Aalto-yliopisto
Sähkötekniikan korkeakoulu
Automaatio- ja systeemitekniikan tutkinto-ohjelma

Ville Piiparinen

Havaintodatan semanttinen mallintaminen ja validointi

Diplomityö
Espoo, 19. tammikuuta 2015

Valvoja: Professori Eero Hyvönen
Ohjaaja: FM Jouni Tuominen

Aalto-yliopisto
Sähkötekniikan korkeakoulu
Automaatio- ja systeemitekniikan tutkinto-ohjelma

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä:	Ville Piiparinen		
Työn nimi:	Havaintodatan semanttinen mallintaminen ja validointi		
Päiväys:	19. tammikuuta 2015	**Sivumäärä:**	60
Professuuri:	Mediatekniikka	**Koodi:**	T-75
Valvoja:	Professori Eero Hyvönen		
Ohjaaja:	FM Jouni Tuominen		

Paikkaan, aikaan ja lajistoon liittyvät havaintoaineistot ovat tärkeitä biologian tutkimuksessa, luonnon monimuotoisuuden hallinnassa, biologian opetuksessa ja harrastustoiminnassa.

Havaintomateriaalin vaihteleva laatu ja siitä johtuva huono luotettavuus aiheuttaa ongelmia kaikissa edellä mainituissa käyttötarkoituksissa.

Tässä tutkielmassa kuvataan kaksi erilaista menetelmää luontohavaintojen validointiin. Tutkimus tehtiin käyttäen kahta erityyppistä tietokantaa lintuhavainnoille, kansalaisten keräämää sekä lintuyhdistysten raportoimaa tietoa. Kummankin menetelmän käyttöä edellä mainittujen tietokantojen kanssa vertailtiin lintuharrastajan näkökulmasta.

Tutkielman tuloksena esitellään mobiililaitteilla toimiva havaintopalveludemonstraattori, joka sopii lintuharrastajien käyttöön.

Asiasanat:	havainnot, luonto, biologia, semantiikka, päättely, luokittelu
Kieli:	suomi

Aalto University
School of Electrical Engineering
Degree Programme of Automation and Systems Technology

ABSTRACT OF
MASTER'S THESIS

Author:	Ville Piiparinen		
Title:	Semantic modelling and validation of observation data		
Date:	January 19, 2015	**Pages:**	60
Professorship:	Media technology	**Code:**	T-75
Supervisor:	Professor Eero Hyvönen		
Instructor:	Jouni Tuominen M.Sc.		

Observation data associated with place, time and the species are important for the study of biology, biodiversity management, biology education and leisure activities.

Poor reliability caused by the varying quality of the observation material cause problems in all the above-mentioned purposes.

This thesis describes two different methods to validate the observations of nature. The study was conducted using two different types of databases for bird observations, observations gathered by citizens and observation reported by ornithology associations. Both of the developed methods were applied for both of the the above-mentioned databases and were compared from a birdwatcher's point of view.

As a result of this study a demonstrator for observation service was made, which is suitable for bird-watchers use.

Keywords:	observations, nature, biology, semantics, reasoning, classifying
Language:	Finnish

Tekijän kiitokset

Kiitos Nina Laurenteelle kannustuksesta sekä toiminnasta biologisen tiedon asiantuntijana ja Mikko Koholle käyttöliittymän implementoinnista. Kiitos myös Rami Aamulehdolle ja Juha Törnroosille kannustuksesta sekä semanttisen webin etsimisestä.

Helsinki, 19. tammikuuta 2015

Ville Piiparinen

Sisältö

Luku 1

Johdanto

Paikkaan, aikaan ja lajistoon liittyvät havaintoaineistot ovat tärkeitä biologian tutkimuksessa, luonnon monimuotoisuuden hallinnassa, biologian opetuksessa ja harrastustoiminnassa.

Näiden yhtenä haasteena on havaintomateriaalin luotettavuus, sillä aineistot sisältävät hyvin eritasoisten tutkijoiden ja harrastajien tekemiä havaintoja. Käytön kannalta haasteena on aineistojen käytön kontekstiherkkyys ja havaitsijoiden vaihteleva tietotaito, joka pitäisi ottaa huomioon eri tilanteissa: havainnot liittyvät aina tiettyyn paikkaan, tiettyyn aikaan ja niitä pitäisi voida hyödyntää mobiilisti paikan päällä kohteiden löytämiseksi, havainnon varmistamiseksi, luotettavien tietojen keräämiseksi ja lisäpalveluiden antamiseksi käyttäjälle. Esimerkiksi lintujen kevät- ja syysmuuton aikana eri lajien havaittavuus vaihtelee nopeasti ja paikkasidonnaisesti.

Käytettävissä on linnuista ja perhosista laajoja pitkäaikaisia aineistoja, joita on koottu pääosin harrastajavoimin. Näitä ovat esimerkiksi Hatikka [34], Tiira.fi [9], Suomen Perhotutkijain Seura ry:n havaintotietokanta sekä The Global Biodiversity Information Facility (GBIF) -aineistot[20]. Näistä tarkemmin osiossa 2.6.

Tutkimuksen tavoitteena on kehittää menetelmiä ja demonstraattori havaintoaineistojen nykyistä monipuolisempaa hyödyntämistä varten. Erityisesti kehitetään ratkaisumalli havaintokantojen hyödyntämiseksi havaintojen varmistamisessa. Ongelmaksi muodostuu, että vain jossain mielessä hienot havainnot kirjataan ja eri henkilöillä ja eri aineistoissa voi olla eri kriteerejä ha-

vaintojen raportoimisessa. Lisäksi aineistot todennäköisesti painottuvat isoihin, kauniisiin ja helpommin tunnistettaviin lajeihin.

Havainnoitsijoiden luotettavuudesta on monessa tapauksessa olemassa arvioita. Tunnettu huippuornitologi on luotettavampi havaitsija kuin koululainen. Hypoteesina on, että luotettavien havaitsijoiden havaintoja mittana käyttäen voidaan arvioida huonosti tunnettujen havainnoitsijoiden havaintojen luotettavuutta, esimerkiksi tunnistaa epäilyttäviä havaintoja (esimerkiksi pääskyshavainto talvella). Tarkastelujen avulla on mahdollista tunnistaa havaintokannassa olevat epäilyttävät havainnot ja parantaa kannan luotettavuutta.

Tässä tutkimuksessa kehitetään lajin tunnistusta tukeva malli, joka hyödyntää olemassa olevaa havaintokantaa ja siihen liittyvää apriori-tietoa, kuten levinneisyyskarttoja, lajistoon liittyviä erityispiirteitä, kuten fenologiaa, tietoa toisiinsa sekaisin menevistä lajeista ja niin edelleen. Edellä mainittuja muuttujia käyttäen tehdään sekä apriori-laskentaan perustuva, havaintopisteistä laskettava levinneisyysaikasarjamalli sekä parempiin tuloksiin tähtäävä bayesilaista päättelyä käyttävä luokitin, joka osaa kertoa havainnon todennäköisyysjakauman eri lajeille.

Ratkaisumallin hyödyllisyyttä testataan demonstraattorilla havaintopalvelusta, jolla on seuraavia toiminnallisuuksia:

Havaitsija voi kysellä paikka- ja aikakontekstissa onko havaittu laji todennäköisesti validi havainto vai ei. Palvelu ohjaa käyttäjän eteenpäin BirdLifen lintuhavainnon ilmoitusjärjestelmään [9] välittäen havainnon tiedot kyseiselle järjestelmälle.

Tämän opinnäytetyön tutkimuskysymykset ovat seuraavat:

1. Miten epäluotettavat havaitsijat voidaan tunnistaa havainnon perusteella?

2. Miten voidaan arvioida havainnon luotettavuutta tiettyyn aikaan tietyssä paikassa?

Näitä varten kehitetään kaksi menetelmää ja niitä testataan empiirisillä kokeilla.

Työssä esitellään aihealueen taustaa, mallinnuksen ja validoinnin menetelmiä sekä esitetään laskennan tulokset ja niiden arviointia. Toteutettu järjestelmä pystyy arvioimaan havainnon luotettavuuden oikein tilanteissa, joissa laskennan pohjana käytetty havaintodata on ollut hyvänlaatuista.

Luku 2

Taustaa

Tässä luvussa esitellään semanttisen webin idea sekä havaintodatan semanttiseen kuvailuun tarkoitettuja ontologioita ja datan validointimenetelmiä. Validointimenetelmistä esitellään erityisesti naiivi Bayes -menetelmä, jota on sovellettu tässä työssä. Lisäksi esitellään suomalaiset lintuhavaintotietokannat, joista tässä työssä käytetty data on peräisin.

2.1 Havaintotietojen esittäminen

Kaksi yleisimmin käytettyä skeemaa biologisen tiedon esittämiseen ovat Darwin Core [57] ja Access to Biological Collections Data (ABCD) [5], joista ensimmäinen on muodostunut standardiksi havaintoaineistojen kuvailussa.

Darwin Core on joukko datan kuvailuun liittyviä standardeja, jotka toimivat Dublin Core -standardin [56] laajennuksina biodiversiteetti-informaatioon liittyvissä sovelluksissa. Darwin Core -XML-skeemassa on määrittelyt lajeihin liittyvien tietojen käsitteille, kuten lajinimi, havaintopaikka ja havaintoaika. Skeeman tarkoituksena on tarjota standardi referenssi biologisen tiedon välitykseen. Darwin Coren ongelmana on sen monitulkintaisuus, se kuvaa tietoa liian yleisellä tasolla.

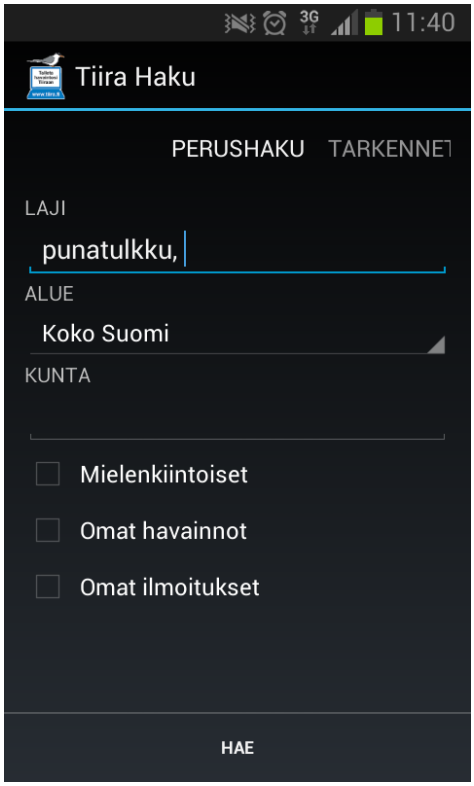
ABCD:llä voidaan kuvata samoja asioita kuin Darwin Corella, mutta se on alunperin kehitetty luonnontieteellisten kokoelmien kuvaamisen skeemaksi.

ABCD:n ongelmana pidetään sen monimutkaisuutta. Se on kattava mutta loppukäyttäjille monimutkainen. Sen käyttäjät eivät useinkaan ole tietotekniikka-alan asiantuntijoita.

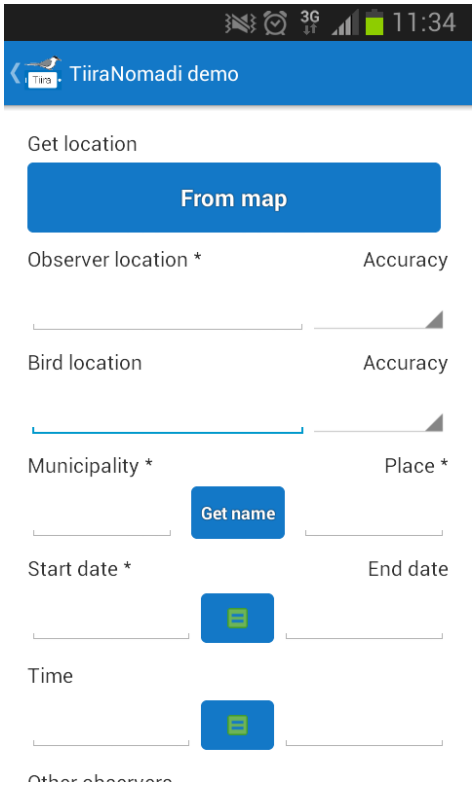
Sovelluspuolella on tehty mobiilisovellus EpiCollect ja toimiva ekosysteemi luontohavaintojen keräämiseen ja analysointiin [4], mutta se ei tue minkäänlaista havaintojen reaaliaikaista validointia. Tämän kehitys näyttää kuitenkin tyrehtyneen. Samoin on kehitetty RB Birds [46], jossa ei ole minkäänlaista havainnon validointia, mutta siinä on Luontoportin [35] kaltainen tuntomerkkeihin perustuva tunnistuspalvelu. Se on kuitenkin ennemminkin tietokirjamainen teos lintulajeista kattavine kuvineen, mutta mobiilisovelluksen muodossa. Italialaiset ovat tehneet blueBill-nimisen mobiilisovelluksen [13], joka toimii vain linnuille mutta ei sisällä levinneisyyskarttoja, ei huomioi habitaatteja eikä ylipäätään sisällä havaintojen validointia. Siinä on kuitenkin kattava tietokanta lintujen äänistä käyttäjien nauhoittamina. Kyseinen sovellus mahdollistaa myös havaintojen jakamisen tekstiviestillä, sähköpostilla tai Facebookin välityksellä.

Näiden lisäksi on tehty iNaturalist-niminen [27] luontohavaintoportaali, jossa voi raportoida luontohavaintoja, tavata muita luontoharrastajia sekä oppia luonnosta. Erityisesti palvelussa on mahdollista pitää kirjaa omista havainnoistaan muun muassa aika- tai paikkakontekstissa (kartta tai päivämäärähaku) ja saada apua luontohavaintojen tunnistamiseen koko yhteisöltä. Tämä apu ei ole reaaliaikaista, mutta se voi olla todella tehokasta yhteisön ollessa suuri ja aktiivinen. iNaturalist-projektilla on myös omat sovellukset eri mobiilikäyttöjärjestelmille, kuten Applen iOS:lle sekä Googlen Androidille.

BirdLife Suomi on kehittänyt mobiilihakupalvelun Tiiran havaintotietokantaan [12]. Se sisältää vain nettisivuilta tutut hakutoiminnallisuudet karttoineen. Tiiran käyttöliittymässä on tekstinsyöttökenttä haetulle lajille sekä alueen valinta. Tämän lisäksi käyttöliittymä sisältää valintaruudut siitä haetaanko lajia mielenkiintoisista lajeista, omista havainnoista tai omista ilmoituksista. Tiira haun käyttöliittymä on esitetty kuvassa 2.1.



Kuva 2.1: Tiira haun mobiilikäyttöliittymä [12]



Kuva 2.2: TiiraNomadin mobiilikäyttöliittymä [17]

Tämän lisäksi Tiiran tietokantaan löytyy maksullinen mobiilisovellus TiiraNomadi [17], joka mahdollistaa havaintojen tekemisen mobiililaitteella. Sovelluksella voi tehdä havaintoja paikka- ja aikakontekstissa, tallentaa havainnot, lähettää havainnon tiedot Tiiran kantaan sekä selata karttoja offline-tilassa.

Käyttöliittymässä on ensimmäisenä valintana sinisellä pyöreäkulmaisella suorakaiteen muotoisella painikkeella käyttäjän sijainnin haku kartalta sekä vaihtoehtoisesti tekstinsyöttökenttä sijainnille. Tämän lisäksi voi erikseen kirjoittaa havaitun linnun sijainnin. Tämän alapuolella ovat tekstinsyöttökenttä kaupungille tai kunnalle, paikalle sekä ajanhetkelle ja muille huomioille. Kaikki tekstinsyöttökentät ovat valkoisella pohjalla. Tämän lisäksi käyttöliittymässä on pienet sinipohjaiset painikkeet päivämäärän haulle kalenterista. Havaitun lajin syöttö ei näy tavallisella älypuhelimella suoraan TiiraNomadin etusivulla, vaan sitä varten on käyttöliittymää rullattava alas. Tästä syystä kyseinen kenttä on leikkautunut pois myös käyttöliittymää esittelevästä kuvasta. TiiraNomadin käyttöliittymä on esitetty kuvassa 2.2.

2.2 Semanttinen web ja ontologiat

Semanttinen web on nykyisen webin laajennus, jossa tiedolle annetaan hyvin määritelty merkitys siten, että se mahdollistaa ihmisten ja koneiden nykyistä tehokkaamman yhteistyön tiedonhallintaan liittyvissä toiminnoissa [7]. Semanttisen webin toiminta perustuu tietoon tiedosta, eli tiedon kuvailuun. Semanttinen yhteentoimivuus koneiden ja myös ihmisten välillä mahdollistetaan käyttämällä ontologioita tiedon kuvailussa.

Ontologiat ovat jonkin tietyn aihealueen käsitteistöjä, jotka kuvaavat formaalisti kyseisen aihealueen käsitteet ja niiden väliset suhteet [22]. Kun kaikki tieto on koneymmärrettävässä muodossa, se voidaan sijoittaa hajautetusti ja sitä voidaan koneellisesti hakea, yhdistää ja käyttää uudelleen, niin että sen semantiikka säilyy.

Ontologiat voidaan linkittää toisiinsa ja toinen ontologia voi täydentää toista. Näin on mahdollista muodostaa kattava ja tarkka tietämys jostain aihealueesta, kunhan vain se on ontologisesti määritelty. Semanttisessa webissä kaikki objektit eli kuvattavat resurssit on yksikäsitteisesti identifioitu URI-tunnisteilla [6]. Näin sekä itse tietoon, tietoon tiedosta eli metatietoon sekä ontologian käsitteisiin voidaan viitata yksikäsitteisesti.

Kun esimerkiksi tietyn alueen linnut tai lintuhavainnot voidaan yksikäsitteisesti nimetä ja linkittää johonkin nimistöontologiaan, voidaan sen jälkeen niitä koneellisesti käsitellessä varmistua siitä, että myös kone ymmärtää yksikäsittei- sesti kyseisen lajin.

2.3 Aineistojen validointi

Aineistoja voidaan validoida monin menetelmin. Yksinkertaisia menetelmiä ovat muun muassa muuttujan arvon rajaus tai muuttujan tyypin tarkistus [50]. Jos aineiston tietyille muuttujille sallitaan vain tietynlaisia arvoja, on aineiston validointi helppoa.

Tilanteissa, joissa aineistoon liittyy tietynlaista epävarmuutta voidaan aineiston arvojen validiudesta tai tarkemmin sanoen todenmukaisuudesta sanoa vain jonkinlainen subjektiivinen todennäköisyys tai algoritmisen laskennan tulos, jossa algoritmi pohjautuu tietynlaisiin oletuksiin. Tällaista aineistoa voidaan validoida tilastollisin menetelmin käyttäen ennustavia mal-

leja tai analyysia. Ne perustuvat matemaattisten yhteyksien löytämiseen opetusaineiston ja käyttöaineiston muuttujien välille [18]. Opetusaineiston muuttuja on validin alkion muuttuja (esimerkiksi paikka koordinaatteina) ja käyttöaineiston muuttuja on vastaava muuttuja luokiteltavassa aineistossa. Metelmän tarkoitus on luokitella uusia muuttujajoukkoja johonkin luokkaan. Luokkia voivat olla esimerkiksi 'roskaposti' tai 'ei roskaposti' [15] tai vaikka Suomessa esiintyvien lintulajien nimet.

Aineiston validointi on monissa tilanteissa tärkeää [2]. Erityisesti tämä koskee luonnontieteellisiä aineistoja [36] [40]. Jo pelkästään biologian tutkimuksen kannalta on tärkeää saada tietoa aineiston oikeellisuuden tilasta.

Wieczorek et al. [55] ovat validoineet kasvihavaintoja laajoilla aineistoilla, mutta menetelmä oli hyvin yksinkertainen ja epävarmat havainnot varmistettiin asiantuntija-arvioin. Arvioinnin perustana käytettiin sitä, onko kyseinen laji esiintynyt varmistettuna havaintona 4x4 kilometrin ruudulla samassa pisteessä viimeisen 35 vuoden aikana.

Tässä työssä esitellään kaksi menetelmää lintuhavaintojen validointiin, jotka ovat laajennettavissa koskemaan muitakin luonnontieteellisiä kohteita.

2.4 Bayesilainen luokitin

Ennustavia malleja käytetään usein tulevaisuuden ennustamiseen esimerkiksi rikostutkinnassa, mutta niitä voidaan käyttää myös minkä tahansa tuntemattoman tapahtuman ennustamiseen [19]. Tähän on monia menetelmiä. Näitä menetelmiä ovat muun muassa naiivi bayesilainen luokitin, joka on ohjatun oppimisen menetelmä, lähin naapuri -algoritmi, tukivektorikone sekä neuroverkot [33].

Lähin naapuri -menetelmä olettaa kaikkien opetusesimerkkien olevan pisteitä n-ulotteisessa avaruudessa. Lähin naapuri on määritelty etäisyyden perusteella, joka on useimmiten euklidinen etäisyys [38]. Neuroverkot koostuvat yksinkertaisista laskentaelementeistä, joilla on monta syötettä ja yksi vaste. Jokaista syötettä vastaa usein painokerroin, jotka määräytyvät verkon oppimisvaiheessa. Tukivektorikone perustuu muuttujien linearisoiintiin ja muuttujien vähentämiseen. Tukivektorikone on myös mahdollista toteuttaa neuroverkoilla.

Tähän opinnäytetyöhön on valittu luokittimeksi (tai ennustimeksi) naiivi bayesilainen päättelijä, koska sen malli on yksinkertainen ja siihen löytyy

hyviä ohjelmointikirjastoja, joita on helppo käyttää.

Todennäköisyyden tulkinnassa vallitsee kaksi pääsuuntausta. Klassisen tilastotieteellisen tulkinnan mukaan tapahtuman todennäköisyys on raja-arvo äärettömyydessä, kun koetta toistetaan useita kertoja. Usein saatetaan kuitenkin puhua todennäköisyydestä tapahtumalle, joka on ainutkertainen ja jonka tilastoitumista ei voida todeta, koska tapahtumaa ei voida toistaa. Voidaan esimerkiksi puhua, mikä on todennäköisyys sille, että tietokone- ja matkapuhelinvalmistaja Apple julkaisee seuraavan sukupolven älypuhelimensa syyskuussa. Useampi ihminen antaa todennäköisesti eri todennäköisyydet edellä mainitulle tapahtumalle, koska heillä on asiasta erilaiset taustatiedot sekä uskomukset. Tästä syystä bayesilaisen tilastotieteen tulkintaa todennäköisyydestä kutsutaan subjektiiviseksi todennäköisyydeksi tai uskomusta kuvaavan asteen mittaluvuksi.

Naiivi Bayes on yksinkertainen tekniikka luokittimen toteuttamiseksi. Se perustuu luokkien määräämiseen opetusaineiston tapauksille, jotka esitetään piirrevektoreina, jotka ovat siis kaikki tapaukseen liittyvät muuttujat. Luokat määräytyvät opetusaineiston perusteella. Esimerkiksi tiettyyn aikaan tietyssä paikassa nähty varis kuuluu luokkaan varis.

Bayesilaista mallia käytetään muun muassa roskapostisuodatuksessa tai tekstin oikolukuohjelmistoissa. Malli ei edellytä aihealueen tarkkaa tuntemusta, kunhan yksittäisestä tilanteesta saa irrotettua tarpeeksi yksilöiviä muuttujia.

Bayesin lauseen [3] mukaan luokkamuuttujalle y ja siihen liittyville riippuville muuttujille x_1:stä x_n:än vallitsee ehdollisen todennäköisyyden kaavaan perustuva lause

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)},$$

jossa y on luokkamuuttuja ja $x_1, \ldots, x_n$ ovat piirrevektorin n piirrettä eli riippuvaa muuttujaa.

Naiivi bayesilainen päättelijä perustuu "naiiviin" oletukseen, että jokainen piirrevektoripari on toisistaan riippumaton. Näin ollen siis

$$P(x_i \mid y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i \mid y)$$

Riippumattomuusoletuksen avulla aiempi kaava yksinkertaistuu muotoon

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Koska $P(x_1, \ldots, x_n)$ on vakio, yksinkertaistuu lause seuraavasti:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

Naiivi bayesilainen luokittelija yhdistää tähän luokittelusäännön. Yleensä tämä sääntö on valita luokka, joka on todennäköisin. Tämän toteuttava luokittelija, bayesilainen luokittelija, on seuraava funktio

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

joka valitsee luokaksi $\hat{y}$ sellaisen luokkamuuttujan y arvon, jonka todennäköisyys $P(y) \prod_{i=1}^{n} P(x_i \mid y)$, saa suurimman arvon.

2.5 Biodiversiteettiaineistot

Biodiversiteetti eli luonnon monimuotoisuus tarkoittaa biologisen elämän monimuotoisuutta. Luonnon monimuotoisuus on vähentynyt huomattavasti viimeisen 40 vuoden aikana, eikä väheneminen ole tasaantunut viime vuosinakaan. WWF:n julkaiseman raportin ”2010 and beyond: Rising to the Biodiversity Challenge” mukaan luonnon monimuotoisuus on vähentynyt vuodesta 1970 vuoteen 2005 27 % [58]. On puhuttu jopa kuudennesta massasukupuutosta. Tähänastisista massasukupuutoista maapallon historiassa tunnetaan 12, joista viittä pidetään erityisen suurena.

Eläinten elinympäristöt pilkkoutuvat, kun rakennetaan moottoriteitä, ostoskeskuksia tai ydinvoimaloita. Ilmaston lämpeneminen osaltaan kiihdyttää luonnon monimuotoisuuden vähenemistä, samoin kuin se muuttaa lajien levinneisyyttä [37]. Kun eläinten elinympäristöjä inventoidaan havaintojen avulla, voivat biologian tutkijat saadaan tarkempia tietoja luonnon tilasta. Näiden havaintojen pohjalta tehdyt päätelmät voivat edesauttaa luonnon monimuotoisuuden säilymistä tulevaisuudessa. Näin tämäkin tutkimus hyödyttää samalla myös biologian tutkimusta.

The Global Biodiversity Information Facility (GBIF) [20] on kansainvälinen järjestö, joka keskittyy laji- sekä biodiversiteettitiedon saattamiseen vapaaseen käyttöön internetissä ja se on perustettu hallitusten toimesta vuonna

2001. GBIFin tarkoitus on edistää lajitiedon avointa saatavuutta ja käyttöä, erityisesti lajien esiintyvyyttä ajassa koko maapallon alueella.

2.6 Lintuhavaintokannat

Yhdistysten, harrastelijoiden tai "kansalaislintutieteilijöiden" havaintoja sisältävät lintutietokannat ovat luonnostaan epäluotettavampia kuin vaikkapa museon ylläpitämät historialliset lajistoseurannat [14]. Tästä johtuen tässä tutkimuksessa kehitettävä analyysimalli on erittäin tarpeellinen.

Kansainvälisistä lintutietokannoista merkittävin on eBird [49]. Se on vuonna 2002 lanseerattu Cornell Lab of Ornithologyn ylläpitämä järjestelmä, joka tarjoaa tietoa lintujen runsaudesta ja levinneisyysalueista erilaisilla spatiotemporaalisilla asteikoilla. eBird sisältää myös paikallisten lintutietokantojen tietoja integroituna omaan järjestelmäänsä.

Suomessa on kaksi merkittäviä lintutietokantapalveluja, Tiira [9] sekä Hatikka [34]. Lintutietokantoja käytetään havaintojen kirjaamiseen ja lukemiseen. Tietokantoihin on vapaa pääsy. Tässä työssä aineistoina käytetään sekä Hatikan että Tiiran lintuhavaintokantoja. Hatikka on Luonnontieteellisen keskusmuseon ylläpitämä luontohavaintotietokanta, jossa kuka tahansa voi ilman rekisteröitymistä kirjata sekä hakea luontohavaintoja. Tiira on BirdLife Suomi ry:n ylläpitämä lintutietopalvelu, jossa on vastaavanlainen havaintokanta linnuille.

Molemmista tietokannoista on mahdollista hakea ja ladata aineistoja GBIF-portaalin [20] kautta. Tiiran lintuhavaintokannasta on otettu vain lintuyhdistysten kirjaamiksi merkittyjä havaintoja, joita voidaan pitää kohtalaisen luotettavina.

Molemmissa kannoissa on havaintoa kohti melko samantyyppisiä muuttujia, joista tärkeimpiä ovat erityisesti aika, paikka geokoordinaatteina sekä havaittu taksoni. Taksonista käytetään tieteellistä nimeä. Useissa tapauksissa havainnosta on kirjattu paikannimi myös selkokielisenä nimenä. GBIF-tietokanta tukee myös monia muita datakenttiä, kuten kuvan URL-osoite tai osin redundantteja kenttiä, kuten koko lajihierarkia sukuja ja heimoja myöten. Nämä kentät olivat kuitenkin käytetyssä datassa tyhjiä. Useista havainnoista puuttuu joitain tärkeitäkin kenttiä tai ne ovat yksinkertaisesti väärin tai osittain väärin. Näistä ongelmista lisää osiossa 6.1.

Luku 3

Aineisto ja menetelmät

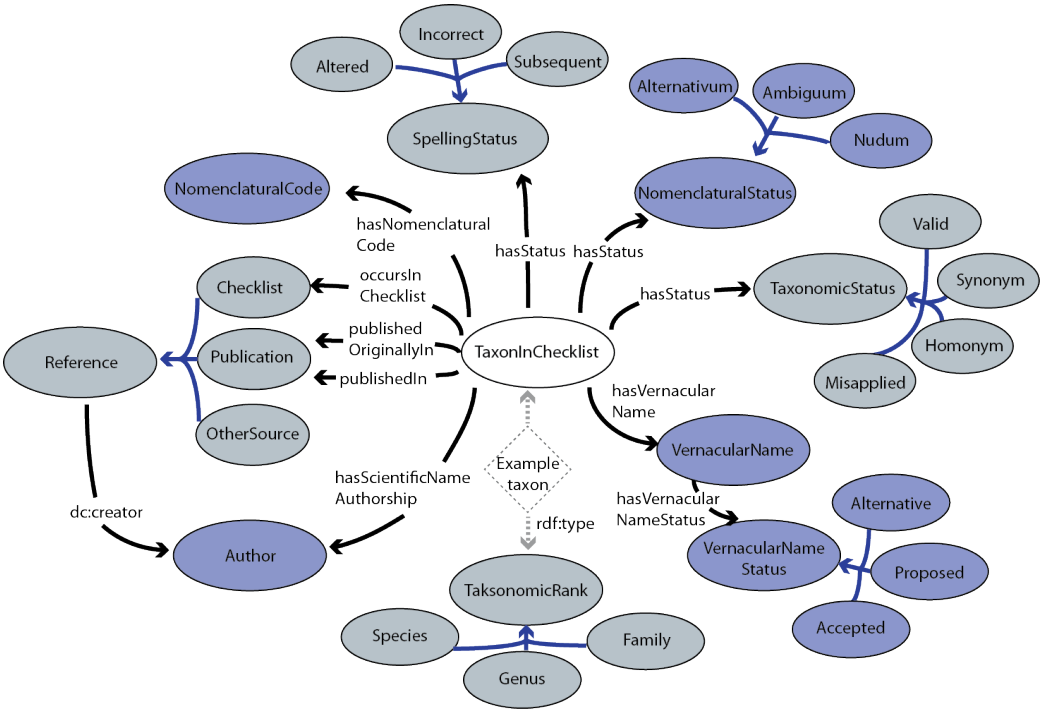
Tässä luvussa kerrotaan, miten aineisto on kerätty ja miten sitä on käsitelty ennen aineistolle suoritettuja laskentoja. Luvussa esitellään myös datan visualisointiin käytetty HAKO-työkalu [30] sekä työssä käytetyt laskenta- ja luokittelumenetelmät.

3.1 Taksonominen metaontologia TaxMeOn

Luonnon monimuotoisuuden hallinta edellyttää heterogeenisen biologisen informaation käsittelyä useista eri lähteistä. Tällaisen tiedon indeksointi, kerääminen ja löytäminen pohjautuu lajien tieteellisiin nimiin. Nämä nimet, niiden väliset suhteet ja niiden kansankieliset nimet muuttuvat ajassa johtuen muun muassa uusista tieteellisistä löydöistä tai erilaisten kansankielisten nimien käytön vakiintumisesta eri kielissä. Tämä kaikki tekee aineistojen integroinnista ja niiden käytöstä hankalaa.

Taksonominen metaontologia TaxMeOn [51] on esitetty kuvassa 3.1. Kuvassa ellipsit kuvaavat eri luokkia, mustat nuolet kuvaavat relaatiota ja sinertävät nuolet kuvaavat alaluokka-suhdetta. Kuvan ellipsit ovat eri värisiä vain esteettisistä syistä. Katkoviivainen kärjellään oleva nelikulmio kuvaa esimerkkitaksonin instanssia. Kuvassa tämän työn kannalta tärkeä luokka on kansankielinen nimi (VernacularName) sekä esimerkkitaksonin (Example taxon)

instanssin tyyppi, joka sitoo taksonin tieteelliseen nimistöön. Kuvan muut ellipsit kuvaavat yksityiskohtaisempia asioita, kuten kansankielisen nimen tilaa (VernacularNameStatus) tai nimen tekijää. Näitä luokkia ei hyödynnetty tässä työssä.

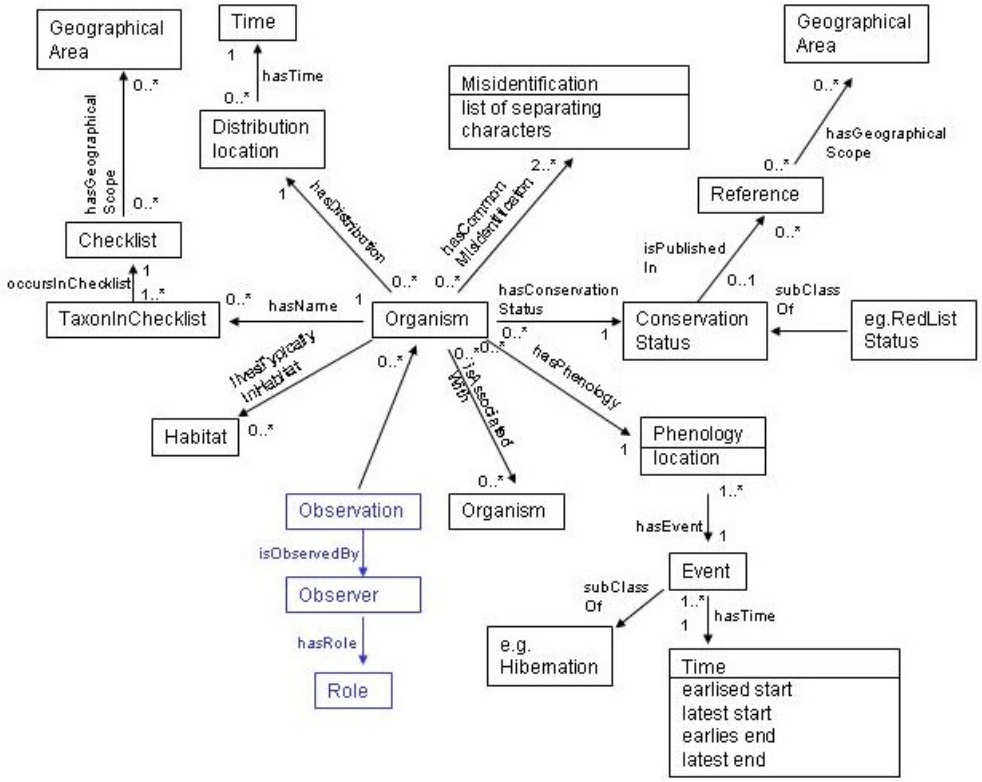


Kuva 3.1: Taksonominen metaontologia TaxMeOn [51]

Metaontologia koostuu kolmesta osasta:

1. Lajilistat
2. Kansankieliset nimet
3. Tieteellinen nimistö

Lajilistoilla voidaan kuvata mitä tahansa lajeja ja niiden lajihierarkiaa. Lajit voidaan liittää tarkasti tieteelliseen nimistöön ja niille voidaan kuvata kansankieliset nimet.



Kuva 3.2: TaxMeOn-laajennos

3.2 TaxMeOn-laajennokset

TaxMeOn-metaontologialle on kehitetty TaxMeOn-laajennos [32], joka mahdollistaa TaxMeOn-metaontologian käyttämisen luontohavaintoihin, koska se mallintaa kaikki taksonia koskevat keskeiset piirteet koneymmärrettävässä muodossa. Tämä on tarpeellista, sillä lajilistaukset eivät suoraan liity semanttisesti havaintoihin. TaxMeOn-laajennos on esitetty kuvassa 3.2. Kuva sisältää luokkia (suorakaiteet) ja niiden suhteita toisiinsa (nuolet). Suhteiden tyyppi on kirjoitettu nuolen viereen ja nuolen eri päihin tieto siitä, montako kyseistä luokkaa suhteen kumpikin pää voi käsittää.

TaxMeOn-laajennos sisältää muun muassa seuraavia luokkia ja niiden suhteita:

1. Levinnäisyysalue (Distribution) kuvaa maantieteellistä aluetta, jossa laji esiintyy. Levinnäisyysalueella on paikka, joka esitetään monikulmiona, joka koostuu WGS84-koordinaateista. Lajit muuttuvat ajallisesti ja siksi levinnäisyysalueeseen voidaan yhdistää ajallinen informaatio.

2. Lajien väärinmääritystä (Misidentification) käytetään osoittamaan yleistä väärinmääritystä läheisien lajien välillä, jotka muistuttavat toisiaan. Sekaisin menevien lajien määrä vaihtelee kahdesta useampaan. Keskeiset pääominaisuudet, jotka erottavat lajit toisistaan, ovat sisällytetty helpottamaan lajien tunnistusta, kun käytetään interaktiivista sovellusta kenttäolosuhteissa.

3. Fenologia (Phenology) kuvaa kasvin tai eläimen vuosittaisen syklin. Fenologia määritellään käyttäen luokkaa Tapahtuma, joka edelleen jakautuu edelleen pienemmiksi alaluokiksi. Tapahtumaan voidaan lisätä temporaalinen (ajallinen) ulottuvuus. Useimmissa tapauksissa lajilla on vain yksi fenologia, mutta malli tukee myös tapauksia, jossa lajilla on useampia fenologioita. Esimerkkinä mainittakoon tilanne, jossa kehittyviä tiloja voi esiintyä useissa paikoissa. Fenologiaan on liitetty spatiaalinen informaatio.

4. Habitaatti (Habitat) on ympäristö, jossa laji elää. Ympäristö(t) kuvaillaan abioottisten ja bioottisten olosuhteiden pohjalta. Yksi laji voi elää useassa eri habitaatissa.

5. Suojelustatus (Conservation status) indikoi onko laji uhanalainen vai yleinen. Tämä informaatio on yleensä saatu lajilistoista (Red List), missä lajit ovat loukiteltu suojelun tarpeen perusteella. Kansalliset Red Listit ovat maantieteellisesti keskittyneitä ja niissä on viittaus spatiaaliseen ontologiaan [44].

6. Laji voi olla riippuvainen jonkin toisen lajin esiintymisestä. Tämä riippuvuussuhde on esitetty isAssociatedWith-suhteella.

7. Luotettavuus-luokka (Reliability) jakaa havaitsijat kolmeen eri rooliin perustuen heidän osaamistasoonsa.

TaxMeOn-metaontologiaa sekä sen laajennosta peilaten ja apuna käyttäen tehtiin tätä työtä varten oma skeema. Käytetty skeema esitellään tarkemmin osiossa 3.4.

3.3 Maailman lintujen suomenkieliset nimet -ontologia

Maailman lintujen suomenkieliset nimet -ontologia on TaxMeOn-metaontologian avulla esitetty ontologia maailman linnuista [52]. Ontologiaa on käytetty tässä työssä suomenkielisten nimien ja hierarkian tiedon käsittelyssä. Ontologia perustuu BirdLifen maailman lintujen suomenkielisten nimien luetteloon [10].

3.4 Aineiston kerääminen ja käsittely

Tässä työssä aineistoiksi valittiin Hatikan [34] sekä Tiiran [9] lintuhavaintoaineistot. Molemmat aineistot ladattiin GBIF-portaalin kautta [20]. Aineistot valittiin pääosin helpon saatavuuden sekä kattavuuden vuoksi. Hatikan aineistosta tehtiin oletus sen olevan epäluotettavampaa johtuen havaitsijoiden suuremmasta kirjoista ja erityisesti vaihtelevasta ammattitaidosta. Aineistot ladattiin CSV-formaatissa (Comma Separated Value), joka on pilkuilla erotettu tekstitiedostoformaatti. Hatikka-aineisto sisälsi noin 30 000 sekä Tiira noin 250 000 havaintoa linnuista.

GBIF-portaalista saadun Hatikan datan kentät ja yhden esimerkkihavainnon kenttien arvot ovat kuvattuna taulukossa 3.1 ja 3.2. Kaikki saman aineiston havainnot noudattivat vastaavaa formaattia. Kentät, joissa esimerkkihavainnon arvot olivat tyhjiä, on jätetty pois taulukosta. Tiiran data oli keskeisiltä osiltaan samannäköistä.

Alkuperäinen Hatikan aineisto sisältää 250 000 havaintoa, joka sisältää myös nisäkkäiden havaintoja. 50 000 lintuahavaintoa sisältävä osuus sisälsi myös joitain Suomen ulkopuolella esiintyviä lajeja. Ilmoitettua lajinimeä ei ollut mitenkään rajoitettu vain Suomessa esiintyville lajeille. Lajisto kattaa koko Suomen, mutta oletettavasti myöskään WGS84-koordinaateille (World Geodetic System) ei ole ollut minkäänlaista tarkistusta. Aineistossa havaittiin ainakin muutamia Suomen ulkopuolisia havaintoja ennen kuin sitä käsiteltiin mitenkään.

Aineisto muutetaan suoraviivaiseen RDF-formaattiin, jota oli helppo käsitellä ja johon valittiin keskeisimpiä ominaisuuksia, joita tultaisiin käyttämään ai-

neiston pohjalta tehtävässä analysoinnissa. Näitä ovat muun muassa paikka, havaitsija, aika ja joitain muita näitä tukevia tietoja. Aineistoon tehdään useita erilaisia suodatuksia, joilla karsitaan pois virheellisiä havaintoja.

Tämän lisäksi kaikki Suomen ulkopuoliset havainnot poistetaan sen perusteella ovatko niiden WGS84-koordinaatit sellaisen suorakaiteen muotoisen kappaleen ulkopuolella, joka sisältää Suomen rajat. Joistain havainnoista kyseiset koordinaatit puuttuvat kokonaan. Jos näin on, mutta kyseinen havainto on ilmoitettu tehdyksi jossain Suomen kunnassa, haetaan Googlen geolokaatiorajapinnasta [21] sille kuuluvat WGS84-koordinaatit. Näin dataa saadaan rikastettua ja sen laatua parannettua. Laskenta perustuu täysin WGS84-koordinaatteihin, jotka sitovat havainnon yksiselitteisesti johonkin pisteeseen maapallolla.

Ajantasainen lista Suomessa esiintyvistä lajeista on saatavilla BirdLifen ylläpitämästä listasta [11]. Listan mukaan Suomessa esiintyy tällä hetkellä 502 erilaista lajia.

Taulukko 3.1: GBIF-portaalin kentät ja esimerkkihavainnon kenttien arvot.

GBIF-kenttä	**Esimerkkihavainnon kentän arvo**
Data publisher	Finnish Museum of Natural History
Dataset	Hatikka Observation Data Gateway
Dataset Rights	All rights reserved by FMNH and creators of the:.
Collector name	HARRI PÄIVÄRINTA
GUID	urn:lsid:luomus.fi:MZH.Hatikka: 395F9117-AD0F-4C4E -8220-4157A6201C5D
Date collected	2006-12-29 00:00:00.0
Institution code	MZH
Collection code	Hatikka
Catalogue No	395F9117-AD0F-4C4E-8220-4157A6201C5D
Basis of record	Observation
Last indexed	2011-03-24 19:00:06.0
Identification date	2006-12-29 00:00:00.0

Aineisto ripustetaan tieteellisen nimen perusteella TaxMeOn-metaontologian avulla kuvattuun AVIO-ontologiaan, josta saadaan luokittelu, voimassaoleva tieteellinen nimi sekä kansankieliset nimet suomeksi tai englanniksi. AVIO-

Taulukko 3.2: GBIF-portaalin kentät ja esimerkkihavainnon kenttien arvot.

GBIF-kenttä	**Esimerkkihavainnon kentän arvo**
Scientific name	Carduelis spinus
Scientif. name (interpreted)	Carduelis spinus
Kingdom	Animalia
Phylum	Chordata
Class	Aves
Country	FI
Country (interpreted)	Finland
Locality	KAANAA-PIRIL
County	Raisio
Continent or Ocean	.
State/Province	.
Region	Northern Europe
Publisher country	Finland
Latitude	60.4586
Longitude	22.1783
Coordinate precision	10000
Cell id	54202
Centi cell id	41
Min depth	0
Max depth	0
Min altitude	0
Max altitude	0
GBIF portal url	http://data.gbif.org/occurrences/231993242
GBIF webservice url	http://data.gbif.org/ws/rest/occurrence/get?key=231993242

ontologiasta on poistettu Suomen linnustoon kuulumattomat lajit. Tämä on tehty erityisesti siksi, että ei ole olemassa mitään ontologiaa, joka mallintaisi vain Suomen linnustoa. Näin ollen muun muassa käyttöliittymän auto-

maattista täydennystä käytettäessä ei voida kysyä palvelimelta pelkästään Suomen lintujen nimiä, vaan joudutaan kysymään koko ontologian taksonilistaus. Tehokkuus- sekä käytettävyyssyistä tyydytään siis rajoittamaan koko ontologia kattamaan vain Suomen lintujen nimet ja hierarkian.

Paikkojen nimet on ripustettu Suomen ajalliseen paikkaontologiaan (SAPO) [26], josta löytyvät kaikki Suomen kunnat muutoksineen viimeisen 150 vuoden ajalta. Ontologiaan sitominen mahdollistaisi esimerkiksi vanhojen havaintojen haun uudempien paikannimien perusteella. Ontologian tarjoamia mahdollisuuksia ei kuitenkaan toteutettu käyttöliittymässä, koska havaintokannan havainnot eivät ole kovin monen vuoden takaa ja siksi niiden ei todettu hyödyntävän lintutieteilijöitä

Näiden ontologioiden pohjalta tehtiin mukautettu malli havainnoille, jossa yksi havainto sisältää taulukossa 3.3 esiintyvät kentät. RDF-muunnettua aineistoa on havainnollistettu kuvassa 3.3.

Kaikki RDF-elementit ovat http://www.hatikka.fi/havainnot/ -nimiavaruuden alla, joka on kuvassa 3.3 ja taulukossa 3.3 lyhennetty hh:ksi. hh:231980154 kuvaa RDF-resurssia, joka on tyyppiä (rdf:type) hh:Observation, joka määrittää, että kyseessä on havainto.

Edellä mainitun lisäksi havainto saa tekstimuotoisen rdfs:label-kentän, jossa on havainnon tieteellinen nimi, paikka, sekä vuosi. Havainnon kerääjä on määritetty resurssiksi, jotta samasta havaitsijasta saadaan yksikäsitteinen esitys, johon voidaan viitata. Havainnon paikka on määritelty WGS84-koordinaateilla ja sen lisäksi on tehty resurssi hh:county, joka liittää havainnon Suomen kuntaan tai kaupunkiin.

Havainnon päivämäärä on mallinnettu päivämäärille tarkoitetulla xsd:date-tietotyypillä. Havainto sidottiin AVIO-lintuontologiaan sen yksikäsitteistä lajintunnistusta varten. Se on määritetty sekä hh:scientific_name että rdf:type -ominaisuuksilla. Tämä resurssi on kuvassa nimellä bio:FMNH_381659.

Kuvassa esiintyvät hh:linearTime ei ole oleellinen kenttä, vaan se liitettiin aineistoon visualisoinnin helpottamista varten, koska käytetty visualisointityökalu ei aluksi osannut laittaa aikaa lineaariseen järjestykseen.

Havainnot siis liitetään edellä mainittuun lintuontologiaan. Tämän lisäksi niitä rikastetaan joidenkin TaxMeOn-laajennosten mukaisesti. Lajeihin myös liitetään niiden yleisimpiä tuntomerkkejä. Tuntomerkit saadaan lataamalla Luontoportin lintutunnistuspalvelusta kaikkia tuntomerkkihakuja vastaavat lintulistaukset ja niiden perusteella jokaiselle TaxMeOn-lajilistauksesta löytyvälle lajille lisätään niiden tuntomerkit. Myös tuntomerkit mallinnetaan RDF-resursseiksi ja niille tehdään Luontoporttia vastaava yksinkertai-

Taulukko 3.3: Tehdyn havaintoskeeman kentät ja niiden sallitut arvot

kenttä	**arvot**
rdf:type	hh:Observation
rdf:type	viittaus AVIO-ontologian resurssiin
rdf:label	tieteellinen nimi, paikka ja vuosi tekstinä
hh:collector	havaitsijaresurssi
hh:country	kunta tai kaupunki
hh:date_collected	tyyppiä xsd:date oleva päivämäärä
hh:scientific_name	viittaus AVIO-ontologian resurssiin
wgs84_pos:lat	WGS84-tyyppiä oleva leveysastekoordinaatti
wgs84_pos:long	WGS84-tyyppiä oleva pituusastekoordinaatti

```
hh:231980154
      rdf:type hh:Observation , hh:day01month03year2008 , bio:FMNH_381659 ;
      rdfs:label "Carduelis chloris, Espoo, 2008" ;
      hh:collector hh:sassiolli ;
      hh:county hh:espoo ;
      hh:date_collected "2008-03-01"^^xsd:date ;
      hh:linearTime hh:day01month03year2008 ;
      hh:scientific_name bio:FMNH_381659 ;
      wgs84_pos:lat "60.1676" ;
      wgs84_pos:long "24.7452" .
```

Kuva 3.3: Yksi havainto RDF-muodossa, kirjoitettuna Turtle-notaatiota [1] käyttäen

nen skeema, joka sisältää niiden keskinäisen hierarkian ja hierarkiatasot.

Laji myös määritellään yleiseksi sen perusteella löytyykö se Luontoportista vai ei. Tätä listaa täydennetään lintukirjoista löytyvien tietojen perusteella [31] [39]. Tuntomerkkien lisäksi Luontoportista saadaan myös lajien habitaatit eli elinympäristöt ja ne lisätään lajeille. Yhdellä lajilla voi olla usempia elinympäristöjä. Näin olleen lajin elinympäristöjä voidaan verrata käyttäjän syöttämiin ja sen perusteella tehdä tulkintoja joko suoraan tai osana bayesilaista analyysiä, josta lisää alaluvussa 2.4. Luontoportista löytyy tietoja vain yleisimmille lajeille, joita on 256 kappaletta. Muille lajeille elinympäristöt lisätään lintukirjojen tietojen perusteella [31] [39]. Elinympäristöt voitaisiin myös ripustaa EU:n habitaattimääritykseen [47] ja sitä kautta taata yhteen-

sopivuus kansainvälisten aineistojen kanssa.

Havaintokannassa olevien havaintojen havaitsijat ovat alkuperäisessä aineistossa tekstimuotoisina kenttinä. Havaitsijoiden luotettavuuden arvioinnin siemenarvoksi valitaan lintututkinnon [25] suorittaneet havaitsijat. Tiedot lintututkinnon suorittajista löytyvät eri lintuyhdistysten sivuilta [24]. Näiden perusteella havaitsijat jaetaan luotettaviin sekä muihin. Havaitsijat tunnistetaan pelkästään nimen perusteella eikä tähän käytetä mitään heuristiikkaa, poislukien etu- ja sukunimen järjestyksen sivuuttaminen. Ei ole myöskään mitään triviaalia keinoa etu- ja sukunimen erottamiseen.

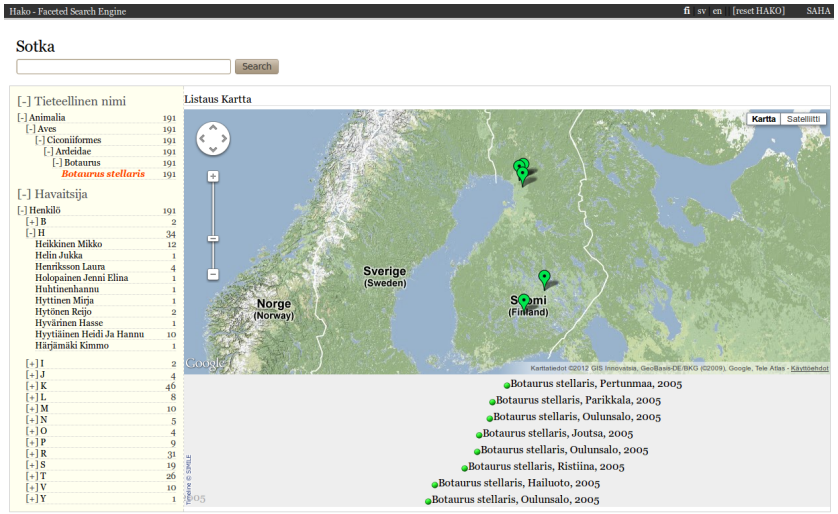
3.4.1 Datan visualisointi

Tiedon ollessa mallinnettu semanttisesti, semanttisen webin työkalut ja sovellukset voivat hyödyntää sitä ilman mitään lisämuunnoksia tai -vaivaa. Aaltoyliopiston semanttisen laskennan tutkimusryhmässä on kehitetty useita semanttisen webin työkaluja sekä sovelluksia. Yksi näistä on HAKO-hakukone [30], joka on kevyt hakukone semanttisesti mallinnetulle tiedolle, joka tukee useita samanaikaisia hakufasetteja. HAKO tukee myös Googlen karttanäkymää ja RDF:n selausta. Hakukone on SAHA-metadataeditorin [54] rinnalle tehty haku- ja visualisointityökalu. SAHA:sta ja HAKO:sta kerrotaan lisää alaluvussa 4.2.

RDF-aineisto ladataan HAKO:on selainkäyttöliittymän avulla. HAKO konfiguroidaan ensimmäisellä käyttökerralla. HAKO:n konfigurointi-ikkunasta valitaan mitä objektien ominaisuuksia halutaan käyttää hakusuodatuksissa, samoin mitä *instansseja* halutaan näyttää selainikkunassa. Kuvassa 3.4 näkyy HAKO:n käyttöliittymä silloin, kun sinne on ladattu lintuaineistot ja niiden skeeman sisältävä RDF-data. Koska havainnot ovat sidottuja TaxMeOn-lajilistauksiin, HAKO osaa näyttää koko lajihierarkian kaikille havainnoille sekä havaintojen lukumäärän laji- tai muulla hierarkiatasolla.

Käyttöliittymä mahdollistaa hierarkian avaamisen tai supistamisen tarpeen mukaan. Tämä näkyy kuvan vasemmassa reunassa keltaisella pohjalla. Karttanäkymä on suurennettavissa ja panoroitavissa sekä harmaalla pohjalla oleva aikajana on myös liikuteltavissa. Aikajanalla olevat tieteelliset nimet ovat linkkejä kartalla näkyviin pisteisiin.

Koska RDF-aineistossa on semanttisesti annotoituja WGS84-koordinaatteja, HAKO näyttää oletusarvoisesti Google Maps -karttanäkymän ja sijoittaa havainnot siihen. Näkymä toteuttaa Google Maps -karttapalvelun kaikki perustoiminnallisuudet, kuten zoomaus, panorointi, satelliittikuva ja niin edel-



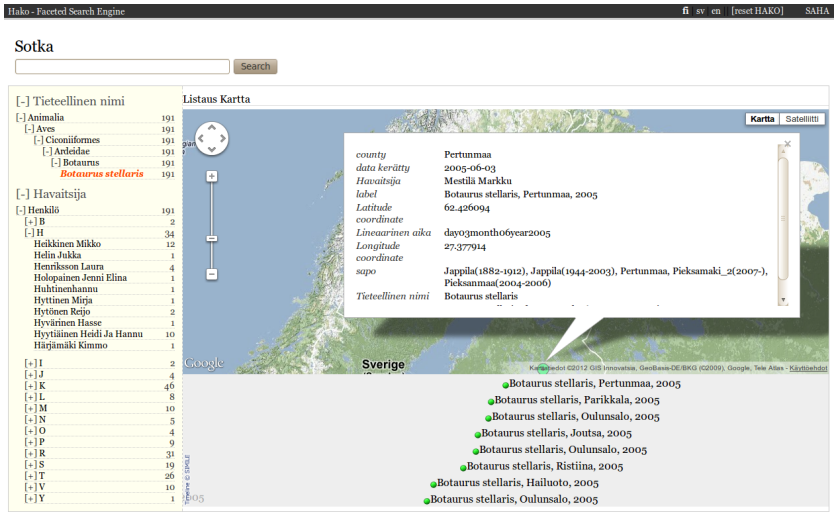
Kuva 3.4: HAKO-hakukone

leen. Tämän lisäksi kartan havaintopisteet toimivat linkkeinä ja näyttävät semanttisesti annotoituja tietoja havainnoista. Tätä on havainnollistettu kuvassa 3.5.

Tämän lisäksi karttapohjan alla on dynaaminen aikajana, jota voi liikuttaa. Tällöin havaintopisteet päivittyvät kartalle sen perusteella osuvatko ne aikajanan näkyvälle välille. Aikajanaa voidaan myös zoomata sisään ja ulos jolloin voidaan muuttaa tarkastelun kohteena olevan aikavälin pituutta. Näin voidaan esimerkiksi käsin aikajanaa liikuttamalla simuloida yhden tai useamman lajin havaintojen muutosta ajan suhteen. Tämä antaa käsitystä esimerkiksi siitä, miten havainnot muuttuvat vaikkapa kevät- tai syysmuuton yhteydessä. HAKO:n karttavisualisointi osaa piirtää sekä pisteitä että useista pisteistä koostuvia monikulmioita.

3.4.2 Ajallinen ja paikallinen ulottuvuus

Kaikki havaintodata on keskeisesti riippuvaista ajasta ja paikasta. Näitä muuttujia tutkimalla voidaan löytää muun muassa eri lajien muuttoaaltoja. Havaintopisteiden ympärille piirretyn pienimmän mahdollisen monikulmion menetelmä on kansainvälisesti hyväksytty menetelmä eri lajien levin-



Kuva 3.5: HAKO-hakukone, havainnon tiedot

neisyyksien tutkimiseen [16]. Yksinkertaistettua versiota tästä menetel-mästä käytetään tässä työssä, jossa pisteiden ympärillä rajataan nelikulmio. Kaikkien luotettavien havaitsijoiden havaintojen, jotka ovat tietystä lajista kuukauden aikana, ympärille piirretään nelikulmio, joka kuvaa lajin mahdollista oleskelualuetta kyseisen kuukauden haarukassa.

Nämä arvot interpoloidaan joka päivälle edeltävän ja seuraavan kuukauden arvoista. Havaintojen määrää ei tässä pidetä hyvänä mittarina lajin havaitsemistodennäköisyydestä, sillä vain tavanomaisisti lajeista on runsaasti havaintoja. Uuden havainnon tilanteessa voidaan tutkia kyseisen havainnon koordinaattien osumista tälle lajille lasketun nelikulmion sisälle. Näiden nelikulmioiden käyttäytymistä voidaan myös tutkia HAKO-visualisaattorilla. HAKO-hakukoneesta lisää seuraavassa luvussa.

3.4.3 Bayesilaisen luokittimen käyttö havaintojen validoinnissa

Huolimatta naiivin bayesin yliyksinkertaistetusta mallista se toimii luokittimena hyvin [60]. Bayesilaista luokitinta käytetään demonstraattorissa havaintojen luokittelemiseksi tietyksi lajiksi tiettyjen muuttujien perusteella.

Luku 4

Havaintopalvelu

Tässä luvussa kerrotaan kuinka datan suodatuksen ja esikäsittelyn jälkeen päästään data-analyysistä laskennan tuloksiin. Tässä esitellään kahden erilaisen mallin toteutus haluttujen tietojen laskentaan.

4.1 Laskenta

Havaintoja tukevaa informaatiota on mahdollista lisätä dataan joko laskemalla tämä informaatio etukäteen ja päivittämällä kantaan tai käyttämällä reaaliaikaista laskentaa. Ensimmäisessä menetelmässä dataan lasketaan etukäteen tietoa lintujen levinnäisyysalueista, jotka sitten iteroidaan ajallisesti koko kuukauden käsittäviksi. Toisessa menetelmässä havaintokannan muuttujista luodaan bayesilainen todennäköisyysjakauma, josta voidaan reaaliajassa kysyä testihavainnon todennäköisyysjakauma eri luokille perustuen bayesilaiseen todennäköisyysmalliin. Menetelmien tuloksien arvioinnista lisää seuraavassa luvussa. Yu et al. [59] ovat esitelleet menetelmän, joka perustuu samantapaiseen havaitsijoiden luotettavuuden mallintamiseen.

4.1.1 Monikulmioiden generointi

Lintujen havaintoalueiden laskemiseksi päätettiin havaintojen pohjalta generoida monikulmioita, jotka määrittävät alueen, jolla tiettyä lajia on havaittu. Monikulmiot ovat koordinaattipisteiden rajaamia alueita kartalla. Järjestelmä voi näin olleen katsoa onko käyttöaineiston havainto tietyltä lajilta, opetusaineiston pohjalta lasketun monikulmion sisällä. Tästä voidaan päätellä havainnon olevan luotettava.

Monikulmioiksi valittiin yksinkertaisuuden vuoksi vain nelikulmio. Tämä liitetään havaintoon hasPolygon-ominaisuuden avulla. Ominaisuuteen liitetään nelikulmion koordinaatit. Nämä ovat liitettyjä kuhunkin lajiin, niille päiville kun lajia on nähty Suomessa. Laskenta on toteutettu Java-ohjelmana, koska Javalle löytyy tehokas Jena-kirjasto [48], jolla voi helposti käsitellä suuren määrän RDF-dataa. Nelikulmioiden laskennat kiinnitettään kuhunkin kuukauteen niin että edellisen kuun 15 viimeisen päivän sekä kyseisen kuun 15 ensimmäisen päivän havainnot luotettavilta havaitsijoilta otetaan huomioon. Nämä havainnot ympäröidään siten pienimmällä mahdollisella nelikulmiolla. Tämä nelikulmio liitetään kyseisen kuun ensimmäisen päivän nelikulmioksi. Kaikille kuun muille päiville arvot iteroidaan ottamalla kyseisen päivän kummallakin puolella olevat nelikulmiot ja laskemalla niiden koordinaateista painotettu keskiarvo, painokertoimen ollessa se kuinka kyseinen päivämäärä sijoittuu suhteessa raja-arvojen päivämääriin. Näin olleen uuden havainnon luotettavuutta voidaan arvioida vertaamalla onko havainto kyseiseltä päivältä edellisten vuosien saman päivän nelikulmion sisällä ja siten päätellä että kyseessä on todennäköinen havainto.

Koska havaintodataa ei välttämättä ole jokaiselle päivällä, on iterointi syytä tehdä edellä mainitusti, jotta saadaan järkeviä arvoja, jotka ovat lineaarisessa suhteessa toisiinsa. Havaitsijaa pidetään luotettavana, jos hän on suorittanut lintututkinnon. Havaitsijan luotettavuutta voidaan myös arvioida sillä ovatko hänen havaintonsa edellä mainitun menetelmän generoimien nelikulmioiden sisällä. Tällä tavalla voidaan iteratiivisesti menetelmää toistamalla parantaa havaintokannan laatua luokittelemalla useampia havaitsijoita luotettaviksi. Kun uusia luotettavia havaitsijoita ei enää synny, on menetelmä saavuttanut päätepisteensä eikä se voi enää parantaa kannan laatua. Tähän dataan perustuen voidaan myös arvioida kaikkien havaintojen laatua koko kannassa.

Iteraation tuloksena saadaan RDF-dataa, joka sisältää jokaisen Suomessa nähtävän lajin hasPolygon-ominaisuuden, eli nelikulmiomallin lajin havainnoista. Tämä data voidaan ladata SAHA-metadataeditoriin tai muuhun

SPARQL-palveluun [43], josta havaintopalvelun käyttöliittymä voi käydä kysymässä tietyn lajin tietylle päivällä liitettyä nelikulmiota ja tutkia onko uusi havainto tämän alueen sisäpuolella.

4.1.2 Bayesilainen luokitin

Bayesilainen päättelijä toteutettiin Python-ohjelmointikielellä käyttäen Scikit learn -nimistä kirjastoa [41]. RDF:n prosessointiin käytettiin RDFLib-ohjelmointikirjastoa [45]. Kieleksi valitsin Pythonin, koska osaan sitä hyvin ja sille löytyy helppokäyttöisiä kirjastoja, joita hyödyntämällä voi kirjoittaa tehokkaita ohjelmia.

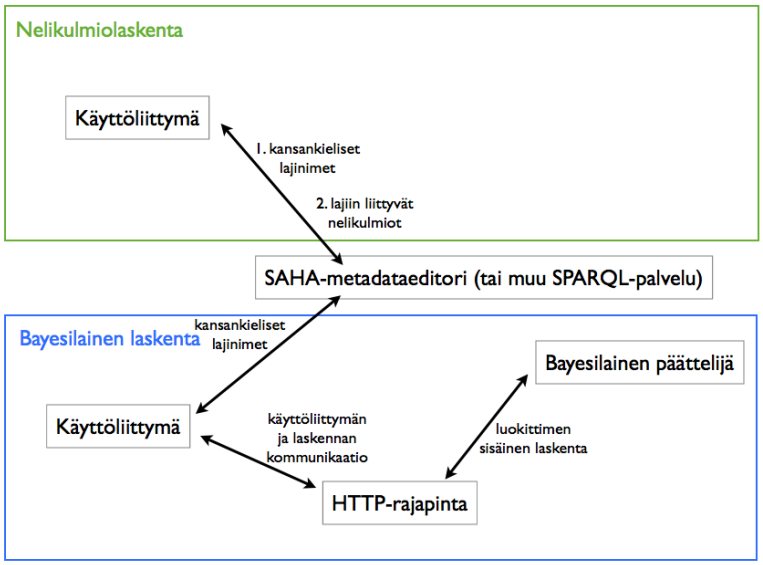
Tässä menetelmässä yritetään luotettavista havainnoista (luotettavien havaitsijoiden havainnoista) saada irti mahdollisimman monta yksilöivää muuttujaa. Näitä muuttujia ovat päivämäärä ja geokoordinaatit. Näiden pohjalta opetetaan bayesilaisen mallin mukainen luokka (luokiksi valitaan taksonit eli lajit) kyseisillä arvoilla. Kun mallille sitten tarjotaan testattavan havainnon vastaavat muuttujat, se laskee todennäköisyysjakauman eri luokille (lajeille). Tätä menetelmää iteratiivisesti toistamalla kaikille havaintokannan havainnoille voidaan tietokannan havaitsijoiden luotettavuutta arvioida ja täydentää tietokantaan. Myöskin tämä malli toimii tehokkaimmin, kun iteraatio päättyy siihen, että uusia luotettavia havaitsijoita ei enää löydetä.

Seuraavassa osiossa esitellään, miten kaikki osat toimivat yhdessä muodostaen demonstraattorin, joka toteuttaa halutut toiminnallisuudet.

4.2 Demonstraattori

Toteutettu demonstraattori sisältää seuraavat osat: 1. Laskennan 2. Palvelinrajapinnan 3. Käyttöliittymän. Näiden lisäksi SAHA-metadataeditoria voidaan käyttää monikulmiolaskennan tulosten editoimiseen sekä jakamiseen käyttöliittymäsovellukselle, esimerkiksi suomenkielisten lajinimien automaattiseen täydentämiseen. HAKO-hakukonetta voidaan käyttää havaintojen visualisointiin tutkimalla havaintoja kartalla erilaisten hakufasettien mukaan rajattuina. Nämä tulokset ovat kuvattu seuraavissa alaluvuissa. Järjestelmän eri osien suhde toisiinsa nähdään kuvassa 4.1.

Kuvassa on esitetty kahden eri menetelmän toteutuksen järjestelmäkaavio.



Kuva 4.1: Kahden eri demonstraattorin järjestelmäkaavio

Järjestelmän eri komponentit on kuvattu harmailla suorakaiteilla, joiden sisällä on komponentin nimi. Musta nuolet kuvaavat järjestelmän eri komponenttien välistä tiedonsiirtoa. Nelikulmiolaskentaan pohjautuvan demonstraattorin komponentit ovat vihreän suorakaiteen sisällä ja bayesilaiseen laskentaan pohjautuvan demonstraattorin sinisen suorakaiteen sisällä. Kummallekin toteutukselle yhteinen komponentti eli SPARQL-palvelu on erillään suorakaiteiden välissä.

Laskennan tuloksena saatu data ladataan SAHA-metadataeditoriin, joka toimii SPARQL-palveluna. Täältä nelikulmiolaskennalla tehdyn toteutuksen käyttöliittymä hakee tiedot sekä lintujen kansankielisistä nimistä, että lasketuista nelikulmioista.

Bayes-laskentaan perustuva järjestelmä toimii muuten samalla tavoin, mutta laskennan tulokset se hakee erillisen HTTP-rajapinnan (Hypertext Transfer Protocol) kautta, joka on yhteydessä bayesilaiseen päättelijään.

4.2.1 SAHA-metadataeditori

Valkeapää et al. ovat kehittäneet SAHA-metadataeditorin [54], johon voidaan ladata RDF-dataa, jossa sitä voidaan editoida ja jossa on SPARQL-palvelu, johon voidaan tehdä kyselyitä liittyen RDF-dataan. Demonstraattorin käyttöliittymä käyttää ontologiapalvelinta muun muassa suomenkielisten lajinimien automaattiseen täyttöön. Nelikulmiolaskentaan perustava demonstraattori käyttää ontologiapalvelinta myös uuden havainnon arviointiin, kysymällä palvelimelta lajin havaintonelikulmiota (hasPolygon -ominaisuutta) ja vertailemalla kyseisen havainnon koordinaatteja siihen.

4.2.2 HAKO-hakukone

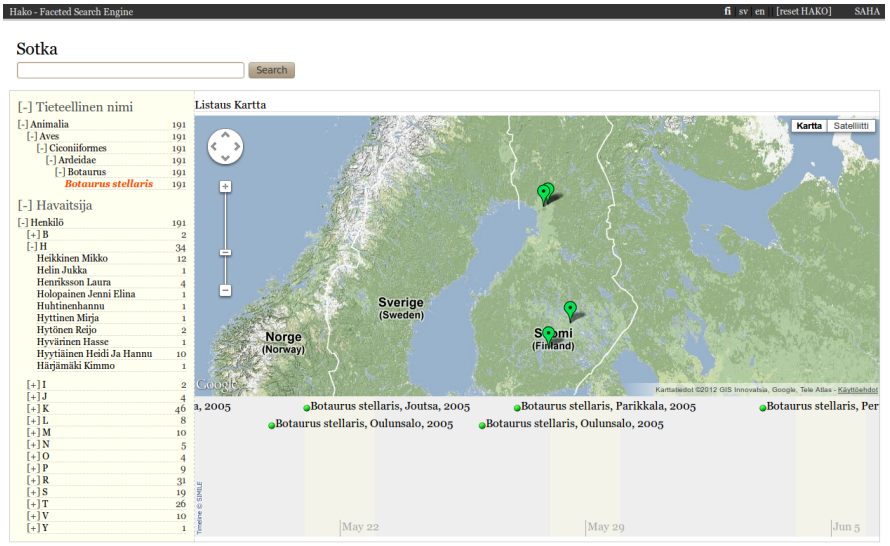
Kurki et al. ovat kehittäneet HAKO-hakukoneen, joka on fasettihakukone sekä visualisaattori [30] SAHA:an ladatulle RDF-datalle. Karttavisualisaatio näyttää havainnot kartalla ja valittaessa myös kontekstivalikon havainnosta. Karttavisualisaatioon liittyy myös aikajanahaku, jossa havaintoja voidaan aikarajalla rajata. Aikajanaa voidaan myös zoomata, kuten kuvassa 4.2.

Kuvan vasemmassa laidassa näkyvän fasettihaun avulla voidaan visualisaatiota rajata muun muassa taksonin koko hierarkian tai havaitsijan suhteen, tulevaisuudessa myös esimerkiksi linnun tuntomerkkien tai paikkojen suhteen tai habitaattien suhteen. Näistä ei kuitenkaaan ole vielä tarpeeksi dataa saatavilla.

4.2.3 HTTP-rajapinta

Bayesilaisen luokittimen toimintaa tukemaan kirjoitettiin HTTP-rajapinta, jota demonstraattorin käyttöliittymä voi käyttää bayesilaisten mallien antamia todennäköisyyksiä varten. Palvelin toteutettiin Python-ohjelmointikielellä käyttäen CherryPy -nimistä kirjastoa [23]. Käyttöliittymän ja HTTP-rajapin- nan kommunikaatio käyttää JSON-formaattia [28] tiedon mallintamiseen.

CherryPy mahdollistaa monisäikeisen HTTP-palvelimen toteuttamisen helposti. Käytännössä havaittiin että 0,005 suuremmat luokkatodennäköisyydet olivat käytännön näkökulmasta uskottavia. Luku 0,005 saatiin tekemällä alustavia laskentoja luotettaviksi tunnetuille havainnoille ja siitä arvioimalla



Kuva 4.2: HAKO-hakukone, fasettihaku, aikajanaa zoomattu

sopivin raja-arvo.

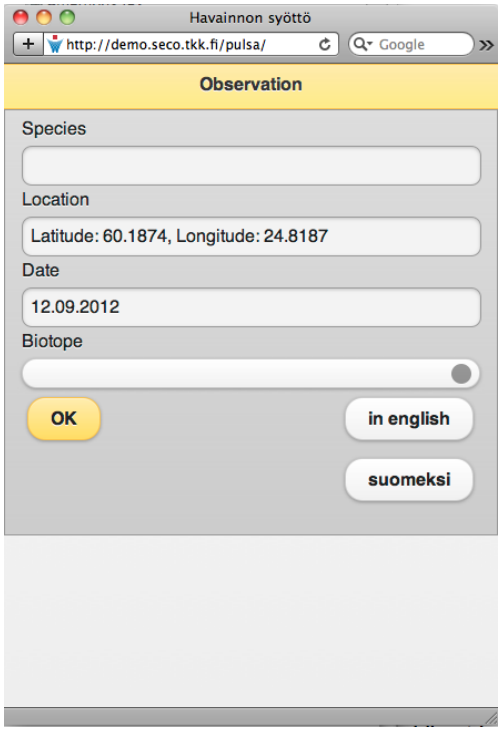
4.2.4 Käyttöliittymä

Käyttöliittymän ulkoasu tehtiin toimimaan tavallisilla nykyaikaisilla älypuhelimilla. Käyttöliittymän etusivu on esitetty kuvassa 4.3. Käyttöliittymä hakee käyttäjän sijainnin geokoordinaatit Googlen Geolocation -rajapinnasta [21]. Tämän toiminta on esitetty kuvassa 4.3. Geolocation-rajapinta palauttaa paikan geokoordinaatit paikan sijainnille varattuun kenttään kuten kuvassa 4.4 nähdään.

Käyttöliittymästä tehtiin kaksi erilaista versiota, jotka eroavat käytännössä vain siinä, miten havainnon todennäköisyys esitetään käyttäjälle. Bayesilaisen laskennan tapauksessa käyttäjälle esitetään luokkatodennäköisyys kyseiselle lajille. Nelikulmiolaskentaan perustuvassa versiossa käyttäjälle esitetään vain onko laji validi vai ei, perustuen siihen onko havainto nelikulmion sisällä vai ei. Käyttöliittymä sisältää myös biotoopin valintaan liittyvän painikkeen, mutta ominaisuutta ei käytetty lopullisessa versiossa, sillä biotooppitieto ei

ollut saatavissa riittävälle määrälle lajeja.

Kuva 4.3: Mobiilikäyttöliittymä, sijainnin automaattinen haku

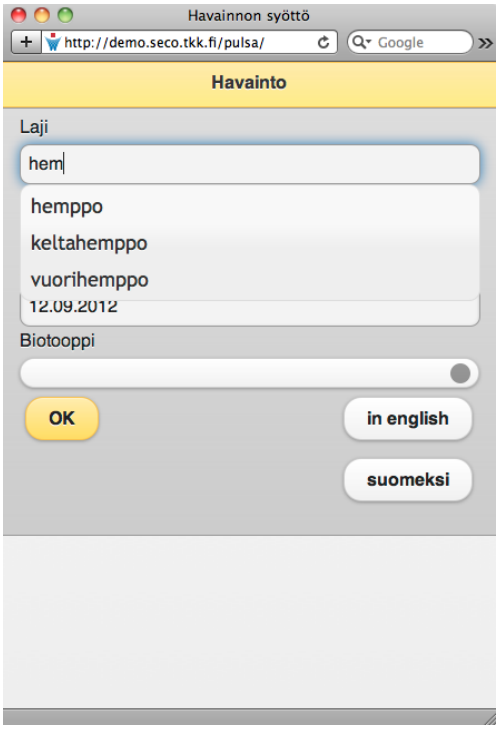


Kuva 4.4: Mobiilikäyttöliittymä, paikannuksen tulos

Käyttöliittymän on tehnyt Mikko Koho tukemaan demonstraattorin toimintaa. Käyttöliittymä on tehty käyttäen Twitterin mobiilisovellusten kehittämiseen tarkoitettua Bootstrap-ohjelmointikirjastoa [53]. Käyttöliittymä käyttää SAHA-metadataeditoria kansankielisten lajinimien esittämiseen sekä lajinsyötön automaattiseen täydentämiseen. Tämä on esitetty kuvassa 4.5. Demonstraattorin nelikulmiolaskentaan perustuva demonstraattori käyttää kyseistä palvelinta myös havainnon lajiin liittyvän nelikulmion hakemiseen.

Demonstraattorin käyttöliittymän toinen versio käyttää laskennan tulosten hakemiseen Pythonilla toteutettua HTTP-rajapintaa. Käyttöliittymän kieltä voidaan vaihtaa suomen ja englannin välillä. Käyttöliittymä sisällyttää tuloksiin myös linkit Wikipediaan sekä Luontoporttiin, jos kyseisen lintulajin tiedot löytyvät palveluista. Käyttöliittymä antaa tulokseksi lajin todennäköisyyden kaikkien lajien joukosta kyseisenä aikana sekä kyseisessä pisteessä. Todennäköisyys esitetään lukuarvona ja se on ehdollinen todennäköisyys sille,

että kyseessä on ehdotettu laji, ehtojen ollessa paikan ja ajan määrittävät muuttujat. Uskottavalle havainnolle määritettiin, että sen todennäköisyyden on oltava yli 0,005.



Kuva 4.5: Mobiilikäyttöliittymän automaattinen täydennys lajinimille

Kuva 4.6: Mobiilikäyttöliittymä kertoo laskennan tuloksen

Käyttöliittymä varoittaa helposti sekaisin menevistä lajeista, jos tämä tieto on ontologiassa kyseille lajille ilmoitettu. Edellä mainitut toiminnot on havainnollistettu kuvassa 4.6.

Luku 5

Arviointi

Mallia kehitettäessä on syytä yrittää välttää liikaa monimutkaisuutta sekä varmistua siitä, ettei biologisen mallin kehittäminen ja soveltaminen ole rajoittavaa tai johda virhepäätelmiin kuten monissa biologisten mallien analysoinneissa on käynyt [29].

Seuraavissa osioissa arvioidaan ratkaisumallin toimintaa normaaleja käyttötilanteita vastaavissa tilanteissa sekä tekaistuilla arvoilla. Myös laskennan vaikutusta havaintokantojen laatuun pohditaan. Ratkaisumalleja tarkastellaan niin, että käytettäväksi lintukannaksi valitaan sekä Hatikan että Tiiran aineistot erikseen. Molemmilla aineistoilla arvioinnit tehdään kumpaakin laskentamallia käyttäen. Lisäksi selvitetään tunnistaako järjestelmä muuttoaaltoja. Lopuksi arvioidaan järjestelmän toimintaa tutkimuskysymysten näkökulmasta.

5.1 Havaintokantojen laatu

Taulukoissa 5.1, 5.2, 5.3 ja 5.4 on esitetty käytetyn lintutietokannan tila koskien luotettavien havaitsijoiden ja havaintojen määrää laskennan alussa, kun luotettavan havaitsijan tunnistamiseen on käytetty pelkästään lintututkinnon suoritusta, ja lopussa, kun laskennan iteraatio on suoritettu loppuun.

Taulukoissa luotettavaksi havaitsijaksi määritellään havaitsija, jonka havain-

noista vähintään yksi kullakin menetelmällä luokitellaan luotettavaksi. Nelikulmiolaskennan tapauksessa tämä tarkoittaa havaitsijan havainnon osumista lasketun nelikulmion sisään ja Bayes-laskennan tapauksessa lajin luokkatodennäköisyyden olevan yli raja-arvon 0,005. Luotettavaksi havainnoiksi määritellään havainto, joka on luotettavan havaitsijan havaitsema.

Lopputilanne kuvaa molempien laskentojen tilannetta, jossa iteraatiot ovat saavuttaneet saturaatiopisteensä, eikä uusia luotettavia havaitsijoita enää pystytä löytämään. Hatikan tietokannassa siemenarvona havaitsijan luotettavuudelle kummassakin laskentatapauksessa on käytetty vain lintututkinnon suorittaneiden listaa. Tiiran tapauksessa havaitsijoina ei ollut yksittäisiä henkilöitä vaan lintuyhdistyksiä, joita oli 20 kappaletta. Tämän lisäksi joukossa oli tuntemattoman havaitsijan tekemiä havaintoja. Molemmilla laskentamenetelmillä kaikki tuntemattomien havaitsijoiden tekemät havainnot tulivat iteraation päätteeksi luokiteltua luotettaviksi.

Taulukko 5.1: Hatikan havaintokannan alku- ja lopputila käytettäessä nelikulmiolaskentaa

Tilanne	**Havaitsijat**	**Luotettavat havaitsijat**
Alku	1500 kpl	224 kpl
Loppu	1500 kpl	1411 kpl
Tilanne	**Havainnot**	**Luotettavat havainnot**
Alku	30423 kpl	7147 kpl
Loppu	30423 kpl	30249 kpl

Taulukko 5.2: Hatikan havaintokannan alku- ja lopputila käytettäessä bayesilaista päättelijää

Tilanne	**Havaitsijat**	**Luotettavat havaitsijat**
Alku	1500 kpl	224 kpl
Loppu	1500 kpl	716 kpl
Tilanne	**Havainnot**	**Luotettavat havainnot**
Alku	30423 kpl	7147 kpl
Loppu	30423 kpl	9063 kpl

Taulukko 5.3: Tiiran havaintokannan alku- ja lopputila käytettäessä nelikulmiolaskentaa

Tilanne	**Havaitsijat**	**Luotettavat havaitsijat**
Alku	21 kpl	20 kpl
Loppu	21 kpl	21 kpl
Tilanne	**Havainnot**	**Luotettavat havainnot**
Alku	249998 kpl	206377 kpl
Loppu	249998 kpl	249998 kpl

Taulukko 5.4: Tiiran havaintokannan alku- ja lopputila käytettäessä bayesilaista päättelijää

Tilanne	**Havaitsijat**	**Luotettavat havaitsijat**
Alku	21 kpl	20 kpl
Loppu	21 kpl	21 kpl
Tilanne	**Havainnot**	**Luotettavat havainnot**
Alku	249998 kpl	206377 kpl
Loppu	249998 kpl	249998 kpl

Tiiran havaintotietokanta vaikuttaa olevan itsessään aika koherentti eli luotettavuuden arviointi iteroimalla johtaa melkein koko kannan tulevan luotettavaksi. Laskennan käyttökelpoisuus riippuu pitkälti siitä miten monta lajia tietokanta ylipäätään sisältää ja miten havainnot ovat painottuneet. Tiiran havaintokannassa oli lopputilanteessa vain 21 havaitsijaa, sillä tuntemattomien havaitsijan havainnot yhdistettiin yhden tuntemattoman havaitsijan alle.

Hatikan kannan tapauksessa lajeja koko kannassa ei ylipäätään ole niin paljon esillä kuin Tiiran tapauksessa ja siten laskennan tuloksetkin usean lajin osalta jäävät vajavaisiksi. Näistä lisää seuraavassa osiossa. Käytettäessä nelikulmiolaskentaa huomataan, että Hatikan tietokannasta luotettavaksi havaitsijoiksi arvioidaan melkein koko tietokannan havaitsijat, kun taas Bayes-laskennan tapauksessa luku jää 716:een. Tämä voi kertoa nelikulmiolaskennan ylimalkaisuudesta siinä suhteessa, että kunkin lajin havaintonelikulmioiksi tulee suuret alueet, jos havaintoja on kootusti suurelta alueelta. Laskettuja alueita

ei siis nelikulmiolaskennan tapauksessa pilkota pienempiin.

5.2 Kenttäarviointi

Kenttätestaus toteutettiin lintuharrastaja Mikko Kohon avustuksella Laajalahden lintutornilla. Tunnistimme Kohon kanssa kymmenen eri lintua. Koho tarjosi myös listan linnuista, joita tällä alueella ei tähän aikaan nähdä. Näitäkin oli kymmenen kappaletta. Lajit ovat taulukoituna taulukossa 5.5.

Havaintoja kaikista testatuista lajeista oli Tiiran havaintokannassa yhteensä ainakin tuhat kappaletta, useista yli 10 000. Kaikista havaituista lajeista, joita Hatikan tietokannan pohjalta tehty laskenta ei tunnistanut oikeiksi, ei ollut ainuttakaan havaintoa koko tietokannassa. Lajeista, joita alueella ei pitäisi nähdä, oli myös Tiiran datassa ainakin tuhat havaintoa per laji. Hatikan datassa ei kyseisistä lajeista ollut yhtään havaintoja, pois lukien laulurasta, joita oli 54 kappaletta.

Taulukko 5.5: Laajalahden lintutornin havainnot, Laajalahti, Espoo

Havaittu laji	**Laji, jota alueella ei tähän aikaan nähdä**
laulujoutsen	pajulintu
kanadanhanhi	haarapääskynen
kyhmyjoutsen	räystäspääsky
merilokki	sitruunavästäräkki
räkättirastas	leppälintu
tukkasotka	lapasorsa
mustarastas	laulurastas
telkkä	pikkusirri
varis	pikkusieppo
isokoskelo	punajalkaviklo

Kumpikin versio laskentamallista antoi näiden suhteen lupaavia tuloksia. Datalähteissä sen sijaan oli eroja. Hatikan tietokantaa käytettäessä kumpikaan järjestelmä ei tunnistanut puoliakaan havainnoista oikeiksi.

Tiiran tietokantaa käyttämällä saatiin molemmissa menetelmissä todella hy-

viä tuloksia. Esimerkiksi nelikulmiolaskentaan perustuva malli Tiiran datalle antoi kaikille kymmenelle havainnolle oikean tuloksen eli luokitteli ne oikeiksi. Vastaavasti vääristä havainnoista se luokitteli yhdeksän kymmenestä oikein eli epävalideiksi havainnoiksi. Hatikan mallissa kummallakin laskentatavalla positiivisten tulos oli sama: vain kolme tunnistettiin valideiksi havainnoiksi.

Bayesilaiseen laskentaan perustuva malli luokitteli Tiiran dataan perustuen myös kaikki oikein, mutta vääristä havainnoista myös yhden virheellisesti validiksi. Se oli tosin lähellä raja-arvoa, jolla se olisi luokiteltu epävalidiksi.

Taulukoissa 5.6. ja 5.7. ovat kuvattuina kaikki nelikulmiolaskennan tulokset. Bayesilaisen laskennan vastaavat tulokset ovat kuvattuina taulukoissa 5.8. ja 5.9. Bayesilaisen laskennan tapauksessa validina havaintona pidetään havaintoa, jossa sen saama luokkatodennäköisyys on suurempi tai yhtä suuri kuin 0,005. Tämä luku määritettiin päättelemällä etukäteislaskentojen pohjalta.

Taulukko 5.6: Laajalahden lintutornin nelikulmiolaskennan tulokset, havainnot

Havaittu laji	**Hatikan havaintokanta**	**Tiiran havaintokanta**
laulujoutsen	ei validi	validi
kanadanhanhi	validi	validi
kyhmyjoutsen	ei validi	validi
merilokki	ei validi	validi
räkättirastas	ei validi	validi
tukkasotka	validi	validi
mustarastas	ei validi	validi
telkkä	validi	validi
varis	ei validi	validi
isokoskelo	ei validi	validi
Yhteensä oikein	3/10	10/10

Taulukko 5.7: Laajalahden lintutornin nelikulmiolaskennan tulokset, väärät havainnot

Väärä havainto	**Hatikan havaintokanta**	**Tiiran havaintokanta**
pajulintu	ei validi	ei validi
haarapääskynen	ei validi	ei validi
räystäspääsky	ei validi	ei validi
sitruunavästäräkki	ei validi	ei validi
leppälintu	ei validi	ei validi
lapasorsa	ei validi	validi
laulurastas	ei validi	ei validi
pikkusirri	ei validi	ei validi
pikkusieppo	ei validi	ei validi
punajalkaviklo	ei validi	ei validi
Yhteensä oikein	10/10	9/10

Taulukko 5.8: Laajalahden lintutornin Bayes-laskennan tulokset, havainnot

Havaittu laji	**Hatikan havaintokanta**	**Tiiran havaintokanta**
laulujoutsen	ei validi (0.000)	validi (0.062)
kanadanhanhi	validi (0.011)	validi (0.006)
kyhmyjoutsen	ei validi (0.000)	validi (0.011)
merilokki	ei validi (0.000)	validi (0.006)
räkättirastas	ei validi (0.000)	validi (0.032)
tukkasotka	validi (0.029)	validi (0.017)
mustarastas	ei validi (0.000)	validi (0.006)
telkkä	validi (0.055)	validi (0.023)
varis	ei validi (0.000)	validi (0.005)
isokoskelo	ei validi (0.000)	validi (0.040)
Yhteensä oikein	3/10	10/10

Taulukko 5.9: Laajalahden lintutornin Bayes-laskennan tulokset, väärät havainnot

Väärä havainto	**Hatikan havaintokanta**	**Tiiran havaintokanta**
pajulintu	ei validi	ei validi
haarapääskynen	ei validi	ei validi
räystäspääsky	ei validi	ei validi
sitruunavästäräkki	ei validi	ei validi
leppälintu	ei validi	ei validi
lapasorsa	ei validi	validi
laulurastas	ei validi	ei validi
pikkusirri	ei validi	ei validi
pikkusieppo	ei validi	ei validi
punajalkaviklo	ei validi	ei validi
Yhteensä oikein	10/10	9/10

5.3 Muuttoaaltojen tunnistaminen

Koska käytämme lintuhavaintokantoja kahdesta eri aineistosta, on mielenkiintoista nähdä onko näissä eroja. Erityisesti koska Hatikan aineisto on kansalaisten keräämää sekalaista aineistoa ja Tiiran taas lintuyhdistysten kirjaamia havaintoja, pitäisi niiden olla eri tavalla painottunutta. Tiirassa esimerkiksi on juuri muuttojen ajalta enemmän ja parempia havaintoja.

Järjestelmien testaamiseksi valittiin kymmenen lajin lista Suomessa esiintyvistä lintuharrastajan näkökulmasta mielenkiintoisista muuttolinnuista [8]. Lintujen muuttotiedot tarkistettiin Luontoportista [35]. Tiedot ovat taulukoituna taulukkoon 5.10 Tämän lisäksi valittiin tarkastelukaupungiksi Helsinki, joka on suosittu muuttolintujen tarkkailussa. Tämän jälkeen järjestelmään ajettiin tekaistu havainto kustakin lajista viikon välein koko vuoden ympäri. Havaintomäärä vuodessa oli yhteensä 48, sillä jokaisesta kuukaudesta valitaan neljä havaintopäivää. Tämä oli luokittimien toteutuksen kannalta yksinkertaisinta testata.

Koska havaintotapauksia tulee tässä tapauksessa runsaasti, on kustakin lajis-

Taulukko 5.10: Tavanomaisia muuttolintuja Suomessa

Laji	**Muutto**
kiuru*	palaa maalis-toukokuussa, lähtee syys-marraskuussa
töyhtöhyyppä	palaa maaliskuussa, lähtee touko-heinäkuussa
kottarainen*	palaa maalis-huhtikuussa, lähtee syys-marraskuussa
mustavaris*	palaa maalis-toukokuussa, lähtee syys-lokakuussa
uuttukyyhky*	palaa maalis-huhtikuussa, lähtee elo-marraskuussa
vihervarpunen**	palaa maalis-toukokuussa, lähtee syys-lokakuussa
urpiainen**	palaa maalis-huhtikuussa, lähtee loka-marraskuussa
punarinta**	palaa huhti-toukokuussa, lähtee syys-lokakuussa
hippiäinen**	palaa maalis-huhtikuussa, lähtee syys-lokakuussa
koskikara***	palaa loka-marraskuussa, lähtee maalis-huhtikuussa

(*) Tähdellä merkityt lajit talvehtivat harvoin myös Suomessa

(**) Kahdella tähdellä merkityt lajit talvehtivat usein myös Suomessa

(***) Koskikara talvehtii Etelä-Suomessa

ta kussakin kaupungissa esitetty vain ajat vuodesta, jolloin lajia olisi järjestelmien mukaan todennäköistä nähdä. Helsingin muuttotulokset ovat taulukoissa 5.11, 5.12, 5.13 ja 5.14.

Helsingille laskettuna Hatikan tietokannan tapauksessa nelikulmiolaskenta ennustaa oikein vain yhden lajin kymmenestä, eikä sekään osu tarkalleen oikein. Kyseessä on kiuru ja sille jää paluumuuttoon muutaman päivän pituinen vaje.

Samalle Hatikan tietokannalle laskee Bayes-laskenta vastaavasti vain yhden kymmenestä oikein ja sekään ei mene aivan oikein.

Tiiran tietokannan tapauksessa molemmat menetelmät toimivat paremmin. nelikulmiolaskenta antaa ylimalkaisempia tuloksia ja ehdottaa että monien muuttolintujen tapauksessa että niitä on mahdollista nähdä koko vuoden. Monia muuttolinnuista on toki mahdollista nähdä Suomessa myös talviaikana mutta ne ovat harvinaisempia. Kesällä kyseisistä lajeista ei taas ole niin paljon havaintoja, koska niitä ei pidetä niin mielenkiintoisina.

Tiiran tietokantaa käytettäessä töyhtöhyyppä on nelikulmiolaskennan tapauksessa ainut laji, jolle paluumuuton ennustaminen osuu oikein. Koski-

Taulukko 5.11: Helsingin koko vuoden havaintolaskelma nelikulmiolaskennalla, Hatikka

Laji	**Todennäköistä nähdä**
kiuru	1.2 - 23.3. ja 1.4. - 23.4. ja 1.5. - 1.12.
töyhtöhyyppä	Ei näy koko vuonna
kottarainen	Ei näy koko vuonna
mustavaris	Ei näy koko vuonna
uuttukyyhky	Ei näy koko vuonna
vihervarpunen	Ei näy koko vuonna
urpiainen	Ei näy koko vuonna
punarinta	Ei näy koko vuonna
hippiäinen	Ei näy koko vuonna
koskikara	Ei näy koko vuonna

Taulukko 5.12: Helsingin koko vuoden havaintolaskelma nelikulmiolaskennalla, Tiira

Laji	**Todennäköistä nähdä**
kiuru	Koko vuoden
töyhtöhyyppä	1.3. - 23.11.
kottarainen	Koko vuoden
mustavaris	1.8. - 1.7.
uuttukyyhky	Koko vuoden
vihervarpunen	Koko vuoden
urpiainen	Koko vuoden
punarinta	Koko vuoden
hippiäinen	Koko vuoden
koskikara	1.9.-1.5. ja 1.6. - 1.8.

kara on näistä lajeista kiinnostava, sillä se muuttaa Etelä-Suomeen talveksi. Nelikulmiolaskenta ei ennusta tätä oikein, mutta on kuitenkin oikeilla jäljillä.

Taulukko 5.13: Helsingin koko vuoden havaintolaskelma Bayes-laskennalla, Hatikka

Laji	**Todennäköistä nähdä**
kiuru	1.1. - 23.9.
töyhtöhyyppä	Ei näy koko vuonna
kottarainen	Ei näy koko vuonna
mustavaris	Ei näy koko vuonna
uuttukyyhky	Ei näy koko vuonna
vihervarpunen	Koko vuoden
urpiainen	Koko vuoden
punarinta	Ei näy koko vuonna
hippiäinen	Ei näy koko vuonna
koskikara	Ei näy koko vuonna

Taulukko 5.14: Helsingin koko vuoden havaintolaskelma Bayes-laskennalla, Tiira

Laji	**Todennäköistä nähdä**
kiuru	1.1. - 15.5.
töyhtöhyyppä	1.1. - 7.10.
kottarainen	Koko vuoden
mustavaris	1.1. - 7.2.
uuttukyyhky	1.1. - 30.5.
vihervarpunen	1.10. - 30.3.
urpiainen	15.11. - 15.3.
punarinta	7.7. - 15.4.
hippiäinen	23.8. - 30.3.
koskikara	1.12. - 23.3.

Tiiran tietokannassa Bayes-laskenta sen sijaan ennustaa jotkut muuttoaalloista oikeaan suuntaan. Näitä ovat kiurun paluumuuton loppuosa, vihervarpunen, urpiainen, punarinta, hippiäinen sekä koskikara. Kesäajalta näille ei

tule todennäköistä havaintoa, mutta tämä johtuu siitä että Tiiran tietokannassa kyseisiä lintuja ei ole havaittu kesäaikaan, koska kyseisiä havaintoja ei pidetä lainkaan mielenkiintoisina.

Kiinnostavaa on että koskikaran muuttoaallot ennustuvat aivan oikein Bayes-laskentaa käyttämällä Tiiran tietokannalle. Koskikara tulee Etelä-Suomeen talvehtimaan Pohjois-Suomesta, Ruotsista sekä Norjasta.

5.4 Tulosten arviointi

Arviointiluvun viimeinen osio päättyy tulosten arviointiin tutkimuskysymysten näkökulmasta. Tässä kohtaa tutkimuskysymyksiin voidaan vastata.

1. Miten epäluotettavat havaitsijat voidaan tunnistaa havainnon perusteella?

Luotettavien havaitsijoiden tunnistaminen kehitetyillä menetelmillä näyttää johtavan hyviin tuloksiin käytettäessä suurta määrää hyvälaatuista dataa. Datan ollessa huonolaatuista, vaikuttaa se vain luotettavien havaitsijoiden määrään, mikä johtaa myös mahdollisesti luotettavien havaitsijoiden luokittelemisen epäluotettaviksi.

2. Miten voidaan arvioida havainnon luotettavuutta tiettyyn aikaan tietyssä paikassa?

Käytetyt menetelmät riippuvat paljon datan laadusta. Jos data on alun perin hyvänlaatuista ja sitä on paljon, näyttää se luokittelevan myös suuren määrän havaintoja luotettaviksi. Pieni määrä tuottaa myös pienen määrän luotettavia havaintoja, mutta se ei sinänsä ole kattavaa. Tämä havainto voidaan perustella jo Hatikan datalla ja vielä paremmin Tiiran datalla, jossa havainnoista osuu oikein jopa kymmenen kymmenestä. Bayesilaisen laskennan tapauksessa myös joidenkin lajien muuttoaaltojen reunat (kevät- tai syysmuutto) nousevat esiin.

Luku 6

Pohdintaa

Tässä luvussa kerrotaan minkälaisia ongelmia nousi esiin tai oli tunnistettavissa työn eri vaiheissa niin aineistossa, laskennassa kuin tulosten esittämisessäkin. Lopuksi esitetään ideoita työn jatkokehitykselle.

6.1 Aineiston ongelmat

Aineisto on haastavaa, koska se on vaihtelevanlaaduista, siinä on puutteita muun muassa eri kenttien sisällöissä sekä sitä ei ole mitenkään semanttisesti annotoitu. Vaikka semanttisessa mallissa paikannimet, lajinnimet tai vaikka päivämäärät ovat yksikäsitteisesti määritettyjä ontologioiden tai standardien avulla, alkuperäisessä aineistossa niitä ei ole mitenkään rajattu.

Yksi aineistossa ongelmia aiheuttava tekijä on havaitsijoiden perusteellisempi tunnistaminen. Samaa havaitsijaa ei tunnisteta samaksi, koska tämä on antanut nimensä usealla eri tavalla tai eri järjestelmät ovat muokanneet sitä erilaiseen muotoon. Tämä ongelma ilmenee erityisesti koostettaessa aineistoa useasta eri lähteestä. Eri lähteissä havaitsijoiden nimet on voitu kirjata eri tavalla ja myös samassa järjestelmässä on voinut olla mahdollista ilmoittaa nimensä eri tavalla. On myös mahdollista että usealla havaitsijalle on sama nimi. Näitä tuli silmämääräisesti vastaan ainakin muutama kappale.

Powell et al. [42] ovat kirjoittaneet paperin tekijännimien tunnistamisesta

ja yhdistämisestä. Tekniikkaa voitaisiin soveltaa tässä. Vaikka havainnon tekijät pystyttäisiinkin tunnistamaan suku- tai etunimen perusteella samoiksi, löytyy silti Suomesta useita Matti Virtasia.

Myöskään paikannimien tarkkuudesta tai kirjoitusasun tarkkuudesta ei ole varmuutta, eivätkä esimerkiksi kahta samaa paikkaa tarkoittavaa kirjoitusasua välttämättä tunnisteta samoiksi. Paikannimien validoinnissa voitaisiin käyttää Googlen Geolocation -rajapintaa, jolla saataisiin suurempi varmuus paikannimen tunnistamiseen oikein.

Kaikissa edellämainituissa tilanteissa ongelmia aiheuttaa myös käytetyn merkistön koodaus. Tämä on erityisesti ongelmallista Suomen kielessä, jossa ä, ö sekä å-kirjaimet eivät välttämättä näytä samoilta eri merkistöissä.

Lisäämällä luotettavuuden määräävää dataa voitaisiin kummassakin laskentamallissa päästä parempiin tuloksiin. Nyt havaitsijan luotettavuuden määritykseen on käytetty vain lintututkinnon suorittaneiden listaa. Suuri osa semanttisen webin ongelmista liittyy edelleen siihen vaiheeseen kun dataa, joka ei alunperin ole semanttisesti annotoitu, yritetään muuttaa semanttiseen malliin.

6.2 Laskennan ongelmat

Molempiin tässä työssä käytettyihin laskentamalleihin liittyy myös omanlaisiaan ongelmia. Nelikulmiolaskentamalli on käytännössä liiankin yksinkertainen. Dataa ei välttämättä ole saatavilla riittävästi, erityisesti tietyille lajeille ja tietyillä aikaväleillä. Koska juuri näiltä tärkeiltä tai kiinnostavilta aikaväleiltä ei ole riittävästi dataa tarjolla, täytyy järjestelmän interpoloida havaintokarttoja. Malli tekee niin karkean approksimaation että se saturoi huomattavan suuren alueen kunkin lajin todennäköiseksi esiintymisalueeksi.

Parannusta tähän toisi esimerkiksi jonkinlaisen tarkemman monikulmion käyttäminen, sekä näiden kulmioiden jako pienempiin keskenään irrallisiin osiin. Näille voitaisiin myös rajatilanteissa interpoloida alueita, jotka pienevät lineaarisesti ajan suhteen. Näissäkin ongelmaksi muodostuisi tiettyjen pisteiden seuraaminen tai ryhmittely. Mahdollisesti jonkinlainen sääkartoissa käytetty tiheys- tai lämpökartta voisi toimia tarkoituksenmukaisesti.

Bayesilaisen mallin tarkkuutta lisäisi huomattavasti jos eri lajeista saataisiin riittävästi tuntomerkkitietoja, esimerkiksi siipien kärkiväli, jotka olisivat jossain lineaarisessa yhteydessä toisiinsa. Myös värit voitaisiin määritellä

väriympyrän mukaisessa järjestyksessä jolloin lähekkäisille väreille saataisiin algoritminen läheisyys.

Ongelmia voinee aiheuttaa myös laskennan tehokkuus. Edellä mainitussa mallissa, jossa laskennan tulokset tallennetaan semanttiseen dataan, voi tästä datasta haettaessa esiintyä viiveitä. Bayesilaiselta luokittimelta kysyttäessä 30 000 havainnon järjestelmässä haku kesti vain alle 0.01 sekuntia kerralta.

Ongelmia aiheuttanee myös tiedon esitys käyttäjälle. Hänen vastuulleen jää todennäköisyysjakauman tulkinta. Ratkaisuvaihtoehtona voitaisiin esimerkiksi pyytää kaikkien luokkien luokkatodennäköisyydet tietylle havainnolle. Laittamalla nämä suuruusjärjestykseen ja katsomalla löytyykö kyseinen laji esimerkiksi kärkikymmeniköstä antaisi paremman kuvan todellisesti esiintymistodennäköisyydestä. Periaatteessa riittää tuntea kyseiselle lintuaineistolle ominainen 0,005 marginaali, jota suuremmat luokkatodennäköisyydet olivat merkitseviä. Kaikkein yksinkertaisin vaihtoehto olisikin ehkä siis vain kertoa havainnon olevan joko validi tai epävalidi, perustuen sen luokkatodennäköisyyteen.

6.3 Tulosten esittämiseen liittyvät ongelmat

Pelkän todennäköisyystiedon tai muun kaksiarvoisen muuttujan esittäminen käyttäjälle ei tällaisessa tapauksessa ole välttämättä järkevää. Todennäköisyyden tapauksessa arvot voitaisiin esimerkiksi normittaa jollekin asteikolle, kuten vaikka 1-10. Näin käyttäjä saisi itse valinnanvaraa päättelyyn siitä onko havainto todella järkevä vai ei.

Nelikulmiolaskennan tapauksessa ei edellä mainitun kaltaista päättelyä voi helposti tehdä. Kuitenkin jos tässä käytettäisiin esimerkiksi sääkartoista tuttuja intensiteettialueita tai muunlaista laskentaa, jossa lintulajien esiintyminen laskisi lineaarisesti tai jonkin muun mallin mukaan tarkastelupisteen loitontuessa havaintopisteistä, voitaisiin edellä mainittuja ongelmia lieventää.

6.4 Jatkokehitys

Jatkossa kehitystyö tulisi suunnata biologisten entiteettien levinnäisyyden mallintamiseen ja mallien kehittämiseen. Tulevaisuudessa muun muassa lin-

nuilla on enenevässä määrin etäluettavia renkaita tai muita nykyteknologian mahdollistamia seurantatyökaluja, jotka mahdollistavat lintujen seuraamisen täysin uudella tasolla.

Myös nykyistä toimivampia menetelmiä syklisen tai ajallisen luontodatan mallintamiseen tarvitaan. Maailmanlaajuisesti saatavaa geodataa voitaisiin tiiviimmin ja tehokkaammin ripustaa olemassa oleviin ontologioihin, jotta voitaisiin varmistua paikanmääritysten oikeellisuudesta.

Alueen sääolosuhteita tutkimalla voisi myös rikastuttaa mallia sopivaan suuntaan. Tulevaisuuden matkapuhelimissa on ilmanpaine- sekä muita sensoreita, jotka mahdollistavat havaintohetken sääolosuhteiden tuntemisen. Nämä tekijät saattavat vaikuttaa lintujen käyttäytymiseen.

Samoin suurella määrällä paikallista tietoa, saatuna esimerkiksi joltain lintuhavaintoasemalta, voitaisiin dataa rikastuttaa.

Luku 7

Yhteenveto

Internetissä esitettävään tietoon liittyy koko ajan enemmän semantiikkaa ja semanttisia ominaisuuksia. Näitä ei kuitenkaan ole sovellettu juurikaan luonnontieteen kentällä, erityisesti luontohavaintoihin liittyen. Taksonominen metaontologia TaxMeOn tarjoaa mahdollisuuksia tämän tiedon esittämiseen ja semanttiseen kuvailemiseen ja siten yhteentoimivampien palvelujen tai ohjelmistorajapintojen tuottamiseen ja tarjoamiseen luonnontieteen piirille.

Tässä työssä esiteltiin kaksi erilaista ratkaisumallia luontohavaintojen validoinnin ongelmaan. Ensimmäinen malli on lähtökohdiltaan täysin ontologioihin perustuva ja siinä lasketaan etukäteen jokaiselle lajille havaintokarttoja. Malli on karkean yksinkertainen eikä siten anna kovin tarkkaa kuvaa todellisesta tilanteesta.

Toinen ratkaisumalli pohjautuu bayesilaiseen todennäköisyyden tulkintaan ja luottaa muuttujien suhteisiin ja näiden keskinäiseen ehdolliseen todennäköisyyteen ja toimii reaaliajassa. Tämä malli ennustaa hieman tarkemmin satunnaisen havainnon todennäköisyyttä olla tietty taksoni eli lintulaji. Tämä ratkaisumalli avaa myös mahdollisuuksia bayesilaiseen luokittelemiseen millä tahansa muulla luonnontieteen alueella, koska mallin laskenta ei ole riippuvainen kyseisestä aihealueesta, pelkästään tapa jolla muuttujia kyseisestä tilanteesta määritetään.

Kumpikin malli avasi uusia uusia mahdollisuuksia sekä myös herätti kysymyksiä lähestymistapojen ongelmista. Työ jätti jälkeensä demonstraattorin, jota halukkaiden on mahdollista jatkokehittää semanttisen webin tutkimuk-

sen puitteissa. Kyseessä on tutkimuskysymys, jossa vaikuttaa samanaikaisesti suuri määrä muuttujia, jotka ovat kuitenkin tiettyyn rajaan asti hallittavissa ja hyödynnettävissä.

Työ myös näytti toteen semanttisen annotoinnin ja ontologiamallien vahvuuden, joskin käytettävä data on oltava mallinnettu tehokkaasti ontologioiden mukaisesti. Todellisuudessa juuri tämä vaihe tuottaa suurimmat haasteet eri lähteistä tulevan informaation yhdistämiseksi.

Kirjallisuutta

[1] Turtle - terse RDF triple language, W3C team submission, 2008. See: http://www.w3.org/TeamSubmission/turtle/.

[2] Adomavicius, G., and Tuzhilin, A. User profiling in personalization applications through rule discovery and validation. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 1999), KDD '99, ACM, pp. 377–381.

[3] Bayes, M., and Price, M. An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions (1683-1775)* (1763), 370–418.

[4] BBCNEWS. Mobile app sees science go global. `http://news.bbc.co.uk/2/hi/science/nature/8258501.stm`, 2009. [Saatavilla; viitattu 12. joulukuuta 2014].

[5] Berendsohn, W. Access to Biological Collection Data. ABCD Schema 2.06 . ratified TDWG Standard. TDWG Task Group on Access to Biological Collection Data, BGBM, Berlin. `http://news.bbc.co.uk/2/hi/science/nature/8258501.stm`, 2005. [Saatavilla; viitattu 12. joulukuuta 2014].

[6] Berners-Lee, T., Fielding, R., and Masinter, L. Uniform Resource Identifier (URI): Generic Syntax. RFC 3986 (Standard), Jan. 2005.

[7] Berners-Lee, T., Hendler, J., and Lassila, O. The semantic web. *Scientific American 284*, 5 (2001), 28–37.

[8] BirdLife Suomi. Lintuharrastuksen alkeet - Vuodenkierto lintujen seurassa. `http://www.birdlife.fi/lintuharrastus/lintuharrastuksen_alkeet_12.shtml`. [Saatavilla; viitattu 12. joulukuuta 2014].

[9] BirdLife Suomi. Lintuhavainnnot. `http://www.birdlife.fi/havainnot/index.shtml`. [Saatavilla; viitattu 12. joulukuuta 2014].

[10] BirdLife Suomi. Maailman lintujen suomenkieliset nimet. `http://www.birdlife.fi/lintuharrastus/nimisto/`. [Saatavilla; viitattu 12. joulukuuta 2014].

[11] BirdLife Suomi. Suomessa tavatut lintulajit. `http://www.birdlife.fi/havainnot/rk/suomessa_tavatut_lintulajit.shtml`. [Saatavilla; viitattu 27. marraskuuta 2014].

[12] BirdLife Suomi. Tiira Haku. `https://play.google.com/store/apps/details?id=com.pschumi.tiira&hl=fi`. [Saatavilla; viitattu 12. joulukuuta 2014].

[13] blueBill Mobile. Tidalwave. `http://bluebill.tidalwave.it/mobile/`. [Saatavilla; viitattu 12. joulukuuta 2014].

[14] Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., and Mace, G. M. Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biol 8*, 6 (06 2010), e1000385.

[15] Bratko, A., Filipič, B., Cormack, G. V., Lynam, T. R., and Zupan, B. Spam filtering using statistical data compression models. *The Journal of Machine Learning Research 7* (2006), 2673–2698.

[16] Burgman, M. A., and Fox, J. C. Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conservation 6* (2 2003), 19–28.

[17] Citynomadi. Tiiran havainnoijille mobiilisovellus. `https://citynomadi.com/fi/tiiranomadi`. [Saatavilla; viitattu 12. joulukuuta 2014].

[18] Dickey, D. A. *Introduction to Predictive Modeling with Examples. Statistics and Data Analysis Global Forum.* Carolina State U. , Raleigh, NC, 2012.

[19] Finlay, S. *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*. Business in the Digital Economy. Palgrave Macmillan, 2014.

[20] GBIF. Gbif.org: Homepage. `http://www.gbif.org/`. [Saatavilla; viitattu 12. joulukuuta 2014].

[21] Google. The Google Maps Geolocation API. `https://developers.google.com/maps/documentation/business/geolocation/`. [Saatavilla; viitattu 27. marraskuuta 2014].

[22] Gruber, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition 5* (1993), 199–220.

[23] Hellegouarch, S. *CherryPy Essentials: Rapid Python Web Application Development Design, Develop, Test, and Deploy Your Python Web Applications Easily*. Packt Publishing, 2007.

[24] Helsingin Seudun Lintutieteellinen Yhdistys Tringa ry. harrastajatutkinnon suorittaneiden nimilyhenteet. `http://www.tringa.fi/harrastajatutkinnon-suorittaneiden-nimilyhenteet/`. [Saatavilla; viitattu 12. joulukuuta 2014].

[25] Helsingin Seudun Lintutieteellinen Yhdistys Tringa ry. Harrastajatutkinto. `http://www.tringa.fi/harrastajatutkinto/`. [Saatavilla; viitattu 12. joulukuuta 2014].

[26] Hyvönen, E., Tuominen, J., and Kauppinen, T. Representing and utilizing changing historical places as an ontology time series. In *Geospatial Semantics and Semantic Web: Foundations, Algorithms, and Applications* (2011), N. Ashish and A. Sheth, Eds., Springer-Verlag. Book chapter.

[27] iNaturalist. iNaturalist - A community for naturalists. `http://www.inaturalist.org`. [Saatavilla; viitattu 12. joulukuuta 2014].

[28] The JSON Data Interchange Format. Tech. Rep. Standard ECMA-404 1st Edition / October 2013, ECMA, Oct. 2013.

[29] Kliebenstein, D. J. Model misinterpretation within biology: phenotypes, statistics, networks, and inference. *Frontiers in plant science 3* (2012).

[30] Kurki, J., and Hyvönen, E. Collaborative metadata editor integrated with ontology services and faceted portals. In *Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece* (June 2010), CEUR Workshop Proceedings, http://ceur-ws.org/.

[31] Laine, L. *Suomalainen lintuopas*. WSOY, Helsinki, 2002.

[32] Laurenne, N., Tuominen, J., and Hyvönen, E. Semantic modelling of biological a priori knowledge for validation of observational data, 2012. Julkaisematon käsikirjoitusrunko.

[33] Lawera, M. Predictive inference: An introduction. *Technometrics 37*, 1 (1995), 121–121.

[34] Luonnontieteellinen keskusmuseo. Hatikka - Havaintopäiväkirja. `http://hatikka.fi`. [Saatavilla; viitattu 12. joulukuuta 2014].

[35] Luontoportti. LuontoPortti. `http://www.luontoportti.com`. [Saatavilla; viitattu 12. joulukuuta 2014].

[36] Martinez-Meyer, E. Climate change and biodiversity: Some considerations in forecasting shifts in species' potential distributions. *Biodiversity Informatics 2*, 0 (2005).

[37] Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Synthesis*. Island Press, Washington, D.C., June 2005.

[38] Mitchell, T. *Machine Learning*. McGraw Hill, 1997.

[39] Mullarney, K. *Lintuopas : Euroopan ja Välimeren alueen linnut*. Otava, Helsingissä, 1999.

[40] Ottaviani, D., Lasinio, G. J., and Boitani, L. Two statistical methods to validate habitat suitability models using presence-only data. *Ecological Modelling 179*, 4 (2004), 417 – 443.

[41] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[42] Powell, J., Collins, L., Eberhardt, A., Izraelevitz, D., Roman, J., Dufresne, T., Scott, M., Blake, M., and Grider, G. "At scale"author name matching with Hadoop/MapReduce. *Library Hi Tech News 29*, 4 (2012), 6–12.

[43] Prud'hommeaux, E., and Seaborne, A. SPARQL Query Language for RDF. Latest version available as `http://www.w3.org/TR/rdf-sparql-query/`, January 2008.

[44] Rassi, P., Hyvärinen, E., Juslén, A. , Mannerkoski, I. (toim.). *Suomen lajien uhanalaisuus - Punainen kirja 2010*. Ympäristöministeriö ja Suomen ympäristökeskus, 2010.

[45] RDFLib. A Python library for working with RDF. `http://code.google.com/p/rdflib/`. [Saatavilla; viitattu 27. marraskuuta 2014].

[46] Roa-Villescas, M., and Orozco-Alzate, M. Bird in the hand: An electronic field guide app for bird watchers. *Potentials, IEEE 31*, 2 (April 2012), 8 –14.

[47] Romao, C. Interpretation manual of european union habitat. version eur 15, 1996.

[48] Seaborne, A. Jena, a Semantic Web Framework. `http://wiki.apache.org/incubator/JenaProposal`. [Saatavilla; viitattu 12. joulukuuta 2014].

[49] Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., and Kelling, S. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation 142*, 10 (2009), 2282 – 2292.

[50] Taylor, S., and Bogdan, R. *Introduction to research methods*. New York: Wiley, 1984.

[51] Tuominen, J., Laurenne, N., and Hyvönen, E. Biological names and taxonomies on the semantic web – managing the change in scientific conception. In *8th Extended Semantic Web Conference (ESWC2011)* (June 2011).

[52] Tuominen, J., Laurenne, N., Koho, M., and Hyvönen, E. The Birds of the World Ontology AVIO. In *The Semantic Web: ESWC 2013 Satellite Events*, P. Cimiano, M. Fernandez, V. Lopez, S. Schlobach, and J. Völker, Eds., vol. 7955 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 300–301.

[53] Twitter Inc. Bootstrap, from Twitter. `http://twitter.github.com/bootstrap/`. [Saatavilla; viitattu 12. joulukuuta 2014].

[54] Valkeapää, O., and Hyvönen, E. Semantic annotation with browser-based annotation tool saha, July 17 2006. Demo paper, 1st Asian Semantic Web Conference (ASWC2006).

[55] Van Landuyt, W., Vanhecke, L., and Brosens, D. Florabank1: a grid-based database on vascular plant distribution in the northern part of Belgium (Flanders and the Brussels Capital region). *PhytoKeys 12*, 0 (05 2012), 59–67.

[56] Weibel, S., Kunze, J., Lagoze, C., and Wolf, M. Dublin Core Metadata for Resource Discovery. RFC 2413, Sept. 1998.

[57] Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., and Vieglais, D. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE 7*, 1 (01 2012), e29715.

[58] WWF. 2010 and Beyond. `http://www.footprintnetwork.org/images/uploads/CBD_2010_and_Beyond.pdf`, 2010. [Saatavilla; viitattu 12. joulukuuta 2014].

[59] Yu, J., Wong, W.-K., and Hutchinson, R. A. Modeling experts and novices in citizen science data for species distribution modeling. In *ICDM* (2010), G. I. Webb, L. Bing, C. Zhang, D. Gunopulos, and X. Wu, Eds., IEEE Computer Society, pp. 1157–1162.

[60] Zhang, H. The Optimality of Naive Bayes. In *FLAIRS Conference* (2004), V. Barr and Z. Markov, Eds., AAAI Press.