COMPUTATIONAL SCIENCE AND TECHNIQUES
Special Issue for Research Innovations Fundamentals 2011
Volume 1, Number 1, 2013, 1-10

Publisher: Klaipėda University
http://journals.ku.lt/index.php/CST
Online ISSN: 2029-9966

BALTIJOS JŪROS DUMBLIŲ ANALIZĖ NAUDOJANT BAJESINĖS ERDVINĖS STATISTIKOS METODUS

Eglė Baltmiškytė, Kęstutis Dučinskas
Statistikos katedra, Klaipėdos universitetas, Herkaus Manto g. 84, LT-92294, Klaipėda
egle.baltmiskyte@gmail.com, kestutis.ducinskas@ku.lt

Anotacija. Erdvinė statistika – tai viena iš statistikos mokslo krypčių, apimanti erdvėje pasklidusių duomenų statistinę analizę. Šiandien erdvinė statistika plačiai taikoma ekologijoje, geografijoje, astronomijoje, vaizdų analizėje, klimato pokyčių tyrimuose. Statistinei duomenų analizei vis plačiau taikomi Bajesinės statistikos metodai. Šiame straipsnyje sudaromas erdvinių duomenų modelis dumblių kiekiui prognozuoti, kuriame priklausomas kintamasis šakotasis banguolis, nepriklausomi kintamieji – gylis, smėlis, gargždas, rieduliai. Naudojant Bajesinės erdvinės statistikos metodus *OpenBUGS* modeliavimo aplinkoje įvertinami nežinomi modelio parametrai, sudaroma Bajesinio krigingo prognozės aposteriorinis pasiskirstymas.

Pagrindiniai žodžiai: Bajeso statistika, erdvinė statistika, Bajeso krigingas, *OpenBUGS*, Gibbs imties sudarymo metodas.

Įvadas

Technikos pažangai suteikiant galimybę atlikti daug sudėtingų skaičiavimų, įvairius statistinius uždavinius tapo įmanoma spręsti pasitelkiant kompiuterius. Prieš sprendžiant statistinius uždavinius pirmiausia yra surenkami duomenys. Jie renkami tam tikrą laiko periodą ir gali būti koreliuoti. Jų tyrimui taikoma laiko eilučių analizė. Tačiau praktikoje duomenys renkami ne tik skirtingais laiko momentais, bet ir skirtingose erdvės vietose. Natūralu manyti, kad duomenys surinkti tam tikrose erdvės vietose koreliuoja. Taikant įvairius metodus šiuos duomenis nagrinėja erdvinė statistika. Uždavinių sprendimui statistiniais metodais vis dažniau naudojami atsitiktiniai dydžiai. Siekiant, kad vis daugiau uždavinių būtų galima spręsti panaudojant kompiuterius, taikomi įvairūs metodai ir algoritmai. Vieni svarbiausių iš tokių metodų yra Markovo grandinių Monte Karlo (MCMC) metodai. Šių metodų panaudojimas leidžia modeliuoti fizikines ir matematines sistemas, kai tikslių rezultatų neįmanoma gauti naudojant deterministinį metodą. Straipsnį sudaro įvadas, skyrius apie erdvinius duomenų modelius, skyrius apie Bajeso metodą, skyrius apie realizuotą duomenų modelį ir išvados.

1. Erdvinių duomenų modeliai

Tariama, kad $s \in \Re^d$ yra d -matės Euklido erdvės taškas, kuriame atliekamas stebėjimas. Natūralu, kad galimas stebėjimas $Z(s)$, erdvės taške s yra atsitiktinis dydis. Tuomet matematinis erdvinių duomenų modelis yra atsitiktinis laukas

$$\{Z(s): s \in D\},$$

kur $D \in \Re^d$ yra erdvinių indeksų aibė, o atsitiktinio lauko realizacija žymima $\{z(s): s \in D\}$.

Atsitiktinis laukas $\{Z(s): s \in D\}$ - tai rinkinys atsitiktinių dydžių, apibrėžtų vienoje tikimybinėje erdvėje $(\Omega, \mathcal{F}, \mathcal{P})$ ir įgyjančių reikšmes erdvėje $\mathcal{B}$. Erdvė $\mathcal{B}$ dažnai yra vadinama būsenų erdve. Pagrindinė savybė, kurią turi tenkinti atsitiktinis laukas, tai stacionarumas (Dučinskas and Šaltytė-Benth, 2003). Praktikoje dažniausiai taikomos skaitinės charakteristikos yra vidurkis ir kovariacija. Todėl tikslinga nagrinėti platesnę nei griežtai stacionarių atsitiktinių laukų klasę.

Atsitiktinis laukas $\{Z(s): s \in D\}$ vadinamas stacionariu, jei $E\{Z(s)^2\} < \infty$ visiems $s \in D$, $E\{Z(s)\} \equiv \mu$ visiems $s \in D$, $C(s_1, s_2) = C(s_1 - s_2)$ visiems $s_1, s_2 \in D$.

Tokie laukai taip pat dar vadinami homogeniškais, stacionariais plačiąja prasme arba silpnai stacionariais atsitiktiniais laukais. Griežtai stacionarūs laukai yra ir stacionarūs, tačiau priešingas tvirtinimas bendru atveju negalioja. Tik Gauso atsitiktinių laukų stacionarumas reiškia griežtą stacionarumą, nes jie pilnai aprašomi vidurkiu ir kovariacine funkcija (Dučinskas and Šaltytė-Benth, 2003).

Taigi požymio Z stebėjimo taške $s \in D$ modelis yra:

$$Z(s) = \mu(s) + \varepsilon(s), \qquad (1)$$

kur $E\{Z(s)\} \equiv \mu(s)$ - vidurkio funkcija, $\varepsilon(s)$ - erdvinė paklaida. Šiame straipsnyje atskirai aptariami vidurkio ir erdvinės paklaidos modeliai.

1.1. Vidurkių modeliai

Aptariami trys plačiausiai taikomi vidurkio modeliai. Pastovaus vidurkio modelis yra

$$E\{Z(s)\} \equiv \mu = const$$

visiems $s \in D$. Tokiu atveju $\boldsymbol{Z}_n$ vidurkių vektorius

$$\boldsymbol{\mu}_n = \mu \mathbf{1}_n,$$

kur $\mathbf{1}_n$ yra n-matis , iš vienetų sudarytas vektorius. Trendo paviršiaus modelis, tai vidurkio priklausančio nuo stebėjimo taško koordinačių, modelis. Taškui $s_i = (x_i, y_i) \in D$, $i = 1, 2, ..., n$, šis modelis užrašomas tokiu būdu (Dučinskas and Šaltytė-Benth, 2003):

$$\mu(s_i) = \sum_{\substack{t+u \le k \\ 0 \le t,u \le k}} \lambda_{tu} x_i^t y_i^u ,$$

kur λ_{tu} yra modelio parametras, k apibrėžia modelio, arba kitaip – trendo paviršiaus eilę t.y. nulinės eilės $(k = 0)$ paviršius atitinka pastovaus vidurkio atvejį, tiesinis, arba pirmos eilės $(k = 1)$, paviršius generuoja plokštumą ir t.t. Taigi aukštesnės eilės generuoja sudėtingesnius paviršius.

Regresiniame modelyje vidurkis taške $s_i \in D$, $i = 1, 2, ..., n$, apibrėžiamas kaip funkcija nuo q aiškinamųjų kintamųjų (regresorių) (Dučinskas and Šaltytė-Benth, 2003):

$$\mu(s_i) = \boldsymbol{x}^T(s_i)\boldsymbol{\beta}, \qquad (2)$$

kur $x^T(s_i)=\left(1, x_1(s_i), x_2(s_i),..., x_q(s_i)\right)$ yra $(q+1)$-matis regresorių vektorius, $\boldsymbol{\beta}=\left(\beta_0, \beta_1, \beta_2..., \beta_q\right)^T$ yra $(q+1)$-matis regresijos parametrų vektorius.

1.2. Kovariacijų modeliai

Stebėjimų vektoriaus $\boldsymbol{Z}_n$ kovariacijų matricą galima tiesiogiai sudaryti naudojant kovariacinės funkcijos išraišką, jos gali būti tik tam tikros formos. Bendruoju atveju, šios funkcijos $C(\cdot)$ erdvėje R^d turi būti teigiamai apibrėžtos, t.y. $\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j C\left(s_i - s_j\right) \geq 0$, bet kokiam baigtiniam erdvinių lokacijų $\{s_i : i=1,...,m\}$ skaičiui ir realiems skaičiams $\{a_i : i=1,...,m\}$. Jei laukas yra izotropinis, jo kovariacinė funkcija priklauso tik nuo vektoriaus $h=s_i-s_j$ ilgio $|h|$ ir nepriklauso nuo jo krypties, t.y. $C(h)=C_0\left(\|h\|\right),\ h\in R^d$ (Cressie, 1993). Pagrindinės kovariacinės funkcijos yra sferinės ir Matern klasės funkcijos. Šiame darbe naudojamos šios Matern klasės funkcijos (Cressie, 1993):

Eksponentinė kovariacinė funkcija:

$$C(h)=\begin{cases}\sigma^2 \exp\left(-\dfrac{h}{\phi}\right), & jei\ h>0\\ \tau^2+\sigma^2, & jei\ h=0.\end{cases} \qquad (3)$$

Gauso kovariacinė funkcija:

$$C(h)=\begin{cases}\sigma^2 \exp\left(-\dfrac{h^2}{\phi^2}\right), & jei\ h>0\\ \tau^2+\sigma^2, & jei\ h=0.\end{cases} \qquad (4)$$

kur ϕ – pločio parametras, σ^2 – dalinio slenksčio parametras, τ^2 – grynuolio parametras. Sudarius modelį, norint jį taikyti reikia įvertinti nežinomus vidurkių ir kovariacinės funkcijos parametrus. Vienas iš būdų įvertinti nežinomus parametrus yra Bajeso metodas.

2. Bajeso metodas

Bajeso statistikos populiarumas pastaraisiais metais labai išaugo dėl spartaus skaičiavimo technikos vystymosi ir atitinkamų efektyvių skaičiavimo algoritmų sukūrimo. Tarkime turime modelį, aprašytą (1) formule ir žinomas jo vidurkio ir kovariacinės funkcijos pavidalai. Tokiu atveju turimas modelis su nežinomais vidurkio koeficientais, kovariogramos slenksčio ir pločio parametrais. Nežinomi parametrai pažymimi θ. Bajeso metodologijoje nežinomo parametro tikroji reikšmė laikoma atsitiktine. Remiamasi prielaida, kad nežinomas parametras θ yra atsitiktinis dydis srityje Θ, kurio pasiskirstymo funkcija $\Pi(\theta)$. Ji atspindi apriorinę informaciją apie θ. $\Pi(\theta)$ ir $\pi(\theta)=\Pi'(\theta)$

vadinami atitinkamai apriorine pasiskirstymo funkcija ir aprioriniu pasiskirstymo tankiu. Jei sritis Θ yra aprėžta, tai tolygusis skirstinys srityje Θ yra „neinformatyvus“, t.y. jis aprašo situaciją, kai jokios apriorinės informacijos nėra.

Aposterioriniu pasiskirstymu vadinamas atsitiktinio parametro θ sąlyginis pasiskirstymas, kai žinomos atsitiktinio dydžio $Z(s)$ su pasiskirstymo funkcija F_Z stebėjimų $Z(s)$ reikšmės. Jis apskaičiuojamas remiantis Bajeso formule (King et al., 2010):

$$\pi(\theta|Z)=\frac{f(Z|\theta)\pi(\theta)}{\int f(Z|\theta)\pi(\theta)d\theta}, \qquad (5)$$

čia Z yra stebėtas atsitiktinis dydis, $f(Z|\theta)$ – imties tikėtinumas (tankis), $\pi(\theta)$ – apriorinis parametro θ tankis. Erdvinių duomenų modelis užrašytas (1) formule perrašomas, taip:

$$Z(s)=\mu(s)+\omega(s)+e(s),$$

čia $\mu(s)$ – vidurkio modelis, $\omega(s)$ – erdvinė paklaida su nuliniu vidurkiu ir kovariacine funkcija C, $e(s)$ – baltas triukšmas, $e(s)\sim iid\ N(0,\tau^2)$. Daroma prielaida, kad $Z(s)$ yra Gauso atsitiktinis dydis t.y.

$$Z\big|\theta\sim MVN\left(\mu,\sigma^2H(\phi)+\tau^2I\right),$$

čia μ – vidurkio modelis, $H(\phi)$ – koreliacijų funkcija, ϕ – pločio parametras, σ^2 – dalinio slenksčio parametras, τ^2 – grynuolio parametras. Nežinomų parametrų įverčiai gali būti gaunami iš aposteriorinio pasiskirstymo, kuris pagal (5) formulę yra:

$$\pi(\theta|Z)\propto f(Z|\theta)\pi(\theta).$$

Tipiškai nepriklausomi apriorinai pasiskirstymai yra parenkami skirtingiems parametrams t.y.:

$$\pi(\theta)=\pi(\beta)\pi(\sigma^2)\pi(\tau^2)\pi(\phi).$$

Kai norima gauti išvadas apie konkrečius parametrus, reikia gauti jų marginalinius aposteriorinius pasiskirstymus. Pavyzdžiui, taškinis įvertis arba pasitikėjimo intervalas β parametrui gaunamas iš (Banerjee et al., 2004)

$$\begin{aligned}p(\beta|y)&=\int\int\int p(\beta,\sigma^2,\tau^2,\phi|y)d\sigma^2d\tau^2d\phi\\&\propto p(\beta)\int\int\int f(y|\theta)\,p(\sigma^2)\,p(\tau^2)\,p(\phi)d\sigma^2d\tau^2d\phi. \qquad (6)\end{aligned}$$

Tokiems integralams skaičiuoti dažnai naudojami MCMC metodai arba kiti skaitinio integravimo būdai. Nežinomo parametro aposteriorinį pasiskirstymo įvertinį gavus iteraciniu būdu, nežinomo parametro β įverčiu laikomas aposteriorinio pasiskirstymo vidurkis arba mediana.

Erdvinę struktūrą galima modeliuoti tiesiogiai, tačiau tai galima atliki ir netiesiogiai, sąlygojant Z procesą erdvinių atsitiktinių efektų $W=(w(s_1),...,w(s_n))^T$ vektoriumi. Tokiu būdu sudaromas hierarchinis erdvinių duomenų modelis. Hierarchinio modelio esmė yra tai, kad kelių atsitiktinių dydžių bendrą pasiskirstymą galima išreikšti jų sąlyginių pasiskirstymų sandauga:

$$f\left(data, process, parameters\right)$$
$$\propto f\left(data\middle|process, parameters\right)\times f\left(process\middle|parameters\right)\times f\left(parameters\right).$$

Sudarytas hierarchinis duomenų modelis atrodo taip (Clark and Gelfand, 2006):

$$Z\middle|\theta,W \sim MVN\left(\mu+W,\tau^2 I\right) \quad (7)$$

$$\text{ir } W\middle|\sigma^2,\phi \sim MVN\left(0,\sigma^2 H\left(\phi\right)\right), \quad (8)$$

kur μ – vidurkio modelis, $H\left(\phi\right)$ – koreliacijų funkcija, ϕ – pločio parametras, σ^2 – dalinio slenksčio parametras, τ^2 – grynuolio parametras, W – atsitiktinių efektų vektorius. Modelio specifikacija užbaigiama pridedant apriorinius pasiskirstymus β ir τ^2 bei σ^2 ir ϕ. Pastarieji du gali būti laikomi hyperparametrais. Bajeso statistikoje visi efektai laikomi atsitiktiniai. Taigi sąlygiškumas nuo W – tai tik papildomas erdvės parametras, nuo kurio $Z\left(s\right)$ tikėtinumo funkcija yra sąlyginė. Parametrų erdvė padidėja nuo θ iki $\left(\theta,W\right)$. Tokiu atvejų nežinomų parametrų aposteriorinis pasiskirstymas yra:

$$\pi\left(\mu,\sigma^2,\phi,\tau^2\middle|Z\right)\propto\int f\left(Z\middle|\mu,\tau^2,W\right)\pi\left(\mu,\tau^2\right)f\left(W\middle|\sigma^2,\phi\right)\pi\left(\sigma^2,\phi\right)dW.$$

Įvertinus nežinomus modelio parametrus atliekama prognozė. Prognozei taikomi įvairūs krigingo metodai: paprastas, ordinarus ar universalus krigingas. Kai nežinomų parametrų reikės vertinamos Bajeso metodu, prognozės metodas vadinamas Bajeso krigingu. Turint θ aposteriorinį pasiskirstymą, prognozės $Z\left(s_0\right)$ aposteriorinis pasiskirstymas naujame erdvės taške s_0 apibrėžiamas taip (Banerjee et al., 2004):

$$f\left(Z_0\middle|Z\right)=\int f\left(Z_0,\theta\middle|Z\right)d\theta=\int f\left(Z_0\middle|Z,\theta\right)\pi\left(\theta\middle|Z\right)d\theta. \quad (9)$$

MCMC metodai gali vėl būti pritaikyti (9) formule apibrėžto aposteriorinio pasiskirstymo įverčiams gauti. Prognozės $Z\left(s_0\right)$ įvertis arba pasitikėjimo intervalas gali būti apskaičiuojamas tokiu metodu, kaip ir nežinomų parametrų įvertinimo atveju.

Atlikus duomenų prognozę, taikomi įvairūs skaičiavimo metodai prognozės tikslumui įvertinti. Šiame straipsnyje prognozės tikslumui įvertinti atliekamas kryžminės patikros metodas, kai iš pradinių duomenų pašalinamas vienas narys ir likusių duomenų pagrindu atliekama pašalintojo nario prognozė. Šis metodas taikomas keliems taškams ir po to skaičiuojamas prognozės tikslumas. Prognozės tikslumui įvertinti skaičiuojamos šios statistikos:

šaknis iš vidutinės kvadratinės prognozės paklaidos (RMSPE – Root Mean Squared Prediction Error) (Rossiter, 2004)

$$RMSPE=\sqrt{\frac{\sum_{i=1}^{m}\left(\hat{Z}\left(s_i\right)-Z\left(s_i\right)\right)^2}{m}}, \quad (10)$$

vidutinis absoliutus nuokrypis (MAD – Mean Absolute Deviation)

$$MAD=\frac{\sum_{i=1}^{m}\left|\hat{Z}\left(s_i\right)-Z\left(s_i\right)\right|}{m}, \quad (11)$$

vidutinis kvadratinis nuokrypio santykis (MSDR – Mean Squared Devation Ratio):

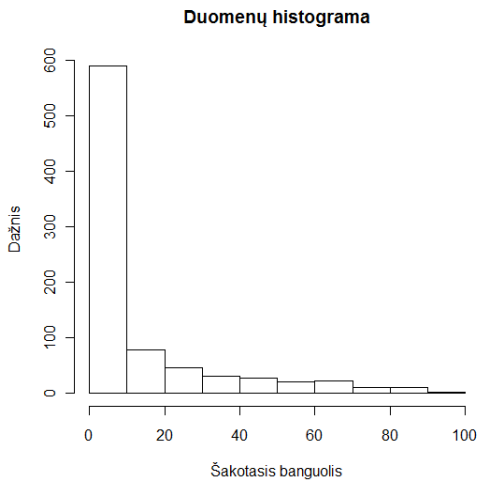
$$MSDR = \frac{\sum_{i=1}^{m}\left[\frac{\left(\hat{Z}(s_i)-Z(s_i)\right)}{\hat{\sigma}(s_i)}\right]^2}{m}, \qquad (12)$$

kur $\hat{Z}(s_i)$ – prognozės reikšmė, $Z(s_i)$ – stebėta reikšmė, $\hat{\sigma}(s_i)$ – prognozės standartinis nuokrypis. Parametrų įverčių skaičiavimas – iteracinis procesas, kurio metu pasirenkamos pradinės įverčių reikšmės ir, taikant tam tikrą skaičiavimo metodą, keičiamos tol, kol stabilizuojasi. Šiame straipsnyje naudojama *OpenBUGS* modeliavimo programa, kurioje nežinomų parametrų įverčiai naudojant imties tikėtinumo funkciją apskaičiuojami taikant MCMC metodus.

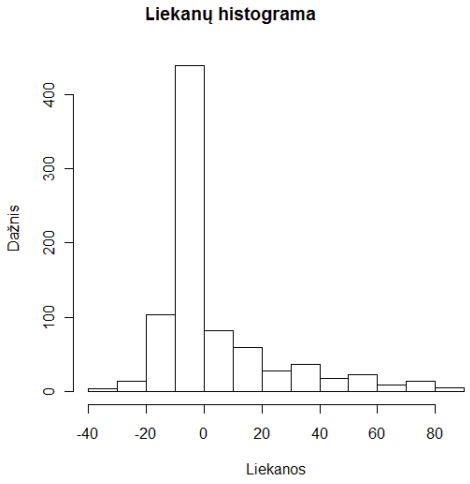
3. Realizuotas duomenų modelis

OpenBUGS yra nemokama statistinio modeliavimo programinė įranga, sukurta projekto BUGS (Bayes inference Using Gibbs Sampling) iniciatyva ir skirta Bajesinei analizei naudojant MCMC metodus (Thomas et al., 2011). *OpenBUGS* naudoja tris MCMC sudarymo metodus, straipsnyje naudojamas Gibbs imties sudarymo metodas atsitiktinių kintamųjų su sudėtinga tankio funkcija generavimui (Clark and Gelfand, 2006). Metodo pagrindinė idėja yra tai, kad atlikus pakankamai daug imituotų stebėjimų, galima gauti tikslų pasiskirstymo vaizdą. *OpenBUGS* modeliavimo aplinkoje galima sudaryti daug įvairių erdvinių duomenų modelių, įvertinti jų koeficientus ir padaryti išvadas. Šiame straipsnyje sudaromas erdvinių duomenų modelis dumblių kiekiui prognozuoti, kuriame priklausomas kintamasis šakotasis banguolis, nepriklausomi kintamieji – gylis, smėlis, gargždas, rieduliai.

Statistiškai išanalizuojami duomenys, patikrinamas nagrinėjamų duomenų normališkumas. Duomenų ir liekanų, kurios gautos iš duomenų pašalinus trendą, histogramos pateiktos 1 ir 2 paveiksluose.



1 pav. Duomenų histograma.



2 pav. Liekanų histograma.

Iš duomenų pašalinus trendą histogramos matyti, kad liekanos gerai aproksimuojamos normalinio skirstinio tankiu. Nagrinėjamiems duomenims sudaromas hierarchinis erdvinių duomenų modelis apibrėžtas (7) (8) formulėmis:

$$Z\left|\theta,W\right. \sim MVN\left(\mu+W,\tau^2 I\right);$$

$$W\left|\sigma^2,\phi\right. \sim MVN\left(0,\sigma^2 H\left(\phi\right)\right).$$

Vidurkio modelis apibrėžiamas (2) formule, t.y.:

$$\mu(s)=\beta_1 x_g(s)+\beta_2 x_r(s)+\beta_3 x_{gr}(s)+\beta_4 x_s(s)$$

kur x_g – gylis, x_r – rieduliai, x_{gr} – gargždas, x_s – smėlis. Koreliacijų matricos $H\left(\phi\right)$ pavidalas priklauso nuo pasirenkamos kovariacinės funkcijos pavidalo, apibrėžto (3) ir (4) formulėmis. Nežinomiems modelio parametrams apibrėžiami neinformatyvūs aprioriniai pasiskirstymai: $\beta \sim N\left(0.0;1.0E-5\right)$, $\tau^2 \sim G\left(0.01;0.01\right), \sigma^2 \sim G\left(0.1;0.1\right)$, $\phi \sim G\left(0.1;0.1\right)$.

Sudaryti du duomenų modeliai, kurie skiriasi tik kovarinės funkcijos pavidalu. Modeliai realizuoti *OpenBUGS* modeliavimo aplinkoje. Sudarius modelį su eksponentine kovariacine funkcija, apibrėžta (3) formule, gautos pagrindinės nežinomų parametrų statistikos: vidurkis, standartinis nuokrypis mediana, procentiliai, pateiktos *„Node statistics“* lentelėje 3 paveiksle.

Node statistics

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
beta_exp[1]	-0.2377	0.1632	0.02735	-0.5169	-0.2507	0.06939	2000	601
beta_exp[2]	0.1176	0.02048	0.00282	0.07891	0.1176	0.1558	2000	601
beta_exp[3]	0.1951	0.02872	0.004796	0.1393	0.195	0.2492	2000	601
beta_exp[4]	-0.01634	0.01855	0.003107	-0.05565	-0.01632	0.01655	2000	601
p_exp	260.8	29.92	3.438	207.9	258.2	327.4	2000	601
tau.precsq_exp	72.28	7.506	0.9684	59.54	72.05	88.21	2000	601
taup.spsq_exp	267.3	27.89	3.152	218.3	266.1	324.7	2000	601

3 pav. Eksponentinio modelio pagrindinės nežinomų parametrų statistikos.

3 paveiksle *„mean“* arba *„median“* stulpeliuose pateiktas reikšmes galima laikyti parametrų įverčiais. Čia β_1 įvertis gaunamas imant jo marginalinio tankio, kuris apibrėžtas (6) formule, vidurkį arba medianą. Remiantis šiame paveiksle pateiktais duomenimis galima įvertinti, ar dominantys parametrai reikšmingai skiriasi nuo 0, t.y. patikrinamos hipotezės:

$$H_0 : \hat{\theta} = \theta_0$$

$$H_1 : \hat{\theta} \neq \theta_0 .$$

Hipotezės patikrinamos naudojant pasitikėjimo intervalus. Pasirenkamas reikšmingumo lygmuo $\alpha = 0,05$ ir apskaičiuojamas 95% pasitikėjimo intervalas, t.y. 2,5% ir 97,5% procentiliai. Šio modelio kiekvienam nežinomam parametrui įvertinti procentiliai pateikti 3 paveiksle. H_0 negalima atmesti, kai $\theta_0 = 0$ priklauso tam intervalui. *„Node statistics“* lentelėje matoma, kad yra intervalų, kurie apima 0, tai reiškia, kad hipotezės, jog jie lygūs nuliui, negalima atmesti. Sudarant modelį šiuos parametrus reikia atmesti. Pagal 3 paveiksle pateiktus duomenis sudaromas duomenų vidurkio ir kovariacinės funkcijos modeliai:

$$\mu(s) = 0{,}12 \cdot x_r(s) + 0{,}20 \cdot x_{gr}(s)$$

$$C(h) = \begin{cases} 266{,}1 \cdot \exp\left(-\dfrac{h}{258{,}2}\right), & jei\ h > 0 \\ 338{,}15 & jei\ h = 0. \end{cases}$$

Įvertinus, nežinomus modelio parametrus atlikta pasirinktų penkių taškų, kurių išmatuotos reikšmės yra $\left(z(s_{0_1}), z(s_{0_2}), z(s_{0_3}), z(s_{0_4}), z(s_{0_5})\right) = (20, 40, 0, 0, 0)$, prognozė. Taikant kryžminės patikros metodą suprognozuotos reikšmės pateikiamos 4 paveiksle.

Node statistics

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
Z0_exp[1]	21.02	7.099	0.3997	8.196	20.81	35.0	2000	601
Z0_exp[2]	43.07	6.051	0.2916	31.18	43.25	54.29	2000	601
Z0_exp[3]	0.7821	6.13	0.3552	-11.18	0.7713	12.61	2000	601
Z0_exp[4]	-0.2329	6.957	0.345	-13.36	-0.4388	14.55	2000	601
Z0_exp[5]	1.872	9.025	0.3322	-15.92	1.733	18.4	2000	601

4 pav. Eksponentinio modelio prognozės statistikos.

4 paveiksle reikšmės pateiktos *„mean“* ir *„median“* yra laikomos prognozės įverčiais, kurie yra gaunami skaičiuojant įvertintų aposteriorinių pasiskirstymo funcijų, apibrėžtų (9) formule, atitinkamai vidurkį arba medianą. Aposteriorinį pasiskirstymą, apibrėžtą (9) formule, yra sudėtinga gauti tiesiogiai integruojant, todėl tam taikomi skaitiniai metodai. Šiame straipsnyje taikomas Monte Karlo metodas, tokiu būdu (Banerjee et al., 2004):

$$\hat{p}\left(z_0 \middle| z, X, x_0\right) = \frac{1}{G}\sum_{g=1}^{G} p\left(z_0 \middle| z, \theta^{(g)}, x_0\right),$$

kur $p\left(z_0 \middle| z, \theta^{(g)}, x_0\right)$ – yra aposteriorinis prognozės pasiskirstymo tankis g-ojoje iteracijoje, $\theta^{(g)}$ – nežinomų parametrų įverčiai g-ojoje iteracijoje, G – iteracijų skaičius. Atlikus pognozę pasrinktuose taškuose, prognozės tikslumui įvertinti apskaičiuojamos statistikos, apibrėžtos (10), (11) ir (12) formulėmis. Gauti rezultatai pateikiami 1 lentelėje:

1 lentelė. Eksponentionio modelio prognozės statistikos.

	RMSPE	MAD	MSDR
Modelis su eksponentine kovariacine funkcija	1,73	1,40	0,08

Sudaromas kitas modelis, kuris nuo ankščiau aprašyto skiriasi tik kovariacinės funkcijos pavidalu. Pasirenkama Gauso kovariacinė funkcija, apibrėžta (4) formule, gautos pagrindinės nežinomų parametrų statistikos: vidurkis, standartinis nuokrypis mediana, procentiliai, pateikiamos *„Node statistics“* lentelėje 5 paveiksle.

Node statistics

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
beta_gau[1]	-0.1784	0.1721	0.01597	-0.5072	-0.1844	0.165	15000	5001
beta_gau[2]	0.1921	0.02492	0.001919	0.1431	0.1917	0.2384	15000	5001
beta_gau[3]	0.2344	0.03092	0.002994	0.1693	0.2369	0.2943	15000	5001
beta_gau[4]	0.002331	0.01718	0.001492	-0.03492	0.002067	0.03526	15000	5001
p_gau	47.28	1.318	0.1093	44.72	47.32	49.84	15000	5001
tau.precsq_gau	56.5	5.938	0.4314	45.75	56.22	68.56	15000	5001
taup.spsq_gau	297.6	21.82	0.8981	257.5	296.2	343.0	15000	5001

5 pav. Gauso modelio pagrindinės nežinomų parametrų statistikos.

5 paveiksle *„mean“* arba *„median“* stulpeliuose pateiktas reikšmes galima laikyti parametrų įverčiais. Įtraukiant tik reikšmingus parametrus pagal pateiktus duomenis sudaromas duomenų vidurkio ir kovariacinės funkcijos modeliai:

$$\mu(s) = 0{,}19 \cdot x_r(s) + 0{,}24 \cdot x_{gr}(s)$$

$$C(h) = \begin{cases} 296{,}2 \cdot \exp\left(-\left(\dfrac{h}{47{,}32}\right)^2\right), & jei\ h > 0 \\ 352{,}42 & jei\ h = 0. \end{cases}$$

Įvertinus, nežinomus modelio parametrus atlikta pasirinktų penkių taškų prognozė, kurių išmatuotos reikšmės yra $\left(z\left(s_{0_1}\right), z\left(s_{0_2}\right), z\left(s_{0_3}\right), z\left(s_{0_4}\right), z\left(s_{0_5}\right)\right) = (20, 40, 0, 0, 0)$. Taikant kryžminės patikros metodą, suprognozuotos reikšmės pateiktos 6 paveiksle.

Node statistics

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
Z0_gau[1]	23.19	11.9	0.2019	-0.3244	23.27	46.64	5000	15001
Z0_gau[2]	32.51	6.52	0.3014	19.64	32.59	45.03	5000	15001
Z0_gau[3]	-0.0674	6.488	0.3919	-12.63	-0.03469	12.74	5000	15001
Z0_gau[4]	0.1353	7.918	0.1064	-15.31	0.1243	15.77	5000	15001
Z0_gau[5]	3.622	11.03	0.09564	-18.01	3.468	25.48	5000	15001

6 pav. Gauso modelio prognozės statistikos.

6 paveiksle reikšmės pateiktos *„mean“* ir *„median“* yra laikomos prognozės įverčiais. Jie taip pat gaunami skaičiuojant įvertintų aposteriorinių pasiskirstymo funkcijų, apibrėžtų (9) formule, atitinkamai vidurkį arba medianą. Atlikus prognozę pasirinktuose taškuose, prognozės tikslumui įvertinti apskaičiuojamos statistikos, apibrėžtos (10), (11) ir (12) formulėmis. Gauti rezultatai pateikti 2 lentelėje:

2 lentelė. Gauso modelio prognozės statistikos.

	RMSPE	MAD	MSDR
Modelis su Gauso kovariacine funkcija	3,94	2,86	0,30

Duomenims pritaikius modelį su eksponentine kovariacijų funkcija, MAD reikšmė yra santykinai artimesnės 0, o RMSPE reikšmė yra santykinai mažesnė, negu modelio su Gauso

kovariacine funkcija, todėl modelis su eksponentine kovariacijų funkcija duomenis aprašo geriau. Baltijos Jūros dumblių kiekiui prognozuoti tinkamesnis modelis su eksponentine kovariacijų funkcija.

Išvados

1. Išanalizavus erdvinių duomenų modelius nustatyta, kad nežinomų parametrų įvertinimui galima naudoti Bajeso metodą.
2. Išnagrinėjus erdvinių duomenų modelių savybes, nustatyta, kad tokie modeliai tinka dumblių kiekio Baltijos jūroje prognozės realizacijai *OpenBUGS* modeliavimo aplinkoje.
3. Siekiant nustatyti, koks modelis geriausiai tinka dumblių kiekio Baltijos jūroje prognozei, sudaryti du modeliai su skirtingomis kovariacijų funkcijomis (Gauso ir eksponentine).
4. Atlikti skaičiavimai rodo, kad modelis su eksponentine kovariacijų funkcija Baltijos jūros dumblių kiekio prognozę atlieka tiksliau.

Literatūra

Dučinskas, K., Šaltytė-Benth, J. (2003). Erdvinė statistika. Klaipėda: Klaipėdos universiteto leidykla.

Cressie, N. A. C., (1993). Statistics for Spatial Data. New York: John Willey & Sons, Inc.

King, R., Morgan, B. J. T., Gimenez, O., Brooks, S. P. (2010). Bayesian Analysis for Population Ecology. Boca Raton.

Banerjee, S., Carlin, B. P., Gelfand, A. E. (2004). Hierarchical Modeling and Analysis for Spatial Data. Boca Raton.

Clark, J. S., Gelfand, A. E. (2006). Hierarchical Modelling for Environmental Sciences. New York.

Rossiter, D. (2004). An Introduction to Applied Geostatistics: a lecture note. International Institute for Geo-Information Science and Earth Observation (ITC). Enschede.

Thomas, A., Best, N., Lunn, D., Arnold, R., Spiegelhalter, D. (2011). OpenBUGS User Manual. http://www.openbugs.info/Manuals/Manual.html

BALTIC SEA ALGAE ANALYSIS USING BAYESIAN SPATIAL STATISTICS METHODS

Eglė Baltmiškytė, Kęstutis Dučinskas

Summary

Spatial statistics is one of the fields in statistics dealing with spatialy spread data analysis. Recently, Bayes methods are often applied for data statistical analysis. A spatial data model for predicting algae quantity in the Baltic Sea is made and described in this article. Black Carrageen is a dependent variable and depth, sand, pebble, boulders are independent variables in the described model. Two models with different covariation functions (Gaussian and exponential) are built to estimate the best model fitting for algae quantity prediction. Unknown model parameters are estimated and Bayesian kriging prediction posterior distribution is computed in *OpenBUGS* modeling environment by using Bayesian spatial statistics methods.

Key words: Bayesian statistics, spatial statistics, Bayesian kriging, *OpenBUGS*, Gibbs sampling.