STOCHASTIC MODELS AND SOFTWARE TECHNOLOGIES IN ECOLOGICAL RESEARCH

Evgeniya Nikolova, Burgas Free University, enikolova@bfu.bg

Abstract: In the past years with the increaseing computing power and use of software solutions, Bayesian statistics has become a powerful alternative to traditional statistics. This paper presents the use of the Bayesian method in ecological research and environmental decision-making. Additionally, attention is to new technologies as a tool for collecting and analyzing data, as well as to the capabilities of new real-time monitoring technologies on the Black Sea.

Key words: Bayesian statistics, ecological software, real-time Black Sea monitoring

СТОХАСТИЧНИ МОДЕЛИ И СОФТУЕРНИ ТЕХНОЛОГИИ В ЕКОЛОГИЧНИТЕ ИЗСЛЕДВАНИЯ

Евгения Николова, Бургаски свободен университет, enikolova@bfu.bg

I like to think
(it has to be!)
of a cybernetic ecology
where we are free of our labors
and joined back to nature,
returned to our mammal
brothers and sisters,
and all watched over
by machines of loving grace.
Richard Brautigan

Във всички съвременни дейности набирането, съхраняването, обработката на информацията и извличането на полезни резултати от анализа ѝ е ключов процес. Екологията е натоварена с предоставянето на информационна подкрепа за решения за политиката в областта на околната среда с широки обществени последици. Много екологични проучвания използват анализа на голям броя данни, за да се стигне до биологично значими изводи. Търсенето на бързи отговори насочва еколозите към използването на Бейсов статистически анализ и специално разработен за техните нужди софтуер. Бейсовата статистика дава възможност за заключения при непълна информация.

Ключови области, в които информационните технологии може да са в помощ в процеса на опазване на околната среда са: знания и разбиране за околната среда, биологичното разнообразие и тяхната взаимовръзка с хората; насърчаване на осведомеността и чувствителността в отделните хора и общности за околната среда, биоразнообразието и неговото значение; насърчаване на хората и общностите да оценяват околната среда и да ги импулсира да се включват активно в процес на подобряване и защита на средата за подобряване на собственото им препитание; да изгради у хората умения да идентифицират, прогнозират, предотвратяват и решават екологични проблеми и да ги направи способни да използват ограничени ресурси по устойчив начин. Съвременните

софтуерни технологии предоставят възможности за разработване на екологични софтуерни продукти, които могат да се справят със сложните изисквания, противоречивите потребителски необходимости и развиващите се структури от данни.
Икономическото значение на Черно море, което произтича от откритите в шелфа находища на полезни изкопаеми, запасите от минерални суровини, разнообразието на флората и фауната, специфичните климатични условия и географско положение, както и факта, че голяма част от външната търговия се извършва чрез морския транспорт, поставя на дневен ред въпросите по неговото опазване и произтичащата от това необходимост за разработване на уеб базирани платформи за мониторинг в реално време.
Целта на настоящата публикация без да претендира за изчерпателност е да представи ролята на Бейсовата статистика и софтуерните решения в екологични научните изследвания и използването на този инструментариум при мониторинг в реално време на Черно море.

Бейсовият метод като статистически инструмент за еколози

Много екологични проучвания използват анализа на голям броя данни, за да се стигне до биологично значими изводи. В съвременната статистика има две парадигми: честотна статиситка и Бейсова статистика. Задачата, която се поставя е да се направи извод за стойността на неизвестен параметър на генералната съвкупност по данни на представителна извадка. Бейсовият метод е важен статистически инструмент, който се използва от еколозите, защото дава възможност за заключения при непълна информация и представлява много по-различен подход към науката, отколкото честотната статистика. Смисълът на му е в разглеждането на неизвестния параметър като случайна величина с някаква плътност на разпределение по отношение на някаква вероятност. В него се предполага, че в момента на провеждане на експермента неизвестния параметър се избира случайно съгласно тази плътност. За представителната извадка $x=(x_1,...,x_n)$ и неизвестният параметър μ се разглежда съвместната функция на разпределение/съвместната плътност. Прилагайки формулата на Бейс, условната плътност на параметъра при условие наблюдаваните данни $f(\mu|x)$ е пропорционална на $f(x|\mu)f(\mu)$. Членът $f(x|\mu)$ е функцията на правдоподобие – съвместната плътност на наблюденията. Членът $f(\mu)$ е априорната плътност на параметъра μ, който съдържа предварителната информация за стойностите на параметъра μ. Условната плътност $f(\mu|x)$ е апостериорно разпределение на μ. Апостериорното разпределение посредством формулата на Бейс обединява експерименталната и неекспериментална информация за стойностите на параметъра μ.
В стаята статия „Bayesian inference in ecology“ [3] Аарон М. Елисън прави обзор на поредица от статии, в които Бейсовите заключения се използват за екологични научните изследвания и вземането на решения в областта на околната среда за периода от 1996 г. до 2003 г.. Той представя таблица на публикациите в основните екологични списания: American Naturalist; Journal of Ecology; Ecology; Ecological Monographs; Journal of Animal Ecology; Oikos; Journal of Applied Ecology; Oecologia; Ecological Applications; Conservation Biology; Ecology Letters, използващи този подход. От нея се вижда, че използавнето му в екологични научни изследвания стартира през 1996 г. с поредица статии в областта на околната среда и продължава с прилагането му за моделиране на динамиката на единични видове, прогнозиране на разпространението на популации, на растежа и изчезването на популации, на промени в структурата на мета-популация на фрагментирани области, за оценка на богатството на видовете от географски или логистично малки извадки, или в отговор на промяна на околната

среда, за оценка на въздействието върху околната среда. Като причина за ограниченото използване на този инструментариум се посочва липсата на удобен софтуер.
Но през последното десетилетие се очертава постоянна възходяща тенденция в използването на Бейсови методи в екологичните изследвания. Това подтиква Хутен и Хобс да публикуват „A guide to Bayesian model selection for ecologists“ през 2015 г. [5], който да служи като справочник за еколозите, за да могат да разберат по-добре възможностите на този подход и да могат подбират подходящ модел, отговарящ на целите на техните изследвания.
Теория на статистическите решения и йерархичното моделиране на видове популации могат да се посочат като примери за прилагане на Бейсовия подход в екологичните изследвания. Теория на статистическите решения (statistical decision theory, SDT) като количествена рамка, с помощта на която да се анализира използването на информация от организмите, е представена от Дал и неговите съавтори през 2005 г. [1, 2]. Тя се основава на използването на Бейсови методи, за да се отговори на въпроса как се адаптират организмите въз основа на личния си опит и еволюционната история (т.е. генетичната информация) при наличие на нова информация. SDT се занимава само с въпроса за актуализирането на информацията. През 2010 ван Джилс [4] за първи път прави реалистично теоретично представяне на това поведение на организмите. Като използва правилото на Бейс той прогнозира потенциалната стойност на модела за пространствено ограничено търсене на храна, когато тя се търси в пространствено автокорелирана заобикаляща среда.
Йерархичното моделиране на видове популации (Hierarchical Modelling of Species Communities, HMSC) е обща рамка за модерен анализ на данните от популации [9], която обхваща както класически подходи като модели за едновидово разпределение, така и модерни инструменти. Рамката на HMSC е реализирана като йерархичен Бейсов модел за съвместно разпределение на популациите и е имплементирано като R- и Matlab пакети, които позволяват изчислително ефективни анализи на големи обеми от данни. Бейс йерархичен модел е статистически модел, който на различни нива оценява параметрите на постериорното разпределение с помощта на Бейсов метод. Тази рамка улеснява формулирането на хипотези, основаващи се на данни, относно процесите, които структурират популациите. Тя осигурява едновременни изводи за популациите и общността, преодолява проблеми на моделирането при оскъдни данни, и е изчислително ефективна като е в състояние да анализира както малки серии от данни, така и големи набори от данни. Авторите на модела предлагат списък от актуални въпроси от екологията на общността и описват как HMSC може да бъде приложено, за да се получат отговори.
Прилагането на тези техники на практика е възможно при използването на статистически софтуер, графичен софтуер и по-специфични програми за анализ на данни и използването на новите технологии за събиране на данни.

Софтуерни технологии за събиране и обработка на екологични данни
Еколозите все повече генерират и споделят огромен обем данни. Такива данни могат да служат за увеличаване на съществуващите бази от данни, като могат да бъдат използвани за синтез като метаанализ, за параметризиране на модели и за проверка на резултатите от изследванията, т.е. възпроизводимост на изследването. Големи обеми от екологични данни могат да бъдат лесно достъпни чрез институции или хранилища за данни, които са най-широко достъпни и могат да служат като ядро на екологичния анализ. Екологичните данни се използват и извън контекста на научните изследвания и се използват за вземане на решения, управление на природните ресурси, образование и други цели. Например, в геномните изследвания развитието на международни бази

данни като бази данни на нуклеинова киселина в Европейската лаборатория по молекулярна биология (EMBL), Gen Bank и DNA база данни на Япония се оказа полезно при идентифицирането на структурата, функцията и историята на гените и протеините; проучванията на биоразнообразието и биогеографията изискват подробни данни за условията на околната среда и разпределението на видовете.
Събирането на данни е само първата стъпка. Обработването и анализирането на много гигабайта данни от различни източници изисква нови инструменти и техники, преди екологичното намерение или планирането на опазването да започне. Все по-често учените-еколози използват език за програмиране като най-често това са езикът R (cran.r-project.org) и Python (python.org). Езикът R се превърна в стандарт за анализ и визуализация на данни сред много еколози, въпреки че не е създаден за обработка на много големи масиви от данни, нито пък има пълна геопространствена функционалност. В същото време има пакети, които ускоряват обработката ("renjin", "riposte"), подобряват управлението на паметта ("bigmemory") и интелигентно обработват геопространствените данни ("raster", "rgdal"). Езикът Python предлага по-голяма скорост, по-добро управление на паметта, може да функционира като интеграционен инструмент за целия работен процес, предлага изключително бърза обработка и анализ на геопространствени данни. Паралелната обработка е техника, която драстично намалява времето за обработка, като използва всички налични процесори на компютър или стотици до хиляди процесори в изчислителен клъстер. Независимо дали се използва персонален компютър или високопроизводителен изчислителен клъстер, пакетите "foreach" за R и "multiprocessing" или "mpi4py" за Python са добра отправна точка. Преходът към по-голямо разчитане на кода се дължи от една страна на увеличаването на количеството и видовете данни, използвани в екологичните проучвания, от друга страна, на подобренията в изчислителната мощност и софтуера.
Повечето еколози сега обикновено пишат код като част от своите лабораторни, полеви или моделиращи изследвания. Обикновено кодът е написан на програмни езици R и Python и се използва от еколозите за голямо разнообразие от задачи, включително манипулиране, анализиране и графично представяне на данни. Ползата от този преход към анализи на базата на код е, че кодът дава точен запис на извършеното, което улеснява възпроизвеждането, адаптирането и разширяването на съществуващите анализи.
Научният код може да бъде разделен на две основни категории - код за анализ и научен софтуер. Кодът за анализ е код, който се използва за коригиране на грешки в данните, симулиране на резултатите от модела и провеждане на статистически анализи. По-голямата част от кода, написан за екологични проучвания, е код за анализ. Научният софтуер е по-общ и е предназначен за използване в много различни проекти. Разработването на екологичен софтуер става все по-често и софтуерът все повече се признава за научен продукт. Като примери могат да се посочат BIOMstat (Statistical Analysis For Biologists) [15], NTSYSpc (Numerical Taxonomy System) [16] и MetaWin [17].
BIOMstat е статистически пакет за MSWindows, който извършва статистически анализи, използвани в биологичните и биомедицинските науки като дискрептивни статистики, честотен анализ, дисперсионен анализ, регресионен анализ, корелационене анализ, непараметрична статистика.
NTSYSpc е програма за многовариантен анализ на данни. Може да се използва за откриване на модел и структура в многомерни данни. Някои от функциите му са определяне на сходства и несъответствия, клъстерен анализ, координационен анализ, дисперсионни и регресионни анализи, интерактивна графика.

MetaWin е програма, която дава възможност за обобщаване на резултатите от множество независими проучвания, използваща мета-анализ, за сравняване на няколко групи с помощта на кумулативни ефекти, оценяване на размера на ефекта във всяко индивидуално изследване, оценяване на качеството на идентифицираните изследвания.
Използването на компютърно генерирани модели за симулиране на екологични събития може да осигури по-добро разбиране на екосистемите и предлага подобрени прогностични възможности на мениджърите по опазване и управление на околната среда. Компютърното моделиране започва да влияе върху екологичната теория. Например, екосистемната свързаност е само един пример за сложен екологичен проблем, с който компютърното моделиране се справя със значителен успех. Компютрите позволяват симулации на експерименти, чиято реализация в реално време или пространство не биха били възможни и това довежда до подобряване на ландшафтната екология. Компютърната симулация на сложни системи помага на еколозите да разберат по-добре естеството на взаимодействията, които оказват влияние върху разнообразието и динамиката на екосистемите.
В статията си [6] от 1997 г. Кломп, Грийн и Фрай стигат до заключението, че по онова време повечето еколози не са запознати с наличните инструменти за моделиране и липсата на разбиране за пространственото моделиране от тяхна страна оставя много данни, недостатъчно използвани, въпреки че през последните 20 години моделирането е било отразено все повече в екологичната литература.
За да изследват текущото състояние на кода в екологичните списания, Мислън и съавторите му [8] идентифицират списания чрез търсене в Journal Citation Reports с помощта на следните термини за търсене: "Екология" за категория, "2013" за година, "SCIE" (Science Citation Index) и SSCI (Social Sciences Citation Index) за рецензирани издания и "Web of Science". Първите 96 списания са прегледани дали авторите споменават код или софтуер в контекста на научните изследвания. Резултатите, които са получили към 1 юни 2015 г. показват, че повече от 75% от списанията по екология не споменават научния код в ръководството за автора. От списанията, в които се споменава научен код, само 14% изискват да бъде предоставен кода, а 38% изискват предоставянето на данни. Много малка част от списанията (7%) са създали специална секция за софтуерни реализации или са добавили софтуерните реализации към списъка с опции на съществуващи раздели.
Резултатите ясно показват необходимостта от разработването на лесен за използване екологичен софтуер. Като пример за такъв може да се посочи статистическият и ГИС (Географска информационна система, Geographical Information System, GIS) инструмент за Windows Biota, който е написан за еколози, за да подпомогне анализа на пространствените и времеви данни [12]. През 2016 група учени предлагат свободен софтуер с отворен код за мета-анализ и мета-регресия OpenMEE: Open Meta- analyst for Ecology and Evolution с лесен за използване графичен потребителски интерфейс, който дава на изследователите еколози достъп до разнообразните и усъвършенствани статистически функционалности, предлагани в R, без да изискват познаване на R програмирането [10, 13].

Новите технологии за мониторинг в реално време на Черно море

Ограниченият достъп до надеждни временни редове от екологични, статистически и социално-икономически източници за мониторинг е основна пречка за разработването на политики и вземане на решения в областта на опазване на Черно море. За решаването на тези проблеми е разработена уеб базирана платформа ENVIROGRIDS, която да позволи откриването и достъпа до важна екологична информация за региона [7]. Тази платформа обхваща: земни, климатични и демографски сценарии;

хидрологията и свързаната с нея уязвимост; както и плажната ерозия. Всеки набор от данни е получен с помощта на съвременни инструменти от наличните мониторингови данни чрез използване на подходящи методи за валидиране. Платформата използва различни статистически процедури като метода Делта за редуциране на данните, Капа статистика, Фъзи Капа за валидиране на модела, Sequential Uncertainty Fitting program за анализи на несигурността.

Най-съвременните информационни технологии за автоматизирана обработка и визуализация на данни онлайн са използвани за разработка на Географската информационна система на Черно море (Geographical Information System GIS) [11]. В тази система за картографиране се използва Mapserver, за съхранение на данните се използва MySQL DBMS, а PHP и Python се използват за достъп до данни, тяхната обработка и обмен. Първата иновативна информационна система за мониторинг в реално време на Черно море е МИС на Черно море [14]. Тя е реализирана като уеб GIS приложение, интегрирана с SharePoint и поддържа множество специализирани уеб приложения. Системата осигурява функционалност за интеграция, обработка, анализ и визуализация на информация; възможност за обмен на данни с вътрешни и външни системи в реално време. Тя интегрира и обработва данни от сензори за автоматично измерване на параметрите на морската вода и атмосферния въздух, както и данни от Европейската програма „Коперник“ посредством услуги в реално време.

References

[1] Dall, S. R. X., Defining the concept of public information, Science 308, 2005, 353-354.

[2] Dall, Sasha R.X., Luc-Alain Giraldeau, Ola Olsson, John M. McNamara, David W. Stephens, Information and its use by animals in evolutionary ecology, Trends Ecol. Evol., 20, 2005, 187-193.

[3] Ellison Aaron M., Bayesian inference in ecology, Ecology Letters, 7, 2004, 509–520.

[4] van Gils, J. A., State-dependent Bayesian foraging on spatially auto-correlated food distributions, OIKOS, vol. 119, Issue2, February 2010, 237-244.

[5] Hooten M. B., N. T. Hobbs, A guide to Bayesian model selection for ecologists, Ecological monographs, vol. 85, Issue 1, February 2015, 3-28.

[6] Klomp N. I., D. G.Green, G. Fry, Roles of technology in ecology, In N. Klomp, & I. Lunt (Eds.), Frontiers in ecology: Building the links, Oxford: Elsevier, 1997, 299-309.

[7] Lehmann Anthony, Yaniss Guigoz, Nicolas Ray, Emanuele Mancosu, Karim C. Abbaspour, Elham Rouholahnejad Freund, Karin Allenbach, Andrea De Bono, Marc Fasel, Ana Gago-Silva, Roger Bär, Pierre Lacroix, Gregory Giuliani, A web platform for landuse, climate, demography, hydrology and beach erosion in the Black Sea catchment, Scientific Data, 4 July 2017, 1-15.

[8] Mislan K.A.S., Jeffrey M.Heer, Ethan P.White, Elevating The Status of Code in Ecology, Trends in Ecology & Evolution, vol. 31, Issue 1, January 2016, Pages 4-7

[9] Ovaskainen Otso, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, Nerea Abrego, How to make more out of community data? A conceptual framework and its implementation as models and software, Ecology Letters, vol. 20, Issue 5, May 2017, 561-576

[10] Wallac Byron C., Marc J. Lajeunesse, George Dietz, Issa J. Dahabreh, Thomas A. Trikalinos, Christopher H. Schmid, Jessica Gurevitch, OpenMEE: Intuitive, open- source software for meta- analysis in ecology and evolutionary biology, Methods in Ecology and Evolution, vol. 8, Issue 8, August 2017, 941-947.

[11] http://blacksea.grid.unep.ch
[12] http://www.ecostats.com/web/Biotas
[13] http://www.cebm.brown.edu/openmee/index.html
[14] https://misbs.bgports.bg/bg/climate-and-weather
[15] https://biostat-2009.soft32.com
[16] http://ntsyspc.software.informer.com
[17] www.metawinsoft.com