8

Design af kurset "Statistical Analysis of Spatial and Observational Ecological Data in R"

Michael Krabbe Borregaard

Natural History Museum of Denmark
University of Copenhagen

Introduction

Jeg har valgt at skrive om designet på et kursus i statistik og dataanalyse og brugen af R, et statistisk programmeringssprog, på kandidat-delen på biologi. Kurset var slået op som et 7.5 ECTS-point kursus med mig som kursusansvarlig og primær underviser. Desværre var der kun én enkelt tilmeldt studerende, og kurset blev aflyst. Jeg har derfor ikke mulighed for at inddrage noget empiri hvor jeg prøver mine idéer af i praksis. Den praktiske del af opgaven består altså i at jeg har forsøgt at implementere ideerne fra Universitetspædagogikum i en konkret plan for undervisningen. I forhold til den pædagogiske litteratur tager jeg først og fremmest udgangspunkt i lærebogen (Ulriksen, 2014).

Hvad er problemet, hvorfor er det relevant og hvem er det et problem for?

Statistik er et fag som de fleste biologer ikke bryder sig særligt meget om, og for mange er motivationen lav for at tilegne sig de færdigheder der ligger i det. Jeg tror en del af årsagen er at mange studerende ikke har en selvstændig interesse i statistik, og det bliver ikke kommunikeret tydeligt nok til dem at her er et fag der giver dem de kompetencer de har brug for til at kunne planlægge og udføre et selvstændigt forskningsprojekt. Altså at dataanalyse er en af de centrale færdigheder for at kunne lave det arbejde de drømmer om. De fleste studerende opdager halvvejs gennem specialet

at de mangler statistiske færdigheder, og på det tidspunkt er der nogle der tilegner sig dem selvstændigt, mens andre kommer i vanskeligheder.

Det centrale problem jeg ønsker at arbejde med er at statistik i meget høj grad er præget af overfladelæring for de studerende. Det er meget almindeligt at de studerende kommer væk fra et statistikkursus med nogle konkrete opskrifter på at kunne lave bestemte analyser, en fornemmelse af at de har hørt begreberne, og en god portion forvirring (eller frustration, for de studerende der har et stort ønske om at forstå dybere). Dette baseret på mit overordnede indtryk, da jeg ikke har lavet specifikke kvalitative interviews med studerende.

Det er mit indtryk at selve undervisningen i statistik tit er anlagt så den opmuntrer til overfladelæring. Statistiske metoder præsenteres ofte med en liste af antagelser der skal være opfyldt for at man kan bruge en test, som de studerende forventes at lære udenad eller kunne slå op i. Hver type test præsenteres med en liste over hvilke tilfælde den kan bruges. Det implicitte billede er at statistisk dataanalyse består i, billedligt talt, at gå ind i et arkiv med tests, identificere den rette hylde, og så tage den rigtige test ned og anvende den. Det er imidlertid et misvisende billede af hvad statistisk dataanalyse er, der ikke hjælper de studerende når de når til at skulle bruge det til deres egne selvstændige projekter. Det beskrevne problem med overfladelæring er derfor et problem for de studerende, og faktisk et ganske stort problem. Overfladelæring styrker ikke de studerendes motivation, og giver en fornemmelse af at faget er kedeligt. Samtidig vil de ofte senere i studiet have brug for en selvstændig dybdeforståelse for at kunne anvende det de har lært i deres egne projekter, og forsinkelser i dataanalyse kan give problemer for studerende der har brug for at færdiggøre deres projekter til tiden. Samtidig er data-analyse en af de kompetencer som mange studerende vil få brug for efter studiet, ikke mindst hvis de vælger at gå forsker-vejen.

Hvad er baggrunden for problemet

Problemet med overfladelæring har flere årsager:

Stoftrængsel

Statistisk dataanalyse er et meget omfattende område i sig selv, og blot at lære brugen af de værktøjer som vil være relevante for de fleste biologistuderende, i teori og praksis, må nødvendigvis tage lang tid. Faget er i høj

grad baseret på at opnå en færdighed, og for at opnå den færdighed kræves der en ret bred tilegnelse af teknikker. Det får underviserne til at anlægge et højt undervisningstempo, for at sikre at de når hele vejen rundt om stoffet. Min gamle statistiklærer, som har designet det nuværende statistik-fag på biologi, startede kurset med at sige: *"Vi tilstræber et tempo i undervisningen, der er som en bus der kører fra stoppestedet lige før I er kommet til det. Vi kører så langsomt at de fleste ikke når at give op og gå ind på fortovet, men så hurtigt at ingen rent faktisk indhenter bussen, sætter sig ind, og slapper af på sædet"*. Jeg kan huske at den dedikation til at køre fra de studerende gjorde indtryk på mig.

Læringsbasis

En anden vanskelighed er at den teori de skal lære ikke bygger naturligt oven på viden fra andre dele af studiet, og det derfor ikke er nemt at identificere zone of nearest learning. For at få en grundlæggende forståelse af de ofte avancerede statistiske teknikker prøver man i undervisningen ofte at gennemgå den underliggende matematik, men det kan forværre vanskeligheden, da de studerende tit ikke har den nødvendige basis for at forstå f.eks. matrixberegninger og linear algebra. Forsøget på at etablere en grundforståelse går derfor hen over hovedet på dem.

Relevans

Endelig kan det være svært for de studerende at se at statistikken er relevant for dem, idet de tit har kurset før de starter på specialearbejdet, og derfor ofte ikke har arbejdet mere dybtgående med egne data. De enkle øvelser og eksempler der anvendes i undervisningen ligger tit langt fra de biologiske spørgsmål de er optagede af, og det kan derfor være svært for de studerende at se hvad de skal bruge det til. Det motiverer de studerende til at lære med henblik på at bestå eksamen snarere end for at tilegne sig stoffet, hvilket også fører til overfladelæring.

Hvad kan man gøre for at afhjælpe det?

Helt overordnet kan man arbejde med en kultur baseret på overfladelæring på flere forskellige måder:

Man kan reducere problemet med stoftrængsel ved at udvælge kraftigt i stoffet, og acceptere at der er elementer, der absolut er relevante for de studerende, der ikke bliver behandlet i undervisningen. Det kræver at man identificerer eksemplars, altså enkelte teknikker og sager der kan give en mere overordnet indsigt i faget som sådan. Dataanalyse har en både lodret og vandret vidensstruktur, vandret idet forskellige analyser lægger sig ved siden af hinanden og supplerer hindanden, lodret idet mange teknikker bygger oven på de samme grundlæggende forudsætninger og overvejelser. Teknikken er her at udvælge såkaldte eksemplars, der giver de studerende mulighed for selv at gøre sig de grundlæggende overvejelser, men også viser bredden af teknikker.

Man kan også basere læringen på problem-baseret undervisning, dvs. løsning af opgaver og samarbejde mellem studerende. Dette er den type af undervisning der er lagt vægt på at vi arbejder med i Universitetspædagogikum. Ved at stille friere opgaver øger man nødvendigheden for de studerende af at få en dybdeforståelse af stoffet, og motiverer dem til at tilegne sig den forståelse. Ved at arbejde i dybden med konkrete biologiske problemer tydeliggør man også forbindelsen mellem dataanalysen og de øvrige fag på studiet.

En større udfordring er hvordan man håndterer at de studerende har meget forskelligt indgangsniveau, og at en del, i øvrigt dygtige, studerende ikke har tilstrækkelige matematisk viden til at forstå den grundlæggende matematik bag de statistiske teknikker. Mit forslag, som jeg er usikker på om virker, er at eksperimentere med at prøve at undervise de studerende i matematikken på en mere indirekte og intuitiv måde – for eksempel ved at bruge snore og farvede sten til at visualisere forudsætninger og teknikker.

Hvad har jeg gjort konkret?

For at imødekomme problemet med overfladelæring har jeg valgt en strategi der baserer sig på problem-baseret læring, hvor de studerende indlærer ved at arbejde med økologiske problemstillinger der ligger tæt på dem der indgår i undervisningen i øvrige fag. Mit mål er, ikke at køre en bus der kører fra de studerende, men snarere en hjælpevogn som dem der kører i Tour de France, der sørger for at give vejledning og en opmuntring på vej op ad Alpe d'Huez. Kerne-indsigten som jeg ønsker at overgive til de studerende er at data-analyse er en kreativ del af den videnskabelige proces, og at der findes mange måder at gribe det an på. Derfor må al analyse tage udgangs-

punkt i at stille sit videnskabelige spørgsmål så præcist så muligt, og så lade analysen afspejle det, snarere end automatisk at vælge en bestemt test.

Udvælgelse af stof

De studerende arbejder med 3 økologiske datasæt, der bliver ved med at dukke op i forskellige undervisningstimer, til at illustrere forskellige elementer. Datasættene er: Et datasæt med artsrigdommen af forskellige dyregrupper fra forskellige øer fra hele verden (fra en artikel af Bunnefeld & Phillimore, Ecography 2010); et datasæt med skove i Volzhsko-Kamsky nationalparken (fra statistikbogen ”Analyzing Ecological Data” af Zuur et al., 2007); og et datasæt med artsudbredelser og fylogenier for kolibrier, der er en del af mit lab’s datasamling. Data er udvalgt så de kan bruges til at svare på mange forskellige spørgsmål, og er rimeligt eksotiske og biologisk spændende. Disse data fungerer som en form for eksemplars.

Der er en risiko som jeg dog er bekymret for ved at anvende de samme data flere gange. Hvis de foretager lignende (men tiltagende sofistikerede) statistik på de samme data kan det skabe forvirring over hvornår de laver noget der er forkert, eller i hvert fald ikke optimalt for et datasæt.

Evaluering

Jeg integrerer formativ evaluering i undervisningen. Idet udgangspunktet for kurset er at give de studerende en færdighed til at foretage selvstændig analyse, og at få erfaring med at tage valg og vurderinger i analyseprocessen, giver det ikke mening at afslutte kurset med en eksamen baseret på at kunne svare rigtigt på en række spørgsmål. I stedet sikrer jeg constructive alignment mellem målet og kursets elementer ved at lægge fokus på selvstændigt arbejde i undervisningen, og ved at evalueringen foregår som en aflevering af 3 projektrapporter i løbet af kurset. Rapporter afleveres i RMarkDown, som er et format der kombinerer grafik, computerkode og formateret tekst. RMarkDown er i sig selv et læringsmål for kurset. Rapporterne skrives 2 og 2 baseret på en selvstændig dataanalyse med et økologisk spørgsmål, og der afsættes en undervisningstime til mundtlig feedback til individuelle grupper efter hver rapport.

Arbejde med artikler

Et andet element som skal øge relevansen og sammenhængen med de øvrige ting de lærer på studiet er arbejde med analyser i publicerede artikler,

gerne artikler der har data og analysekode som supplementary materials. En ofte underkendt funktion af statistiske færdigheder er at kunne vurdere og forstå andres artikler. Der vil hver uge være diskussion og gennemgang i grupper på 4 af en artikel, baseret på nogle spørgsmål stillet på forhånd af mig. De vil bla. blive bedt om at identificere tvivls-spørgsmål, uudnyttede muligheder og evt svagheder. Undervejs vil eleverne 2-3 af gangene herefter prøve at gentage analysen fra artiklen, og evt inddrage de muligheder de har identificeret. En af disse gange vil de skulle skrive en rapport, der er en af de 3 der nævnes ovenfor.

Fri arbejdsform og gruppearbejde

Som beskrevet ovenfor vil en del arbejde bestå i gruppe-arbejde, og der lægges vægt på at opgaverne stilles så de studerende må læse i hjælpefiler – eller i hjælpefora på internettet – for at kunne besvare opgaverne fuldt ud. Dette er den samme proces mange specialestuderende gennemgår senere, hvor de ikke nødvendigvis kan få god feedback på det analytiske. Her har jeg også én bekymring: at de studerende har svært ved at følge den mindre strukturerede form og forfalde til facebook m.m. når det bliver vanskeligt at se det næste skridt.

Hands-on

Endelig vil jeg prøve at håndtere det vanskelige problem at statistik bygger på relativt kompleks matematik som der ikke er mulighed for at lære de studerende at forstå til bunds med mindre de allerede har en stærk matematisk baggrund. Jeg tror man kan opnå en del ved at lade de studerende få ting i hænderne. F.eks måle vægten på en række småsten, og lave en fordeling og gøre det samme med sneglehuse (sikkert normal, hvis det er samme art, log-normal hvis det er forskellige arter), og bruge det som udgangspunkt for en diskussion af fordelinger (eller af, fra bunden – hvad ville de studerende gøre hvis de skulle udvikle en test for om den ene art havde større huse end den anden). Eller at bruge en pind og elastikker på søm til at udtrykke hvordan least squares (cirka) virker når man fitter en trendlinje. Eller at have små stykker plastik (hvis man kan skaffe det) som man kan lægge oven i hindanden for at illustrere additiv effekt af variable i en multipel regression.

Er der dokumentation for at det fungerer?

Jeg har ikke pt. nogen dokumentation for at mine idéer vil fungere. Jeg har dog det indtryk, bla. fra lærebogen, at der er bred empiri der understøtter at gruppearbejde, problem-baseret læring, og en satsning på dybde-forståelse er en effektiv undervisningsform. Nu blev kurset aflyst i år, men det giver mig tid til at udvikle idéerne endnu mere konkret, og evt at opsætte en prøveballon med nogle af de her elementer i forbindelse med noget af den anden undervisning jeg skal lave på andre kurser i det kommende år.

Hvilke overvejelser gør jeg nu? Er der flere perspektiver?

Der er masser af muligheder for at gøre det mere spændende. Det kunne være f.eks. spændende at lade de studerende generere et datasæt for holdet (højde, køn, geografisk baggrund, familiebaggrund etc.) og så også bruge det datasæt gennemgående. Det kræver dog nok et hold på mindst 30 personer at få data nok til det, så det er et perspektiv på længere sigt. Perspektivet er at blive ved med at udvikle kurset. Jeg forestiller mig at skruer man det rigtigt sammen kan man lave et rigtigt sjovt og interessant kursus, der samtidig giver de studerende baggrund for at lave et bedre speciale end de ellers kunne have gjort – og måske med mindre frustration.

Referencer

Ulriksen, L. (2014). *God undervisning på de videregående uddannelser* (1. udg.). Frydenlund.

A Learning outcomes

Statistical Analysis of Spatial and Observational Ecological Data in R (REcoStat)

Education

MSc Programme in Biology

Content

The course aims at giving biology students the tools to perform independent data analysis for projects in ecology, and to understand and critically debate statistical data analysis in published papers. The main tool used in the course is the scientific programming language R, which is the de facto standard for ecological data analysis. The course focuses primarily on comparative analyses and observational data, which pose different challenges than designed experiments. The exercises will give the students an overview of the tools and packages available for analyzing and visualizing ecological data. Students will work independently on data analysis projects that focus on building the competence to do independent data analysis projects. Students are recommended to take the course during or prior to beginning their MSc thesis project.

Learning outcomes

After completion of the course, the student is expected to be able to:

Knowledge:

- describe the basic elements of the R programming language
- detail the statistical methods available for analysis of observational data
- describe the issue of pseudoreplication and autocorrelation, and detail the possible methods to deal with it
- know functions implemented in the R packages vegan for community ecological data analysis, nodiv and ape for working with macroecological data with phylogenetic trees, sp and raster for spatial data, ggplot2 for data visualization and dplyr for manipulation of data sets
- describe the difference between frequentist and bayesian statistics

Skills:

- use R to load data sets and do basic data analysis tasks
- program simple functions and simulations

- use the R documentation to find solutions for coding problems
- produce basic figures, such as scatter plots, histograms and bar plots for data visualization
- test and summarize statistical models of ecological data
- identify the assumptions of statistical tests and test if they are met
- use standard linear regression, and derived techniques, such as spatial linear models, generalized linear models with different error families, phylogenetic regression, and random effects.
- use the Rmarkdown syntax to produce a lab log of the analytical processes in a statistical analysis

Competences:

- work independently to perform statistical analyses in ecology
- understand the biological background and significance of different statistical tests and outcomes
- critically debate and replicate analyses in published research papers
- identify the right packages and tools for their data and problem
- identify and acquire the necessary knowledge to conduct novel types of analysis

Teaching and learning methods

The teaching will consist of class-room teaching that blends lectures with practical exercises. The practical exercises will be supplemented with discussions of analyses and discussions of some published papers. Three times during the course, the students will work in groups to prepare a lab report detailing a statistical data analysis, written in Rmarkdown.

Academic qualifications

The students should be familiar with basic statistical terms, such as variance, mean, normal distribution, common linear regression, significance and hypothesis testing. No previous experience with R or statistical software is assumed. Students familiar with R must expect to experience some repetition, as this constitutes an important element in the course.

Exam

Credit: 7.5 ECTS
Type: Assesment based on participation and the delivered reports.
Marking: passed/not passed
Censorship: No external censorship