Matematisk beskrivelse af Dansk Vandløbsplante Indeks

Notat fra DCE - Nationalt Center for Miljø og Energi

Dato: 25. august 2015

Søren E. Larsen og Annette Baattrup-Pedersen

Institut for Bioscience
Aarhus Universitet

Rekvirent:
Naturstyrelsen
Antal sider: 14

Faglig kommentering:
Peter Wiberg-Larsen
Kvalitetssikring, centret:
Poul Nordemann Jensen

AARHUS
UNIVERSITET
DCE – NATIONALT CENTER FOR MILJØ OG ENERGI

Tlf.: 8715 0000
E-mail: dce@au.dk
http://dce.au.dk

Indhold

Indledning

Som et led i implementeringen af EU's vandrammedirektiv er der udviklet et indeks til vurdering af økologisk tilstand fra plantesammensætningen i danske vandløb: *D*ansk *V*andløbs*P*lante *I*ndeks (DVPI; se Baattrup-Pedersen & Larsen 2013; Søndergaard et al. 2013). DVPI beregnes på baggrund af en artsliste og arternes dækningsgrader ved hjælp af en prædiktionsmodel, der klassificerer vandløbet i en tilstandsklasse med en tilhørende EQR værdi.

I udviklingen af DVPI er eksperters viden om plantesamfund i danske vandløb anvendt, herunder deres kendskab til effekten af menneskeskabte påvirkninger på vegetationen. I alt 6 eksperter har uafhængig af hinanden vurderet plantedata fra 1.244 vandløbsstrækninger og henført disse i en økologisk tilstandsklasse. Ud over artslister og dækningsgrader fra vandløbstrækningerne har eksperterne i forbindelse med deres arbejde også haft oplysninger om vandløbenes størrelse (bredde og dybde) og vandets alkalinitet som er væsentlige naturlige plantefordelende faktorer i danske vandløb.

Efterfølgende analyser viste at der var god overensstemmelse mellem eksperternes vurdering af økologisk tilstand. Derfor blev ekspertvurderingerne anvendt til at træne en statistisk model i at genfinde mønstre i plantesammensætningen svarende til de mønstre eksperterne tillagde værdi i deres tilstandsfastsættelse. Den statistiske model blev efterfølgende afprøvet og grænsefastsættelsen mellem tilstandsklasserne blev justeret, således at der var overensstemmelse til de interkalibrerede værdier.

Formålet med dette projekt er at give en fuldstændig og komplet faglig dokumentation af alle de matematiske og statistiske beregninger, der ligger til grund for udviklingen af DVPI, samt for beregningen af DVPI for en artliste som ikke har været anvendt i udviklingen af indekset. Det er aftalt med Naturstyrelsen, at dokumentationen skal leveres i form af en faglig rapport, således at det er muligt senere og ved anvendelse af rapporten, at beregne et planteindeks ud fra en given artsliste. Den faglige dokumentation vil endvidere også indeholde beskrivelse af, hvordan planteindekset oversættes til en EQR (Ecological Quality Ratio) værdi.

Metodeoversigt

Udviklingen af DVPI og tilhørende EQR værdi bestod af en række adskilte vurderinger og beregninger med anvendelse af et antal forskellige matematiske og statistiske metoder. Fra etablering af empiriske data til definition af DVPI bestod udviklingen af følgende 6 trin:

1. Specialist vurdering af 1244 artslister fra danske vandløbsstrækninger som resulterer i en a priori klassifikation af disse.
2. Anvendelse af DCA (Detrended Correspondence Analysis) på dette store dataset til ordination af disse strækninger på baggrund af deres kvantitative plantesammensætning.
3. Valg af et træningsdatasæt, som er en delmængde af de 1244 artslister anvendt af specialisterne.
4. QDA (Quadratic Discriminant Analysis) baseret på DCA scores og træningsdatasæt. Giver som resultat en superviseret klassifikationsmodel, som etablerer en metode til prædiktion af sandsynligheder for at falde i en af de fem klasser anvendt ved specialisternes evaluering.
5. Metode til beregning af DVPI_EQR, som er EQR værdien hørende til den økologiske tilstandsklasse.

6. Metode til beregning af DVPI ud fra DVPI_EQR.

I det følgende gennemgås disse punkter og det beskrives matematisk og statistisk, hvilke beregninger der er nødvendige i de enkelte trin.

Specialist vurderinger

I udviklingen af DVPI er som udgangspunkt anvendt data fra i alt 1244 vandløbsstationer, dels fra det nationale overvågningsprogram NOVANA Ferskvand, dels det tidligere overvågningsprogram NOVA samt diverse data fra forsknings- og rådgivningsprojekter. De anvendte data er beskrevet i detaljer i Baattrup og Larsen (2013) i afsnit 2.1 – 2.3. I alt 6 uafhængige specialister skulle a priori klassificere vandløbene i fem forskellige økologiske tilstandsklasser: høj, god, moderat, ring og dårlig økologisk kvalitet. Specialisternes a priori klassifikation skulle baseres på artslister, dækningsgrader, oplysninger om bredde og dybde og vandets alkalinitet samt eksisterende viden om, hvordan uforstyrrede plantesamfund ser ud samt hvilke ændringer, der sker som følge af antropogene påvirkninger. A priori klasserne blev anvendt til at træne klassifikationsmodellen i at genkende vandløbstilstand ud fra artssammensætning og dækningsgrad.

Specialisterne anvendte en række forskellige kriterier til at fordele vandløbsstrækningerne i de 5 forskellige klasser og de kriterier kan ses i bilag 2 i Baattrup-Pedersen og Larsen (2013). Hovedparten af strækningerne blev af specialisterne klassificeret til at have økologisk tilstandsklasse 3, dvs. moderat økologisk tilstand. Ligeledes blev mange strækninger klassificeret til klasse 2 eller 4. Meget få blev klassificeret til klasse 1 eller 5. Enigheden i specialisternes klassifikation blev beregnet via et indeks *I* (Baattrup-Pedersen og Larsen, 2013). Indekset antog værdier mellem 0,61 og 0,82. Dette skal sammenholdes med at 0.06 angiver størst uenighed og 1 angiver fuldkommen enighed. Overensstemmelsen blev derfor vurderet til at være tilstrækkelig god til at anvende eksperternes vurdering til træning af prædiktionsmodellen.

DCA

En Detrended Correspondence Analysis (DCA) er anvendt på data fra alle vandløbsstrækninger for at kunne repræsentere variationen i vegetationens sammensætning i form af få floristiske gradienter/ordinationsakser. Prøvefelternes (vandløbsstrækningernes) placering langs de 3 vigtigste ordinationsakser blev anvendt som forklarende variable i klassifikationsmodellen. Det var derved vigtigt at inddrage alle data, det vil sige alle vandløbsstrækninger da man således opnår en robust og stabil ordination ved anvendelsen af DCA.

DCA er en multivariat statistisk metode, som anvendes til at beregne vinkelrette gradienter i store data matricer, hvor der forefindes mange prøver og mange arter, men hvor matricen er sparsom, hvilket betyder at den indeholder en del nulværdier fordelt rundt omkring i matricen. Dette er typisk for plantedata, fordi ikke alle arter er registreret på alle vandløbsstrækninger (prøver). DCA er implementeret som en iterativ algoritme som beregner 3 eller flere hovedakser, som hver især typisk har en økologisk betydning.

DCA er en videreudvikling af metoden Correspondence Analysis (CA) eller Reciprocal Averaging (RA) som den også kaldes. DCA blev udviklet i 1979 af M. O. Hill (Hill og Gauch, 1980) og anledningen til udviklingen var, at CA

gav en række uheldige resultater. Blandt andet havde CA en tendens til at producere en gradient som repræsenterede samples i en buet form. DCA løser dette problem ved at anvende "detrending".

Metoden er som nævnt implementeret som en iterativ algoritme i forskellige software pakker. Først og fremmest i pakken PC-ORD (McCune og Mefford, 2011), men også i CANOCO og S-PLUS. DCA kan ligeledes køres i den gratis pakke i R via "decorana()" funktionen i "vegan" biblioteket. I arbejdet med udviklingen af DVPI er PC-ORD anvendt.

DCA har en tendens til at give sjældne arter for stor betydning. Derfor er det en god ide, at bruge muligheden for at nedtone disse ved at anvende "downweight rare species" som er en option i de fleste software pakker. På grund af metodens iterative natur er det ikke muligt at give en detaljeret beskrivelse, da iterationerne afhænger af data, men metodens beregningstrin i algoritmen kan formelt beskrives.

I det følgende vil vi anvende følgende matematiske notation. Lad *A* betegne en *nxp* matrice, som repræsenterer data fra de 1244 (*n*) vandløbsstrækninger som de 6 specialister har klassificeret. Matricen *A* indeholder dækningsgrader for *p* (194) forskellige plantearter på de *n* strækninger.

DCA starter ud med algoritmen fra RA (Hill, 1973). Denne algoritme har følgende trin

1. Vælg et vilkårligt sæt af scores $y_j, (j = 1,2,\cdots,p)$ for arterne.

2. Standardiser disse artscores til middelværdi nul og længde 1 ved at anvende vægte defineret ved artsmatricen *A*

$$\sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}y_j = 0\,, \sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}y_j^2 = 1.$$

3. Ordiner strækninger ved anvendelser af vægtede gennemsnit sådan at score for strækning i, x_i, er middelscoren af de arter som optræder på strækning *i*

$$x_i = \frac{\sum_{j=1}^{p} a_{ij}y_j}{\sum_{j=1}^{p} a_{ij}}, i = 1,2,\cdots,n.$$

4. Ordiner nu arter, sådan at score for hver art er den vægtede middel score for de strækninger, hvor arten er registreret på. Lad $y_j^{'}, (j = 1,2,\cdots,p)$ betegne de nye artscores

$$y_j' = \frac{\sum_{i=1}^{n} a_{ij}x_i}{\sum_{i=1}^{n} a_{ij}}, j = 1,2,\cdots,p.$$

5. Anvend de nye artscores $y_j^{'}, (j = 1,2,\cdots,p)$ som basis scores og vend tilbage til trin 2. Anvend denne iterative algoritme indtil begge sæt af scores er stabiliseret og kun ændrer sig minimalt fra en iteration til den næste. Den aktuelle stoppe regel vil være bestemt af hvilken software pakke man anvender.

Vektoren $y_j, (j = 1,2,\cdots,p)$ kaldes i RA for "forsøgs" vektoren og hvis der gælder at $y_j = y_j^{'}, (j = 1,2,\cdots,p)$ så er $y_j, (j = 1,2,\cdots,p)$ en egenvektor som er et begreb fra matematisk algebra. Så med andre ord går algoritmen ud på at beregne et sæt artscores, som er en egenvektor til artsmatricen.

Denne algoritme giver første ordinationsakse i en RA ordinationen. For at beregne anden akse i ordinationen skal man bruge samme algoritme igen men tilføje følgende trin lige efter trin 3.

3a. Ortogonaliser med hensyn til første akse ved at fratrække værdien af en lineær regression på den første akse. Hvordan man lige gør dette er ikke vigtigt for beskrivelsen af DCA, som det vil fremgå lidt senere. Hvis $x_i^{'}, (i = 1,2,\cdots,n)$ er scores for strækninger på første akse, så er ortogonalitet defineret ved reglen

$$\sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}x_i^{'}x_i = 0.$$

Dette er ikke ortogonalitet som defineret i den traditionelle matematiske formel fra algebra, men en vægtet ortogonalitet, hvor vægte er indgangene i artsmatricen *A*.

For at beregne ordinationsakse 3 skal der ortogonaliseres på både akse 2 og 1, og så videre for restende ordinationsakser.

Ovenstående er den komplette beskrivelse af RA ordination. I DCA ændres trin 3a til at være en detrending samt en reskalering i stedet for en ortogonalisering. I figur 3 i artiklen af Hill og Gauch (Hill og Gauch, 1980) er detrending metoden gengivet. Beskrevet med ord gør detrending følgende: Den første ordinationsakse (akse 1) deles op i et antal segmenter med ens længde (antallet af segmenter er 26 i PC-ORD). I hver segment recentreres sample scores på akse 2 sådan at de har middelværdi nul for det givne segment. Til sidste foretages en rescalering af de detrended artscores, ved at gøre de vægtede varianser af artscores lig 1 i hvert af segmenterne.

En mere detaljeret beskrivelse af det DCA program, der er i PC-ORD, er givet i Hill (1979). Funktionen "decorana" i R benytter en anden form for detrending.

Træningsdatasæt

Den model der er blevet anvendt til udviklingen af DVPI er en superviseret klassifikationsmodel. Der findes et antal forskellige statistiske metoder til klassifikation af diskrete data og ved udviklingen af DVPI er der anvendt kvadratisk diskriminantanalyse. Da vi anvender superviseret klassifikation til udvikling af en model til vurdering af økologisk tilstand, er der behov for a priori klassificerede plantedata til at træne modellen med. Træningsdata udvælges som en delmængde af de 1244 vandløbsstrækninger, og skal bestå af strækninger indenfor alle økologiske tilstandsklasser. De valgte træningsdata består af ca. 30% af de 1244 strækninger og de er udvalgt blandt de strækninger, hvor der er størst enighed blandt specialisterne i deres klassifikationer.

Det er vigtigt for udviklingen af et vellykket planteindeks, at træningsdata udvælges med omhu, således at modellen giver estimater, som passer godt med virkeligheden målt på et datasæt, hvor virkeligheden kendes bedst muligt. Vi har været igennem flere iterationer mht. træningsdatasættet for at finde den delmængde af de 1244 stationer, som gav det bedste modelfit. I Baattrup-Pedersen og Larsen (2013) afsnit 2.7 beskrives træningsdata.

Kvadratisk diskriminations analyse

I udviklingen af DVPI har vi som allerede nævnt anvendt en såkaldt superviseret klassifikation som den statistiske metode.

Den superviserede klassifikationsmodel er bygget op omkring den parametriske statistiske metode, benævnt kvadratisk diskriminant analyse (Quadratic Discriminant Analysis, QDA). Det er en statistisk klassifikationsmetode som kan anvendes til at separere 2 eller flere klasser af objekter. I tilfældet med DVPI vil den superviserede klassifikationsmodel forudsige, hvilken tilstandsklasse en vandløbsstation har størst floristisk lighed med ud fra dets sammensætning af arter. Metoden bygger på a priori klassificerede vandløb i de 5 økologiske tilstandsklasser høj, god, moderat, ringe og dårlig, og klassifikationsmodellen vil efterfølgende kunne benyttes til at beregne sandsynligheden for et givent vandløbs tilhørsforhold til hver a priori klasse, som direkte kan anvendes i fastsættelse af økologisk tilstand på en skala fra 0 til 1.

Før det er det relevant at beskrive QDA (Venables og Ripley, 1999), er det nødvendigt at gennemgå andre statistiske begreber. Det drejer som om begreber tilhørende metoder fra beslutningsteorien: Tab, risiko, Bayes klassifikationsmodel, "plug-in" klassifikationsmodel og lineær diskriminant analyse (Ripley, 1996). Men først skal der beskrives den notation, der skal anvendes i gennemgangen af den superviserede klassifikation.

Lad T være en vektor af længde m som indeholder a priori økologisk tilstandsklasse for de m vandløbsstrækninger som er blevet udvalgt som træningsdata. Med andre ord er T blevet bestemt af de 6 specialister på baggrund af en del af de 1244 plantelister som udgør matricen A. Værdierne i T er enten 1,2,3,4 eller 5.

Der vælges 3 ordinationsakser fra DCA, og det man skal have fat i er scores/koordinater til de m vandløbsstrækninger på disse 3 akser. Disse betegnes samlet X_T. Med andre ord er X_T en mx3 matrice. Som beskrevet under afsnittet om DCA, så benævnes de 3 akser normalt som gradienter og er alle vinkelrette på hinanden. De data der dermed skal være input til den superviserede klassifikationsmodel er (X_T,T) og målet med modellen er at være i stand til at estimere klassen for nye observationer/plantelister, når vi kun kender X, det vil sige placeringen af den nye strækning som kan beregnes via arternes placering langs de 3 DCA gradienter. Så man kan sige, at økologisk klasse er responsvariablen og de 3 gradienter er de forklarende variable i modellen. I matematisk sprogbrug er målet med modellen: Estimer så godt som muligt Y givet værdien X for en ny planteliste, hvor Y er den økologiske klasse og værdimængden for y er $\mathcal{Y} = \{1,2,3,4,5\}$.

Nu er den basale notation på plads og selve modelapparatet kan så beskrives. Først skal man beskrive hvilken tabsfunktion L man vil benytte sig af i modeludviklingen. En tabsfunktion skal beskrive, hvilket tab eller hvilken fejl modellen begår ved estimering af en ny klasse, det vil sige hvor god estimatet er. Her skal følgende tabsfunktion benyttes

$$L(Y,\hat{Y}) = \begin{cases} 0, Y = \hat{Y} \\ 1, Y \neq \hat{Y} \end{cases}$$

hvor $\hat{Y}$ er estimatet for Y. Denne tabsfunktions kaldes også misklassifikationen.

For en given tabsfunktion L er risikoen for en klassifikationsmodel givet ved det forventede tab

$$R(\hat{Y}) = \mathbb{E}\left(L(Y, \hat{Y})\right)$$

hvor nu $\hat{Y} = \hat{Y}(X)$ er en funktion af den tilfældige forklarende variabel X. Ideelt så ønsker man at få en klassifikationsmodel, som minimerer risikoen. Den teoretiske risiko er ukendt, men da man har m vandløbsstrækninger i træningsdatasættet, så kan man benytte den empiriske risiko. Så derfor skal man forsøge at minimere den empiriske risiko defineret ved

$$R_m(\hat{Y}) = \mathbb{E}_m\left(L(Y, \hat{Y})\right) = \frac{1}{m}\sum_{i=1}^{m} L(Y_i, \hat{Y}_i).$$

Nu man kan man stille spørgsmålet: Hvad er den bedste klassifikationsmodel hvis den statistiske fordeling for (X,Y) er kendt? Det er Bayes klassifikationsmodellen, som forklares i det følgende. Fordelingen f for den tilfældige forklarende variabel X kan skrives som

$$f(X) = \sum_{k=1}^{5} f_k(X)P(Y = k),$$

hvor, for $k = 1,2,3,4,5$, a priori sandsynligheder for klasser er $P(Y = k) = \pi_k$ og $f_k(X)$ er de betingede fordeliger for X givet $Y = k$. Bayes klassifikationsmodellen $\hat{Y}$ er den model, der minimerer risikoen med misklassifikationen som tabsfunktion. Teoretisk kan vi skrive risikoen som

$$\begin{aligned} R(\hat{Y}) &= \mathbb{E}[L(Y, \hat{Y}(X))] \\ &= \mathbb{E}\big[\mathbb{E}[L(Y, \hat{Y}(x))|X = x]\big] \\ &= \int_{\mathcal{X}} \mathbb{E}\big[\mathbb{E}[L(Y, \hat{Y}(x))|X = x]\big] f(x)dx, \end{aligned}$$

hvor $\mathcal{X}$ er mængden af værdier X kan antage. For Bayes klassifikationsmodellen er det nok at minimere $\mathbb{E}[L(Y, \hat{Y}(x))|X = x]$ for hver værdi x. Med andre ord givet $X = x$, så skal $\hat{Y}(x) \in \{1,2,3,4,5\}$ vælges sådan, at det betingede tab er mindst muligt. Nu er

$$\mathbb{E}\left[L\left(Y, \hat{Y}(x)\right)|X = x\right] = \sum_{k=1}^{5} L\left(Y, \hat{Y}(x)\right) P(Y = k|X = x).$$

Vælges $\hat{Y}(x) = l$, hvor $l \in \{1,2,3,4,5\}$, så er

$$\mathbb{E}\left[L\left(Y, \hat{Y}(x)\right)|X = x\right] = 1 - P(Y = l|X = x).$$

Bayes klassifikationsmodellen vælger den klasseværdi, som har den største a posteriori sandsynlighed, det vil sige

$$\begin{aligned} \hat{Y}(x) &= \arg\max_{k=1,2,3,4,5} P(Y = k|X = x) \\ &= \arg\max_{k=1,2,3,4,5} \frac{\pi_k f_k(x)}{\sum_{k=1}^{5} \pi_k f_k(x)} \\ &= \arg\max_{k=1,2,3,4,5} \pi_k f_k(x). \end{aligned}$$

Man er normalt i en situation, hvor man hverken kender a priori sandsynlighederne π_k og de betingede fordelinger f. Men med givne estimater $\hat{\pi}_k$ for $\pi_k, k = 1,2,3,4,5$ og $\hat{f}_k(x)$ så er en "plug-in" klassifikationsmodel, den model der vælger klassen defineret ved

$$\hat{Y}(x) = \arg\max_{k=1,2,3,4,5} \hat{\pi}_k \hat{f}_k(x).$$

Både lineær diskriminant analyse (LDA) og QDA er eksempler på "plug-in" klassifikationsmodeller. LDA er mere simpel end QDA og det vil være nyttigt først at beskrive LDA, som er den mest kendte og mest simple form for "plug-in" klassifikationsmodel. I LDA antager man følgende parametriske form for $f_k(x)$, som er fordelingen for *X* givet $Y = k$

$$X|Y = k \sim \mathcal{N}(\mu_k, \Sigma),$$

altså en 3-dimensional Normalfordeling med 5 forskellige middelværdivektorer, men den samme kovariansmatrice. Dermed har man, at

$$P(Y = k|X = x) \propto \pi_k f_k(x)$$
$$\propto \frac{\pi_k}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right\}$$

Da der helt generelt gælder, at

$$\arg\max_{k=1,2,3,4,5} g(x) = \arg\min_{k=1,2,3,4,5} -2\log g(x)$$

for en vilkårlig reel funktion *g*, så skal man vælge klassen *k* sådan, at

$$-2\log P(Y = k|X = x) \propto (x-\mu_k)^T \Sigma^{-1}(x-\mu_k) - 2\log(\pi_k) + konstant$$

minimeres. I formlen afhænger konstanten ikke af klassen *k*.

I statistisk teori kaldes størrelsen $(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)$ for Malahanobis afstanden mellem *x* og μ_k i metrikken (afstandsfunktion i matematisk forstand) givet ved Σ.

Ved at gange udtrykket $(x-\mu_k)^T \Sigma^{-1}(d-\mu_k)$ ud, kan man se, at $-2\log P(Y = k|X = x)$ er proportional med en lineær diskriminant funktion, som deler rummet $\mathcal{X}$ ind i områder med samme klasse estimater ved hjælp af opdelende rette flader.

Man mangler nu, for at definere LDA præcist, at finde gode estimater for $\pi_k, k = 1,2,3,4,5$, $\mu_k, k = 1,2,3,4,5$ og Σ, sådan at LDA kan defineres via en "plug-in" klassifikationsmodel som tidligere nævnt.

A priori sandsynlighederne $\pi_k = P(Y = k)$ er ganske simpelt estimeret ved andelen af strækninger i træningsdatasættet med klassen *k*, det vil sige

$$\hat{\pi}_k = \frac{|\{i: Y_i = k\}|}{m}.$$

Hvis man lader $m_k = |\{i: Y_i = k\}|$ betegne antallet af strækninger med klassen *k*, så er

$$\hat{\mu}_k = \frac{1}{m_k} \sum_{j:Y_k=j} x_j$$

og

$$\hat{\Sigma} = \frac{1}{m}\sum_{k=1}^{5}\sum_{j:Y_k=j}(x_j - \hat{\mu}_k)(x_j - \hat{\mu}_k)^T.$$

Nu er så den bedste klassifikationsmodel, under antagelsen at $X|Y = k \sim \mathcal{N}(\hat{\mu}_k, \hat{\Sigma})$ med "plug-in" estimater for $\mu_k, k = 1,2,3,4,5$ og Σ, lig

$$\begin{aligned}\hat{Y}_{LDA}(x) &= \arg\min_{k=1,2,3,4,5}\left\{(x - \hat{\mu}_k)^T\hat{\Sigma}^{-1}(x - \hat{\mu}_k) - 2\log(\hat{\pi}_k)\right\} \\ &= \arg\min_{k=1,2,3,4,5}\left\{\hat{\mu}_k{}^T\hat{\Sigma}^{-1}\hat{\mu}_k - \hat{\mu}_k{}^T\hat{\Sigma}^{-1}x - 2\log(\hat{\pi}_k)\right\}.\end{aligned}$$

I QDA klassifikationsmodellen antager man, at $X|Y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$, det vil sige at man nu tillader at hver af de 5 klasser har en forskellig kovariansmatrice. Man kan matematisk bevise, at $-2\log P(Y = k|X = x)$ nu kan omskrives til en kvadratisk funktion og derfor omtaler vi denne model som QDA og derfor deles rummet $\mathcal{X}$ ind i områder med samme klasse estimater ved hjælp af opdelende kvadratiske flader.

Den eksakte formel for QDA klassifikationsmodellen er

$$\hat{Y}_{QDA}(x) = \arg\min_{k=1,2,3,4,5}\left\{(x - \hat{\mu}_k)^T\hat{\Sigma}_k^{-1}(x - \hat{\mu}_k) - 2\log(\hat{\pi}_k) + \log(|\hat{\Sigma}_k|)\right\}$$

for hvert punkt $x \in \mathcal{X}$ og hvor "plug-in" estimaterne $\hat{\pi}_k$ og $\hat{\mu}_k$ er som for LDA og

$$\hat{\Sigma}_k = \frac{1}{m_k}\sum_{j:Y_k=j}(x_j - \hat{\mu}_k)(x_j - \hat{\mu}_k)^T.$$

Med hensyn til antagelse om multivariat Normalfordeling, så er det i DVPI tilfældet en fuldt gyldig antagelse da de forklarende variable er konstrueret via en DCA analyse. Både LDA og QDA er implementeret i software pakkerne S-PLUS og R som funktionerne "lda()" og "qda()" i biblioteket "MASS" (Venables og Ripley, 1999). Disse to funktioner danner ved modelestimeringen en generisk funktion kaldet "predict()", således at man kan estimere de 5 a posteriori sandsynligheder for at en planteliste fra en ny vandløbsstrækning tilhører hver af de fem tilstandsklasser. Med andre ord er output fra "predict()" en vektor med de 5 a posteriori sandsynligheder $P(\hat{Y} = k|X = x), k = 1,2,3,4,5$.

Sandsynligheder

Resultatet af den superviserede klassifikationsmodel er en matematisk/statistisk metode til at beregne fem sandsynligheder (P_1, P_2, P_3, P_4, P_5) baseret på en artsliste med dækningsgrader fra en given vandløbsstrækning og DCA scores for arterne i artslisten. De fem sandsynligheder angiver sandsynligheden for, at prøven tilhører hver af de fem tilstandsklasser, og den klasse med den største sandsynlighed blandt de fem er tilstanden for den aktuelle vandløbsstrækning.

EQR

I dette afsnit anvises, hvordan man skal beregne DVPI_EQR, som er EQR værdien hørende til en observeret planteliste for en vandløbsstrækning. Beregningen af DVPI_EQR tager udgangspunkt i de 5 sandsynligheder (P_1, P_2, P_3, P_4, P_5), som angiver sandsynlighederne for at vandløbsstrækningen, beregnet via den superviserede klassifikationsmodel, tilhører en af de 5 klasser, som er defineret a priori af de 6 specialister. DVPI_EQR beregnes ved anvendelse af følgende formel

$$DVPI_{EQR} = \frac{5 \cdot P_1 + 15 \cdot P_2 + 25 \cdot P_3 + 45 \cdot P_4 + 55 \cdot P_5}{50}.$$

Som det kan ses af formlen er den teoretiske mindste værdi for DVPI_EQR lig 0,1 og den maksimale værdi lig 1,1. Giver en planteliste en DVPI_EQR værdi over 1 kan man vælge at nedskrive værdien til 1 da en EQR værdi skal tilhører intervallet [0;1].

Ovenfor viste formel er fremkommet ved et ønske om at få spredt EQR-værdierne mest muligt ud i hele intervallet fra 0 til 1, og som det fremgår af formlen, så er den beregnede EQR værdi en funktion af eksperternes a priori fastlæggelse af tilstanden.

Definition af DVPI

Endelig fastsættes DVPI ud fra DVPI_EQR ved anvendelse af følgende oversættelsestabel:

0 ≤ DVPI_EQR < 0,20: DVPI=1 (dårlig økologisk tilstand)
0,20 ≤ DVPI_EQR < 0,35: DVPI=2 (ringe økologisk tilstand)
0,35 ≤ DVPI_EQR < 0,50: DVPI=3 (moderat økologisk tilstand)
0,50 ≤ DVPI_EQR < 0,70: DVPI=4 (god økologisk tilstand)
0,70 ≤ DVPI_EQR: DVPI=5 (høj økologisk tilstand).

Grænserne mellem tilstandsklasserne er fremkommet i forbindelse med en interkalibreringsproces, hvor 800 europæiske vandløb blev klassificeret med den danske klassifikationsmodel, og denne klassifikation er herefter sammenlignet med et gennemsnit af en række europæiske metoder (Søndergaard et al. 2013). Grænserne mellem de forskellige tilstandsklasser blev harmoniseret, således at grænsefastsættelsen vha. DVPI svarer til den europæiske.

Eksempel

Som en illustration af hvordan man beregner DVPI og DVPI_EQR for en ny planteliste, så antag at vi har en planteliste med procentvis dækning af et antal plantearter registreret efter standardmetoden. Antag endvidere at plantelisten er generaliseret ved anvendelse af en automatiseret procedure, der sikrer kompatibilitet i listerne. Fx. konverteres arter af vandranukel/vandstjerne til de respektive slægter, ligeledes konverteres nogle underartsbestemmelser til artsbestemmelser fx Juncus bulbosus ssp. bulbosus til Juncus bulbosus, nogle krydsninger fx Potamogeton lucens x perfoliatus til Pamogeton lucens og nogle varianter fx Zannichellia palustris var. pedunculata til Zannichellia palustris, fordi disse enten ikke indgår eller kun er fåtalligt i det oprindelige datasæt, og således ikke var del af den udviklede model.

Plantelisten er som følger:

Art	Procentvis dækningsgrad
Bidens cernua	0.00712250712250712
Carex rostrata	0.04985754985754986
Catabrosa aquatica	0.04273504273504274
Elodea canadensis	0.78347578347578339
Equisetum fluviatile	0.17806267806267806
Glyceria maxima	5.87606837606837300
Iris pseudacorus	1.26780626780626780
Mentha aquatica	0.14245014245014243
Persicaria maculosa ssp. maculos	0.07834757834757833
Phalaris arundinacea	1.71652421652421650
Phragmites australis	0.37749287749287747
Poa trivialis	4.52991452991452890
Potamogeton perfoliatus	0.52706552706552701
Ranunculus acris	0.26353276353276350
Ranunculus repens	1.16809116809116830
Rumex obtusifolius	0.04273504273504274
Sparganium emersum	5.70512820512820260
Urtica dioica	0.01424501424501425
Veronica catenata	1.88746438746438820

De resterende 175 arter, der indgår i DVPI er ikke registreret på denne vandløbsstrækning.

Lad os definere følgende størrelser:

$$A_j, j = 1,2,\cdots,p$$

som dækningsgrader for de p arter. Og X er en 3xp matriks med scores for de p arter langs de 3 DCA ordinationsakser. Herefter beregnes følgende 3 størrelser:

$$x_m = \frac{\sum_{j=1}^{p} X_{mj} A_j}{\sum_{j=1}^{p} A_j}, m = 1,2,3.$$

Disse 3 størrelser læses ind i den prædiktive S-PLUS funktion, som hører til den superviserede klassifikationsmodel udviklet med QDA. Med ovennævnte eksempel får man

$$x_1 = 212,27$$
$$x_2 = 228,86$$
$$x_3 = 253,12$$

Model output er 5 sandsynligheder, som angiver sandsynligheden for at artslisten tilhører hver af de 5 økologiske klasser. Modellen giver følgende 5 sandsynligheder for eksemplet: 0,01;0,19;0,72;0,06;0,02.

Nu er det muligt at beregne DVPI_EQR for eksemplet ved at anvende formlen

$$DVPI_{EQR} = \frac{5 \cdot P_1 + 15 \cdot P_2 + 25 \cdot P_3 + 45 \cdot P_4 + 55 \cdot P_5}{50} = 0,494$$

Til sidst får vi at DVPI=3 (moderat økologisk tilstand) ved anvendelse af oversættelsestabellen givet på side 11.

Så strækningen i dette eksempel har den økologiske tilstand moderat, men meget tæt på god økologisk tilstand.

Afslutning

I denne rapport er gennemgået og dokumenteret alle relevante matematiske og statistiske beregninger ved formler, som var nødvendige ved udviklingen af DVPI og på en sådan måde, at de kan anvendes til beregning af DVPI for en ny planteliste. På den her beskrevne baggrund bør det derfor være muligt at gennemføre en beregning af DVPI dog baseret på en vis matematisk indsigt hos vedkommende, der skal gennemføre beregningen samt med adgang til software med matematisk programmerings facilitet. Desuden skal vedkommende have adgang til data fra de 1244 stationer samt vide, hvilke stationer som udgør træningsdatasættet, således at vedkommende kan estimere den prædiktive QDA klassifikationsmodel.

Som nævnt flere steder var det nødvendigt at anvende statistiske software pakker til udviklingen af DVPI og ligeledes nødvendigt at anvende software til beregning af DVPI for en planteliste på en ny vandløbsstrækning. Både PC-ORD og S-PLUS blev anvendt til udviklingen af DVPI og funktioner i S-PLUS er programmeret til beregning af DVPI for nye plantelister. Disse to nævnte software pakker er desværre ikke freeware, men pakke R, som er gratis, kunne også have været anvendt til udviklingen af DVPI, og de S-PLUS funktioner der beregner nye DVPI, kan direkte overføres til R-pakken.

Det anbefales at der udvikles en brugerflade til indekset for at sikre at DVPI beregningen ud fra en planteliste udføres korrekt.

Referencer

Baattrup-Pedersen, A. & Larsen, S.E. (2013) Udvikling af planteindeks i danske vandløb. Vurdering af økologisk tilstand (Fase I). Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 32 s. – Videnskabelig rapport fra DCE - Nationalt Center for Miljø og Energi nr. 60.
http://www.dmu.dk/Pub/SR60.pdf

Hill, M. O. (1973) Reciprocal averaging: an Eigenvektor method of ordination. J. Ecol. 61: 237-249.

Hil, M. O. (1979) DECORANA – A FORTRAN program for detrended correspondence analysis and reciprocal averaging. Ecology and Systematics, Cornell University, Ithaca, New York 14850, 52 pp.

Hill, M. O. og Gauch, H. G. (1980) Detrended correspondence analysis: An improved ordination technique. Vegetatio 42: 47-58.

McCune, B. og M. J. Mefford. 2011. PC-ORD. Multivariate Analysis of Ecological Data. Version 6. MjM Software, Gleneden Beach, Oregon.

Ripley, B. D. (1996) Pattern Recognition and Neural Networks. Cambridge. Cambridge University Press.

Søndergaard, M., Lauridsen, T.L., Kristensen, E.A, Baattrup-Pedersen, A., Wiberg-Larsen, P., Bjerring, R. & Friberg, N. 2013. Biologiske indikatorer til vurdering af økologisk kvalitet i danske søer og vandløb Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 76 s.. - Videnskabelig rapport fra DCE - Nationalt Center for Miljø og Energi nr. 59.

Venables, W. N. og Ripley, B. D. (1999) Modern Applied Statistics with S-PLUS. 3rd Edition. New York. Springer-Verlag.