sotilaallisen aktiivisuuden arvioinnissa”. Kysymystä tarkasteltiin käyttämällä luvussa 4 esiteltyä lähestymistapaa sotilaallisen aktiivisuuden arviointiin kahdessa esimerkkitilanteessa. Esimerkeissä käytetyt aineistot kerättiin avoimista lähteistä.

Luvussa 4 esiteltyä lähestymistapaa käytettiin tämän luvun ensimmäisessä esimerkissä harjoitustoiminnan arviointiin. Kohdemuuttuja oli tyypiltään hypoteesi ja sen todelliset arvot tunnettiin. Osa kohdemuuttujan tunnetuista arvoista poistettiin aineistosta testiä varten ja lähestymistapaa käyttäen saatuja tuloksia verrattiin näihin alun perin tunnettuihin kohdemuuttujan arvoihin. Testissä tarkasteltiin kohdemuuttujan arvioinnin lisäksi malliin valittujen indikaattorien ja vaihtoehtoisten selitysten merkitystä. Kohdemuuttujan posterioriestimaattien mukainen arvio vastasi alun perin tunnettuja kohdemuuttujan arvoja. Validointivaiheessa havaittiin eroja indikaattorien ja vaihtoehtoisten selitysten merkittävyyksissä.

Luvussa 4 esiteltyä lähestymistapaa käytettiin tämän luvun toisessa esimerkissä varuskunnan sotilaallisen aktiivisuuden arviointiin. Kohdemuuttuja oli välimatka-asteikollinen. Käytössä ollut aineisto ei sisältänyt kohdemuuttujan todellisia arvoja. Aineisto ei sisältänyt kaikkia selittävien muuttujien arvoja valitulla ajanjaksolla. Validointivaiheessa tehtiin havaintoja asiantuntija-arvion ja datan yhteneväisyydestä, indikaattorien merkittävyyksistä sekä kohdemuuttujan prioriestimaatin merkityksestä. Kohdemuuttujan posterioriestimaattien mukaista arviota verrattiin lopuksi aineiston ulkopuoliseen dataan ja asiantuntija-arvioon.

Tässä luvussa tarkasteltujen esimerkkien perusteella lähestymistapaa voidaan hyödyntää, vaikka aineistoa olisi käytössä vain vähän. Lähestymistapa mahdollistaa lähteiden merkityksen ja luotettavuuden arvioinnin. Lähestymistavan alussa tehtävä asiantuntija-arvioiden muodostaminen voi olla työlästä. Aikaa voidaan säästää tämän jälkeen, koska ihmisen ei tarvitse esimerkiksi lukea kaikkea aineistoa. Posterioriestimaattien muodostamaa arviota voidaan päivittää havaintojen määrän lisääntyessä palaamatta vaiheeseen 1. Vaiheen 5 tarkastelu voidaan jättää tällöin

suppeaksi, jos RNM-parametrien posterioriestimaatit eivät muutu merkittävästi. Pitkään jatkuvassa saman asiantuntija-arvion hyödyntämisessä validointi tulee toistaa säännöllisesti, koska asiantuntija-arvion oletukset voivat vanhentua. Lähestymistapaa voidaan käyttää esimerkiksi ihmisen huomion suuntaamiseen niihin ajankohtiin, jolloin arvio aktiivisuudesta muuttuu.

Yhteenvetona luvussa 4 esitellyllä lähestymistavalla todettiin olevan tiedustelusovelluksissa hyödyllisiä ominaisuuksia. Lähestymistavan käyttökelpoisuutta sotilaallisen aktiivisuuden arviointiin voidaan testata luvussa tarkasteltujen esimerkkien tapaan. Lähestymistavan testaaminen voi perustua kohdemuuttujan tunnettuihin arvoihin tai tulosten vertaamiseen asiantuntija-arvioon.

6 DISKUSSIO

6.1 Työn metodologinen kontribuutio

Työn pääkontribuutio on luvussa 4 esitelty uusi Bayes-päätelyä ja järjestysasteikollisten solmujen menetelmää hyödyntävä lähestymistapa asiantuntija-arvioiden ja datan yhdistämiseen. Aikaisemmassa kirjallisuudessa järjestysasteikollisten solmujen menetelmän parametrit on määriteltä tunnettuina arvoina, jotka on elisitoitu asiantuntijalta. Tämän työn lähestymistavassa nämä parametrit määritellään satunnaismuuttujina. Asiantuntijalta elisitoidaan parametrien odotusarvot ja tarkkuusarviot, joiden perusteella muodostetaan parametrien priorijakaumat. Priorijakaumat päivitetään Bayes-päätelyä käyttäen datan huomioivaksi posteriorijakaumaksi. Posteriorijakaumasta muodostetaan otos Gibbsin otannalla ja tästä otoksesta lasketaan parametrien posterioriestimaatit. Lähestymistapa mahdollistaa järjestysasteikollisten solmujen menetelmän parametrien estimoinnin osittain tai kokonaan aineiston perusteella.

Järjestysasteikollisten solmujen menetelmää laajennetaan tässä työssä parametrien satunnaismuuttujatulkinnan lisäksi välimatka-asteikollisella muunnoksella ja satunnaislukuihin perustuvalla muunnoksella. Välimatka-asteikollinen muunnos mahdollistaa ehdollisten todennäköisyysjakaumien muodostamisen välimatka-asteikollista dataa käytettäessä. Aikaisemmissa tutkimuksissa välimatka-asteikollinen data on muunnettu järjestysasteikolliseksi ennen ehdollisten todennäköisyysjakaumien muodostamista. Välimatka-asteikollisen muunnoksen osoitettiin voivan tuottaa välimatka-asteikollisella datalla tarkempia tuloksia aikaisemmin käytettyyn datan mitta-asteikon muunnokseen verrattuna. Satunnaislukuihin perustuva muunnos on vaihtoehto ehdollisten todennäköisyyksien muodostamiseen järjestysasteikollisella datalla. Kun tavanomaisessa järjestysasteikollisten solmujen menetelmässä ehdollisten todennäköisyyksien laskennassa kohdattava integraali approksimoidaan hilalla, satunnaislukuapproksimaatioissa hila korvataan satunnaisesti valittavilla pisteillä. Pisteet voidaan valita joko pseudo- tai kvasisatunnaisina lukuina. Satunnaislukumuunnos osoitettiin joissakin tilanteissa tavanomaista ehdollisten todennäköisyyksien laskentatapaa tarkemmaksi.

Kohdemuuttujan ja parametrien estimointi tehdään tämän työn lähestymistavassa samanaikaisesti. Tämä eroaa ohjatusta koneoppimisesta, jossa parametrit estimoidaan ensin opetusdataan perustuen. Ohjattu koneoppiminen vaatii tyypillisesti laajaa ja puutteetonta opetusdataa. Kaiken datan hyödyntäminen sen puutteista huolimatta on tämän työn lähestymistavan vahvuus. Toisaalta tämän työn lähestymistapaa voidaan pitää ohjattua koneoppimista monimutkaisem-

pana, sillä estimoitu malli riippuu satunnaismuuttujista. Lisäksi tämän työn lähestymistavassa estimoitu ja validoitu malli ei ole lopullinen. Aineistoa lisättäessä estimointi ja validointi toutetaan uudestaan. Uusi data voi vaikuttaa edellisiin estimaatteihin ja kerran validoidut parametrit voivat muuttua dataa lisättäessä. Tämä on vahvuus, sillä ajan kuluessa mallienkin on muututtava. Tämän työn lähestymistavan mallia voidaan yksinkertaistaa muuttamalla parametrien tarkkuusarvot arvoon “varma”. Tällöin parametreja ei tulkita satunnaismuuttujina. Malli ei tässä tapauksessa muutu uutta aineistoa lisättäessä.

Tämän työn lähestymistavan vaiheessa 5 sallitaan asiantuntijan alkuperäisen arvion muokkaus. Tämä voi olla hyödyllistä esimerkiksi toisistaan eroavien asiantuntija-arvioiden vertaamiseksi sekä väärinkäsitysten korjaamiseksi, mutta siihen liittyy vaara. Aineiston näkemisen jälkeen asiantuntijalla voi olla kiusaus muokata alkuperäistä arviotaan havaintojen suuntaiseksi. Näin ei tule toimia, koska lähestymistavan tuottaman arvion on tarkoitus kuvata alkuperäisen arvion ja datan yhdistämisen tulos. Alkuperäisen arvion muokkaaminen aineiston suuntaiseksi vähentää ihmisen tekemän arvion tuomaa lisäarvoa.

6.2 Työn pragmaattinen kontribuutio

Työssä selvitettiin julkisten lähteiden perusteella kokemuksia koneoppimisen soveltamisesta tiedusteluanalyysiin. Merkittävimmiksi haasteiksi tunnistettiin datan saatavuus ja luotettavuus, ajankäytölliset haasteet sekä vaikeudet ymmärtää kokeiltujen koneoppimismenetelmien toimintaa. Nämä haasteet tulee huomioida harkittaessa koneoppimismenetelmien hyödyntämistä tiedusteluanalyysissä.

Luvussa 4 esiteltyä lähestymistapaa testattiin sotilaallisen aktiivisuuden arviointiin luvussa 5 kahdella aineistolla. Lähestymistavalla todettiin voitavan tuottaa tarkkoja arvioita sotilaallisesta aktiivisuudesta. Lähestymistavan käyttäminen voi myös parantaa ymmärrystä tarkasteltavasta ilmiöstä, sillä se edellyttää ilmiön riippuvuussuhteiden tarkan määrittelyn ja mahdollistaa saatujen tulosten validoinnin. Validointivaiheessa käytettävä yksi-pois -ristiinvalidointi tuottaa virheestimaatin, jonka perusteella voidaan vertailla vaihtoehtoisia asiantuntija-arvioita. Posterioriestimaattien visualisointi ja herkkyysanalyysi lisäävät lähestymistavan ymmärrettävyyttä ja luotettavuutta. Posterioriestimaattien muodostamiseen kului tarkastelluissa esimerkeissä toimistokäyttöön tarkoitettulla tietokoneella vähintään kymmenen sekuntia ja enintään kaksi minuuttia. Laskenta tehtiin yhdellä prosessorilla. Suurin osa ajasta kului Gibbsin otannan toteuttamiseen. Yksi-pois -ristiinvalidoinnissa estimointi toistetaan jokaiselle kohdemuuttujan tunnetulle arvol-

le, mikä voi johtaa pitkään laskenta-aikaan.

Luvussa 4 esitellyn lähestymistavan ensimmäinen vaihe voi olla työläs, koska siinä on muodostettava tarkasteltavaa ilmiötä kuvaavia asiantuntija-arvioita. Näiden arvioiden muodostaminen vaatii Bayes-päätelyn ja järjestysasteikollisten solmujen menetelmän perusteiden tuntemisen. Lähestymistavan hyödyntäminen voi silti säästää aikaa, koska aineisto käsitellään koneellisesti. Tämä mahdollistaa myös kirjallisten ja ei-kirjallisten lähteiden samanaikaisen hyödyntämisen. Lähestymistapa tukee lähteiden merkityksen ja luotettavuuden arviointia. Lähteiden merkitystä voidaan tarkastella painokertoimilla ja hajontaparametreilla sekä luotettavuutta tarkkuusparametreilla. Lähestymistapaa voidaan käyttää esimerkiksi ihmisen huomion suuntaamiseen niihin ajankohtiin, jolloin arvio aktiivisuudesta muuttuu. Lähestymistavan validointivaiheen avulla voidaan myös tunnistaa arvion muutoksen kannalta merkittävät havainnot ja suunnata ihmisen huomio näihin.

6.3 Työn luotettavuuden arviointi

Työssä esitelty uusi lähestymistapa perustuu todennäköisyyspäättelyä ja järjestysasteikollisten solmujen menetelmää käsittelevään kirjallisuuteen. Tärkeimmät käytetyt lähteet ovat oppikirjat Gelman ym.: “Bayesian Data Analysis” [60], Neapolitan: “Learning Bayesian Networks” [36] ja Robert ym.: “Monte Carlo statistical methods” [61]. Teokset ovat 10–20 vuotta vanhoja, mutta tämä ei vaikuta matemaattisten oppikirjojen sisällön luotettavuuteen. Järjestysasteikollisten solmujen menetelmää tarkasteltiin alkuperäisen artikkelin [30] lisäksi viimeisimpiin vertaisarvioituihin tutkimuksiin [31–34] perustuen.

Järjestysasteikollisten solmujen menetelmään alaluvussa 3.3 esitetyt kaksi muunnosta perustuvat matemaattiseen analyysiin. Satunnaislukuihin perustuvan muunnoksen laskennallisen vaativuuden analyysi perustuu kirjaan Press: “Numerical Recipes: The Art of Scientific Computing” [37]. Muunnoksia testattiin lisäksi keinotekoisilla aineistoilla. Testien asetelmat ja tulokset on esitetty liitteissä 1 ja 2.

Luvussa 2 tehty tarkastelu koneoppimisen tiedusteluanalyysiin soveltamisen kokemuksista perustuu pääosin kolmeen teokseen. Nämä ovat Spielmann: “The Logic of Intelligence Analysis” [41], Stenslie: “Intelligence Analysis in the Digital Age” [42] ja Waltz: “Qualitative Intelligence Analysis: Applied analytic models, simulations and games” [43]. Teokset ovat enintään 10 vuotta vanhoja ja siten edelleen ajankohtaisia. Luvun 2 tarkastelu perustui useissa lähteissä toistuneisiin teemoihin. Näitä havaintoja voidaan siten pitää luotettavina.

Tässä työssä esitellyn lähestymistavan testaamista sotilaallisen aktiivisuuden arviointiin tarkasteltiin luvussa 5. Tarkastelussa käytettiin avoimista lähteistä kerättyjä aineistoja. Luvussa 5 tarkastellaan arvioiden tarkkuuden sijaan havaintoja lähestymistavan käytettävyydestä sekä mahdollisuuksia lähestymistavan testaamiseen sotilaallisen aktiivisuuden arvioinnissa. Aineistot ovat riittäviä tähän tarkoitukseen. Lähestymistavan systemaattinen arviointi todellisissa sovelluksissa ei kuulu tämän tutkimuksen tavoitteisiin.

6.4 Jatkotutkimusaiheita

Tässä työssä esiteltyä lähestymistapaa voidaan kehittää edelleen tutkimalla Bayes-verkon rakenteen päättelystä sekä mallintamalla pitkällä aikavälillä esiintyviä riippuvuuksia. Bayes-verkon rakenteen optimointia voidaan tutkia esimerkiksi alaluvussa 4.6 esiteltyyn virhe-estimaattiin perustuen. Pitkällä aikavälillä esiintyvien riippuvuuksien mallintamiseksi voidaan tutkia esimerkiksi uusia liukuvan keskiarvon painotuksia datan esikäsittelyvaiheeseen. Lähestymistapaa voidaan laajentaa tarkastelemalla syvempiä riippuvuuksia. Tämä tarkoittaa vaihtoehdoisen selityksen esittämistä pelkän isäsolmun sijaan myös lapsisolmuna.

Lähestymistavan hyödyntämistä tulee arvioida systemaattisesti käyttäjäkokemusten perusteella. Tätä varten olisi luotava helppokäyttöinen käyttöliittymä. Käyttäjä voisi piirtää sen avulla suunnatun syklittömän verkon ja valita sen lapsisolmuille RNM-funktiot. Käyttöliittymä voisi esittää käyttäjälle kysymyksiä parametrien ja tarkkuusarvioiden elisitoimiseksi. Käyttöliittymän tulisi mahdollistaa erilaisten aineistojen käyttäminen sekä aineiston päivittäminen uusilla havainnoilla. Käyttöliittymä voisi lisäksi helpottaa validointia esimerkiksi laskemalla virhe-estimaatin automaattisesti.

Estimoinnin laskennallinen vaativuus Bayes-verkon rakenteen suhteen tulisi selvittää. Korkea laskennallinen vaativuus voi hankaloittaa lähestymistavan soveltamista, jos laskenta-aika voi venyä käytännössä esimerkiksi päivien mittaiseksi. Tähän liittyen tulisi tutkia laskennan tehostamista. Gibbsin otanta on lähestymistavan laskennallisesti vaativin osuus. Tämän jakaminen useille prosessoreille on teoriassa mahdollista.

7 JOHTOPÄÄTÖKSET

Tämän tutkimuksen tavoitteena oli selvittää, miten sotilaallisen aktiivisuuden arviointia voidaan tukea dataan ja asiantuntija-arvioihin perustuvalla Bayes-verkolla. Tätä varten työssä selvitettiin, miten koneoppimismenetelmien soveltuvuutta tiedusteluanalyysiin on aikaisemmin arvioitu ja miten sotilaallisen aktiivisuuden arviointia voidaan tukea Bayes-päätelyyn perustuvilla menetelmillä. Työssä selvitettiin myös, miten data ja asiantuntija-arviot yhdistetään ja hyödynnetään järjestysasteikollisten solmujen menetelmään perustuvassa Bayes-verkossa. Lopuksi työssä selvitettiin, miten asiantuntija-arvioihin ja dataan perustuvaa Bayes-verkkoa hyödyntävän lähestymistavan käyttökelpoisuus testataan sotilaallisen aktiivisuuden arvioinnissa. Tutkimuksen päämenetelmä oli mallintaminen.

Tutkimuksen päätuloksena esiteltiin uusi dataa ja asiantuntija-arvioita yhdistävä lähestymistapa. Lähestymistapaa voidaan käyttää tuntemattoman kohdemuuttujan arviointiin, kun saatavilla on asiantuntija-arvioita ja dataa. Tuntematon kohdemuuttuja voi olla tyypiltään hypoteesi tai välimatka-asteikollinen luku. Lähestymistapa perustuu järjestysasteikollisten solmujen menetelmään. Menetelmän parametrit käsitellään olemassa olevaan kirjallisuuteen verrattuna uudella tavalla satunnaismuuttujina. Parametrien käsittely satunnaismuuttujina mahdollistaa niitä koskevien asiantuntija-arvioiden päivittämisen datan perusteella. Parametria koskeva asiantuntija-arvio voidaan myös jättää antamatta, jolloin parametri estimoidaan vain datan perusteella.

Tutkimuksessa esitettiin lisäksi kaksi uutta muunnosta järjestysasteikollisten solmujen menetelmään. Näistä ensimmäinen mahdollistaa ehdollisten todennäköisyysjakaumien muodostamisen suurilla isäsolmujen määrillä aikaisemmin kirjallisuudessa käytettyä tapaa tarkemmin. Toinen muunnos mahdollistaa välimatka-asteikollisen datan käyttämisen järjestysasteikollisten solmujen menetelmässä muuntamatta dataa järjestysasteikolliseksi. Tämän muunnoksen osoitettiin parantavan menetelmän tarkkuutta alun perin välimatka-asteikollista dataa käytettäessä.

Lähestymistapaa sovellettiin tässä työssä tiedusteluanalyysiin. Testit osoittavat lähestymistavan voivan tuottaa tarkkoja arvioita sotilaallisesta aktiivisuudesta. Arvioiden tuottamisen lisäksi lähestymistapa soveltuu ymmärryksen lisäämiseen tarkasteltavasta ilmiöstä. Ymmärrystä lisäävät ilmiön riippuvuussuhteiden tarkka määrittäminen ja tulosten validointi. Validointivaiheessa voidaan esimerkiksi vertailla erilaisten asiantuntija-arvioiden yhteensopivuutta datan kanssa yksi-pois -ristiinvalidoinnin tuottamien virhe-estimaattien avulla.

Tässä työssä esitellyn lähestymistavan ensimmäinen vaihe voi olla työläs. Tämä johtuu tarpeesta

määrittää asiantuntija-arvio etukäteen. Asiantuntija-arvion määrittämisen jälkeen tässä lähestymistapa mahdollistaa aineiston käsittelyn koneellisesti. Lähestymistapaa käyttäen voidaan hyödyntää samanaikaisesti kirjallisia ja ei-kirjallisia lähteitä. Lähteiden merkitystä voidaan arvioida painokertoimien ja hajontaparametrien perusteella sekä niiden luotettavuutta tarkkuusparametrien perusteella. Lähestymistavan käytön aloittamisen työläyden vuoksi se soveltuu parhaiten yksittäisen kohteen pitkäaikaiseen arviointiin. Lähestymistavan tuottamaa arviota voidaan käyttää ihmisen huomion ohjaamiseen merkitykselliseen dataan.

Lähestymistapa vastaa useisiin haasteisiin, joita on kohdattu sovellettaessa koneoppimista tiedusteluanalyysiin. Sen käyttäminen ei vaadi suuria tietomassoja. Kaikki saatavilla oleva data voidaan hyödyntää, koska erillisiä testi- tai validointiaineistoja ei muodosteta. Puuttuvat havainnot eivät estä muun datan käyttöä. Harvoin havaittavia ja siten pelkän aineiston perusteella vaikeasti arvioitavia indikaattoreita voidaan huomioida asiantuntija-arviolla. Lähestymistavan tuottamaa visuaalista Bayes-verkkoa on lisäksi helppo ymmärtää, koska parametrien muutosta voidaan verrata alkuperäiseen asiantuntija-arviioon. Parametrien estimaatteja voidaan analysoida ja tarvittaessa korjata.

Koneoppimisen hyödyntäminen päätöksenteon tukemisessa tulee yleistymään tulevaisuudessa. Sotilassovelluksissa tarvitaan usein perusteltuja analyyskejä epävarmoissa tilanteissa. Tällöin korostuu koneoppimismenetelmän kyky hyödyntää suurten datamassojen sijaan puutteellista dataa. Samoin korostuvat ihmisen mahdollisuudet tarkistaa ja korjata koneoppimismenetelmän toimintaa. Tämä tutkimus vastaa näihin haasteisiin tarjoamalla uuden lähestymistavan, joka mahdollistaa ihmisen ja koneoppimisen välisen yhteistyön.

LÄHTEET

- [1] *Sotilastiedustelu: julkinen katsaus 2021*. Pääesikunta. url: puolustusvoimat.fi/documents/1948673/74055459/PV_sotilastiedustelu_raportti_www_FI_2021.pdf (viitattu 01.03.2023).
- [2] M. Lohse, "The intelligence process in Finland," *Scandinavian journal of military studies*, 2020. DOI: 10.31374/sjms.55.
- [3] R. Hu, J. Granderson, D. Auslander ja A. Agogino, "Design of machine learning models with domain experts for automated sensor selection for energy fault detection," *Applied energy*, vol. 235, s. 117–128, 2019. DOI: <https://doi.org/10.1016/j.apenergy.2018.10.107>.
- [4] A. Silva ja M. Gombolay, "Encoding human domain knowledge to warm start reinforcement learning," *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, nro 6, s. 5042–5050, 2021. DOI: 10.1609/aaai.v35i6.16638.
- [5] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang ja S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Medical image analysis*, vol. 69, s. 101985, 2021, ISSN: 1361-8415. DOI: <https://doi.org/10.1016/j.media.2021.101985>.
- [6] X. Tang, A. Chen ja J. He, "A modelling approach based on Bayesian networks for dam risk analysis: integration of machine learning algorithm and domain knowledge," *International journal of disaster risk reduction*, vol. 71, s. 102818, 2022. DOI: 10.1016/j.ijdr.2022.102818.
- [7] W. Song et al., "Incorporating domain knowledge into machine learning for laser-induced breakdown spectroscopy quantification," *Spectrochimica acta part B: atomic spectroscopy*, vol. 195, s. 106490, syyskuu 2022. DOI: 10.1016/j.sab.2022.106490.
- [8] J. Pearl, "Bayesian networks: a model of self-activated memory for evidential reasoning," teoksessa *Proc. of cognitive science society (CSS-7)*, 1985.
- [9] H. Laine, Kuulustelua, haastattelua vai keskustelua? Käsiteanalyysi elisitaatiosta henkilötiedustelun tiedonhankinnan viitekehyksessä, Pro gradu -tutkielma, Maanpuolustuskorkeakoulu. 2018.
- [10] J. J. Galán, R. A. Carrasco ja A. LaTorre, "Military applications of machine learning: a bibliometric perspective," *Mathematics*, vol. 10, nro 9, 2022. DOI: 10.3390/math10091397.

- [11] P. Hemminki, New brains for the defence system, Pro gradu -tutkielma, Turun yliopisto. 2020.
- [12] O. Zoder, Automated collection of open source intelligence, Pro gradu -tutkielma, Masaryk University. 2020.
- [13] L. A. Morales, Data mayhem versus nimble information: transforming hectic imagery intelligence data into actionable information using artificial neural networks, opinnäytetyö. Air Command and Staff College, Air University. 2017.
- [14] A. K. Bourret, J. D. Wines ja J. M. Mendes, Assessing sentiment in conflict zones through social media, opinnäytetyö, Naval postgraduate school. 2016.
- [15] X. Zeng, A. S. Abumansour ja A. Zubiaga, ”Automated fact-checking: a survey,” *Language and linguistics compass*, vol. 15, nro 10, e12438, 2021. DOI: 10.1111/lnc3.12438.
- [16] S. Agarwal, A. Sureka ja V. Goyal, ”Open source social media analytics for intelligence and security informatics applications,” teoksessa *Big data analytics*, N. Kumar ja V. Bhatnagar, toim., Cham: Springer international publishing, 2015, s. 21–37, ISBN: 978-3-319-27057-9.
- [17] *Bayes’ theorem in the Korean war. CIA intelligence report no. 0605/68.* url: www.cia.gov/readingroom/docs/DOC\0001205738.pdf (viitattu 22.08.2022).
- [18] C. W. Karvetski, D. R. Mandel ja D. Irwin, ”Improving probability judgment in intelligence analysis: from structured analysis to statistical aggregation,” *Risk analysis*, vol. 40, nro 5, s. 1040–1057, 2020. DOI: 10.1111/risa.13443.
- [19] R. J. Heuer, *The psychology of intelligence analysis*. Washington, DC: CQ Press, 1999.
- [20] D. R. Mandel, ”Instruction in information structuring improves Bayesian judgment in intelligence analysts,” *Frontiers in psychology*, vol. 6, 2015. DOI: 10.3389/fpsyg.2015.00387.
- [21] D. R. Mandel, C. W. Karvetski ja M. K. Dhami, ”Boosting intelligence analysts’ judgment accuracy: what works, what fails?” *Judgment and decision making*, vol. 13, s. 607–621, 2018.
- [22] M. Valtorta, J. Dang, H. Goradia, J. Huang ja M. Huhns, ”Extending Heuer’s analysis of competing hypotheses method to support complex decision analysis,” teoksessa *Proceedings of the 2005 international conference on intelligence analysis (IA-05)*, 2005.

- [23] C. W. Karvetski, K. C. Olson, D. T. Gantz ja G. A. Cross, "Structuring and analyzing competing hypotheses with Bayesian networks for intelligence analysis," *EURO journal on decision processes*, vol. 1, nro 3, s. 205–231, 2013. DOI: 10.1007/s40070-013-0001-x.
- [24] L. Kaikkonen, T. Parviainen, M. Rahikainen, L. Uusitalo ja A. Lehtikainen, "Bayesian networks in environmental risk assessment: a review," *Integrated environmental assessment and management*, vol. 17, nro 1, s. 62–78, 2021. DOI: 10.1002/ieam.4332.
- [25] M. Tsagris, "A new scalable bayesian network learning algorithm with applications to economics," *Computational economics*, vol. 57, nro 1, s. 341–367, 2021. DOI: 10.1007/s10614-020-10065-.
- [26] L. van der Stap et al., "The feasibility of a Bayesian network model to assess the probability of simultaneous symptoms in patients with advanced cancer," *Scientific reports*, vol. 12, nro 1, s. 22 295, 2022, ISSN: 2045-2322. DOI: 10.1038/s41598-022-26342-4.
- [27] P. Fuster-Parra, A. García-Mas, F. Ponseti ja F. Leo, "Team performance and collective efficacy in the dynamic psychology of competitive team: a Bayesian network analysis," *Human Movement Science*, vol. 40, s. 98–118, 2015, ISSN: 0167-9457. DOI: 10.1016/j.humov.2014.12.005.
- [28] C. Eldridge, C. Hobbs ja M. Morán, "Fusing algorithms and analysts: open-source intelligence in the age of 'big data'," *Intelligence and national security*, vol. 33, nro 3, s. 391–406, 2018. DOI: 10.1080/02684527.2017.1406677.
- [29] M. J. Barons, S. Mascaro ja A. M. Hanea, "Balancing the elicitation burden and the richness of expert input when quantifying discrete bayesian networks," *Risk analysis*, vol. 42, nro 6, s. 1196–1234, 2022. DOI: 10.1111/risa.13772.
- [30] N. E. Fenton, M. Neil ja J. G. Caballero, "Using ranked nodes to model qualitative judgments in Bayesian networks," *IEEE transactions on knowledge and data engineering*, vol. 19, nro 10, 1420–1432, 2007. DOI: 10.1109/TKDE.2007.1073.
- [31] P. Laitila ja K. Virtanen, "Improving construction of conditional probability tables for ranked nodes in Bayesian networks," *IEEE transactions on knowledge and data engineering*, vol. 28, nro 7, s. 1691–1705, 2016. DOI: 10.1109/TKDE.2016.2535229.
- [32] P. Laitila ja K. Virtanen, "On theoretical principle and practical applicability of ranked nodes method for constructing conditional probability tables of Bayesian networks," *IEEE transactions on systems, man, and cybernetics: systems*, vol. 50, nro 5, s. 1943–1955, 2020. DOI: 10.1109/TSMC.2018.2792058.

- [33] P. Laitila ja K. Virtanen, ”Advancing construction of conditional probability tables of Bayesian networks with ranked nodes method,” *International journal of general systems*, vol. 51, nro 8, s. 758–790, 2022. DOI: 10.1080/03081079.2022.2086541.
- [34] P. Laitila ja K. Virtanen, ”Portraying probabilistic relationships of continuous nodes in Bayesian networks with ranked nodes method,” *Decision support systems*, vol. 154, s. 113 709, 2022, ISSN: 0167-9236. DOI: 10.1016/j.dss.2021.113709.
- [35] P. Laitila ja K. Virtanen, Technical Properties of Ranked Nodes Method, arXiv:2107.12747 [stat.ME]. 2021.
- [36] R. Neapolitan, *Learning Bayesian Networks*. Prentice-Hall, 2003, ISBN: 978-0130125347.
- [37] W. H. Press, S. A. Teukolsky, W. T. Vetterling ja B. P. Flannery, *Numerical recipes: the art of scientific computing*, 3. painos. Cambridge University Press, 2007, ISBN: 0521880688.
- [38] *The data science Venn diagram*. url: drewconway.com/zia/2013/3/26/the-data-science-venn-diagram (viitattu 13. 03. 2023).
- [39] *The essential data science Venn diagram*. url: towardsdatascience.com/the-essential-data-science-venn-diagram-35800c3bef40 (viitattu 13. 03. 2023).
- [40] *EBSCO government & military collection*. url: www.ebsco.com/products/research-databases/military-government-collection.
- [41] K. Spielmann, *The logic of intelligence analysis: why hypothesis testing matters*. Routledge, 2018, ISBN: 9781138601963.
- [42] S. Stenslie, L. Haugom ja B. H. Vaage, *Intelligence analysis in the digital age*. Routledge, 2021.
- [43] E. Waltz, *Quantitative intelligence analysis: applied analytic models, simulations and games*. Rowman & Littlefield publishers, 2014, ISBN: 9781442235861.
- [44] *Google Scholar*. url: scholar.google.fi.
- [45] Y. Zhang, Z. Dai, L. Zhang, Z. Wang, L. Chen ja Y. Zhou, ”Application of artificial intelligence in military: from projects view,” teoksessa *2020 6th international conference on big data and information analytics*, 2020, s. 113–116. DOI: 10.1109/BigDIA51454.2020.00026.

- [46] P. Svenmarck, L. Luotsinen, M. Nilsson ja J. Schubert, ”Possibilities and Challenges for Artificial Intelligence in Military Applications,” teoksessa *Proceedings of the NATO big data and artificial intelligence for military decision making specialists’ meeting*, 2018. url: www.sto.nato.int/publications/STOMeetingProceedings/STO-MP-IST-160/MP-IST-160-S1-5.pdf (viitattu 03.08.2022).
- [47] K. Maharana, S. Mondal ja B. Nemade, ”A review: data pre-processing and data augmentation techniques,” *Global transitions proceedings*, vol. 3, nro 1, s. 91–99, 2022, International conference on intelligent engineering approach (ICIEA-2022). DOI: <https://doi.org/10.1016/j.gltip.2022.04.020>.
- [48] F. Zhuang et al., ”A comprehensive survey on transfer learning,” *Proceedings of the IEEE*, vol. 109, nro 1, s. 43–76, 2021. DOI: 10.1109/JPROC.2020.3004555.
- [49] C. Molnar, *Interpretable machine learning, A guide for making black box models explainable*, 2. painos. 2022. url: christophm.github.io/interpretable-ml-book (viitattu 13.03.2023).
- [50] R. R. Hoffman, S. T. Mueller, G. Klein ja J. Litman, *Metrics for explainable AI: challenges and prospects*, arXiv:abs/1812.04608. 2018.
- [51] D. Gunning, E. Vorm, J. Y. Wang ja M. Turek, ”DARPA’s explainable AI (XAI) program: a retrospective,” *Applied AI Letters*, vol. 2, nro 4, e61, 2021. DOI: 10.1002/ail2.61.
- [52] *State of data science 2020*. url: know.anaconda.com/rs/387-XNW-688/images/Anaconda-SODS-Report-2020-Final.pdf (viitattu 05.08.2022).
- [53] L. Saenen, M. Heyvaert, W. V. Dooren, W. Schaeken ja P. Onghena, ”Why humans fail in solving the Monty Hall dilemma: a systematic review,” *Psychologica belgica*, vol. 58, s. 128–158, 2018.
- [54] J. Bernardo ja A. Smith, *Bayesian theory*. Wiley, 2000, ISBN: 9780470028735.
- [55] *Handbook of Bayesian analysis for intelligence*. 1975. url: www.cia.gov/readingroom/docs/CIA-RDP86B00269R001100080001-0.pdf (viitattu 23.08.2022).
- [56] *Bayes’ theorem for intelligence analysis*, 1970. url: www.cia.gov/static/c9503a45e36af6458699549b8daf5a77/Bayes-Theorem-for-Analysis.pdf (viitattu 23.08.2022).
- [57] G. F. Cooper, ”The computational complexity of probabilistic inference using bayesian belief networks,” *Artificial intelligence*, vol. 42, nro 2, s. 393–405, 1990, ISSN: 0004-3702. DOI: 10.1016/0004-3702(90)90060-D.

- [58] I. Sobol, "On the distribution of points in a cube and the approximate evaluation of integrals," *USSR computational mathematics and mathematical physics*, vol. 7, nro 4, s. 86–112, 1967. DOI: 10.1016/0041-5553(67)90144-9.
- [59] A. Salmerón, R. Rumí, H. Langseth, T. D. Nielsen ja A. L. Madsen, "A review of inference algorithms for hybrid Bayesian networks," *Journal of artificial intelligence research*, vol. 62, nro 1, 799–828, 2018. DOI: 10.1613/jair.1.11228.
- [60] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari ja D. Rubin, *Bayesian data analysis*, 3. painos. Taylor & Francis, 2013, ISBN: 9781439840955.
- [61] C. Robert ja G. Casella, *Monte Carlo statistical methods*, 2. painos. Springer-Verlag, 2004.
- [62] M. Plummer, "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," teoksessa *3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria*, vol. 124, 2003.
- [63] S. Sturtz, U. Ligges ja A. Gelman, "R2WinBUGS: A package for running WinBUGS from R," *Journal of statistical software*, vol. 12, nro 3, s. 1–16, 2005.
- [64] A. Thomas, "BUGS: a statistical modelling package," *RTA/BCS Modular Languages Newsletter*, vol. 2, s. 36–38, 1994.
- [65] D. Fink, *A compendium of conjugate priors*, 1997. url: www.johndcook.com/Compendium\~OfConjugatePriors.pdf (viitattu 28.02.2023).
- [66] P. Damlén, J. Wakefield ja S. Walker, "Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables," *Journal of the royal statistical society series B: statistical methodology*, vol. 61, nro 2, s. 331–344, 2002. DOI: 10.1111/1467-9868.00179.
- [67] D. Spiegelhalter, A. Thomas, N. Best ja D. Lunn, *WinBUGS user manual. Version 1.4*. www.mrc-bsu.cam.ac.uk/wp-content/uploads/manual14.pdf. (viitattu 28.02.2023).
- [68] O. Chapelle, B. Schölkopf ja A. Zien, toim., *Semi-supervised learning*. The MIT Press, 2006, ISBN: 9780262033589.
- [69] *AgenaRisk Software, Version 10.0*. url: bayesianrisk.com/ (viitattu 22.02.2023).
- [70] T. Hastie, R. Tibshirani ja J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2. painos. Springer, 2009, ISBN: 9780387848587.

- [71] P. Koistinen, *Tilastotieteen ja todennäköisyyslaskennan englantilais–suomalainen sanasto*. url: www.math.helsinki.fi/petrin/sanasto/tilastosanasto.html (viitattu 16.03.2023).
- [72] W. J. Conover, *Practical nonparametric statistics*, 3. painos. Wiley, 1999, ISBN: 978-0471160687.
- [73] B. Efron ja T. Hastie, *Computer age statistical inference: algorithms, evidence, and data science*. Cambridge university press, 2016. DOI: 10.1017/CBO9781316576533.

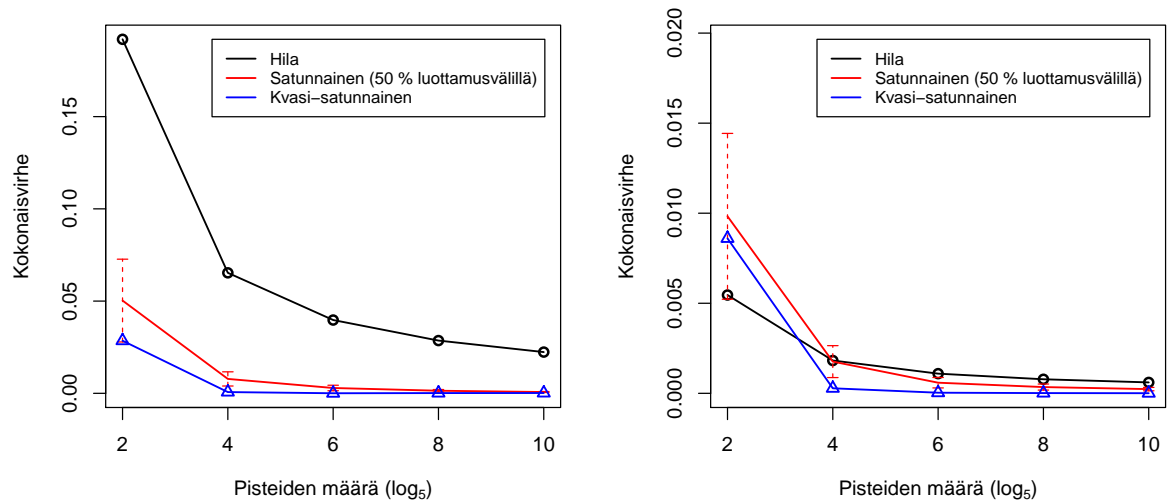
Liite 1: Satunnaislukuihin perustuvien muunnosten testaus

Perusmuotoisessa RNM-menetelmässä ehdolliset todennäköisyydet lasketaan kaavalla (3.5). Alaluvussa 3.3 esitetään tälle vaihtoehtoiset muunnokset käyttäen tavallisia satunnaislukuja sekä kvasi-satunnaislukuja. Näitä kolmea tapaa valita approksimaation perustana olevat pisteet testattiin RNM-mallilla, jossa viisi isäsolmua vaikutti yhteen ja samaan lapsisolmuun WMEAN-funktion mukaisesti. Kullakin solmulla oli viisi mahdollista tilaa. Mittarina jakauman tarkkuudelle käytettiin kokonaisvirhettä $V = \sum_{i=1}^5 |p_i - t_i|$, missä p_i on tuloksena saadun jakauman ja t_i todellisen jakauman antama todennäköisyys tilalle $i \in \{1, \dots, 5\}$. Testi toistettiin kahdesti muuttaen parametrien (x, w, σ) arvoja. Ensin luotiin parametreilla $x = (1, 1, 5, 1, 1)$, $w = \frac{1}{9}(1, 1, 5, 1, 1)$ ja $\sigma = 0.05$ voimakkaasti huipukas jakauma A . Tämän jälkeen käytettiin tasaisempaa jakaumaa B parametreilla $x = (1, 1, 1, 1, 1)$, $w = \frac{1}{5}(1, 1, 1, 1, 1)$ ja $\sigma = 0,2$.

Todellinen ehdollinen jakauma $t = (t_1, \dots, t_5)$ laskettiin satunnaislukumuunnoksella suurta pisteiden määrää käyttäen. Laskenta toistettiin tätä varten useita kertoja pisteiden määrää nostaen, kunnes pisteiden määrän kasvattamisen vaikutus tulokseen jäi pieneksi. Todellinen ehdollinen jakauma päädyttiin näin menetellen laskemaan sekä jakauman A että jakauman B tapauksessa 25^5 satunnaispistettä käyttäen.

Kokonaisvirhe laskettiin testissä eri approksimaatiopisteiden lukumäärillä $N = 2^5, 4^5, \dots, 10^5$. Eksponentti tekee luvuista sopivat viisiulotteisen hilan pisteiden lukumääräksi. Satunnais- ja kvasi-satunnaisluvuilla käytettiin samoja pisteiden määriä vaikka näillä N voitaisiin valita vapaamminkin. Koska tavallisten satunnaislukujen käyttäminen tuottaa satunnaisen kokonaisvirheen, toistettiin testi sen osalta 20 kertaa sekä jakauman A että B tapauksissa. Kuvan 15 tuloksissa esitetään näiden toistojen keskiarvot sekä 50 % luottamusvälit.

Kuvassa 15 esiteyistä tuloksista nähdään, että suurilla N paras tulos saadaan kvasi-satunnaisluvuilla ja heikoin hilalla. Tämä tulos on yhdenmukainen muunnosten teoreettisen vertailun kanssa. Pienemmillä pisteiden määrillä jakauman A kohdalla menetelmien erot korostuvat, mutta jakauman B kohdalla hila tuottaa jopa muita paremman tuloksen. Menetelmien erot pienillä pisteiden määrillä eivät ole käytännössä merkittäviä. Satunnaislukujen tuottama keskimääräinen tulos ei ole merkittävästi kvasi-satunnaislukujen tulosta heikompi, mutta pienillä pisteiden määrillä satunnaislukuihin perustuvan vaihtoehdon hajonta on merkittävän suuri.



Kuva 15: Vasemmalla huipukkaamman jakauman A ja oikealla tasaisemman jakauman B approksimoinnin tulokset. Koska molemmat jakaumat koostuvat viidestä pistetodennäköisyydestä, kokonaisvirhe $V = 0,05$ tarkoittaa keskimäärin yhden prosenttiyksikön poikkeamaa solmun tilojen todennäköisyydessä.

Liite 2: Välimatka-asteikollisen muunnoksen testaus

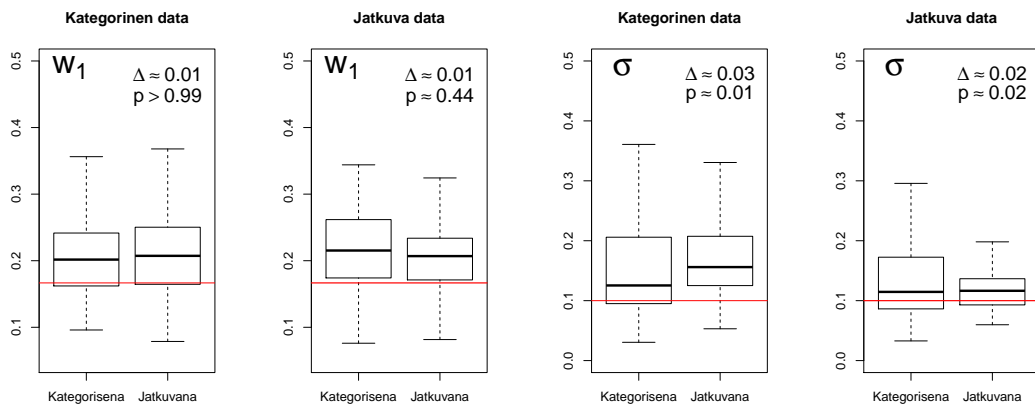
Alaluvussa 3.3 esiteltiin välimatka-asteikollinen muunnos tavanomaiseen RNM-menetelmään. Tässä liitteessä tarkastellaan kokeellisesti datan asteikon muunnoksen vaikutusta luvussa 4 esitellyn lähestymistavan tuottamiin tuloksiin. Tarkastelua varten luotiin järjestysasteikollinen sekä välimatka-asteikollinen aineisto. Aineistojen luomisessa käytetyt parametrien arvot estimoitiin alaluvussa 4.5 kuvatulla tavalla. Estimointi toteutettiin käyttäen sekä alkuperäisiä aineistoja että niiden muunnoksia toiselle asteikolle. Järjestysasteikollisen aineiston estimointi toteutettiin käyttäen satunnaislukuihin perustuvaa kaavaa (3.8). Välimatka-asteikollisen aineiston estimointi toteutettiin käyttäen kaavaa (3.11). Muunnettuun aineistoon käytettiin muunnoksen jälkeistä asteikkoa vastaavaa kaavaa. Molempien kaavojen käyttäminen molempien aineistojen tapauksissa tuotti yhteensä neljä testiä. Jokainen testi toistettiin 100 kertaa.

Kuhunkin testiin simuloitiin kymmenen havainnon suuruinen aineisto. Aineistot luotiin mallista, jossa kaksi isäsolmua vaikuttaa yhteen lapsisolmuun WMEAN-funktion mukaisesti. Parametrit valittiin $w_1 = 1/6$ ja $\sigma = 0,1$. Isäsolmujen arvot simuloitiin järjestysasteikolliseen aineistoon viisiarvoisesta tasajakaumasta ja välimatka-asteikolliseen aineistoon välin $[0, 1]$ tasajakaumasta. Lapsisolmun arvo simuloitiin järjestysasteikolliseen aineistoon tavanomaisen RNM-menetelmän määrittämästä ehdollisesta jakaumasta (3.5). Välimatka-asteikolliseen aineistoon lapsisolmun arvo simuloitiin tiheysfunktion (3.11) määrittämästä ehdollisesta jakaumasta.

Ensimmäisessä testissä parametrit estimoitiin järjestysasteikollisesta aineistosta käyttäen tätä vastaavaa kaavaa (3.8). Toisessa testissä sama aineisto muunnettiin välimatka-asteikolliseksi. Käytetty muunnos järjestysasteikollisesta muuttujasta $y_j \in J$ välimatka-asteikolliseksi muuttujaksi $y_v \in \mathbb{R} \cap [0, 1]$ on $y_v = J/5 - 0,1$, kun $J = \{1, \dots, 5\}$ on järjestysasteikollisen muuttujan arvojen esitys lukuina. Parametrit estimoitiin muunnetusta aineistosta käyttäen välimatka-asteikollisen aineiston kaavaa (3.11). Kolmannessa testissä samaa välimatka-asteikollisen aineiston kaavaa (3.11) käytettiin parametrien estimointiin alun perin välimatka-asteikollisesta aineistosta. Neljännessä testissä alun perin välimatka-asteikollinen aineisto muunnettiin järjestysasteikolliseksi. Käytetty muunnos välimatka-asteikollisesta muuttujasta $y_v \in \mathbb{R} \cap [0, 1]$ järjestysasteikolliseksi muuttujaksi $y_j \in \{1, \dots, 5\}$ on $y_j = 1 + l(y_v/0,2)$, missä l on desimaaliosan alaspäin pyöristävä lattiafunktio. Muunnoksen jälkeen y_j tulkitaan järjestysasteikolliseksi. Parametrit estimoitiin muunnetusta aineistosta käyttäen järjestysasteikollisen aineiston kaavaa

(3.8).

Tulokset antavat empiirisen sadan havainnon jakauman molempien parametrien estimaateille jokaisessa testissä. Asteikon muunnosten tuottamia eroja saman aineiston estimoinnissa voidaan testata kahden otoksen Kolmogorovin-Smirnovin testillä [72]. Testin nollahypoteesin mukaan aineiston muuntamisella ei ole vaikutusta estimaatin jakaumaan. Bonferroni-korjatut [73] p-arvot ovat painon w_1 osalta suuremmat kuin 0,4 ja hajontaparametrin σ osalta pienemmät kuin 0,02. Aineiston muuntamisella vaikuttaa siis olevan tilastollisesti merkitsevä vaikutus hajontaparametrin estimointiin. Kuvasta 16 nähdään että kaikkien estimaattien jakaumien mediaanit ovat lähempänä parametrin todellista arvoa muuntamattomasta aineistosta laskettuna. Hajontaparametrin σ osalta ero on tilastollisesti merkitsevä mutta painon w_1 osalta ei. Yhteenvetona voidaan sanoa estimaattien olevan epätarkempia asteikoltaan muunnettua aineistoa käytettäessä.



Kuva 16: Aineiston asteikon muuntamisen merkitys parametrien estimointiin. Vasemmanpuoleisissa kahdessa paneelissa on esitetty painon w_1 ja oikeanpuoleisissa kahdessa paneelissa hajontaparametrin σ estimaattien empiiriset jakaumat eli otokset. Jokainen laatikko-janakuvi esittää sadan toiston otoksen parametrin estimaatille. Otosten mediaanien välinen ero Δ sekä Bonferroni-korjattu Kolmogorovin-Smirnovin testillä saatu p-arvo nollahypoteesille "otokset on saatu eri jakaumista" on merkitty kuhunkin paneeliin. Parametrien todelliset arvot on esitetty vaakaviivoilla.

Liite 3: Tarkkuusasteikon testaus

Satunnaismuuttujana kuvattavan RNM-parametrin keskihajonta määritetään tarkkuutta kuvaavalla termillä. Nämä on määritelty alaluvussa 4.2. Samoilla termeillä esitetään sekä priori- että posterioriestimaattien tarkkuudet. Prioriestimaatit ovat ennen datan huomioimista muodostettuja asiantuntija-arvioita. Posterioriestimaatit huomioivat prioriestimaatin sekä datan. Datan tulisi korjata väärän suuntaisia prioriestimaatteja lähempänä parametrien todellisia arvoja oleviksi posterioriestimaateiksi. Tässä liitteessä testataan prioriestimaattien korjautumista kohti parametrien todellisia arvoja alaluvun 4.2 mukaisia tarkkuuden termejä käytettäessä.

Testiä varten simuloitiin dataa erilaisista keinotekoisista Bayes-verkoista, joiden alkuperäiset parametrit estimoitiin tämän jälkeen. Jokaisessa Bayes-verkossa kolme isäsolmua vaikutti yhteen ja samaan lapsisolmuun jonkin RNM-funktion mukaisesti. Alkuarvion vaikutuksen esiintuomiseksi parametrien päättely tehtiin jokaisen verkon osalta ensin hyvästä ja tämän jälkeen huonosta prioriestimaatista aloittaen. Hyvässä arviossa prioriestimaatti valittiin parametrin todellisten arvojen mukaisesti. Tämän tarkkuudeksi valittiin “melko tarkka (**)”. Huonossa arviossa kahden parametrin arviot vaihdettiin väärin päin. Myös tämän tarkkuudeksi valittiin “melko tarkka (**)”. Jälkimmäistä prioriestimaattia voidaan kuvailla harhaanjohtavaksi, mutta itsevarmaksi arvioksi. Datan satunnaisvaihtelun vaikutusta päättelyn tarkkuuteen testattiin toistamalla koe WMEAN-funktion osalta sekä kohtalaisen tarkalla että kohinaisella datalla. Kohtalaisen tarkan datan keskihajonnaksi valittiin $s = 0,01$ ja kohinaisen datan keskihajonnaksi $s = 0,1$. Muilla funktioilla kohinataso pidettiin näiden arvojen välissä valitsemalla $s = 0,05$.

Kahden datapisteen vaikutus hyvään arvioon on vähäinen. Huonoa arviota kaksi datapistettä korjaavat selvästi, pois lukien WMIN-funktion tapauksessa. Taulukossa 4 esitetyistä tuloksista nähdään, että erityisen selvä korjaus tapahtuu pienen kohinan WMEAN-testeissä. Kohinan vähäisyys tekee vähäisestäkin datasta merkittävää. Suuren kohinan WMEAN-testeissä ensimmäiset kaksi datapistettä heikensivät hyväkin arviota, koska nämä pisteet ovat kaukana odotusarvoistaan. Kaikissa testeissä parametrit kyetään lopulta päättelemään, kun dataa on käytettävissä tarpeeksi. “Erittäin tarkan (***)” tarkkuuden saavuttaminen vaikuttaa edellyttävän hyvästä arviosta lähdettäessä vähintään 20 datapistettä. Harhaanjohtavasta arviosta aloitettaessa pisteitä tarvitaan enemmän.

WMIN-funktion parametrit vaikuttavat haastavimmilta päätellä oikein, sillä ainoana testin tapauksena niiden harhaanjohtava alkuarvio ei hahmotu selvästi vääräksi vielä 20 datapisteen kohdalla. Tämä voi selittyä sillä, että muista RNM-funktioista poiketen sen painot ovat toisistaan riippumattomat. Näin ollen niiden päättely voi vaatia muihin RNM-funktioihin verrattuna enemmän informaatiota. Johtopäätöksenä posterioriestimaatit tarkentuvat odotetusti datan määrän kasvaessa. Alaluvussa 4.2 määritellyt tarkkuusarvot ovat saavutettavissa järkevällä datamäärällä huonostakin alkuarviosta lähdettäessä. Poikkeuksena WMIN-funktion parametrien tarkka päättely voi vaatia satoja datapisteitä. Satunnaismuuttujien jakaumat voidaan esittää tiiviisti ja ymmärrettävästi taulukon 1 määritelmiä ja notaatiota käyttäen.

Taulukko 4: RNM-menetelmän parametrien arvioiden tarkentuminen datan määrän lisääntyessä. Testatun mallin RNM-funktio on esitetty vasemmanpuoleisimmassa sarakkeessa. Toisessa sarakkeessa oikealta on esitetty datan luomiseen käytetyt todelliset parametrit. Kolmannessa sarakkeessa on prioriarvio. Datapisteiden määriin 2, 20 ja 200 posterioriarviot on esitetty kolmessa oikeanpuoleisimmassa sarakkeessa. Arvioiden tarkkuudet on esitetty tähdillä taulukon 1 määritelmiä käyttäen.

Malli	Parametri	$n = 0$	$n = 2$	$n = 20$	$n = 200$
WMEAN, pieni kohina, hyvä arvio	$w_1 = 0.1$	0.1 **	0.05 **	0.10 ***	0.10 ***
	$w_2 = 0.2$	0.2 **	0.22 *	0.19 ***	0.20 ***
	$w_3 = 0.7$	0.7 **	0.73 **	0.72 ***	0.70 ***
	$\sigma = 0.01$	0.01 **	0.01 **	0.01 ***	0.01 ***
WMEAN, pieni kohina, huono arvio	$w_1 = 0.1$	0.7 **	0.13	0.10 ***	0.10 ***
	$w_2 = 0.2$	0.2 **	0.17 *	0.19 ***	0.20 ***
	$w_3 = 0.7$	0.1 **	0.70 *	0.72 ***	0.70 ***
	$\sigma = 0.01$	0.01 **	0.02 *	0.01 ***	0.01 ***
WMEAN, suuri kohina, hyvä arvio	$w_1 = 0.1$	0.1 **	0.05 *	0.04 **	0.10 ***
	$w_2 = 0.2$	0.2 **	0.12	0.33 *	0.18 ***
	$w_3 = 0.7$	0.7 **	0.83	0.63 *	0.71 ***
	$\sigma = 0.1$	0.1 **	0.11 **	0.12 **	0.11 ***
WMEAN, suuri kohina, huono arvio	$w_1 = 0.1$	0.7 **	0.51	0.12 *	0.11 ***
	$w_2 = 0.2$	0.2 **	0.21	0.30 *	0.18 ***
	$w_3 = 0.7$	0.1 **	0.28	0.58 *	0.71 ***
	$\sigma = 0.1$	0.1 **	0.13 **	0.12 **	0.11 ***
WMIN, hyvä arvio	$w_1 = 2$	2 **	2.0 **	2.0 **	2.1 **
	$w_2 = 4$	4 **	4.0 **	4.0 **	4.1 **
	$w_3 = 5$	5 **	4.9 **	4.9 **	4.9 **
	$\sigma = 0.05$	0.05 **	0.05 **	0.05 ***	0.05 ***
WMIN, huono arvio	$w_1 = 2$	5 **	5.0 *	4.9 *	4.2 **
	$w_2 = 4$	4 **	4.0 *	4.0 **	4.0 **
	$w_3 = 5$	2 **	2.2 *	2.2 *	3.3 *
	$\sigma = 0.05$	0.05 **	0.08 **	0.08 ***	0.07 ***
MIXMINMAX, hyvä arvio	$w_1 = 0.3$	0.3 **	0.27 **	0.29 ***	0.30 ***
	$w_2 = 0.7$	0.7 **	0.73 **	0.71 ***	0.70 ***
	$\sigma = 0.05$	0.05 **	0.04 **	0.04 ***	0.05 ***
WMIXMINMAX, huono arvio	$w_1 = 0.3$	0.7 **	0.64 *	0.35 **	0.31 ***
	$w_2 = 0.7$	0.3 **	0.36 *	0.65 **	0.69 ***
	$\sigma = 0.05$	0.05 **	0.13 *	0.06 **	0.05 ***