2a:1

Organisering av basdata

FÖRFATTARE

Marcus Sundbom, ITM, Institutionen för tillämpad miljövetenskap, Stockholms universitet

2a:1

Organisering av basdata

FÖRFATTARE

Marcus Sundbom, ITM, Institutionen för tillämpad miljövetenskap, Stockholms universitet

INNEHÅLL

Ange sidorna 118–124 om du vill skriva ut detta kapitel.

BAKGRUND

IKEU har genererat en mycket stor mängd mätdata sedan starten 1989. IKEUs data lagras på de laboratorier som analyserar prover men data finns, med viss fördröjning, även tillgängliga hos nationella datavärdar vid SLU (kemi, plankton och bottenfauna), Fiskeriverket (fisk) och IVL (metaller i fisk). Vissa typer av data, t.ex. historiska data samt områdesbeskrivningar finns enbart på IKEUs hemsida. På hemsidan och hos datavärdarna finns alltså det mesta av IKEUs data tillgängliga för allmänheten. Andra kringdata såsom flödesmodelleringar och temperaturloggarnas avläsningar finns på IKEUs årliga CD-skiva som ska innehålla alla data framtagna inom projektet. CDn distribueras till naturvårdsverket och till alla IKEUs medarbetare och är den primära datakällan för internt arbete.

De flesta av IKEU-utvärderingens underlagsrapporter bygger på analyser av insamlade mätdata och ambitionen var att försöka beakta hela det befintliga dataunderlaget för att kunna bistå utvärderingen med ett så fullständigt underlag som möjligt. Det finns emellertid problem med hur vattenkemiska data på CDn är organiserade. Dessa är i huvudsak upplagda på samma sätt som hos den nationella datavärden där data framtagna av IMA och ITM ligger separat och det finns inga rutiner för att koppla ihop dessa data. Detta har i sin tur medfört att allmän vattenkemi och spårmetaller inklusive aluminium ligger i olika filer på CDn. Även data för de olika delprogrammen (referenssjöar, IKEU-sjöar och kalkavslutssjöar) ligger i olika filer på CDn pga hur de är inlagda hos datavärden. Ett särskilt problem utgör historiska data, dvs. data från tiden före IKEU som samlats in från en mängd olika källor inom ramen för ett tidigare specialprojekt (Persson & Wilander, 2002). Äldre vattenkemiska data ligger i enskilda filer för varje sjö och de har ofta sinsemellan olika variabelbeteckningar och enheter, vilka inte heller alltid överensstämmer med dem som används i datavärdens databas. Det säger sig självt att det krävs mycket tid för att sammanfoga alla dessa datafiler till en sammanhängande och kvalitetsgranskad databasfil. Om var och en skulle ge sig på detta vore det slöseri med tid och det finns uppenbar risk att det uppstår fel som är svåra att spåra. Detta är den huvudsakliga anledningen till att detta delprojekt bedömdes vara nödvändigt.

Delprojektet ska alltså betraktas som en tidsbesparande och kvalitetssäkrande serviceinrättning åt övriga projekt inom syntesprojektet 2a, *kemiska och biologiska effekter av kalkning i sjöar*. Syftet är förbättra förutsättningarna för integrerade studier genom att garantera att alla inblandade på ett tidigt stadium får tillgång till källdata, att alla använder samma källdata och att dessa är granskade och rensade från eventuella felaktigheter. Projektet begränsar sig till vattenkemiska data från sjöar. Biologiska data ingår inte utan vi beslöt att dessa hanteras bättre inom respektive delprojekt. I samband med den vattenkemiska datasammanställningen gjordes även en stationsdatabas och en kalkningsdatabas. Den förra sammanställdes av Gunnar Persson och inkluderar en lång rad metadata och områdesdata för varje IKEU-objekt. Kalkningsfilen sammanställdes av Björn Bergquist och innehåller information om alla kända kalkningar i och uppströms de kalkade IKEU-objekten. Kalkningsdata samlades in i samband med ett tidigare specialprojekt (S4-07 Bergquist, 2008).

METOD OCH RESULTAT

En datafil med samtliga vattenkemiska data t.o.m. 2006 levererades den 12 mars 2008 till alla medarbetare inom delprojekt 2a. Ungefär samtidigt distribuerades stationsdatafilen och kalkningsdatabasen. Alla tre levererades som Excel-filer och vikt lades vid att stationsbeteckningar skulle vara enhetliga för att dessa enkelt skulle kunna sammanföras vid behov. Vattenkemifilen innehöll sammanlagt 338177 mätvärden varav 49131 var historiska data. Data fördelar sig på drygt 15000 rader som var och en motsvarar ett unikt mättillfälle och provtagningsdjup. Nedan beskrivs kortfattat vad databasen omfattar, de problem som upptäcktes under arbetet och vilka övervägningar som gjorts för komma runt dessa. Excel-filen finns med på IKEUs senaste CD från hösten 2008.

Beskrivning av datafilen

Excel-arbetsboken innehåller vattenkemidata t.o.m. 2006 från IKEUs olika sjöprogram: IKEU-sjöar, referenssjöar, kalkavslut, extensivsjöar och överkalkningssjöar. De flesta data kommer från IKEU-CDn 2006, inklusive historiska data. För äldre data från Åva-sjöar har även en datafil som sammanställts på ITM används. Denna finns inte på CDn men de flesta av dessa data finns i historiefilerna. Även sådana sjöar som under perioden 1989–2006 utgått ur IKEU-programmet finns med, förutom Nävsjön som bara provtagits vid ett enda tillfälle 1989. Ett omfattande arbete har lagts ned på att sammanfoga de olika filerna. Avvikelser i koordinater, datum, djup och enheter mellan olika filer har korrigerats där det varit möjligt. Relevanta indexkolumner har lagts till för att lätt kunna dela upp sjöar olika grupper. I Tabell 1 nedan finns kortfattade förklaringar till alla variabler/kolumner.

Dubbletter

Jag har eliminerat dubbletter och tripletter, dvs. flera noteringar med samma datum och djup, om de innehöll identiska data. Ibland innehöll dubbletterna deluppsättningar av variabler varav vissa var överlappande med identiska mätvärden och då har jag sammanfogat dem till en och samma observation innan dubbletten togs bort. Det återstår några dubbletter och då handlar det oftast om olika stationer i samma sjö (äldre data). Alla observationer där det finns noterat "inlopp", "inflöde" eller liknande har jag tagit bort eftersom dessa inte representerar sjöns kemi. Det kan fortfarande finnas enstaka dubbletter kvar som jag har missat pga. att djup kan ha noterats olika i de olika ursprungsfilerna (t.ex. 0,1 m och 0,5 m).

Provtagningsdatum

I förvånande många fall har datum för en och samma provtagning varit olika i ursprungsfilerna för allmänkemi (IMA) respektive metaller/aluminium (ITM). Det datum som noterats av IMA har använts och ITMs datumnotering har alltså ignorerats utan att ha kollat upp vilket datum som faktiskt är rätt (vilket i de flesta fall är omöjligt). I praktiken saknar detta dock betydelse eftersom avvikelsen oftast bara uppgick till en eller två dagar. Provtagningsdatum anges dels i separata kolumner för år, månad och dag, dels i en datumformaterad kolumn. De separata kolumnerna är praktiska vid olika aggregeringar men är också viktiga att ha med eftersom referensdatum skiljer sig mellan Windows och Mac och vid klipp och klistrande mellan olika Excel-dokument kan då datumet ändras (med 1462 dagar) utan att det upptäcks.

Ibland saknas "Dag" och då har jag alltid antagit den 15:e eftersom provtagningsanvisningarna säger att prov ska tas ungefär vid månadsmitt. Vid några få tillfällen saknas både månad och dag, då har 1 juli använts. Dessa godtyckligt valda datum finns endast i den datumformaterade kolumnen (rödmarkerade); i kolumnerna för månad och dag har lämnats tomt för dessa ofullständiga datumangivelser.

Provtagningsdjup

Angivet provtagningsdjup kan variera mellan olika datakällor och perioder, framförallt vad gäller ytvattenprov. Vid sammanfogande av allmänkemi och metaller har jag konstruerat ett ytvattenprov genom att sätta alla djup ≤ 1 m till samma värde. Det finns även än djupkolumn med originaldjupet och då har jag använt det djup som noterats i IMAs filer även om det skulle skilja sig från ITMs notering. I många fall saknas djupnotering helt (äldre data). Jag har då utgått från att det handlar om ytprover, vilket också är det mest rimliga.

Värden under detektionsgränsen

Uppmätta halter under detektionsgränsen är i regel noterade som mindre-än-värden (så kallade nondetects) i ursprungsfilerna. Mindre-än-värden har genomgående ersatts med halva detektionsgränsen. Denna ansats kan kritiseras utifrån rent statistiska grunder, men valdes för att den är enkel och har använts tidigare inom IKEU. Syftet med substitutionen var att få enbart numeriska data i de olika kolumnerna, vilket är en förutsättning för standardberäkningar eller för import av data till olika statistikprogram. För de variabler som innehöll nondetects finns en extra (gråskuggad) kolumn med de ursprungliga detektionsgränserna angiven parallellt med de nya numeriska värdena. Dessa kan användas för sådana analyser som tar hänsyn till mindre-än-värden (t.ex. Kaplan-Meyer) och halverade värdena kan då helt enkelt ersättas med motsvarande mindre-än-värden.

Alkalinitet anges som alkalinitet/aciditet där negativa värden motsvarar uppmätt aciditet. Aciditet har dock inte mätts under alla perioder och i ursprungsfilerna har då alkalinitetsbestämningar under detektionsgränsen antingen angetts som ett mindre-än-värde eller som noll. Jag har ersatt mindre-än-värden med halva detektionsgränsen medan noterade nollvärden lämnats oförändrade. Det kan vara nödvändigt att beakta detta vid trendanalys eller andra temporala jämförelser av alkalinitet.

Idiotvärden och uteliggare

Jag har försökt att om möjligt korrigera helt orimliga värden (typ temp = 191°C, pH = 515). Om detta inte gick att göra med säkerhet har orimliga data helt enkelt strukits. Systematiska fel orsakade av uppenbart felaktigt angivna enheter (t.ex. µg istället för mg) har rättats till. Det fanns också avvikande data som var svårare att avgöra orsak till. Det kan handla om t.ex. inmatningsfel, kontamination eller analysfel, men det kan också representera ett "normalt" och korrekt extremvärde. Jag har inte haft kunskaper eller resurser att gå till botten med vart och ett av dessa fall som alltså behållits som de är. Vissa typer av jämförelser är mer känsliga än andra för i vilken grad uteliggare påverkar slutsatserna och jag har överlåtit till medarbetarna inom de olika delprojekten att avgöra hur dessa svårbedömda uteliggare ska hanteras. Se upp med ammonium, fosfat och vissa spårmetaller där risken för kontamination är stor. Historiska data har som förväntat generellt fler misstänkta uteliggare än IKEU-data. Rena felstansningar eller värden angivna med felaktig enhet är inte ovanligt för äldre data. Mina ändringar har markerats med rött i filen.

Metaller

De metallanalyser som främst ska användas är de som analyserats av ITM. Under perioder har metaller analyserats parallellt av IMA. Dessa värden finns med sist i Excel-tabellen med tillägget "IMA" i kolumnbeteckningen.

SLUTSATS

Projektet lyckades uppnå sitt syfte att i ett tidigt skede av utredningsarbetet leverera en komplett vattenkemidatabas för IKEUs sjöar. Under arbetet blottlades vissa brister med hur data är organiserade idag. Det optimala vore om data redan från början, d.v.s. hos datavärden var inlagda så att utdrag av kemidata från olika delprogram och olika källor kunde fås i en och samma fil. Detta är naturligtvis ingen ny insikt utan har diskuterats internt vid flera tillfällen och det saknas inte förslag till hur situationen kan förbättras. De flesta fall där datum inte överensstämde mellan IMAs och ITMs filer har i samband med detta delprojekt ändrats även hos datavärden, vilket bör förenkla arbetet med att förbättra databasens struktur för IKEUs data hos datavärden. Det vore onekligen synd att behöva upprepa detta tidskrävande arbetsmoment gång på gång i framtiden. Även för externa användare vore det bra om IKEU-data gick att komma åt på ett enklare och mer konsekvent sätt.

REFERENSER

Bergquist, B. (2008) *Sammanställning av kalkningsuppgifter för IKEU-sjöar och vattendrag*. Fiskeriverket.

Persson, G. & Wilander, A. (2002) *Allmän vattenkemi före och efter kalkning inom Integrerad KalkningsEffektUppföljning (General water chemistry before and after liming in lakes included in the Integrader Studies of the Effects of Liming in Acidified Waters) in Swedish*. SLU, SLU, Uppsala 2002:8.

TABELL 1. Förklaringar till samtliga kolumnbeteckningar i Excel-databasen över vattenkemi i IKEUs sjöar.

Kolumn	Förklaring
RadID	Löpnummer för att kunna återställa ursprunglig sortering
RadID2	Radnumrering för eget bruk för att kunna referera till ursprungsfiler ifall något måste kontrolleras
Källfil IKEU CD	Den fil på IKEU-CDn som data är hämtat ifrån. "IKEUWEB-Sjöhist" är ett samlingsnamn för Excel-filerna med gamla data som ligger på hemsidan (och CDn).
Källa gamla data	Gunnar Perssons noteringar för de källmaterial som använts för att sammanställa historiefilerna
Delprogram_i_IKEU(nr)	Delprogram enligt stationsdatafilen.
Kalkad	Anger kalkningsstatus: Okalkad/före kalkning = **0**; Efter första kalkning = **1**; Efter sista kalkning, dvs. kalkavslut = **2**
Labkod IMA	IMAs interna stationsbeteckning
StnID ITM	ITMs stations-ID
Namn	Sjönamn
LänNr	Länsnummer
X	Nordkoordinat SMHI
Y	Ostkoordinat SMHI
Pos ID	Beteckning för provtagningsplats. Finns endast för gamla data, för nyare data gäller sjömitt.
År	
Månad	
Dag	
Datum	Datum baserat på År, Mån, Dag. Då uppgifter saknas har ett datum (re)konstruerats (se kommentarer i texten).
Nivå	Provtagningsdjup (m)
Nivå2	Provtagningsdjup (m), ytprov (oftast <= 1 m) ersatta med "Y".
Siktdjup m	
Temp °C	
pH	
Kond25 mS/m	Konduktivitet vid 25°C. Det fanns ett mindre antal äldre data mätta vid andra temp och med andra enheter. Dessa har jag valt att inte ta med.
Alk/Acid	Alkalinitet, Aciditet anges som negativ alkalinitet. Observera att aciditet inte alltid har mätts
ND Alk/Acid mekv/l	Angiven detektionsgräns för Alkalinitet
Syrgas mg/l	
Ca mekv/l	Kalcium
Mg mekv/l	Magnesium
ND Mg mekv/l	Angiven detektionsgräns för Mg
CaMg mekv/l	Kalcium + magnesium när summan angetts i ursprungsfilerna. Ca+Mg är inte uträknat för alla tillfällen som Ca & Mg bestämts separat.
ND CaMg mekv/l	Angiven detektionsgräns för CaMg
Na mekv/l	Natrium
ND Na mekv/l	Angiven detektionsgräns för Na
K mekv/l	Kalium
ND K mekv/l	Angiven detektionsgräns för K
SO4 mekv/l	Sulfat (Både SO4_MC och SO4_IC)
ND SO4 mekv/l	Angiven detektionsgräns för SO4
Cl mekv/l	Klorid

Kolumn	Förklaring
ND Cl mekv/l	Angiven detektionsgräns för Cl
Fluorid mg/l	
NH4-N µg/l	Ammonium
NO2-N µg/l	Nitrit
NO2+NO3-N µg/l	Nitrit + nitrat
ND NO2+NO3-N µg/l	Angiven detektionsgräns för NO2+NO3-N
NO3-N µg/l	Nitrat
Kjeld-N µg/l	Kjeldahl-kväve
Org-N µg/l	Organiskt kväve
Tot-N µg/l	Totalkväve
ND Tot-N µg/l	Angiven detektionsgräns för Tot-N
Tot-N sum µg/l	Totalkväve som summa (NO23+Kjeldahl?). Ej uträknat för alla tillfällen där det är möjligt.
PO4-P µg/l	Fosfat
ND PO4-P µg/l	Angiven detektionsgräns för PO4-P
Övr-P µg/l	Övrig fosfor
Tot-P µg/l	Totalfosfor
ND Tot-P µg/l	Angiven detektionsgräns för Tot-P
Si mg/l	Kisel
ND Si mg/l	Angiven detektionsgräns för Si
Färg mgPt/l	Vattenfärg komparator
ND Färg mgPt/l	Angiven detektionsgräns för Färg
Abs OF 420nm/5cm	Absorbans vid 420 nm och 5 cm kyvett, ofiltrerat
Abs F 420nm/5cm	Absorbans vid 420 nm och 5 cm kyvett, filtrerat
Abs Diff 420nm/5cm	Skillnaden mellan ofilterat och filtrerat
TOC mg/l	Totalt organiskt kol
KMnO4 mg/l	Permanganatförbrukning (Ibland omräknat från COD)
ND KMnO4 mg/l	Angiven detektionsgräns för KMnO4
Kfyll mg/m3	Klorofyll
Turbiditet FNU	Turbiditet (FNU=FTU=JTU)
MN-NK µg/l	Mangan
ND MN-NK µg/l	Angiven detektionsgräns för Mn
FE-NK54 µg/l	Järn
ND FE-NK54 µg/l	Angiven detektionsgräns för Fe
CO-NKX µg/l	Kobolt
ND CO-NKX µg/l	Angiven detektionsgräns för Co
NI-NK60X µg/l	Nickel
CU-NK65X µg/l	Koppar
ND CU-NK65X µg/l	Angiven detektionsgräns för Cu
ZN-NK66 µg/l	Zink
ND ZN-NK66 µg/l	Angiven detektionsgräns för Zn
MO-NK98 µg/l	Molybden
ND MO-NK98 µg/l	Angiven detektionsgräns för Mo
CD-NK114 µg/l	Kadmium
ND CD-NK114 µg/l	Angiven detektionsgräns för Cd
PB-NK678 µg/l	Bly
ND PB-NK678 µg/l	Angiven detektionsgräns för Pb

Kolumn	Förklaring
AL-NA µg/l	Aluminium syralösligt. Till denna kategori har även förts en del äldre ospecificerade Al-data.
ALM-NAD µg/l	Monomert aluminium
ND ALM-NAD µg/l	Angiven detektionsgräns för ALM-NAD
ALO-NAJ µg/l	Stabilt monomert aluminium
ND ALO-NAJ µg/l	Angiven detektionsgräns för ALO-NAJ
ALI-NAJ µg/l	Labilt monomert aluminium, definierad som skillnaden mellan ALM-NAD och ALO-NAJ
ND ALI-NAJ µg/l	Angiven detektionsgräns för ALI-NAD
Anm	Eventuella anmärkningar i ursprungsfilerna
Fe µg/l IMA	Järn analyserat av IMA
Mn µg/l IMA	Mangan IMA
Cu µg/l IMA	Koppar IMA
Zn µg/l IMA	Zink IMA
Al s µg/l IMA	Aluminium syralösligt IMA
Al ICPAES µg/l IMA	Al ICP IMA (>≈AL-NA)
Al ICPKJB µg/l IMA	Al ICP IMA efter katjonbyte (>≈ALI-NAJ)
Cd µg/l IMA	Kadmium IMA
Pb µg/l IMA	Bly IMA
Cr µg/l IMA	Krom IMA
ND Cr µg/l IMA	Angiven detektionsgräns för Cr
Ni µg/l IMA	Nickel IMA
Co µg/l IMA	Kobolt IMA
As µg/l IMA	Arsenik IMA
V µg/l IMA	Vanadin IMA