Linnunpesien sijainnin ennustaminen spatiaalisella tiedonlouhinnalla

Jan Lindstsröm

7. maaliskuuta 2003

1 Johdanto

Spatiaalisessa tiedonlouhinnassa etsitään kiinnostavia ja mahdollisesti hyödyllisiä spatiaalisia malleja (spatial patterns), joita löytyy spatiaalisesta tietokannasta.

Yleisesti käytetty ratkaisumenetelmä spatiaalisessa tiedonlouhinnassa on käyttää klassisia tiedonlouhintamenetelmiä spatiaalisten yhteyksien materialisoinnin jälkeen olettaen että tietopisteet ovat riippumattomia. Tämä oletus rikkoo ensimmäistä maantieteen sääntöä: kaikki tieto on riippuvaista toisistaan mutta lähempänä toisiaan olevat tiedot ovat enemmän riippuvaisia kuin kaukaiset tiedot. Eli spatiaalisesti toisiaan lähellä olevat tiedot vaikuttavat toisiinsa. Spatiaalisessa tilastotieteessä, tilastotieteen tutkimusalue spatiaalisen tiedon analysointiin, tätä kutsutaan *autokorrelaatioksi* (autocorrelation). Tästä syystä klassiset tiedonlouhintamenetelmät usein toimivat heikosti spatiaalisille tietojoukoille, joissa on korkea spatiaalinen autokorrelaatio. Tässä esityksessä esitellään spatiaalisia tilastotieteellisiä menetelmiä, jotka tehokkaasti mallintavat spatiaalista autokorrelaatiota. Menetelmää sovelletetaan lintujen sijaintien ennustamiseen kosteikossa.

2 Sovellusalue

Tarkkojen asuinpaikkamallien saatavuus on tärkeä työkalu sekä villieläinten tutkimukseen että kriittisten asuinpaikkojen ja uhanalaisten lajien suojeluun. Koska taustalla olevan villieläinten ja ympäristön osatekijöiden välinen vuorovaikutusprosessi on mutkikas, tilastollisia menetelmiä käytetään muodostamaan käsitys perustuen maastosta kerätyyn tietoon. Tämän esityksen perustana on spatiaalinen malli suoalueella pesivien lintujen pesintäpaikkoista. Tiedot on kerätty kahdelta suoalueelta (Darr ja Stuble), jotka sijaitsevat Erie-järven rannalla Ohiossa. Alue jaettiin tasaisiin lohkoihin ja jokaiseen lohkoa kuvaavaan tietueeseen talletettiin arvoja useista rakenteellisista ja ympäristöllisistä tekijöistä. Tällaisia ovat esimerkiksi veden syvyys,

hallitsevan kasvillisuuden kestävyysindeksi ja etäisyys avoimeen veteen. Nämä kolme osatekijää ovat tärkeimmät selittävät muuttujat valitussa sovellusalueessa. Jokaiseen lohkoon talletettiin myös tieto siitä sisältikö lohko linnunpesän vai ei. Linnunpesän läsnäolo kuvasi siis riippuvaa muuttujaa.

2.1 Ongelma

Linnunpesien sijaintien ennustaminen on erikoistapaus kaksiluokkaisesta spatiaalisesta luokitteluongelmasta, joka määritellään seuraavasti:

- *Syöte:*
  - Spatiaalinen kehys S sisältäen sijainnit $\{s_1, .., s_n\}$ taustalla olevasta maantieteellisestä tilasta G.
  - Joukko selittäviä funktoita $f_{X_k} : S \rightarrow R^k, k = 1, ..., K$. R^k on selittävien funktioiden mahdollisten arvojen vaihteluväli.
  - Riippuvaisuusfunktio $f_Y : S \rightarrow R^Y$.
  - Perhe $\mathcal{F}$ opintomallifunktioita kuvaten $R^1 \times ... \times R^K \rightarrow R^Y$.
- *Etsi*: Funktio $\hat{f^Y} \in \mathcal{F}$.
- *Tavoite*: Maksimoi luokittelutarkkuus $(\hat{f^Y}, f_Y)$.
- *Rajoitteet*:
  1. Maantieteellinen tila S on useampiuloitteinen Euclininen tila.
  2. Selittävien funktioiden F_{X_k} ja tulosfunktion f_Y arvot eivät välttämättä ole riippumattomia lähellä toisiaan olevista paikoista eli spatiaalista autokorrelaatiota esiintyy.
  3. $R^k \in \mathcal{R}$
  4. $R^Y = \{0, 1\}$.

3 Peruskäsitteet

Olkoon y n:n muuttujan vektori havaintoja ja $n \times m$ matriisi $\underline{X}$ selittävää dataa. Klassinen *lineaarinen regressio* mallintaa y:n ja $\underline{X}$:n suhteen:

$$y = X\beta + \epsilon \tag{1}$$

Missä $X = [1, \underline{X}]$ ja $\beta = (\beta_0, ..., \beta_m)^t$. Yleisesti virhevektorista ϵ oletetaan, että jokainen komponentti on generoitu riippumattomasta, identtisestä ja normaalijakautuneesta todennäköisyysavaruudesta eli $\epsilon_i = \mathcal{N}(0, \sigma^2)$. Riippuvan muuttujan ollessa binäärinen, kuten sovellusalueella onko linnun pesää vai ei, malli muunnetaan logistisen funktion avulla ja riippuva muuttuja tulkitaan todennäköisyydeksi löytää pesä kyseisestä paikasta. Eli, $Prob(y = 1) = \frac{e^{X\beta}}{1+e^{X\beta}}$. Tätä muunnettua mallia kutsutaan logistiseksi regressioksi.

Spatiaalisen autokorrelaation määrittämiseen on useita menetelmiä. Jokaisella menetelmällä on omat vahvuutensa ja heikkoutensa. Tässä esityksessä kuvataan Moran I -menetelmä. Moran I -menetelmä kuvaa spatiaalisen autokorrelaation määrää tietojoukossa.

$$I = \frac{\sum_{i=1, i\neq j}^{n} \sum_{j=1}^{n} W_{ij} y_i y_j}{W \sqrt{Var(y)}} \tag{2}$$

missä W_{ij} on sijaintien i ja j spatiaalinen etäisyys (esimerkiksi käänteinen etäisyys pisteiden i ja j välillä) tai 0 tai 1 kuvaten että i ja j ovat tietyn etäisyysrajan sisällä toisistaan. y_i on havaittu arvo muuttujalle y muunnettuna siten, että mediaani on nolla $(i, j = 1, 2, ..., n)$, W on summa kaikissa muuttujan w_{ij} n^2 arvoista ja $Var(y)$ on otoksen varianssi.