BILAGA D: PRESENTERA OCH ANALYSERA DATA

Analys av vattenkvalitet och vattentekniska processer kräver mätningar. Detta resulterar i data som måste analyseras och kommuniceras med omvärlden på ett begripligt sätt. Här beskriver jag några olika typer av analyser och ger exempel på bra och dåliga sätt att presentera data.

Exempel 1: Presentation av data i en tabell

Låt oss anta att vi har mätt alkalinitet i tre olika sjöar och vi vill presenta resultaten i en tabell. Figur D-1 visar ett dåligt och ett bra exempel på en sådan tabell. Den vänstra tabellen är dålig av två anledningar: antalet värdesiffror och avsaknaden av enheter.

Antalet värdesiffror indikerar precisionen av vår mätning. I den vänstra tabellen har alkaniteten i Anksjön och Fisksjön orimligt många värdesiffror medan Grodsjön bara har en. En generell regel för värden beräknade med multiplikation eller division är att inte ha fler värdesiffror i svaret än det ingående mätvärde med minst antal värdesiffror. Alkalinitet beräknas med följande ekvation (se även kapitel 4):

$$ALK_{TOT} = \frac{Volym_{syra} \cdot Konc_{syra}}{Volym_{vatten}}$$

Om vi antar att volymen syra är 3,52 mL (3 värdesiffror), koncentrationen syra är 0,1052 M (fyra värdesiffror) och volymen vatten är 50,0 mL (3 värdesiffror), så ska den beräknade alkaliniten anges med tre värdesiffror eftersom detta är det minsta antalet värdesiffror hos de ingående mätvärdena. Om ett värde istället beräknas genom addition eller subtraktion av ingående mätvärden så ska antalet decimaler i svaret vara lika många som det ingående mätvärde med minst antal decimaler.

När data presenteras så måste enheter anges. I en tabell är det ofta lämpligt att ange enheter tillsammans med namnet på parametern i översta raden. För att undvika ett stort antal inledande eller avslutande nollor är det fördelaktigt att använda tiopotensform som för alkaliniteten i Grodsjön i högra tabellen.

Dålig

Sjö	Alkalinitet
Anksjön	1,235644
Grodsjön	0,00002
Fisksjön	0,896123

Bra!

Sjö	Alkalinitet (mekv/L)
Anksjön	1,24
Grodsjön	$2{,}01 \cdot 10^{-5}$
Fisksjön	0,896

Figur D-1. Exempel på en bra och en dålig tabell.

Exempel 2: Jämförelse av två dataset

Låt oss anta att vi vill jämföra alkaliniten i Anksjön och Fisksjön och därför har gjort 20 mätningar i varje sjö. Figur D-2 visar tre olika sätt att presentera datan. Det vänstra diagrammet (X) är dåligt av följande anledningar:

- Det finns ingen information om enheter och parameter på y-axeln.
- Endast medelvärden visas och det finns ingen information om variationerna mellan individuella mätningar eller hur många mätningar som gjorts.

Det mittersta och det högra diagrammet (Y och Z) är bra. Vid y-axeln finns namnet på parametern och enheten. I figurtexterna finns information om vad som visas och antalet mätpunkter som

diagrammen baseras på. Variationen mellan individuella mätningar är tydlig genom felstaplar (error bars) som visar standardavvikelsen i Y och låddiagrammen i Z.

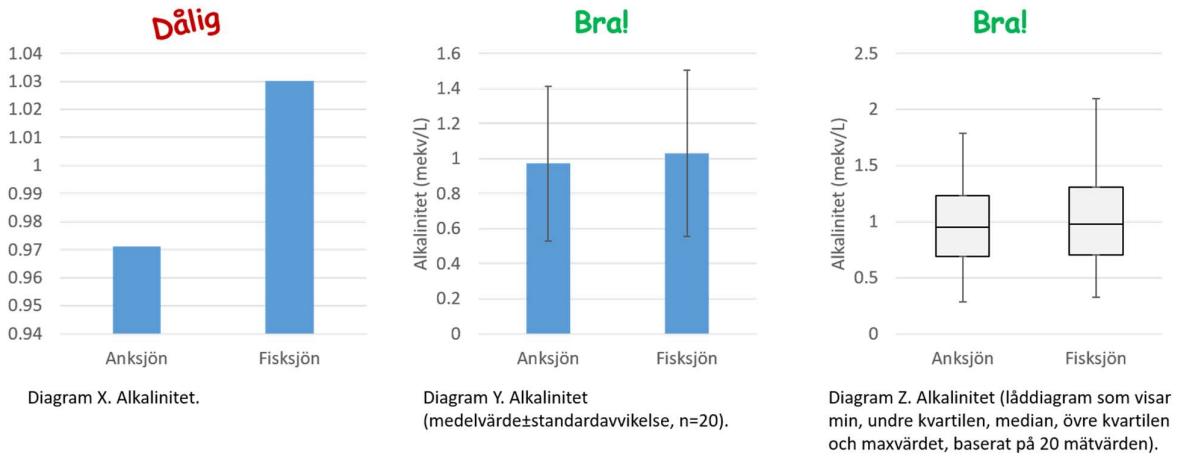


Diagram X. Alkalinitet.

Diagram Y. Alkalinitet (medelvärde±standardavvikelse, n=20).

Diagram Z. Alkalinitet (låddiagram som visar min, undre kvartilen, median, övre kvartilen och maxvärdet, baserat på 20 mätvärden).

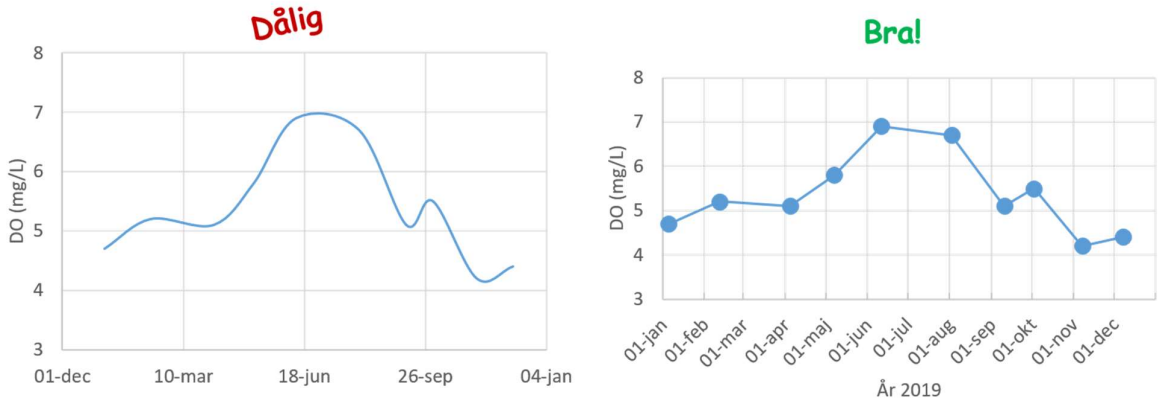
Figur D-2. Exempel på ett dåligt och två bra diagram som jämför två dataset.

Exempel 3: Trender över tid

I detta exemplet antar vi att vi har mätt koncentrationen löst syre vid 10 tillfällen i en sjö under ett år. Vi vill illustrera i ett diagram hur syrekoncentrationen förändras under året. Figur D-3 visar ett bra och ett dåligt diagram. Det vänstra diagrammet är dåligt av följande anledningar:

- Det finns ingen information om när mätningar utfördes.
- Linjen är utjämnad (smoothed) mellan mätpunkterna trots att vi inte har någon information om hur syrekoncentrationen förändras där.
- Skalan på x-axeln är svår att följa.

Det högra diagrammet är bra. Varje mätvärde är tydligt markerat med en punkt. Linjen som binder samman punkterna är rak eftersom vi inte har någon information om hur syrekoncentrationen förändras där (det hade också varit okej att helt utelämna den linjen och bara visa punkterna). Diagrammets x-axel är tydligt indelad med en markering den 1:a varje månad och det finns information om vilket år som avses.



Figur D-3. Exempel på ett bra och ett dåligt diagram som visar en serie av mätningar över tid. DO står för koncentrationen löst syre.

Exempel 4: Testa en hypotes

Låt oss gå tillbaka till exempel 2 där vi jämförde alkaliniteten i Anksjön och Fisksjön. Är det någon skillnad i alkalinitet mellan de två sjöarna? Om vi hade använt oss av diagram X så hade vi kanske dragit slutsatsen att alkaliniteten i Fisksjön är högst eftersom den är 1,03 mekv/L medan alkaliniteten i Anksjön bara är 0,97 mekv/L. Det är dock en felaktig slutsats. I diagram Y och Z ser vi att även om medelvärdet för alkaliniteten är högre i Fisksjön så är det en stor variation mellan individuella mätningar och alkaliniteten verkar vara ungefär lika hög i båda sjöarna.

För att undersöka om det finns en statistiskt signifikant skillnad kan man använda ett t-test. Detta är en statistisk metod för att testa hypoteser. Låt oss anta att vi gör ett oändligt antal mätningar i de två sjöarna och att medelvärdet av detta oändliga antal är den sanna alkaliniteten i varje sjö. Nollhypotesen är att den sanna alkaliniteten i Anksjön är exakt lika stor som den i Fisksjön. I verkligheten har vi dock bara tagit 20 mätvärden i varje sjö. Med t-testet kan vi beräkna sannolikheten att få det resultat som vi fått med våra 20 mätvärden om nollhypotesen skulle vara sann. Denna sannolikhet kallas p-värde. Om p-värdet är under ett kritiskt värde förkastas nollhypotesen. Ofta brukar 0,05 användas som kritiskt värde. Om p-värdet är under 0,05 indikerar det att sannolikheten att observera vår mätdata givet att nollhypotesen är sann, är under 5%. Det är alltså en ganska låg risk att vi felaktigt förkastar nollhypotesen.

t-testet kan användas om mätdatan har normalfördelning samt att variansen och antalet mätpunkter är samma i de två dataseten. Om detta inte är uppfyllt kan man istället använda Welch's t-test. Man kan även använda t-test för att undersöka om medelvärdet för ett dataset är lika med ett givet värde eller om lutningen på en regressionslinje är signifikant. Om Vatten beskriver inte i detalj hur dessa olika statistiska test fungerar. Den informationen går dock hitta på många ställen, t.ex. Wikipedia och hemsidan Real Statistics (http://www.real-statistics.com). Excel och många andra mjukvaror innehåller funktioner som gör det enkelt att göra en mängd olika statistiska test.

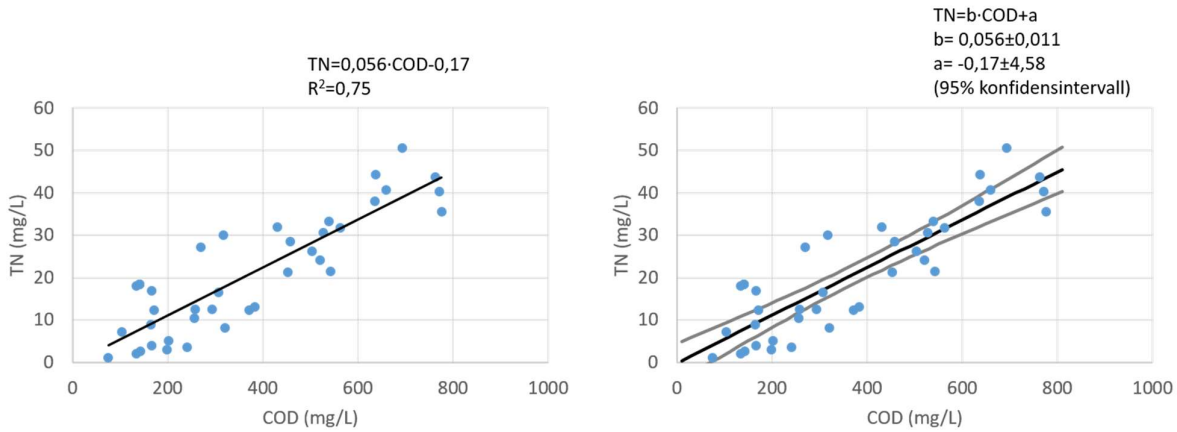
I fallet med alkalinitet i Anksjön och Fisksjön använde jag Excels funktion t.test och beräknade ett p-värde på 0,69. Det betyder att vi inte kan förkasta nollhypotesen och att medelvärdet på alkaliniteten i de två sjöarna troligtvis är ungefär lika stor. Så här skulle man kunna rapporterna resultatet från mätningarna: "Alkaliniteten i Anksjön var 0,97±0,44 mekv/L och i Fisksjön var den 1,03±0,48 mekv/L (medel±standardavvikelse). Det var ingen statistiskt signifikant skillnad i alkalinitet mellan de två sjöarna (p=0,69, t-test, n=20)."

Exempel 5: Undersöka en korrelation

Låt oss anta att vi har 39 mätningar av koncentrationen av COD och totalkväve (TN) i avloppsvatten. Vi frågar oss om det finns en korrelation mellan dessa två parametrar. I Figur D-4 är mätvärdena plottade i ett diagram med COD-koncentration på x-axeln och TN-koncentration på y-axeln. Bara genom att inspektera diagrammet ser vi att det troligtvis finns en linjär korrelation mellan dessa två parametrar. Den svarta linjen i det vänstra diagrammet visar en linjär regressionsanalys av datan. Denna typen av analys är enkel att göra i Excel. Vi får ekvationen för linjen och determinationskoefficienten (R^2). R^2-värdet anger hur stor del av variationerna i y-variabeln (TN) som kan förklaras av variationer i x-variabeln (COD). I det här fall är R^2 lika med 0,75 vilket betyder att 75% av variationerna i TN kan förklaras av variationer i COD.

I det högra diagrammet får vi ytterligare lite mer information. Här har vi räknat ut konfidensintervall för lutningen på regressionslinjen (b) och värdet där linjen skär y-axeln (a). Konfidensintervallet är ett mått på osäkerheten hos parametrarna. Det 95%-konfidensintervall som anges i det högra

diagrammet innebär att det är 95% sannolikhet att det sanna värdet för parametern ligger inom det intervallet. Denna information är värdefull om vi vill testa en hypotes om att det finns en korrelation mellan två variabler. I det har fallet ville vi testa hypotesen att det finns en positiv korrelation mellan TN och COD. Eftersom värdet 0 inte finns inom konfidensintervallet för lutningen (b) så kan vi dra slutsatsen att vår hypotes sannolikt är korrekt. En bra beskrivning av hur man beräknar konfidensintervall (och en massa annan statistik) finns här: http://www.real-statistics.com.



Figur D-4. Korrelation mellan koncentrationen totalkväve (TN) och COD. De blå punkterna visar mätvärden, den svarta linjen är den linjära regressionslinjen, och de grå linjerna i det högra diagrammet visar 95% konfidensintervall för regressionsanalysen.