Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
53 t., 2012, 51–55

ISSN 0132-2818
www.mii.lt/LMR/

Diskrečių skirstinių taikymas duomenų apie gamtos plotų padengimą Bajesinėje analizėje

Kęstutis Dučinskas[1], **Eglė Baltmiškytė**[1], **Martynas Bučas**[1,2]

[1] *Klaipėdos universitetas, Gamtos ir matematikos mokslų fakultetas*
H. Manto g. 84 LT-92294 Klaipėda

[2] *Klaipėdos universitetas, Baltijos pajūrio aplinkos tyrimų ir planavimo institutas*
H. Manto g. 84 LT-92294 Klaipėda

E. paštas: kestutis.ducinskas@ku.lt, egle.baltmiskyte@gmail.com, martynas@corpi.ku.lt

Santrauka. Klasikinės statistikos metodai ne visada leidžia pasiekti norimų rezultatų, todėl ieškoma alternatyvių duomenų analizės metodų. Dėl modernėjančios skaičiuojamosios technikos, statistinei duomenų analizei vis plačiau taikomi Bajesinės statistikos metodai. Šiame darbe sudaromi keli diskrečiųjų skirstinių modeliai įvairių gamtos plotų padengimui analizuoti. Šie modeliai gali būti pritaikyti tokių gamtos plotų, kaip miškų, vandens telkinių dugno, dirvožemio analizei, kai duomenys pateikiami procentiniais dydžiais, išreikštais sveikaisiais skaičiais. Šiame darbe sudaromi modeliai panaudojant kelis skirtingus diskrečiuosius skirstinius, t. y. Puasono ir neigiamai binominį skirstinį. Naudojant Bajesinės statistikos metodus OpenBUGS modeliavimo aplinkoje įvertinami nežinomi modelio parametrai. Atliekama parametrų įverčių patikimumo statistikų analizė, realizuota OpenBUGS modeliavimo aplinkoje. Gamtos plotų padengimo analizei skirtų modelių realizavimui buvo pasirinkti duomenys, aprašantys Baltijos jūros dugno padengimą raudondumbliu šakotuoju banguoliu. Šiame darbe tiriama padengimo priklausomybė nuo kelių regresorių, aprašančių abiotinius ir fizinius veiksnius.

Raktiniai žodžiai: Puasono skirstinys, neigiamai binominis skirstinys, Bajeso statistika, OpenBUGS, MCMC.

Įvadas

Įvairių gamtos plotų, tokių kaip miškų, vandens telkinių dugno, dirvožemio padengimui analizuoti dažnai taikomi tolydieji skirstiniai. Atliekant tokio pobūdžio tyrimus, duomenys paprastai yra surenkami procentiniais dydžiais, išreikštais sveikaisiais skaičiais. Dėl šios priežasties tokių duomenų analizei kartais taikomi diskretieji skirstiniai. Vienas žinomiausių diskrečiųjų skirstinių yra Puasono skirstinys. Viena pagrindinių prielaidų, kurią turi tenkinti duomenys, yra tai, kad duomenų vidurkis ir dispersija turi sutapti. Kai ši sąlyga netenkinama ir dispersija viršija vidurkį, sprendžiama perteklinės dispersijos problema. Vienas iš sprendimo metodų – Puasono skirstinį pakeisti neigiamai binominiu skirstiniu. Vienas pagrindinių statistikos uždavinių yra modelių nežinomų parametrų vertinimas. Vis dažniau šiam uždaviniui spręsti taikomas Bajeso metodas, kuris nežinomą parametro reikšmę laiko atsitiktine. Siekiant, kad vis daugiau uždavinių būtų galima spręsti pasitelkiant modernią skaičiavimo techniką, taikomi įvairūs metodai ir algoritmai. Markovo grandinių Monte Karlo (MCMC) metodų panaudojimas leidžia modeliuoti fizikines ir matematines sistemas,

kai tikslių rezultatų neįmanoma gauti naudojant deterministinį metodą. Straipsnį sudaro įvadas, skyrius apie Bajeso metodą, skyrius apie realizuotą duomenų modelį ir išvados.

1 Bajeso metodas

Dėl naujų skaičiavimo algoritmų, skirtų sudėtingų uždavinių sprendimui pasitelkiant kompiuterius, vis labiau populiarėja Bajeso statistika. Tai statistikos rūšis, kuri remiasi Bajeso metodu, kurio pagrindinė idėja tai, kad modelių nežinomų parametrų tikrosios reikšmės laikomos atsitiktinėmis [2]. Šiame darbe sudaromi Puasono regresijos ir neigiamai binominės regresijos GLM modeliai su logaritmine jungties funkcija, apibrėžta formule [4, 8]:

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_q X_{iq}, \quad i = 1, 2, \ldots, n, \tag{1}$$

čia μ_i – atitinkamo modelio vidurkis taške i, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)'$ – nežinomų parametrų vektorius, X_{ij} – nepriklausomų kintamųjų matricos elementas.

Tokiu atveju turimi du modeliai su nežinomų regresijos koeficientų vektoriais $\boldsymbol{\beta}$. Apriorinis $\boldsymbol{\beta}$ vektorių tankiai gali būti apibrėžiami taip [2]: $\boldsymbol{\beta} \sim N(a_0; b_0 I)$. Šiame darbe apriorinis $\boldsymbol{\beta}$ tankis abiems modeliams pasirenkamas neinformatyvus, su parametrais $a_0 = 0_q$, $b_0 = 10^5$. Parametrų vektoriaus $\boldsymbol{\beta}$ aposteriorinis tankis Puasono regresijos modeliui gali būti apibrėžiamas taip [1]:

$$\pi(\boldsymbol{\beta}|\boldsymbol{Y}) \propto \exp\Bigg\{ \sum_{1 \leqslant i \leqslant n} \big[y_i x_i' \boldsymbol{\beta} - \exp\{x_i' \boldsymbol{\beta}\} \big] - \frac{1}{2} \big[(\boldsymbol{\beta} - a_0)' (b_0 I)^{-1} (\boldsymbol{\beta} - a_0) \big] \Bigg\}. \tag{2}$$

Neigiamo binominio skirstinio atveju aposteriorinis $\boldsymbol{\beta}$ tankis gali būti apibrėžiamas taip:

$$\pi(\boldsymbol{\beta}|\boldsymbol{Y}) \propto \exp\Bigg\{ \sum_{1 \leqslant i \leqslant n} \left[y_i \ln\left(\frac{\exp\{x_i' \boldsymbol{\beta}\}}{\exp\{x_i' \boldsymbol{\beta}\} + k} \right) + k \ln\left(1 - \frac{\exp\{x_i' \boldsymbol{\beta}\}}{\exp\{x_i' \boldsymbol{\beta}\} + k} \right) \right] - \frac{1}{2} \big[(\boldsymbol{\beta} - a_0)' (b_0 I)^{-1} (\boldsymbol{\beta} - a_0) \big] \Bigg\}. \tag{3}$$

Dėl sudėtingai apskaičiuojamos normalizavimo konstantos, tokiems aposteriorinams tankiams skaičiuoti dažnai naudojami MCMC arba kiti skaitiniai metodai. Nežinomo parametro aposteriorinį pasiskirstymo įvertinį gavus iteraciniu būdu, nežinomo parametro $\boldsymbol{\beta}$ įverčiu laikomas aposteriorinio pasiskirstymo vidurkis arba mediana.

Norint patikrinti modelio patikimumą, galima apskaičiuoti DIC (Deviance Information Criterion) statistiką, kuri apibrėžiama taip [5]:

$$\mathrm{DIC} = 2\mathbb{E}\big[D(\theta)|Y\big] - D\big(\mathbb{E}[\theta|Y]\big), \tag{4}$$

čia $D(\theta) = -2 \log f(Y|\theta)$ – nuokrypis, $f(Y|\theta)$ – imties tikėtinumo funkcija, o θ – nežinomų parametrų vektorius. Patikimesniu laikomas modelis su mažesne DIC reikšme.

Gavus aposteriorinius nežinomų parametrų tankius, apibrėžtus (2) ir (3) formulėmis, taip pat buvo atlikta duomenų prognozė, kurios metu taikant Bajeso metodą gaunamas prognozės aposteriorinis tankis [2].

Prognozės tikslumui įvertinti atliekamas kryžminės patikros metodas ir skaičiuojamos šios statistikos [7, 6]: RMSPE (Root Mean Squared Prediction Error) – šaknis iš vidutinės kvadratinės prognozės paklaidos, MAE (Mean Absolute Error) – vidutinė absoliuti paklaida, MSDR (Mean Squared Devation Ratio) – vidutinis kvadratinis nuokrypio santykis. Tiriant prognozės tikslumą skaičiuojant šias statistikas, pageidautina, kad RMSPE statistika būtų kuo mažesnė, MAE statistika būtų artima 0, MSDR statistika artima 1.

Modeliams sudaryti naudojama OpenBUGS modeliavimo programinę įrangą, kurioje taikant MCMC metodus nežinomų parametrų įverčiai apskaičiuojami iteraciniu procesu.

2 Duomenų modelio realizacija

Šiame straipsnyje OpenBUGS modeliavimo aplinkoje sudaromi du diskrečiųjų skirstinių duomenų modeliai aprašantys Baltijos jūros dugno padengimą raudondumbliu šakotuoju banguoliu. Tiriama padengimo priklausomybė nuo kelių regresorių, aprašančių abiotinius ir fizinius veiksnius. Visi kiekybiniai regresoriai standartizuoti. Nepriklausomo kintamojo duomenys pateikti procentiniais vienetais. Nagrinėjamiems duomenims sudarytas Puasono regresijos modelis:

$$Y_i \sim P(\mu_i),$$

čia μ_i – Puasono skirstinio vidurkis. Sudaroma logaritminė jungties funkcija, apibrėžta (1) formule:

$$\begin{aligned}\log(\mu_i) = {} & \beta_0 + \beta_1 x_{i,s} + \beta_2 x_{i,a} + \beta_3 x_{i,c} + \beta_4 x_{i,o} \\ & + \beta_5 x_{i,se} + \beta_6 x_{i,p} + \beta_7 x_{i,sed1} + \beta_8 x_{i,sed2} + \beta_9 x_{i,d},\end{aligned} \tag{5}$$

čia $x_{i,s}$ – dugno nuolydis (laipsniai, slope), $x_{i,a}$ – dugno ekspozicija (didžiausio aukščio kitimo kryptis: Šiaurę, pietryčius ir t. t., aspect), $x_{i,c}$ – dugno nelygumas (indeksas, curvature), $x_{i,o}$ – bangų orbitinis greitis priedugnyje (m/s, orbitalbv), $x_{i,se}$ – Secchi disko gylis (m, secchi), $x_{i,p}$ – akmenuotas dugnas eufotinėje zonoje (1 – yra, 0 – nėra), $x_{i,sed1}$ – dugno nuosėdų tipai ($x = 1$ – žvirgždas; $x = 0$ – kita), $x_{i,sed2}$ – dugno nuosėdų tipai ($x = 1$ – rieduliai $x = 0$ – kita), $x_{i,d}$ – atstumas iki smėlio lauko (m, distosand).

Kadangi apie nežinomus regresijos parametrus jokios išankstinės informacijos nėra, laikoma, kad parametrai yra tarpusavyje nepriklausomi. Pasirenkamas neinformatyvus normalusis pasiskirstymas.

Norint patikrinti, ar Puasono regresijos modelio prielaida tenkinama, skaičiuojamas perteklinės dispersijos parametras, apibrėžiamas formule [3]: $OD = \frac{1}{n-q} \times \sum_{1 \leqslant i \leqslant n} z_i^2$, čia z_i – standartizuota liekana taške i, n – imties dydis, q – regresorių skaičius. Šio parametro įvertis gaunamas lyginant duomenis su priderintu modeliu. Kai įvertis lygus 1, duomenyse perteklinės dispersijos nėra. Atlikus skaičiavimus OpenBUGS aplinkoje gautas įvertis $OD = 7{,}582$ yra gerokai nutolęs nuo 1, todėl daroma prielaida, kad nagrinėjamiems duomenims perteklinės dispersijos problema egzistuoja.

1 lentelė. Nežinomų parametrų įverčiai.

	Puasono	Neigiamas binominis, $k = 1$	Neigiamas binominis, $k = 10$	Neigiamas binominis, $k = 100$	Neigiamas binominis, $k = 1000$
β_0	1,612	0,964	1,48	1,597	1,612
β_1	0,132	0,261	0,114	0,126	0,127
β_2	0,232	0,53	0,276	0,237	0,231
β_3	−0,053	−0,964*	0,207*	−0,102*	−0,166*
β_4	0,123	1,98	0,244	0,13	0,12
β_5	−0,046	2,387	0,009*	−0,044	−0,045
β_6	0,181*	0,994*	0,526*	−0,232*	0,177*
β_7	0,417	1,881	0,551	0,432	0,416
β_8	−0,386*	0,084*	0,708*	0,092*	0,619
β_9	0,379	1,549	0,553	0,404	0,379

Dėl reikšmingo perteklinės dispersijos parametro įverčio, duomenims taikomas neigiamai binominio skirstinio regresijos modelis, su logaritmine jungies funkcija ir vidurkio modeliu apibrėžtu (5) formule:

$$Y_i \sim B_-(\mu_i, k),$$

čia μ_i – neigiamo binominio skirstinio vidurkio parametras, k – neigiamai binominio skirstinio formos parametras.

Sudarant neigiamai binominės regresijos modelį, parametras k laikomas fiksuotu, todėl šiame darbe sudaromi 4 neigiamai binominės regresijos modeliai, su skirtingomis parametro k reikšmėmis. Gauti parametrų įverčiai pateikti 1 lentelėje.

Parametrų įverčiai, pažymėti $*$, gauti nereikšmingi ir į modelį neįtraukiami. Pagal gautus parametrų įverčius sudaryti 5 skirtingi duomenų vidurkio, kuris apibrėžtas (5) formule, modeliai su skirtingais parametrų vektoriaus $\boldsymbol{\beta}$ įverčiais.

Sudarytiems duomenų modeliams apskaičiuotas DIC kriterijus, apibrėžtas (4) formule. Įvertinus nežinomus modelio parametrus, atlikta pasirinktų taškų prognozė taikant kryžminės patikros metodą. Atlikus prognozę pasirinktuose taškuose, prognozės tikslumui įvertinti apskaičiuotos statistikos. Gauti rezultatai pateikti 2 lentelėje.

Neigiamai binominės regresijos modeliui, kai $k = 1$, DIC kriterijaus reikšmė mažiausia. Taip pat šio modelio MSDR reikšmė yra arčiausiai 1. Tačiau kitos įvertintos prognozės statistikos (RMSPE ir MAE) yra labai prastos lyginant su kitais modeliais. Dėl šios priežasties geresniu laikytinas neigiamai binominės regresijos modelis, kai $k = 10$, nes nors šio modelio DIC reikšmė yra tik antra pagal gerumą, tačiau statistikos RMSPE ir MAE nėra tokios prastos, kaip modelio su $k = 1$ atveju.

2 lentelė. Modelių prognozės statistikos.

Modelis	RMSPE	MAE	MSDR	DIC
Puasono regresijos	14,97	9,608	5,6	10490
Neigiamai binominės regresijos, kai $k = 1$	78,5	36,44	1,205	2431
Neigiamai binominės regresijos, kai $k = 10$	17,76	10,96	2,506	5960
Neigiamai binominės regresijos, kai $k = 100$	15,19	9,704	4,933	9476
Neigiamai binominės regresijos, kai $k = 1000$	15,0	9,591	5,547	10380

3 Išvados

Išnagrinėjus diskrečių skirstinių duomenų modelių savybes, nustatyta, kad šie modeliai tinka Baltijos jūros dugno padengimo analizės realizacijai OpenBUGS modeliavimo aplinkoje.

Siekiant nustatyti, koks modelis geriausiai tinka Baltijos jūros dugno padengimo šakotuoju banguoliu analizei, sudaryti du modeliai su skirtingais diskrečiaisiais skirstiniais (Puasono ir neigiamai binominio).

Atlikti skaičiavimai rodo, kad remiantis DIC kriterijumi ir prognozės statistikomis, iš išnagrinėtų modelių Baltijos jūros dugno padengimo šakotuoju banguoliu analizei geriausiai tinka neigiamai binominės regresijos modelis, kai $k = 10$.

Literatūra

[1] A. Gelfand and M. Ghosh. Generalized linear models: A bayesian view. In D.K. Dey, M. Ghosh and S. Ghosh(Eds.), *Generalized Linear Models: A Bayesian Perspective*, pp. 3–22, New York, USA, 2000. Marcel Dekker Press.

[2] A. Gelman, J.B. Carlin and H.S. Stern. *Bayesian data Analysis,* second edition. Chapman and Hall/CRC Press, New York, USA, 2004.

[3] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, New York, USA, 2007.

[4] M.J. Hilbe. *Negative Binomial Regression,* second edition. Cambridge University Press, New York, USA, 2011.

[5] M. Kyung and S.K. Ghosh. Bayesian inference for directional conditionally autoregressive models. *Bayesian Anal.*, **4**(4):675–706, 2009.

[6] R. Webster and M.A. Oliver. *Geostatistics for Environmental Scientists,* second edition. John Wiley and Sons, Ltd, Chichester, West Sussex, England, 2007.

[7] C.J. Willmott and K. Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *J. Art.*, **30**(1):79–82, 2005.

[8] A.F. Zurr, E.N. Ieno, N.J. Walker, A.A. Saveliev and G.M. Smith. *Mixed Effects Models and Extensions in Ecology with R.* Springer Science+Business Media, LLC, New York, USA, 2009.

SUMMARY

Application of the discrete distribution in Bayes analusis of nature area coverage data

K. Dučinskas, E. Baltmiškytė, M. Bučas

Classical statistical methods do not always provide desired results for every situation. Therefore, new alternative methods of data analysis are in demand. As the computational power becomes more modern, Bayes statistical methods are increasingly applied for statistical data analysis. This article describes several discrete models for analyzing nature area coverage. These models can be applied for analysis of such areas as forests, water ponds, soil, etc. when data is provided in integer data in percent. Poisson and negative binomial distributions are used in this article. Unknown parameters of the models were estimated using Bayes statistical methods in OpenBUGS modeling environment. The models of nature area coverage analysis were implemented using the data of Baltic Sea bottom algae coverage. This article analyzes coverage dependence of abiotic and physical factors.

Keywords: Poisson distribution, negative binomial distribution, Bayes statistics, OpenBUGS, MCMC.