ISSN 1893-1170 (online utgave)
ISSN 1893-1057 (trykt utgave)
Mineralproduksjon 4 (2013) B9-B14
MINERALPRODUKSJON
www.norskbergforening.no/mineralproduksjon

Notat

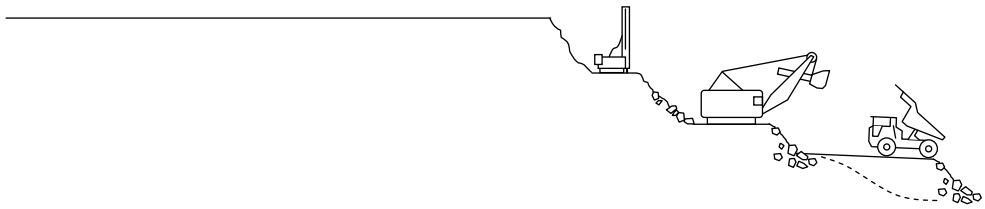
Utfordringer knyttet til statistisk analyse av kompositttdata

Steinar Løve Ellefmo [1,*]

[1] Institutt for geologi og bergteknikk, Sem Sælandsvei 1, 7491 Trondheim

** Korresponderende forfatter: steinare@ntnu.no*

Vi er omgitt av kompositttdata. Denne typen data inneholder bare relativ informasjon og fordrer at man tar forhåndsregler dersom man ønsker å gjennomføre en etterprøvbar og robust statistisk analyse. Dette notatet ser på hva kompositttdata er og utfordringer knyttet til en statistisk analyse av slike data.



1. HVA ER KOMPOSITTDATA?

Kompositttdata er tidligere blitt definert som tilfeldige vektorer av strengt positive komponenter som summeres til en konstant, f.eks. 100, 1 eller en million. Forskningsarbeid gjennomført det siste tiåret har redefinert kompositttdata til å inkludere alle vektorer som representerer andeler av en helhet som bare inneholder relativ informasjon. Dette inkluderer dermed ikke bare deler per million eller prosent, men også molarkompositter.

Typiske eksempler finner man innen mange fagdisipliner (ikke komplett) (CoDaWeb 2013):

- Geologi
    - Geokjemi (bulkkjemi)
    - Mineralkjemi
    - Andeler av ulike sedimenttyper i en sedimentær bergart
    - Andeler ulike mineraler i en bergart
- Medisin
    - Kroppssammensetning (fett, bein, muskler etc)

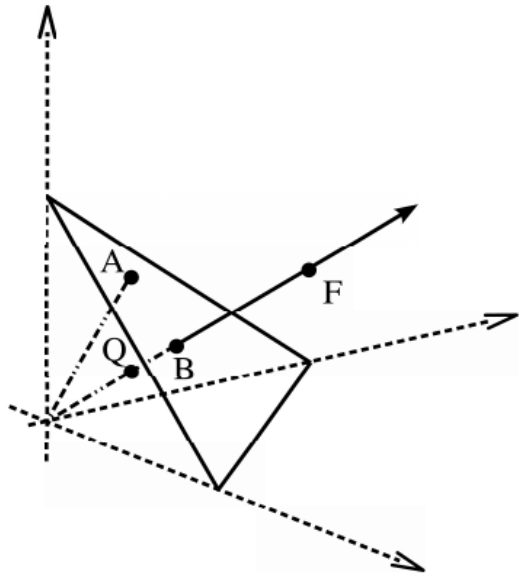
- Kjemi
    - Kjemisk sammensetning
- Økologi
    - Fordeling av ulike dyrearter innen en region

Et eksempel fra en geokjemisk analyse av en bergart er gitt i tabell 1.

Tabell 1. Eksempel på kompositttdata. Kun deler av datasettet som er fremstilt i figur 2 og 3.

Sample	SiO_2	TiO_2	Al_2O_3	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5	Fe_2O_3t	Sum
RT1-1	73,40 %	0,22 %	14,19 %	0,04 %	0,48 %	1,04 %	4,33 %	4,18 %	0,06 %	1,74 %	99,68 %
RT1-10	71,66 %	0,14 %	15,34 %	0,03 %	0,37 %	0,91 %	4,82 %	4,76 %	0,05 %	1,11 %	99,19 %
RT1-16	72,47 %	0,14 %	14,84 %	0,03 %	0,33 %	0,71 %	4,44 %	5,14 %	0,05 %	1,05 %	99,20 %
RT1-17	72,26 %	0,13 %	14,51 %	0,02 %	0,29 %	0,72 %	4,37 %	5,11 %	0,04 %	0,98 %	98,43 %
RT1-19	71,76 %	0,16 %	14,49 %	0,03 %	0,34 %	1,04 %	4,82 %	3,90 %	0,05 %	1,21 %	97,80 %
RT1-20	72,25 %	0,10 %	13,75 %	0,02 %	0,23 %	0,72 %	4,58 %	4,15 %	0,04 %	0,80 %	96,64 %

Som vi ser av tabell 1, er det ikke et krav at radene (vektorene) summeres til 100 % (eller en annen konstant verdi) for at dataene skal anses som kompositttdata. Enkeltelementene i vektorene, inneholder bare relativ informasjon, de sier ingenting om det absolutte innholdet av (i dette tilfellet) oksidene.



Figur 1. Simplexen for et tredimensjonalt system (Pawlowski-Glahn et al. 2007).

Kompositdata defineres på den såkalte simplexen. I et tredimensjonalt system (f.eks. mengde i gram av tre oksider er kvantifisert) vil simplexen være en trekant som vist i figuren 1.

De stiplede linjene i figur 1 angir mengden i gram (eller en annen passende enhet) av hvert oksid. I simplexen er dataene regnet om til %. Punkt Q, B og F i det stiplede koordinatsystemet er tre klart ulike punkt (ulik mengde av hvert oksid). I simplexen er punktene identiske.

Hva har dette å si for vår analyse av slike data?

2. UTFORDRINGER OG MULIGE KONSEKVENSER

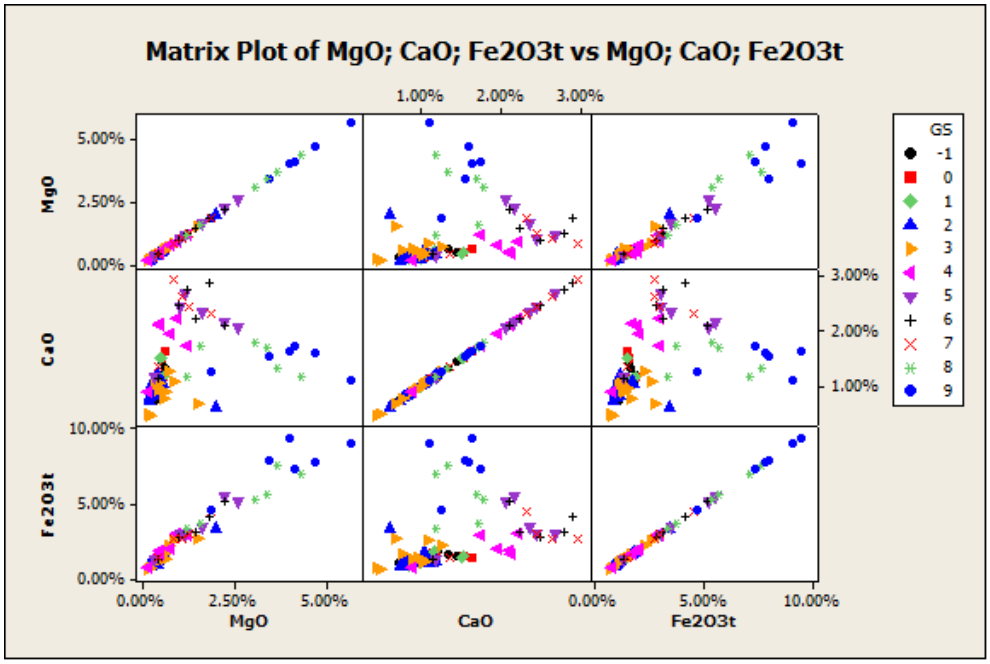
Ofte ønsker man å se på sammenhengen eller korrelasjonen mellom ulike parametere. I et system der vektorene summerer seg til en konstant, vil positive korrelasjoner undertrykkes og negative korrelasjoner forsterkes. Dette følger av den enkle betraktning at dersom et element i vektoren går opp, vil et annet element måtte gå ned, fordi summen skal være konstant. Dette ble beskrevet allerede av Karl Pearson i 1879 (Pearson 1979). Pearson kalte dette «spurios correlations» eller uekte / falske korrelasjoner. Man risikerer med andre ord å identifisere korrelasjoner som ikke er et resultat av f.eks. en geologisk prosess, men som er et resultat av den såkalte lukkingen av datasettet (som finner sted når man tar steget over til et datasett bestående av vektorer som summeres til en konstant).

Når man analyserer geologiske data, ønsker man å studere forholdet mellom et utvalg elementer i datasettet. Da er det vanlig å isolere disse og normalisere de slik at de summeres til 100 %. Man lukker datasettet. Dette gjør man for å kunne plotte dataene i såkalte ternære plott, eller trekantplott. Figur 2 og 3 illustrerer hva som skjer.

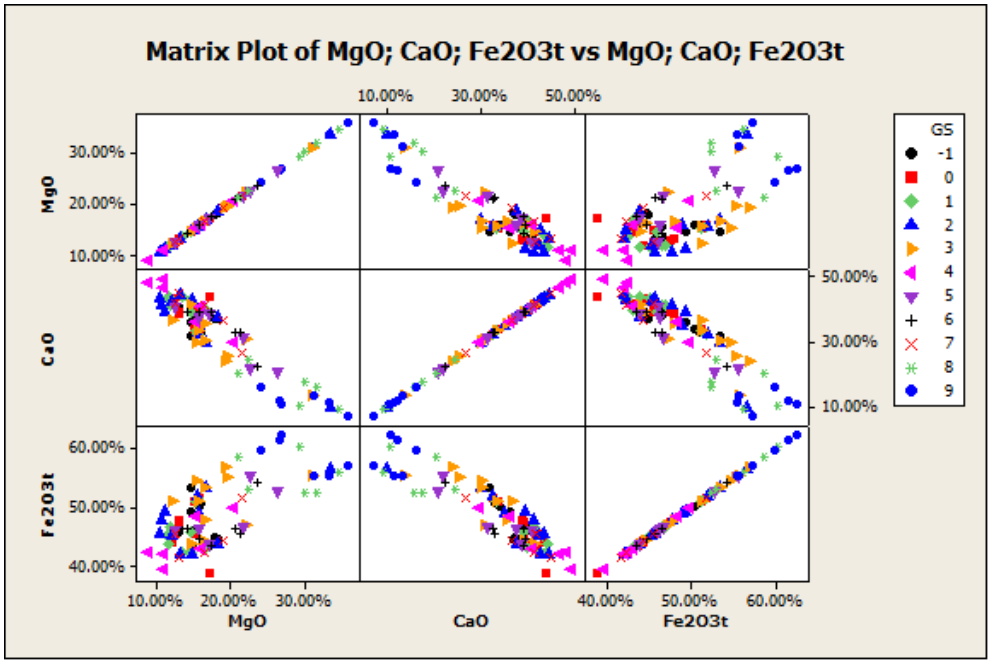
Figur 2 angir sammenhengene mellom MgO, CaO og Fe_2O_3 i det fulle datasettet i tabell 1. Vi ser av figur 2 at det er en klar sammenheng mellom MgO og Fe2O3, at det er mangel på sammenheng mellom MgO og CaO og mellom CaO og Fe2O3.

Isolerer man de tre parametrene og normaliseres slik at elementene summeres til 100 % og lager et tilsvarende plott ser man nye sammenhenger. Disse vises i figur 3.

Man kan nå tydelig se at det er en signifikant sammenheng mellom alle parameterne. Korrelasjonen mellom MgO og Fe_2O_3 er tydelig positiv, mens de andre korrelasjonene tydelig negative. Denne korrelasjonen er kun et resultat av lukkingen.



Figur 2. Sammenhengen mellom MgO, CaO og Fe_2O_3 i det fulle datasettet.



Figur 3. Sammenhengen mellom MgO, CaO og Fe_2O_3 i deldatasettet.

En statistisk analyse av data bør gi samme resultat uavhengig om man studerer hele datasettet eller kun ser på delkompositter. I dette tilfelle er det åpenbart at konklusjonene ville blitt annerledes om man studerer sammenhengene i figur 2 eller i figur 3.

Skulle man f.eks. ønske å etablere sammenhengen mellom bulkgeokjemi og mineralogi, vil man ha tre nivåer av kompositdata, alle med falske korrelasjoner. For det første vil bulkgeokjemien være kompositdata. For det andre vil andelen av de ulike mineralene være kompositdata. For det tredje vil mineralkjemien være kompositdata. Hvordan skal man forvente å finne reelle sammenhenger med så mye falsk korrelasjon?

3. FORSLAG TIL LØSNING

Forslag til løsning på disse utfordringene kom med arbeidet til Aitchison (1986). Han foreslår en transformasjon av dataene der hvert element i vektorene divideres med andelen av et av elementene. Transformasjonen kalles alr-transformasjonen (ligning 1):

$$\mathrm{alr}(\mathbf{x}) = \left[\ln\frac{x_1}{x_D}, \ln\frac{x_2}{x_D}, ..., \ln\frac{x_{D-1}}{x_D}\right] = \mathbf{y}. \quad (1)$$

Der n er antall element eller variable går man dermed fra n antall element til n-1 antall element. Analysen av et slikt transformert datasett er uavhengig av skala (skalainvariant), den er uavhengig av permutasjoner (rekkefølgen spiller ingen rolle) og det spiller ingen rolle om man ser på kompositter eller delkompositter.

Senere arbeid har resultert i to andre, alternative transformasjoner gitt i ligning 2 og 3.

$$\mathrm{clr}(\mathbf{x}) = \left[\ln\frac{x_1}{g(\mathbf{x})}, \ln\frac{x_2}{g(\mathbf{x})}, \ldots, \ln\frac{x_D}{g(\mathbf{x})}\right] = \boldsymbol{\xi}. \quad (2)$$

$$y_i = \sqrt{\frac{i}{i+1}}\ \ln\left[\frac{g(x_1, \ldots, x_i)}{x_{i+1}}\right], \quad i = 1, 2, \ldots, D-1, \quad (3)$$

Clr-Transformasjonen i ligning 2 gir n nye element, mens ilr-transformasjon i ligning 3 gir n-1 nye element.

De nye datasettene man får ved å bruke disse transformasjonene kan man analysere med tradisjonelle statistiske teknikker som regresjonsanalyse, variansanalyse og clusteranalyse. De falske korrelasjonene er løst opp. Dessverre kan det være utfordrende å tilbaketransformere resultatene, samt at tolkningen av analyseresultatene kan være krevende.

4. ETTERORD

Potensielt ligger det vesentlige feller man kan falle i når det gjelder analysen av komposittdata. Institutt for geolog og bergteknikk ved NTNU vil søke å finne ut av når spesielle hensyn må tas og når man dermed må benytte noen av de presenterte transformasjonene.

REFERANSER

CoDaWeb 2013: http://www.compositionaldata.com//pages/about-us.php

Pawlowsky-Glahn, V et. al. 2007: Lecture notes on compositional data analysis. Retrieved from http://dugi-doc.udg.edu/bitstream/handle/10256/297/CoDa-book.pdf?sequence=1, 7/10-13

Pearson K., 1897: Mathematical Contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurements of organs: Proc. R. Soc., v. 60, p. 489-498

Aitchison J., 1986: The Statistical Analysis of Compositional Data. Chapman and Hall