Håndtering af større datasæt i Excel

Ulrik Gerdes
Klinisk Biokemisk Laboratorium :: Århus Universitetshospital

Formålet med denne artikel er at pege på nogle faciliteter i Excel, som erfaringsmæssigt anvendes alt for lidt i det kliniske biokemiske miljø, selvom de er yderst velegnede til de typer af opgaver vi typisk løser med brug af regneark. Det gælder især opgaver der indebærer håndtering af større datasæt, fx med mere end 500 poster og 10 forskellige variable, eller meget store sæt, fx med mere end 50.000 poster.

Artiklen bør være særligt inspirerende, hvis du jævnligt sidder ved din computer og udfører påfaldende ensformige rutiner med at udvælge, kopiere og flytte tekster, data og formler rundt mellem forskellige regneark og faner, mens du konstant kæmper for at bevare overblikket. Eller hvis du hyppigt sidder og må lave tidsrøvende korrektioner af mange opstillinger rundt omkring i flere regneark, fordi du har måttet tilføje eller fjerne nogle data, eller fordi du har fundet fejl i nogle formler, som du har kopieret og anvendt mange steder. Med andre ord: Hvis du oplever, at det er dig der arbejder for computeren, og ikke omvendt — som det naturligvis bør være!

Mit hovedbudskab er at det heldigvis er let at få vendt om på tingene: Du skal bare koncentrere dig om at samle og strukturere dine data på en sådan måde, at du med enkle midler kan lade Excel overtage de fleste rutiner med udvælgelser, opstillinger, sammenregninger og meget mere.

Artiklen indeholder ikke konkrete instrukser, men de er meget nemme at finde andre steder (se det sidste afsnit).

Lidt om forskellige Excel-versioner

De fleste i miljøet anvender sandsynligvis Excel 2003, men jeg ved, at der stadig er en del der anvender Excel 2000 (eller ældre), som ikke har alle de funktioner jeg omtaler nedenfor. Omvendt er vi vistnok kun nogle få som p.t. er opdateret til Microsoft Office System 2007.

De velkendte programmer har bl.a. fået en helt ny brugergrænseflade og anvender et åbent filformat, det såkaldte Office Open XML (OOXML). De har også fået en række nye funktioner og faciliteter, og for Excel 2007 er én af de interessante, at man nu råder over 1 million rækker og 16.000 kolonner i et ark (mod kun 65.000 rækker og 256 kolonner i de tidligere versioner), dvs. at man nu kan håndtere *meget* store datasæt i Excel.

Noget helt grundlæggende om datahåndtering i Excel

Lad være med…

- Selv at sortere og fordele dine data (råmaterialet) fra et givet projekt i forskellige filer, faner og/eller tabeller.
- At kopiere data til forskellige filer, eller (værre) at kopiere dele af dine data til forskellige filer, faner og/eller tabeller.

Og gør i stedet sådan hér!

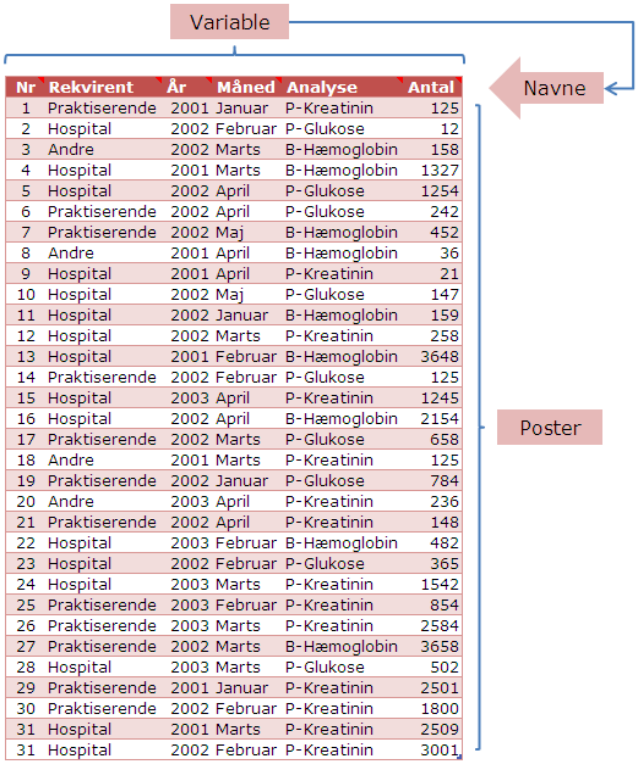
Anbring alle dine data i en samlet blok ét og kun ét sted, og tilføj variable som karakteriserer informationerne på alle leder og kanter.

For at Excel kan håndtere data mest effektivt, skal de nemlig findes i en sådan samlet blok (matrix), med én kolonne for hver variabel og én række for hver post, fraset den øverste række i blokken, som skal indeholde

navnene på variablene. Det er samme struktur som fx anvendes i databaser og statistikprogrammer, og ligheden er bestemt ikke et tilfælde.

Figur 1 viser et eksempel på en sådan blok, med konstruerede data for antal rekvisitioner af 3 analyser i 3 kategorier af rekvirenter i månederne januar til maj i 3 år.

Figur 1. Strukturen i en datamatrix



Nr	Rekvirent	År	Måned	Analyse	Antal
1	Praktiserende	2001	Januar	P-Kreatinin	125
2	Hospital	2002	Februar	P-Glukose	12
3	Andre	2002	Marts	B-Hæmoglobin	158
4	Hospital	2001	Marts	B-Hæmoglobin	1327
5	Hospital	2002	April	P-Glukose	1254
6	Praktiserende	2002	April	P-Glukose	242
7	Praktiserende	2002	Maj	B-Hæmoglobin	452
8	Andre	2001	April	B-Hæmoglobin	36
9	Hospital	2001	April	P-Kreatinin	21
10	Hospital	2002	Maj	P-Glukose	147
11	Hospital	2002	Januar	B-Hæmoglobin	159
12	Hospital	2002	Marts	P-Kreatinin	258
13	Hospital	2001	Februar	B-Hæmoglobin	3648
14	Praktiserende	2002	Februar	P-Glukose	125
15	Hospital	2003	April	P-Kreatinin	1245
16	Hospital	2002	April	B-Hæmoglobin	2154
17	Praktiserende	2002	Marts	P-Glukose	658
18	Andre	2001	Marts	P-Kreatinin	125
19	Praktiserende	2002	Januar	P-Glukose	784
20	Andre	2003	April	P-Kreatinin	236
21	Praktiserende	2002	April	P-Kreatinin	148
22	Hospital	2003	Februar	B-Hæmoglobin	482
23	Hospital	2002	Februar	P-Glukose	365
24	Hospital	2003	Marts	P-Kreatinin	1542
25	Praktiserende	2003	Februar	P-Kreatinin	854
26	Praktiserende	2003	Marts	P-Kreatinin	2584
27	Praktiserende	2002	Marts	B-Hæmoglobin	3658
28	Hospital	2003	Marts	P-Glukose	502
29	Praktiserende	2001	Januar	P-Kreatinin	2501
30	Praktiserende	2002	Februar	P-Kreatinin	1800
31	Hospital	2001	Marts	P-Kreatinin	2509
31	Hospital	2002	Februar	P-Kreatinin	3001

Når du markerer en celle i en sådan blok, kan Excel automatisk 'se helheden i data', hvorefter du umiddelbart kan begynde at sortere, filtrere og analysere data med pivottabeller og andre værktøjer.

Orden!

Jeg vil dog varmt anbefale, at du først bruger tid på at ordne forskellige ting. Med lidt træning tager det typisk kun 10-15 minutter, selv for større datasæt, og du sparer ofte dig selv (og ikke mindst andre) for en masse ærgrelser og besvær på længere sigt:

- Fortæl Excel at blokken af data er en tabel (i Excel 2003 hedder det en liste), og give tabellen et navn, fx 'Data_Tabel'. Det medfører bl.a. at afgrænsningen bliver dynamisk, så du fx kan indsætte eller slette rækker og kolonner uden at skulle revidere formler, funktioner og programkode der er knyttet til indholdet. Det medfører også at data kan importeres direkte af andre programmer ved hjælp af forespørgsler i Structured Query Language (SQL) og tilsvarende (se nedenfor).
- Formatér tabellen så indholdet fremtræder ordentligt og ensartet. Det er ikke primært et spørgsmål om æstetik, men om overskuelighed, herunder mulighederne for at opdage fejl og uhensigtsmæssigheder. Hvis det er data du skal dele med andre, vil de dog også sætte pris på at få tingene serveret på en måde, der ikke giver kortslutninger på nethinderne.

- Beskriv indholdet i variablene, og anført også andre relevante informationer, fx om oprindelsen af data. Det kan gøres meget enkelt ved at indsætte en kommentar i cellerne med variablenes navne. Selvom det kan tage lidt tid, så vil du (og især andre) på et senere tidspunkt have stor glæde af indsatsen med denne journalføring.
- Kontrollér dine data for strukturelle fejl, fx dubletter, stavefejl, malplacerede punktummer, kommaer eller mellemrum (som konverterer tal til tekst eller omvendt, eller som korrumperer datoangivelser). Fejl kan som bekendt være særdeles destruktive og meget drilagtige. Excel har en indbygget facilitet til fejlkontrol, som fx markerer en celle, hvis typen af indholdet afviger fra indholdet i de øvrige celler i samme kolonne. Du kan også bruge filtreringer, sorteringer, betinget formatering og diagrammer til at spore fejl, fordi de ses som 'besynderligheder' i diverse sammenhænge. Hvis data skal indtastes, så brug i øvrigt faciliteten datavalidering for at undgå fejl: Du kan styre hvad det er muligt at fylde i cellerne, herunder fx at knytte cellerne til rullelister med givne valgmuligheder.

Pivottabeller

Dette er et særdeles effektivt værktøj, som bliver brugt alt for lidt! Det ved Microsoft (og andre) godt, og i Excel 2007 er aktiveringen af pivottabeller gjort mere synlig i brugergrænsefladen.

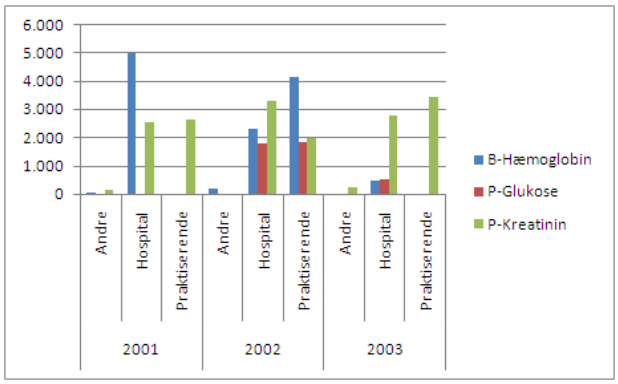
Pivot betyder »(en) tap, hvorom noget drejer sig«. Pivottabeller er et lynhurtigt og særdeles effektiv værktøj til at udvælge, opstille og analysere alle slags data på alle leder og kanter. Og som du kan opdatere med et enkelt museklik, hvis blokken eller indholdet i dine data er ændret.

Pivotdiagrammer er en valgmulighed, som knyttes til tabellerne, så du samtidig kan se resultaterne i grafisk form.

Figur 2 viser et eksempel på en pivottabel med et diagram, baseret på de

Figur 2. En pivottabel med et tilhørende diagram

Sum af Antal	Analyse			
Rækkeetiketter	B-Hæmoglobin	P-Glukose	P-Kreatinin	Hovedtotal
2001	5.011		5.281	10.292
Andre	36		125	161
Hospital	4.975		2.530	7.505
Praktiserende			2.626	2.626
2002	6.581	3.587	5.207	15.375
Andre	158			158
Hospital	2.313	1.778	3.259	7.350
Praktiserende	4.110	1.809	1.948	7.867
2003	482	502	6.461	7.445
Andre			236	236
Hospital	482	502	2.787	3.771
Praktiserende			3.438	3.438
Hovedtotal	**12.074**	**4.089**	**16.949**	**33.112**



viste data i Figur 1.

Der er beregnet en sum af antallene af de tre analyser (kolonner), fordelt på år og derefter rekvirentkategorier (rækker). Med lidt øvelse kan et sådan resultat fremtrylles på 15-20 sekunder, og hvis du fx vil bytte om på år og rekvirenter, eller vende tabellen 90 grader, så tager det 5-10 sekunder.

Det er let at lære at bruge pivottabeller

Hvis dine data er samlet i en blok (matrix, tabel) som beskrevet ovenfor, er det er ikke særlig svært at lære at bruge pivottabeller. Brugere af SPSS eller andre statistikprogrammer i samme vægtklasse vil have let ved det, da fidusen med at kunne pivotere tabeller fx også findes i tilknytning til den procedure der hedder 'krydstabulering'.

Det kræver naturligvis nogen øvelse og studier af diverse finurligheder, hvis man vil udnytte pivottabellers fulde potentiale. Der findes nogle udmærkede online kurser på Microsofts website (se nedenfor).

Jeg kunne nævne mange eksempler på hvad pivottabeller kan bruges til indenfor klinisk biokemi, men vil nøjes med at nævne nogle egenskaber jeg ofte bruger:

- Man kan gruppere (kategorisere) indholdet i variable der indeholder dato og/eller tidspunkt. Det betyder, at man på få sekunder kan genere pivottabeller og -diagrammer med antal, middelværdier m.m. af analyseresultater sammenregnet for år, måned, uge etc.
- Man kan styre sorteringen i en pivottabel, så der fx fås en tabel hvor de højeste antal, middelværdier m.m. findes øverst.
- Man kan have mange pivottabeller knyttet til de samme data, dvs. at man kan arbejde med en serie 'standardopsætninger' til et givet formål, fx vedrørende produktionsstatistik.
- Man kan pivotere databaser (tabeller) der har karakter af oversigter som overvejende indeholder tekster, fx projektoversigter og indholdsfortegnelser.

Analyse af eksterne data med pivottabeller

De data der skal anvendes i en pivottabel kan udmærket findes et andet sted end i det aktuelle regneark, og behøver end ikke at findes i en Excel-fil. Man kan fx oprette en pivottabel som læser indholdet i en Access-database eller andre SQL-baserede databaser. Du kan derfor pivotere indholdet i kæmpestore datasæt med flere hundrede millioner poster.

Det er en funktion der fx er meget velegnet til diverse produktions- og forbrugsstatistikker, især hvis man har mulighed for at koble Excel op til databasen i sit laboratorieinformationssystem.

Håndtering af data fra eksterne kilder

Excel kan læse og importere eksterne data fra en lang række forskellige kilder og filtyper, men jeg vil kun omtale de mest almindeligt anvendte.

Lidt generelt

Når du henter eksterne data, så bliver der (som hovedregel) automatisk oprettet en såkaldt 'Forespørgsel' fra Excel til kilden. Den udformes i Microsoft Query og bliver indbygget i det regneark du arbejder med (den kan eventuelt også gemmes i en separat kommandofil). Det har nogle store fordele:

- Hvis de eksterne data ændres, kan du med et enkelt klik opdatere alle de data du har hentet ind i dit regneark.
- Hvis de indhentede data er et udtræk af en større database, og du ønsker at ændre udtrækket, dvs. at udvælge flere eller færre poster og/eller variable, kan du bare redigere forespørgslen og derefter opdatere.

Data i Excel-filer

Du kan naturligvis kopiere og indsætte data fra andre (eksterne) Excel-filer, men det kan være farligt at 'klone' et datasæt på den måde, da det medfører en høj risiko for at indholdet i forskellige filer ikke er identisk efter et stykke tid. Det kan undgås ved at bruge formler der læser data i den eksterne fil på forskellige måder, men løsningen er tung at danse med, hvis det drejer sig om store datasæt, og kan også give problemer, hvis der flyttes rundt på data.

Jeg vil anbefale at bruge forespørgsler, som især er lette at anvende, hvis data i findes samlet i en navngivet blok (tabel, liste) i den eksterne Excel-fil. Udover at undgå de ovennævnte problemer, er det også en effektiv løsning, hvis du kun skal bruge dele af det eksterne datasæt. I Excel 2007 bliver de hentede data automatisk konverteret til en tabel, hvilket er meget bekvemt af forskellige grunde.

Data i simple tekstfiler

Import af tekstfiler, hvor indholdet i de enkelte variable er adskilt med kommaer eller andre specialtegn, er hyppigt anvendt indenfor klinisk biokemi, fordi mange laboratorieinformationssystemer ikke kan producere andre filtyper.

Det fungerer ofte også udmærket, hvis man er omhyggelig og sikrer sig, at data altid bliver korrekt formateret i Excel, fx så datoangivelser bliver opfattet som datoer, og CPR-numre bliver opfattet som tekst og ikke tal.

Hvis du har tilbagevendende opgaver med import af en bestemt type tekstfiler, kan det betale sig bruge en VBA makro der styrer importen og/eller at indstille forespørgslen fra Excel, så du kan opdatere alle data i et regneark ved blot at vælge en ny tekstfil.

Data i Access-filer

Du kan uden problemer kopiere data fra Access til Excel, og vice versa, men jeg vil varmt anbefale at bruge forespørgsler. Dels af de grunde der er nævnt ovenfor vedrørende Excel-filer, og dels fordi en Access-database kan indeholde mange flere poster end Excel og kan have en kompliceret struktur, så du typisk har brug for kun at hente udvalgte data.

Excel har indbygget valgmuligheden 'Hent fra Access', men det er bedre at bruge forespørgsler i Microsoft Query, fordi de er mere fleksible og kan redigeres.

Et analyseværktøj til statistik

Excel har et lille tilføjelsesprogram der kan udføre diverse gængse statistiske analyser, fx ANOVA, korrelationsanalyser og lineær regression. Hvis du ikke kan finde programmet, skyldes det højst sandsynligt, at du ikke har aktiveret det i indstillingerne af Excel.

Det er let gjort, og når du alligevel er i gang, så aktivér også diverse andre tilføjelsesprogrammer, især 'Analysis ToolPack', som øger antallet af tilgængelige formler i Excel.

Programmering i VBA

Man kan programmere alle funktioner i Excel med det tilhørende Visual Basic for Applications (VBA). Det kan være ekstremt tidsbesparende at programmere diverse rutiner (og er også forbundet med langt færre fejl), især hvis man har større, tilbagevendende og rutineprægede opgaver, fx med at lave løbende produktionsstatistikker.

Det er nemt at indspille makroer i VBA til at kunne afvikle mindre rutiner, mens større, samlede og hurtigkørende løsninger kræver et dybere kendskab til VBA, og eventuelt ekstern konsulentbistand.

Et par andre fiduser

- Hvis du konstruerer projektmapper som andre skal anvende, og gerne vil skrive vejledninger til anvendelsen, så indsæt Word-dokumenter som objekter i Excel-filerne. Det er meget lettere end at skrive tekster direkte i Excel. Du kan også indsætte billeder, lyd og videosekvenser, hvis det er relevant.
- Hvis du ændrer indholdet i nogle celler, fx fordi du har fundet fejl, så indsæt kommentarer i cellerne, med dato og beskrivelse af ændringerne. Det kan siden spare dig (og andre) for unødvendige og ofte tidsrøvende efterforskninger.
- Udover at kunne styre hvad der kan indsættes i en celle (datavalidering), kan du også låse et regneark, eller dele af det, så andre ikke kan ændre indholdet.

Hvis du vil lære mere

Denne artikel kan downloades i flere gængse dokumentformater fra www.kliniskbiokemi.net, så du fx kan linke direkte til de websites jeg har anført nedenfor. Word 2007 dokumentet indeholder også den Excel 2007 projektmappe jeg har brugt til illustrationerne.

Hjælpefunktionen i Excel

Der findes faktisk mange gode instrukser og eksempler, især hvis man aktiverer adgangen til hjælp på Internettet. Her kan man ofte linke sig videre fra mindre artikler til uddybende artikler, online kurser m.m.

Oversigtsartikler og online kurser

Du kan via www.microsoft.dk finde et væld af artikler og online kurser om brugen af Excel, og de er ofte er af høj kvalitet.

Bøger (papir)

Der findes mange bøger om brugen af Excel på forskellige niveauer, fx Excel 2007 Plain and Simple, —Step-by-Step, —Inside-Out, —For Dummies, —How to do Everything, —Bible, og —Developers Guide.

Det er typisk værker på mellem 500 og 1.000 sider, som koster mellem 200 og 500 kr. De er vældig gode at have i huset, både til at bladre lidt rundt i på må og få (for inspirationens skyld), til systematisk læsning og som håndbøger.

Man kan finde og bestille diverse bøger på www.bookworld.dk eller andre tilsvarende sites.

Kurser

Der findes efterhånden mange udbydere af kurser i brugen af Excel på forskellige niveauer. De professionelle kurser kan findes via

www.microsoft.dk. Kurser koster typisk omkring 5.000 kr. per person for 2 dage.