Data analysis and Geostatistics

Short Course on the use of statistical techniques in the geosciences

Vincent van Hinsberg • McGill University

Geotop Short Course in Data Analysis and Geostatistics

Vector methods, PCA, FA and PLS

A few words of caution….

Eigenvector and clustering methods are extremely powerful aids in understanding your data, and the underlying processes that control the variability in your study

However;

"Principal component analysis belongs to that category of techniques, including cluster analysis, in which appropriateness is judged more by performance and utility than by theoretical considerations"

Davis, 3rd ed., 2002

And;

Eigenvector methods require there to be multidimensional correlations in the data set with meaningful causation —> if these are absent, they are not going to magically appear, and eigenvector methods are useless.

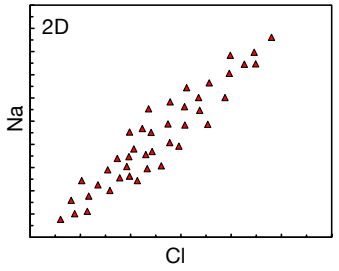
Also, if they are present in 2-D, there is no added value in multi-D —> you look for hidden directions in your data in eigenvector methods

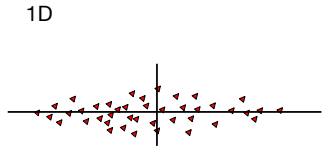
Eigenvector methods

Two main techniques: principle component and factor analysis

both techniques perform a transformation of the data to allow for easier interpretation through:

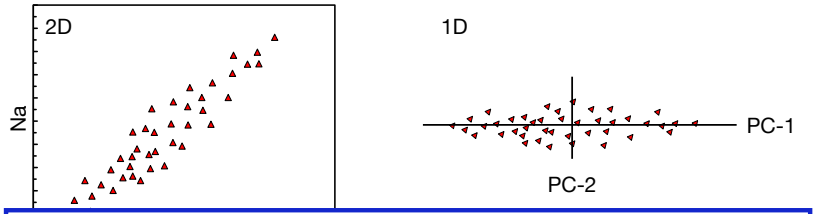
- reduction of variables
- suggestion of underlying processes





Eigenvector methods

System can be transformed to its principle components



The data have been re-cast into a new coordinate system where the axes are the principal processes operating on your data + noise

majority of the variance in this system resides in PC-1 and PC-2 can be interpreted as just the scatter/noise in the data:

dimensionality of the system reduced from 2-D to 1-D
without losing any information !

Eigenvector methods - PCA and FA

General notes on Principle Component Analysis and Factor Analysis

- Principle components are the principle non-correlated directions in your data set (maximized variation along, minimized variation perpendicular to PC)
- Nowadays datasets with 50 to 100 variables are not uncommon. A reduction to a much smaller number of unrelated variables (the Factors) makes it much easier to mine such a dataset
- In PCA all variance is redistributed to new PCs, resulting in the same number of PCs as original variables.
- In FA, only those PCs that are informative are retained and the remainder is discarded as noise. The PCs can also be rotated to simplify interpretation.
- Strictly speaking PCA is a mathematical transformation of your data that retains all information, whereas FA is an interpretive model of your data.
- In reality, most software package call both PCA

Principle component analysis - PCA

Especially useful in multi-D space

have already seen an example of this approach when looking at DFA:

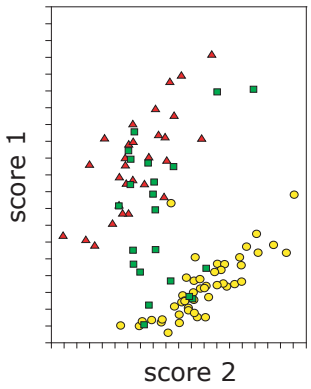
5-D replaced by 2 vectors that allow you to recognize clustering in the data:

this was not obvious in original data

However;

in this case the data are correlated in this 2-D vector space, whereas PCs are not allowed to be correlated

Principle components are the principle non-correlated directions in your data set



Principle component analysis - PCA

So what do we do in principle component analysis ?

look in the data for vectors that have maximum variance along them
(i.e. a strong correlation/covariance)

as all variables that display the same correlation/covariance are grouped together, the trend they describe cannot be shared with any other PC

PC-1 = vector that explains most of the variance
the strongest direction in your data

PC-2 = vector that next explains most of the residual variance

PC-3 = vector that next explains most of the residual variance

etc

so PC-1 explains more variance than any single original variable and therefore, PC-n explains less variance than a single variable (noise)

Principle component analysis - PCA

So what do these principle components look like ?

PC: $PC_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + a_{14}X_4 + a_{15}X_5 + ...$
$PC_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + a_{24}X_4 + a_{25}X_5 + ...$
$PC_3 = a_{31}X_1 + a_{32}X_2 + a_{33}X_3 + a_{34}X_4 + a_{35}X_5 + ...$
$PC_4 = a_{41}X_1 + a_{42}X_2 + a_{43}X_3 + a_{44}X_4 + a_{45}X_5 + ...$
.....

X_i = original vars
a_{ij} = coefficients that relate the original vars to the PCs

Note that to satisfy multi-non-colinearity some of the a-coefficients have to be 0

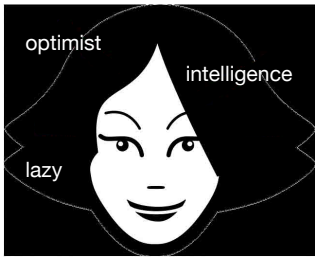
in matrix notation:

PC = A X

$$\begin{vmatrix} PC_1 \\ PC_2 \\ PC_3 \\ PC_4 \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{vmatrix} \cdot \begin{vmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{vmatrix}$$

Principle component analysis - PCA

Another way to explain PCA: psychological questionnaires



A psychologist want to know your intelligence, whether you are extroverted, a pessimist, etc.

Have to work this out from indirect questions that correlate with the variable that you are interested in (e.g. intelligence, optimism, etc)

Many questions lead to a small number of ultimate variables

$$\begin{vmatrix} PC_1 \\ PC_2 \\ PC_3 \\ PC_4 \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{vmatrix} \cdot \begin{vmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{vmatrix}$$

matrix A tells you how to score the answers

Principle component analysis - PCA

The transformation matrix A is what you want to obtain

the matrix that translates the original variables to line up with the principle directions in the data: the PCs

so, it redistributes the variance of the original variables over the PCs, maximizing it for PC_1: $Var(PC_1) = max$
it ensures that the PCs are uncorrelated: $Cov(PC_i\text{-}PC_{i+1}) = 0$

Note that we are only translating the data to a new coordination system: no info loss !

The matrix A is obtained from the covariance or the correlation matrix

when all variables are equivalent (e.g. all wt%, all ppm, etc)

when mixing variables (e.g. ppm + wt% + pH)

Principle component analysis - PCA

Link with correlation makes sense (I hope):

all variables that are correlated define one trend in the data so they should be combined in one PC, and this PC and its component variables should have an insignificant correlation with all remaining variables and PCs

e.g. 5 variables with the following correlation matrix:

	1	2	3	4	5
1	-	0.85	0.14	0.23	0.78
2	-	-	0.21	0.19	0.95
3	-	-	-	0.9	0.25
4	-	-	-	-	0.13
5	-	-	-	-	-

strong correlations:

1 & 2
1 & 5
2 & 5
3 & 4

weak correlations:

1 & 3
1 & 4
2 & 3
2 & 4
5 & 3
5 & 4

so, this dat set has two PCs, with low correlation between them

Principle component analysis - PCA

Strong reduction in dimensionality: 5D to 2D

this allows for much easier data visualization and (hopefully) interpretation

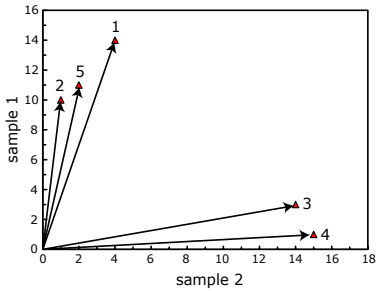
+

it may point to two underlying processes, affecting a different set of vars

a good way to represent this is to plot it in variable space

Now you get two clusters of variables and these are your PCs

So in a way, PCA is cluster analysis on your variables



Principle component analysis - PCA

So how do we obtain the transformation matrix from cor/cov ?

have to determine the eigenvectors in the correlation or covariance matrix

these are the weights that relate the original variables to the PC vectors

and scale these so that the variance of a PC equals 1

pure mathematical transformation

$$\mathbf{PC} = \mathbf{U}\,\mathbf{L}^{-0.5}\,\mathbf{X}$$ in matrix notation

- **PC**: this is the PC vector i.e. the new variables
- **U**: the eigenvector matrix of the cor/cov matrix
- $\mathbf{L}^{-0.5}$: the diagonal eigen-value matrix of the cor/cov matrix, i.e. the scale factor
- **X**: these are the original variables, i.e. the input

input: original vars

operator: eigenvector matrix scaled by eigenvalues

output: PC vector

Principle component analysis - PCA

An example for thermal spring data from the Rockies:

	Si	Al	Cu	Zn	pH	T
PC-1	0.6	-0.2	0.1	0.3	-0.4	0.9
PC-2	0.3	0.7	-0.2	-0.1	0.6	-0.1
PC-3	0.1	-0.1	0.9	0.8	-0.7	0.4
PC-4	-0.2	0.2	-0.3	0.1	-0.4	-0.2
....						

the coefficients are the factor loading:

the correlation between the original variables and the PCs

they display a clear grouping of variables

PC-1: Si and T - as T increases the solubility increases
PC-2: Al and pH - unclear, clay effect? speciation?
PC-3: Ca, Zn, -pH and ±T - low pH + high conc. base metals: sulfides
PC-4: no clear associations - residual noise ?

You get as many PCs as there are original variables, but not all will be meaningful.

Principle component analysis - PCA

The variance in the original variables is redistributed;

PC-1 will have a variance greater than a single original variable (it explains more variance in the data set than a single original variable)

so, subsequent PCs will eventually explain less variance than a single original var

such PCs can generally be ignored thereby reducing the dimensionality

but where should we put the boundary?

the eigenvalues show you how much variance a PC explains compared to the original variables and this value can therefore be used to define a cut-off:

- all eigenvalues less than 1 are insignificant (generally too restrictive)
- use a scree plot (PC-number versus eigenvalue) - where there is a kink in this plot: boundary - use all PCs up to this point and one beyond
- maximum likelihood method - the goodness-of-fit of the factor model is iteratively tested using the X^2 test and additional factors are calculated from the residual covariance/correlation matrix only if it fails the X^2 test

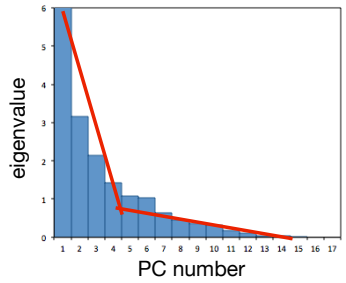
Restricting the number of PCs: FA

The variance in the original variables is redistributed;

PC-1 will have a variance greater than a single original variable (it explains more variance in the data set than a single original variable)

so, subsequent PCs will eventually explain less variance than a single original var

such PCs can generally be ignored thereby reducing the dimensionality



The cut-off can be determined in a scree-plot

Principle component analysis - PCA

To facilitate interpretation the resulting PCs are commonly rotated in multi-dimensional space

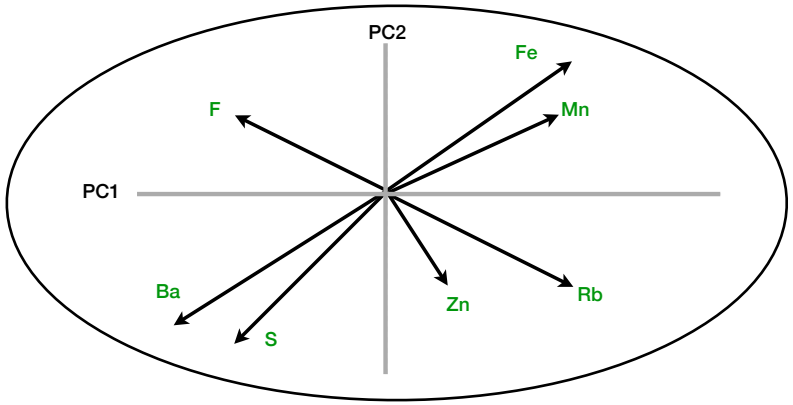
The most popular technique is Varimax rotation:

"rotation to maximize the variance of the squared loadings within each column of the loadings matrix"

this rotation results in the correlations with the original variables to be either large or small, so it enhances the contrast, producing PCs that are highly correlated with a few original variables and very weakly with the rest:

much easier to interpret, because it is immediately clear which variables are important and that in turn can directly point to the underlying process

Varimax rotation



A Factor Analysis example using PAST

The Sn-W mineralized granite of Regoufe (Portugal)

6 km^2
55 samples

After cleaning
15 elements

All elements log-transformed

A Factor Analysis example

The Sn-W mineralized granite of Regoufe (Portugal)

NR	XCO	YCO	F	ZR	SR	CE	BA	B	LI	W	NB	RB	U	TH	TA	AS	CS	SN
1	94	36	3520	29	53	14.6	24.6	540	758	11.8	38	715	5	1.62	14.28	2.6	48	44
2	94	45	3020	43	34	28.1	94.0	640	785	15.5	31	624	4	4.81	10.26	5.4	73	48
3	81	46	3800	43	33	28.8	108.7	280	647	14.43	31	669	7	5.2	11.51	206.4	68	54
4	75	40	3320	27	26	15.2	38.5	330	627	18.66	36	683	1	2.69	12.72	65.1	58	57
5	60	35	2040	31	25	21.2	76.0	860	317	11.02	19	542	11	3.69	9.13	84.7	29	34
6	77	31	2920	29	29	17.2	58.3	360	592	16.78	37	748	8	2.83	17.86	80.5	66	69
7	56	134	4020	19	32	16.2	18.0	70	442	16.71	43	869	20	1.75	19.9	2592.3	44	70
8	65	127	2840	19	15	9.3	32.0	50	423	16.75	53	782	21	1.43	20.05	1045.0	45	64
9	81	130	3600	21	38	17.9	39.0	40	517	19.85	44	889	8	2.23	15.51	463.8	56	76
10	69	118	3200	41	33	25.9	94.3	200	686	29	27	674	10	4.49	11.12	119.7	85	63
12	96	107.5	6880	20	100	8.3	38.0	40	488	10.74	51	840	13	1.17	25.95	256.2	51	101
13	88	112	3800	19	24	12.9	28.9	40	594	10.89	53	935	13	1.83	26.8	210.2	60	74
14	58	115	6160	38	32	21.7	76.0	310	520	37.85	32	601	13	3.75	10.61	588.1	50	60
15	45	113	4280	45	30	27.1	99.9	320	666	11.1	28	638	7	5	8.11	57.5	71	49
16	19	120	4520	17	34	8.1	11.4	40	526	10.01	45	822	13	0.79	20.12	46.5	46	72
17	24	116	4240	26	36	18.0	43.1	120	333	13.12	37	675	18	2.52	11.2	235.3	50	57
18	93	100	3520	24	34	13.4	23.7	40	440	9.93	43	812	8	1.6	18.94	279.4	39	59
19	92	93	4960	19	43	12.7	20.0	120	630	12.5	50	923	7	1.17	23.64	571.8	62	72
20	81	80	6440	18	37	9.4	11.8	50	586	12.67	46	895	13	1.21	25.6	288.8	55	77
21	79	69	3240	40	34	26.9	101.0	40	397	22	33	612	18	4.73	8.43	1010.7	38	58
22	79	60	3300	41	37	23.0	129.9	560	684	9.9	35	679	11	4.2	13.57	16.9	102	55
23	92	57	3720	43	32	24.0	81.0	340	832	10.75	33	703	6	4.75	15.53	58.8	113	63

A Factor Analysis example - PAST

The Sn-W mineralized granite of Regoufe (Portugal)

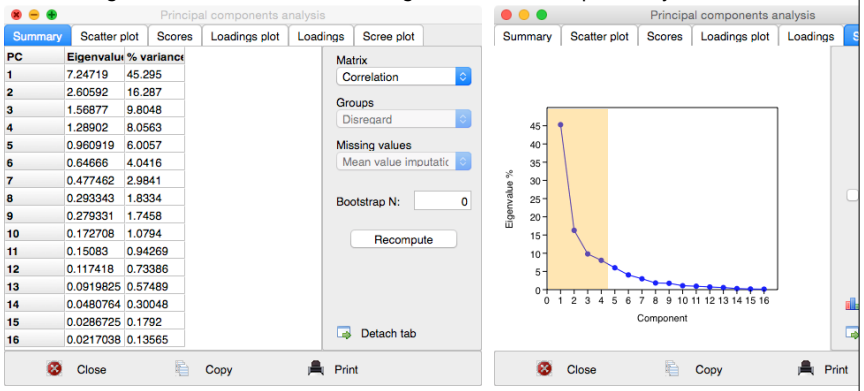
The correlation matrix

	F	ZR	SR	CE	BA	B	LI	W	NB	RB	U	TH	TA
F		-0.50221	0.30731	-0.54151	-0.42093	-0.28885	0.19875	0.15289	0.46813	0.58401	0.012455	-0.53583	0.50934
ZR	-0.50221		0.055172	0.95109	0.86482	0.4027	0.13749	-0.056488	-0.71756	-0.78795	0.012894	0.95715	-0.76939
SR	0.30731	0.055172		0.062507	0.066011	-0.082219	-0.076446	-0.18118	0.041973	0.0027489	0.014244	-0.0051581	0.10591
CE	-0.54151	0.95109	0.062507		0.86814	0.35716	0.03861	-0.037515	-0.72012	-0.77759	0.07226	0.96249	-0.77617
BA	-0.42093	0.86482	0.066011	0.86814		0.42212	0.056067	0.079535	-0.76312	-0.75126	0.10691	0.88669	-0.77485
B	-0.28885	0.4027	-0.082219	0.35716	0.42212		0.11569	-0.10935	-0.57752	-0.55081	-0.20643	0.40807	-0.53463
LI	0.19875	0.13749	-0.076446	0.03861	0.056067	0.11569		-0.059215	0.076383	0.2815	-0.63521	0.081594	0.10293
W	0.15289	-0.056488	-0.18118	-0.037515	0.079535	-0.10935	-0.059215		-0.12996	-0.0084492	0.15031	-0.044045	-0.096932
NB	0.46813	-0.71756	0.041973	-0.72012	-0.76312	-0.57752	0.076383	-0.12996		0.81782	-0.030427	-0.7507	0.89426
RB	0.58401	-0.78795	0.0027489	-0.77759	-0.75126	-0.55081	0.2815	-0.0084492	0.81782		-0.18664	-0.79942	0.88746
U	0.012455	0.012894	0.014244	0.07226	0.10691	-0.20643	-0.63521	0.15031	-0.030427	-0.18664		0.0092227	-0.11037
TH	-0.53583	0.95715	-0.0051581	0.96249	0.88669	0.40807	0.081594	-0.044045	-0.7507	-0.79942	0.0092227		-0.79443
TA	0.50934	-0.76939	0.10591	-0.77617	-0.77485	-0.53463	0.10293	-0.096932	0.89426	0.88746	-0.11037	-0.79443	
AS	0.20773	-0.32656	-0.093352	-0.21183	-0.24763	-0.34152	-0.24679	0.29603	0.26883	0.32194	0.46611	-0.30133	0.22816
CS	0.072648	0.29065	-0.036181	0.25418	0.29177	0.10383	0.75412	-0.09801	-0.079188	0.080846	-0.37248	0.30078	-0.067257
SN	0.73552	-0.59186	0.15903	-0.60059	-0.47931	-0.47726	0.28084	0.15559	0.56749	0.78645	0.019273	-0.5946	0.62395

A Factor Analysis example - PAST

The Sn-W mineralized granite of Regoufe (Portugal)

PC eigenvalues: how much of the original variance is captured by each PC



A Factor Analysis example - PAST

The Sn-W mineralized granite of Regoufe (Portugal)

A variable's communality tells you how much of its variance is explained by your factors. In this case, for 4 factors:

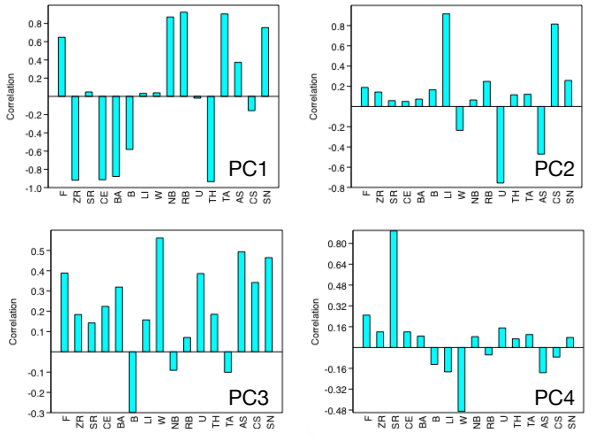
Communalities

	Initial	Extraction
F	1.000	.728
ZR	1.000	.944
SR	1.000	.698
CE	1.000	.905
BA	1.000	.901
B	1.000	.675
LI	1.000	.899
W	1.000	.710
NB	1.000	.790
RB	1.000	.923
U	1.000	.624
TH	1.000	.935
TA	1.000	.860
AS	1.000	.723
CS	1.000	.820
SN	1.000	.873

A Factor Analysis example - PAST

The Sn-W mineralized granite of Regoufe (Portugal)

Variable loadings



A Factor Analysis example - PAST

The Sn-W mineralized granite of Regoufe (Portugal)

Summary | Scatter plot | Scores | Loadings plot | Loadings

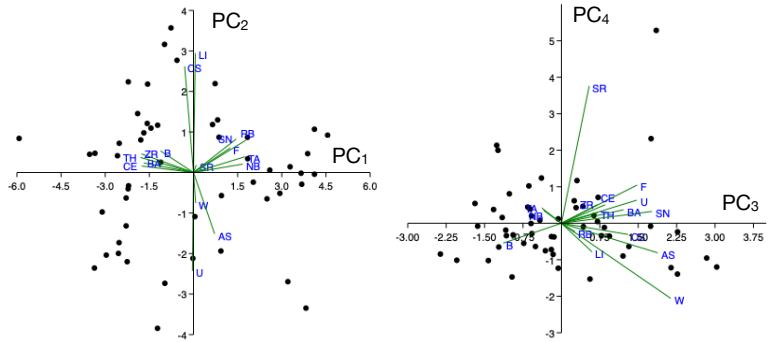
	PC 1	PC 2	PC 3	PC 4
2	0.82047	1.2982	-1.6875	0.54203
3	-2.2156	2.2375	-0.19651	-0.72032
4	-1.5725	1.2107	0.83885	-0.11722
5	0.64699	1.1837	-0.96398	-1.4718
6	-2.5503	-1.9959	-2.3628	-0.84471
7	0.86607	0.86924	-0.16158	-0.85254
8	3.8211	-3.347	2.8373	-0.95335
9	3.2064	-2.6937	0.55837	-1.5285
10	2.5915	0.058918	0.78231	-0.34281
11	-1.2186	1.1672	2.1411	-1.2177
12	5.0685	0.46625	1.8525	5.2839
13	3.8793	0.46543	-0.14352	-0.37514
14	0.04771	-1.0882	3.0333	-1.1993
15	-1.8972	1.454	0.42957	-0.093339
16	3.6772	-0.28991	-0.61238	0.39277
17	0.93084	-1.9372	0.25046	0.61637
18	2.4969	-0.64977	-1.1553	0.15791
19	4.1089	1.0677	0.28669	0.42668
20	4.6143	0.41045	0.71639	0.70874
21	-0.9729	-2.7342	2.2684	-0.23563
22	-1.559	2.1812	0.63526	0.21126
23	-0.76501	3.5679	0.93892	-0.35468

The scores for each sample on the different factors

A Factor Analysis example - PAST

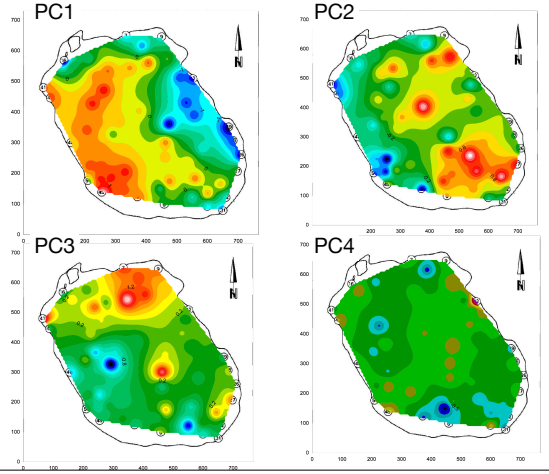
The Sn-W mineralized granite of Regoufe (Portugal)

Bi-plot: combines the data loadings on the PCs and the variable scores on the PCs



A Factor Analysis example

The Sn-W mineralized granite of Regoufe (Portugal)



Tentative interpretation:

PC1: inverse of degree of greisinisation and albitisation

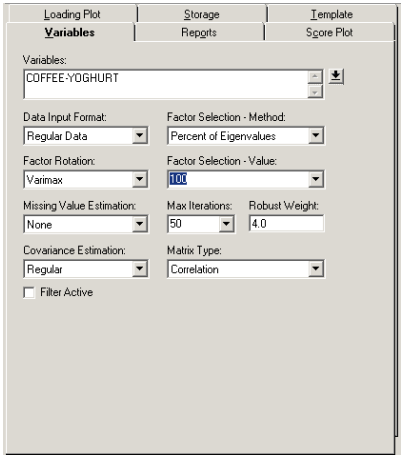
PC2: deuteric alteration

PC3: ore-related elements

PC4: different levels within the granite magma body

Factor Analysis vs. Cluster Analysis

the data set: the eating habits of Europe



variables: the original variables as input

rotation: you can tell NCSS to perform a PC rotation such as Varimax or none

missing values: if you have these you have to tell NCSS how to deal with them: row-wise exclusion, replace by mean or estimate from correlations

matrix type: correlation or covariance

factor selection: start by selecting % eigenvalues and setting this to 100%: gives you all PCs. Can then decide that only first 4 are meaningful and change this

Principle component analysis - PCA

output: the eating habits of Europe

Descriptive Statistics Section

Variables	Count	Mean	Standard Deviation	Communality
Coffee	16	77.5	25.76561	1.000000
Nescafe	16	39.25	23.14735	1.000000
Tea	16	78.5	18.54005	1.000000
Sweetener	16	17.1875	11.02252	1.000000
Biscuits	16	60.875	19.18637	1.000000
Pack_soup	16	49	15.42725	1.000000
Tin_soup	16	18.4375	20.2154	1.000000
Frozen_fish	16	21.875	13.98034	1.000000
Frozen_veg	16	15.875	12.78997	1.000000
Fresh_apples	16	66.8125	17.58112	1.000000
Tin_fruit	16	41.9375	23.25645	1.000000
Jam	16	55.1875	22.59268	1.000000
Garlic	16	42.3125	34.67702	1.000000
Butter	16	75.8125	20.91002	1.000000
Margerine	16	69	26.73076	1.000000
Olive_oil	16	54.1875	28.8426	1.000000
Yoghurt	16	20.625	18.34076	1.000000

Principle component analysis - PCA

output: the eating habits of Europe

Correlation Section

Variables	Variables Coffee	Nescafe	Tea	Sweetener	Biscuits	Pack_soup	Tin_soup	Frozen_fish	Frozen_veg	Fr
Coffee	1.000000	-0.451482	-0.154771	0.267487	-0.106133	-0.302228	-0.238387	0.409759	0.315186	
Nescafe	-0.451482	1.000000	0.290961	0.228696	0.259018	0.727154	0.506803	-0.291814	-0.064515	
Tea	-0.154771	0.290961	1.000000	0.669251	0.209905	0.441225	0.523398	0.320735	0.478507	
Sweetener	0.267487	0.228696	0.669251	1.000000	0.186738	0.226212	0.495064	0.617516	0.814964	
Biscuits	-0.106133	0.259018	0.209905	0.186738	1.000000	0.313521	0.548802	0.029017	0.232484	
Pack_soup	-0.302228	0.727154	0.441225	0.226212	0.313521	1.000000	0.244120	-0.243882	-0.085481	
Tin_soup	-0.238387	0.506803	0.523398	0.495064	0.548802	0.244120	1.000000	0.189154	0.474401	
Frozen_fish	0.409759	-0.291814	0.320735	0.617516	0.029017	-0.243882	0.189154	1.000000	0.905160	
Frozen_veg	0.315186	-0.064515	0.478507	0.814964	0.232484	-0.085481	0.474401	0.905160	1.000000	
Fresh_apples	0.241139	0.480764	0.133454	0.423680	0.538291	0.480530	0.335071	0.030005	0.285694	
Tin_fruit	-0.100075	0.694902	0.546490	0.646054	0.672017	0.613927	0.740128	0.269813	0.530707	
Jam	-0.385434	0.383235	0.538513	0.410780	0.431766	0.366096	0.729504	0.088306	0.339465	
Garlic	0.110990	0.033616	-0.593807	-0.520098	-0.351444	0.015203	-0.546658	-0.322249	-0.473994	
Butter	-0.140384	0.154920	0.302403	0.094169	0.455585	0.114699	0.133634	0.043929	0.186117	
Margerine	-0.174813	0.387019	0.534448	0.351163	0.233069	0.319930	0.300657	0.031041	0.148003	
Olive_oil	0.022293	0.075116	-0.480293	-0.432305	-0.139580	-0.049143	-0.183091	-0.111702	-0.197097	
Yoghurt	0.313610	0.497244	0.003725	0.196254	0.011225	0.409263	0.036074	-0.259675	-0.144018	

Principle component analysis - PCA

output: the eating habits of Europe

Bar Chart of Absolute Correlation Section

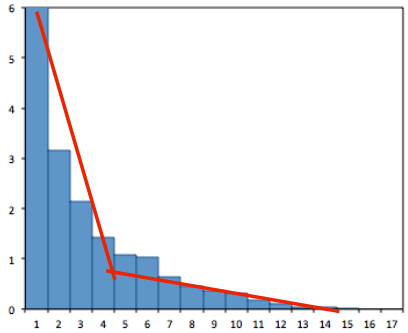
Variables	Variables Coffee	Nescafe	Tea	Sweetener	Biscuits	Pack_soup	Tin_soup	Frozen_fish	Frozen_veg
Coffee									
Nescafe									
Tea									
Sweetener									
Biscuits									
Pack_soup									
Tin_soup									
Frozen_fish									
Frozen_veg									
Fresh_apples									
Tin_fruit									
Jam									
Garlic									
Butter									
Margerine									
Olive_oil									
Yoghurt									

Principle component analysis - PCA

output: the eating habits of Europe

Eigenvalues

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	6.020081	35.41	35.41	|||||||
2	3.156737	18.57	53.98	||||
3	2.139819	12.59	66.57	|||
4	1.425082	8.38	74.95	||
5	1.078415	6.34	81.29	||
6	1.037413	6.10	87.40	||
7	0.642829	3.78	91.18		
8	0.465207	2.74	93.92		
9	0.368342	2.17	96.08		
10	0.319876	1.88	97.96		
11	0.172239	1.01	98.98		
12	0.106809	0.63	99.60		
13	0.037875	0.22	99.83		
14	0.026913	0.16	99.99		
15	0.002364	0.01	100.00		
16	0.000000	0.00	100.00		
17	0.000000	0.00	100.00		



now rerun the routine for up to 4 principle components

Principle component analysis - PCA

output: the eating habits of Europe

Descriptive Statistics Section

Variables	Count	Mean	Standard Deviation	Communality
Coffee	16	77.5	25.76561	0.825747
Nescafe	16	39.25	23.14735	0.826203
Tea	16	78.5	18.54005	0.712119
Sweetener	16	17.1875	11.02252	0.885670
Biscuits	16	60.875	19.18637	0.638357
Pack_soup	16	49	15.42725	0.712475
Tin_soup	16	18.4375	20.2154	0.687004
Frozen_fish	16	21.875	13.98034	0.848950
Frozen_veg	16	15.875	12.78997	0.936613
Fresh_apples	16	66.8125	17.58112	0.781884
Tin_fruit	16	41.9375	23.25645	0.929247
Jam	16	55.1875	22.59268	0.706079
Garlic	16	42.3125	34.67702	0.839802
Butter	16	75.8125	20.91002	0.386488
Margerine	16	69	26.73076	0.404393
Olive_oil	16	54.1875	28.8426	0.682256
Yoghurt	16	20.625	18.34076	0.874905

Principle component analysis - PCA

output: the eating habits of Europe - coefficients

Bar Chart of Absolute Eigenvectors after Varimax Rotation

	Factors			
Variables	Factor1	Factor2	Factor3	Factor4
Coffee	||	||||||	||||||||||	|||
Nescafe	|||||	||||||||			
Tea	|||||||	||	|||	||||
Sweetener	|||||||	|||||	|||||	||||
Biscuits	|||||	|||			||||||||
Pack_soup	|||||	|||||||			|||
Tin_soup	|||||||			|||	||||
Frozen_fish	|||	||||||||	||||	|||
Frozen_veg	||||||	|||||||	||||	||||
Fresh_apples	|||||	||||	|||||||	
Tin_fruit	|||||||	|||	|||	||||
Jam	|||||||			|||||		
Garlic	||||||	|||||	||||||	|||
Butter	|||	||			||||||||
Margerine	|||||			||||	||||
Olive_oil	||||	||||	||||	||||||||||
Yoghurt	||	||||||	||||||||	|||||||

Principle component analysis - PCA

output: the eating habits of Europe - correlations

Bar Chart of Absolute Factor Loadings after Varimax Rotation

	Factors			
Variables	Factor1	Factor2	Factor3	Factor4
Coffee	||||||||||||||	||||	||||||	||||||
Nescafe	|||||||	||||||||||||	|||||	|||||||
Tea	||||	|||	||||||||||||||	||||
Sweetener	||||||||||||	||||||	||||||||||	||
Biscuits			||||	||||	||||||||||||
Pack_soup	|||||||	||||||||||||	|||||	|||||
Tin_soup	||	|||	||||||||||	||||||||||
Frozen_fish	||||||||||||||||	||||||	||||	|||
Frozen_veg	||||||||||||||||	||	|||||||	||||||
Fresh_apples	||||||	||||||||||||||		|||||||
Tin_fruit	|||||	||||||||||	|||||||	||||||||||||
Jam	||	||	||||||||||||	|||||||
Garlic	||||	|||||	||||||||||||||||	|||
Butter							||||||||||||
Margerine	||	||||	||||||||||	|||
Olive_oil	|||	||	||||||||||||	||||||
Yoghurt	||	||||||||||||||||	||	|||||

Principle component analysis - PCA

output: the eating habits of Europe

Factor Structure Summary after Varimax Rotation

	Factors		
Factor1	**Factor2**	**Factor3**	**Factor4**
Frozen_fish	Yoghurt	Garlic	Biscuits
Frozen_veg	Fresh_apples	Tea	Tin_fruit
Coffee	Nescafe	Olive_oil	Butter
Sweetener	Pack_soup	Jam	Tin_soup
	Tin_fruit	Sweetener	
		Margerine	
		Tin_soup	

Principle component analysis - PCA

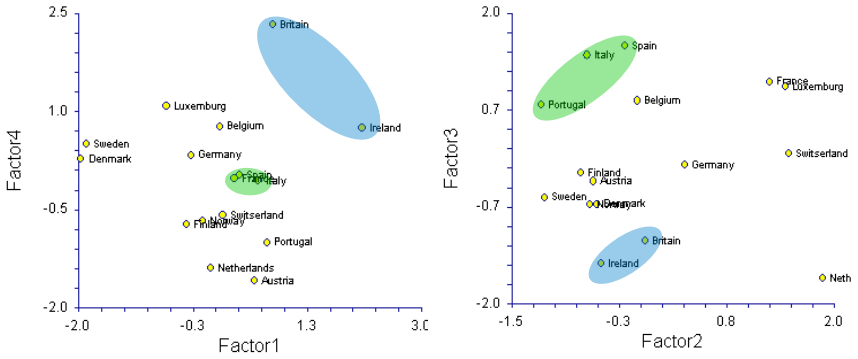
output: the eating habits of Europe - new row transformation

Factor Score after Varimax Rotation

	Factors			
Row	**Factor1**	**Factor2**	**Factor3**	**Factor4**
1	-0.3644	0.3788	-0.0904	0.3360
2	0.6165	-0.6782	1.4195	-0.0599
3	0.2665	1.2995	1.0593	-0.0254
4	-0.0786	1.8754	-1.6456	-1.3972
5	0.0605	-0.1348	0.7957	0.7732
6	-0.7204	1.4754	0.9845	1.0832
7	0.8281	-0.0470	-1.1348	2.3280
8	0.7444	-1.1724	0.7406	-1.0016
9	0.5602	-0.6078	-0.3117	-1.5889
10	0.0995	1.5130	0.0692	-0.5877
11	-1.8878	-1.1434	-0.5359	0.4995
12	-1.9707	-0.5770	-0.6248	0.2745
13	-0.1946	-0.6459	-0.6321	-0.6754
14	-0.4299	-0.7455	-0.1961	-0.7278
15	0.3406	-0.2631	1.5489	0.0244
16	2.1301	-0.5269	-1.4462	0.7452

Principle component analysis - SIMCA

output: the eating habits of Europe



PLS-R and PLS-DA

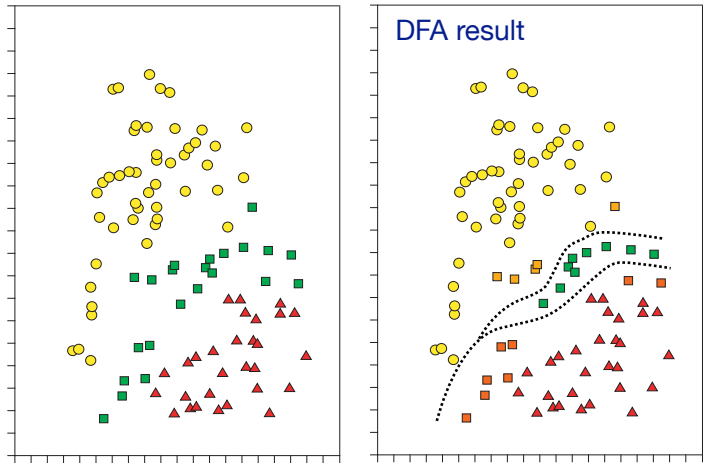
An extension to eigenvector methods with a dependent variable

PCA and FA re-cast the **independent variable** matrix into a new coordinate system aligned with the directions of **maximum variance** with the aim of separating noise from information, reducing the dimensionality of your data, and identify processes

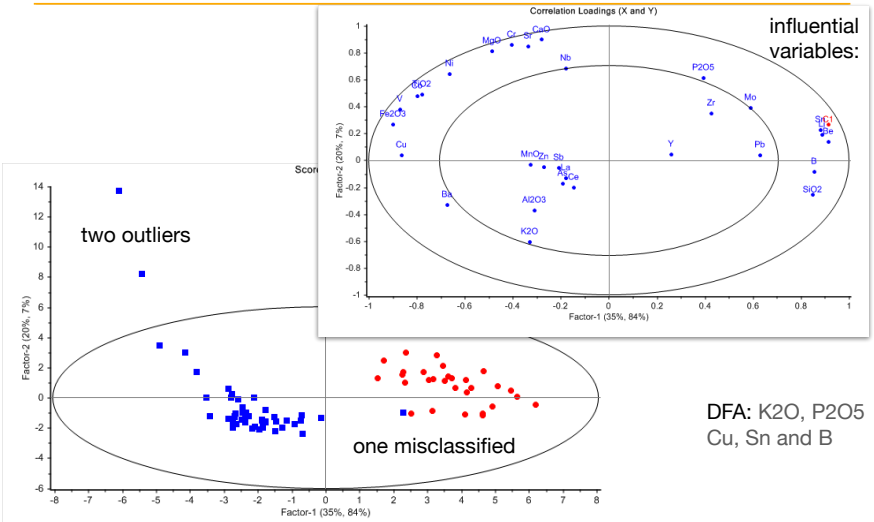
PLS-R and PLS-DA re-cast the **independent variable** matrix into a new coordinate system aligned with a **dependent variable** ($Y = f(\mathbf{X})$) with the aim of classification (-DA) or quantification of a regression model (-R), for example for calibration.

> You can of course do a DA or R based on the original variables, but you here make the assumption that there are directions in your data that better line up with Y than the original variables —> you obtain those from a PCA-style transformation of your data

PLS-DA example with the Unscrambler



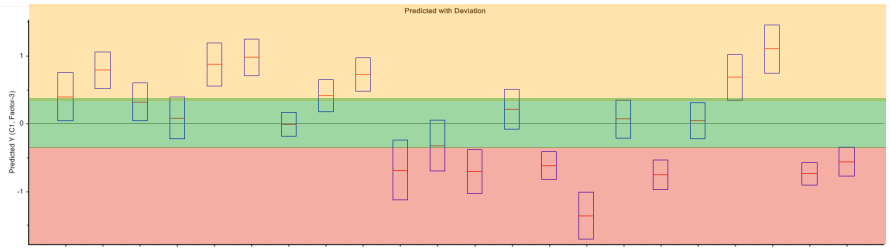
PLS-DA example with the Unscrambler



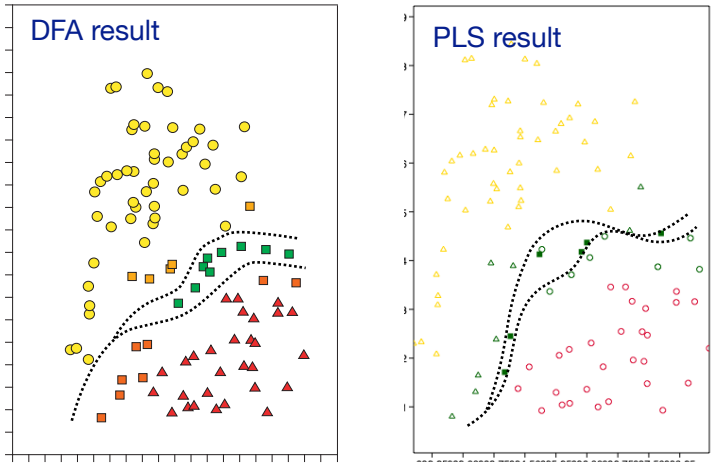
DFA: K2O, P2O5 Cu, Sn and B

PLS-DA example with the Unscrambler

predictions from the PLS model for the unknowns:



PLS-DA example with the Unscrambler

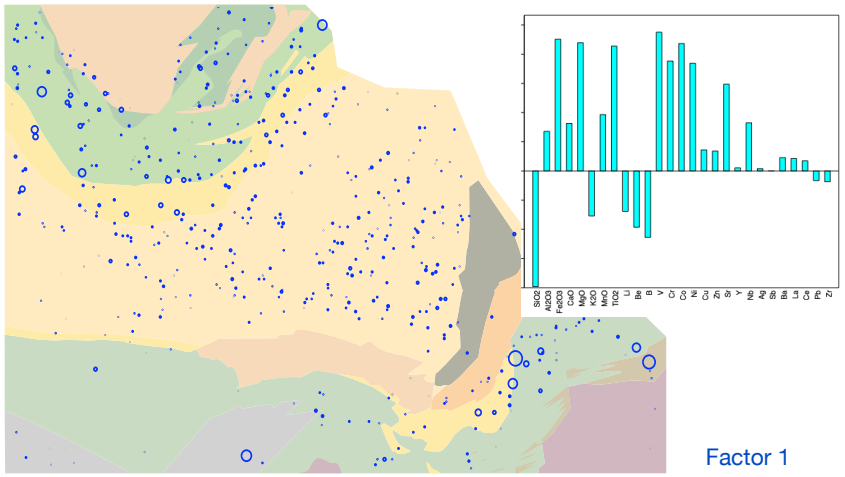


Geotop Short Course in Data Analysis and Geostatistics

Spatial analysis of data

FA - processes in Massif Central dataset

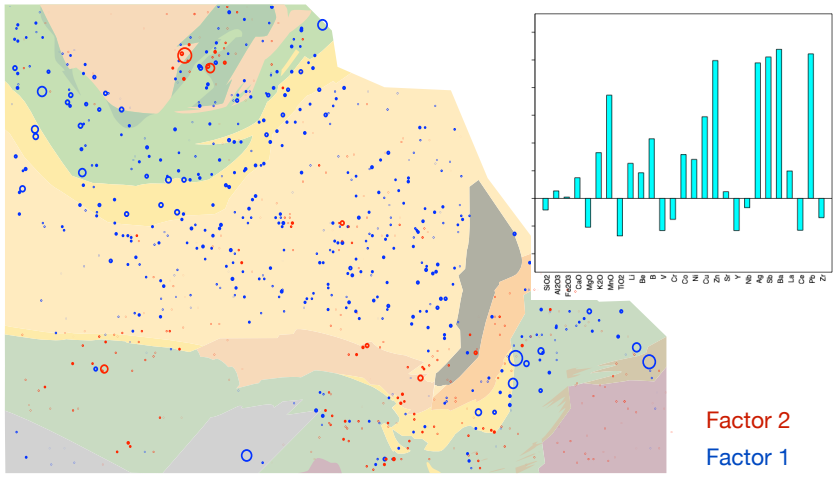
Loadings show the importance of that factor at each location



Factor 1

FA - processes in Massif Central dataset

Loadings show the importance of that factor at each location

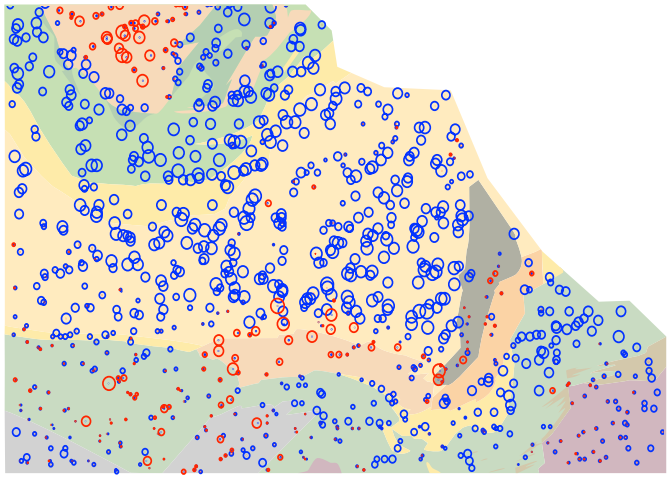


Factor 2

Factor 1

Clustering - groups in Massif Central dataset

Fuzzy cluster assignment shows spatial grouping of samples

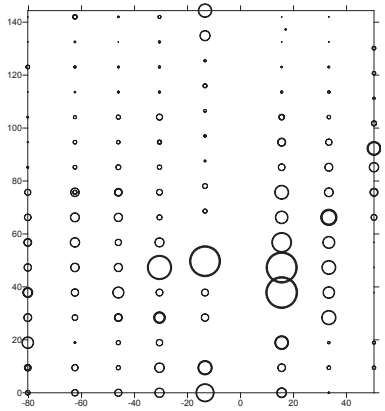


cluster 8

cluster 4

Plotting data on maps: bubble plots

Data are plotted at their spatial coordinates with a symbol whose size represents the value of the data point



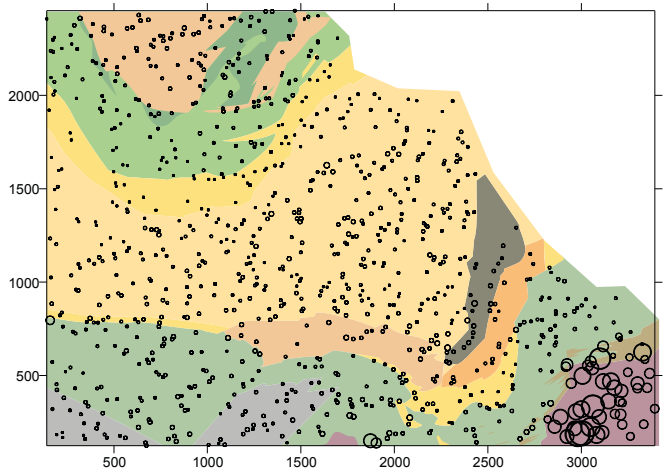
Can apply exactly the same tools as used on the element map:

adjust contrast, isolate features and perform data transformations

can also overlay these bubbles on another layer, such as a topo map, geol map, stream map etc

Plotting data on maps: bubble plots

Stream sediments as a reflection of the local geology: Beryllium

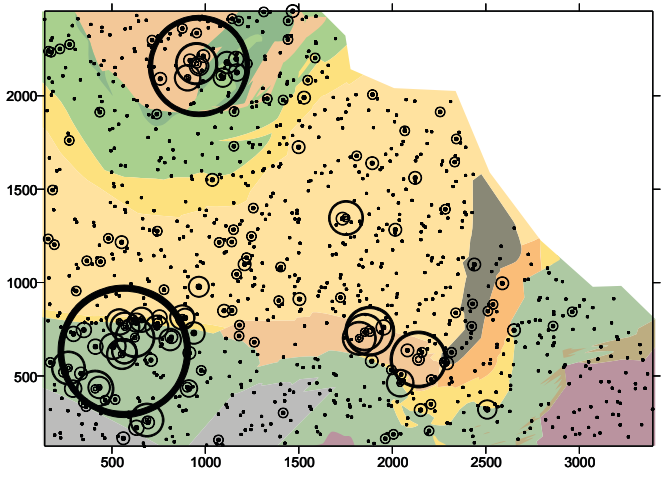


Be concentrations without processing:

sometimes it just works!

Plotting data on maps: bubble plots

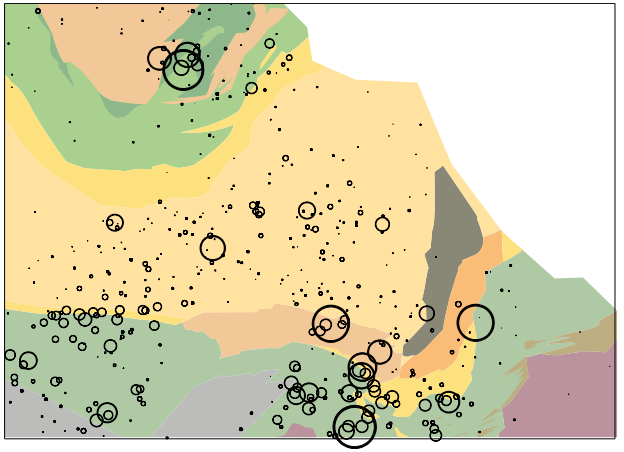
Silver concentrations: working with a non-normal distribution



Ag log-transformed scale

Plotting data on maps: bubble plots

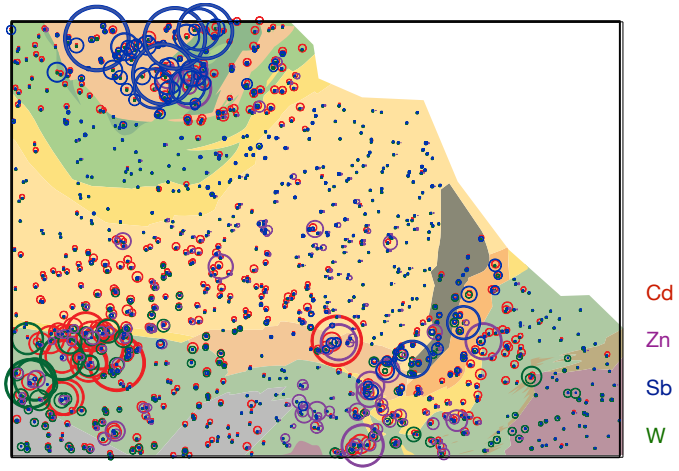
Don't have to plot all the data in the dataset: applying a cut-off at low values will highlight interesting samples, whereas a high cut-off removes outliers



Zn, only data with > 50 ppm

Plotting data on maps: bubble plots

Looking for element associations by combining bubble plots



Plotting data on maps

Combining elements by using multi-coloured bubble plots is useful, but fast becomes confusing: can lead you to miss interesting samples

Can also calculate such associations beforehand and plot them directly:

- Sb + Zn
- Sb / Zn

Or you can apply logical rules to the data before plotting:

- plot Sb if S > 200 ppm
- if SiO_2 > 60 wt% then plot K / Zr

Note that such properties are calculated much easier and faster in programs designed for such calculations: e.g. Excel or Quattro Pro

Plotting data on maps: QGIS and BC dataset

The BC survey makes a digital version of its geological map available onto which you can plot their geochemical data: need a GIS package (qgis.org)

Download the geol map as a shape file here: https://www2.gov.bc.ca/gov/content/industry/mineral-exploration-mining/british-columbia-geological-survey/publications/digital-geoscience-data

make a new file in QGIS, go to project > properties > CRS and set the coordinate system of the file to 3005

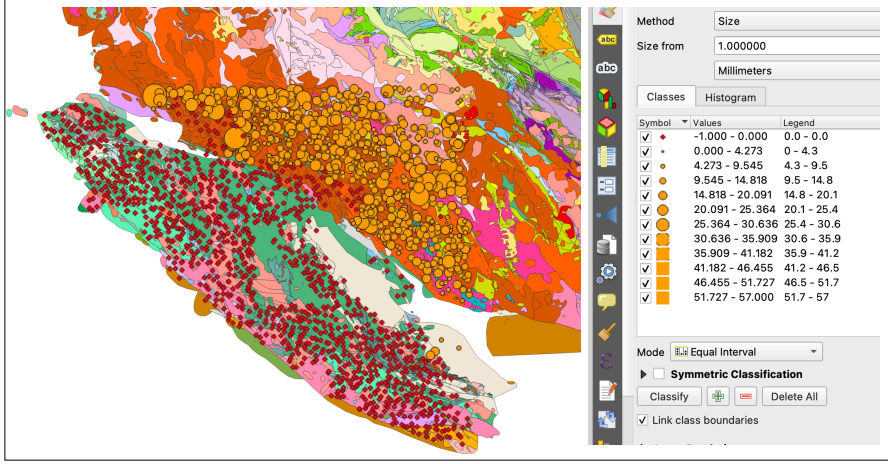
Drag the .shp file into the layers panel. To get the correct colours, go to layer properties > symbology > categorized > style > load style > open file: open the .qml file.

To get your data in, export the excel file as a .csv. Then in QGIS > add layer > add delimited text layer > open your .csv file. Make sure longitude and latitude are selected as x and y fields and set the CRS to 4326 (WGS 1984)

To do fun stuff: click on symbology > graduated > method:size > value:Co > mode:equal interval > classify > apply. You now have a bubble plot for Co

Plotting data on maps: QGIS and BC dataset

The BC survey makes a digital version of its geological map available onto which you can plot your geochemical data: need a GIS package (qgis.org)



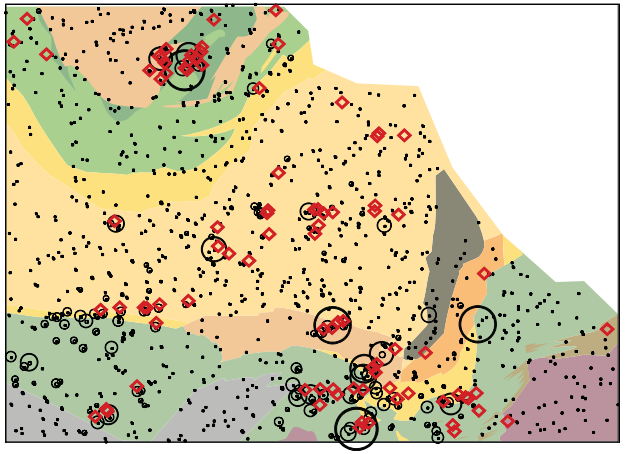
Plotting data on maps

Not limited to plotting data, but can also plot derived properties such as the mean, median, standard deviation, etc

and not just values, but also other observations:
geol code / vegetation / mode in multi-modal distribution

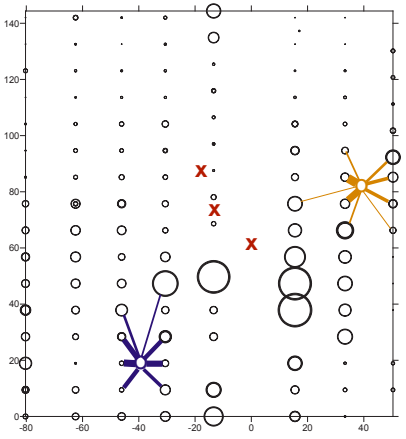
Plotting data on maps: bubble plots

Plotting processed data - standard deviation: the variability at a sample site



Spatial data visualization

To be able to calculate contours and surfaces: interpolation



need to know the concentration at any point in the sampling space to be able to draw smooth contours:

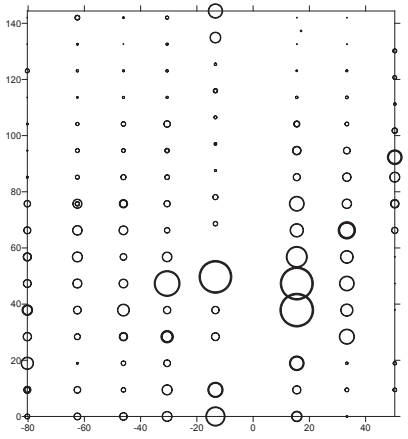
interpolate between values

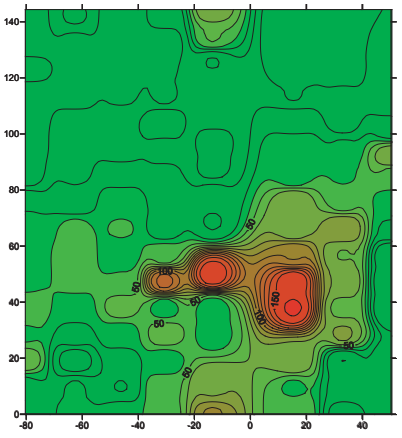
interpolation on As content grid;

- x nearest neighbour
- o radius technique: 1/r
- o radius technique: $1/r^2$

Spatial data visualization

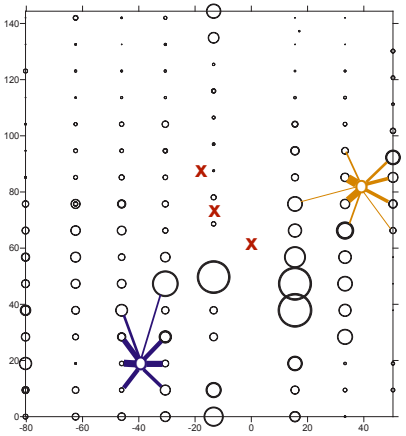
Results of different interpolation techniques:





Spatial data visualization

To be able to calculate contours and surfaces: interpolation



interpolation on As content grid;

- x nearest neighbour
- o radius technique: 1/r
- o radius technique: $1/r^2$

main issue: what samples should be included in the interpolation:

what should the maximum radius be?

Interpolation radius

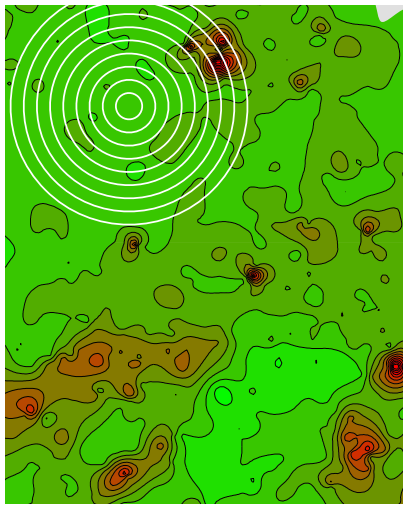
Spatial data have a very useful property:

adjacent samples should be most similar, whereas samples that are far apart can be distinctly different, or:

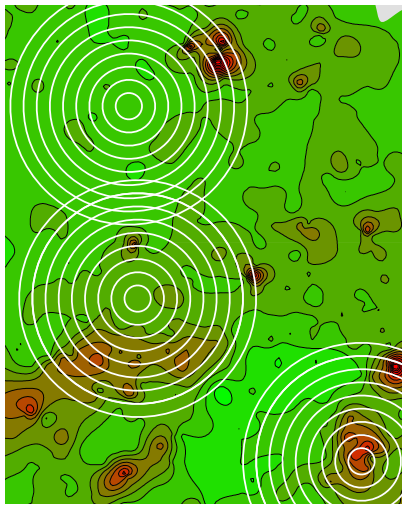
- the variance for a small interpolation radius is small, as the variance between adjacent samples is small
- the variance increases as the interpolation radius increases (i.e. as samples further away from the point of interest are included)
- at some radius the variance will no longer increase as we have reached the overall variance, which is called the “regional variance”

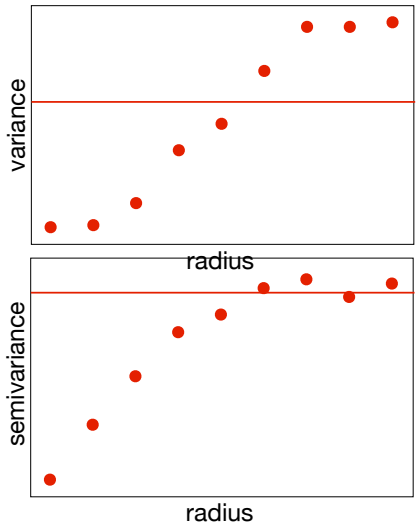
including values beyond the regional variance radius is pointless as such samples do not contain any information on the value at the point of interest

Interpolation radius

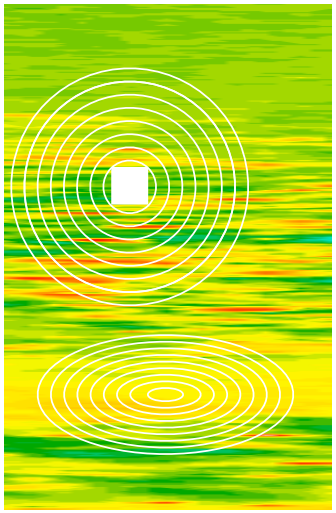


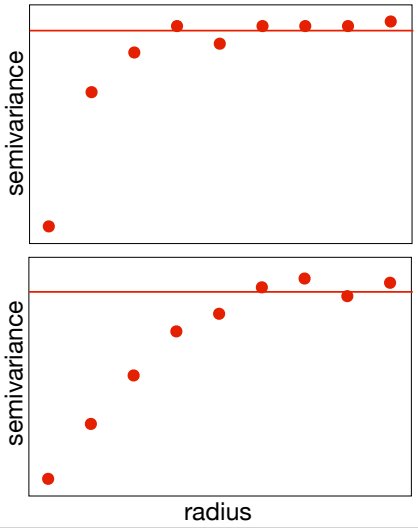
Interpolation radius





Interpolation radius





Semivariance and semivariograms

This concept is semivariance and is shown in a semivariogram

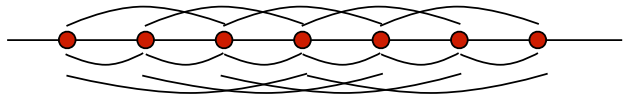
semivariance: the variance between samples a specified interval or distance apart

as the interval increases, the semivariance will approach the total variance of the data set, so it is a spatially controlled partial variance of the data

$$\gamma_h = \frac{\sum (z_i - z_{i+h})^2}{2(n - h)}$$

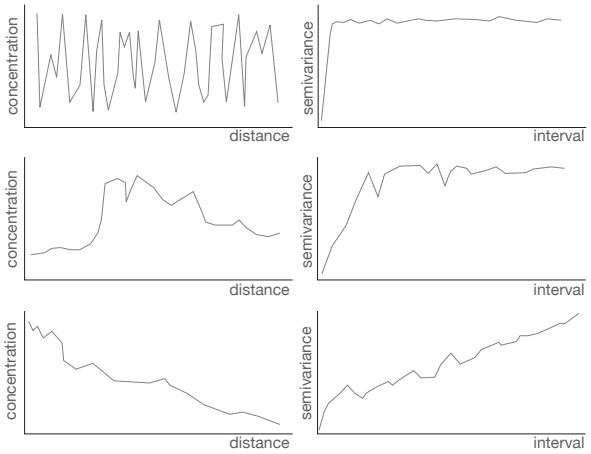
with: γ = semivariance for interval h
n = total number of samples
z_i = value at position i

as h increases, the relatedness of the samples decreases and the variance will therefore increase:



Semivariance and semivariograms

plotting the semivariance against h: semivriogram



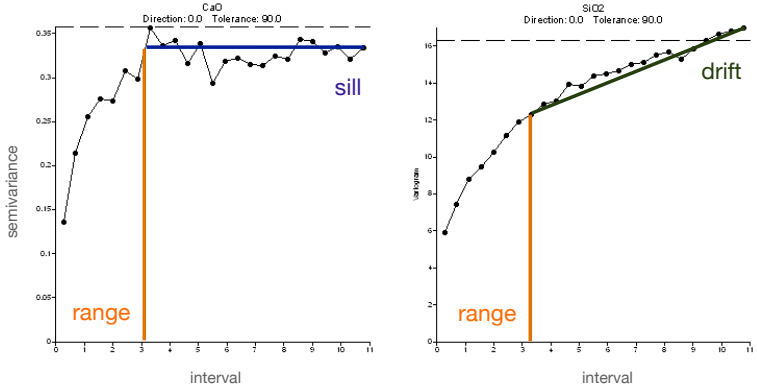
no relation with distance: random

gradual changes in concentration

continuous variation with distance: trend

Semivariance and semivariograms

properties of a semivariogram :



the range is the interval within which there is similarity between the samples

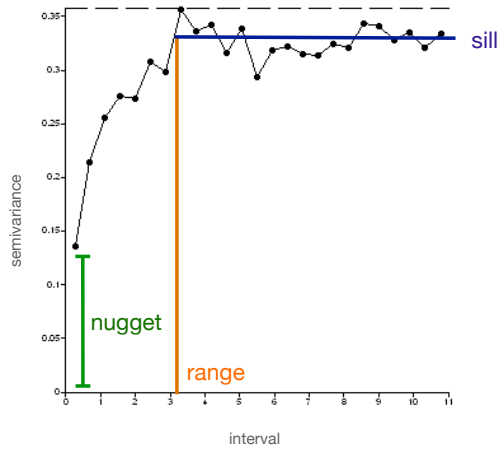
Semivariance and semivariograms

Semivariograms provide our maximum radius criterion: only samples that fall within the range are included in interpolation

before we continue, a few notes:

- semivariograms have to be determined for each variable as each has its own range: interpolation has to be performed separately as well
- semivariograms are generally different for different spatial directions (N, SW, etc). Such anisotropy can point to an underlying geological phenomenon such as layering or a fault control on conc. This can be corrected for either manually by stretching the coordinate system perpendicular to the main axis, or automatically by kriging software
- most semivariagrams have an apparent cut-off at zero distance that has a semivariance ≠ 0. This is called the nugget effect and is caused by sample heterogeneity (= field duplicate variance)

Nugget effect in semi-variograms



There is always some uncertainty at a given sample site, which you could quantify by taking field duplicates.

This sample site variance is the “nugget” in a semivariogram (in essence the variance at zero distance)

Every element will have such a nugget, but the effect is strongest for elements that are heterogeneously distributed, such as gold present as nuggets in a sediment because we use mean + var

Using semivariogram information: kriging

The interpolation technique that employs the range information as obtained from semivariograms is called kriging

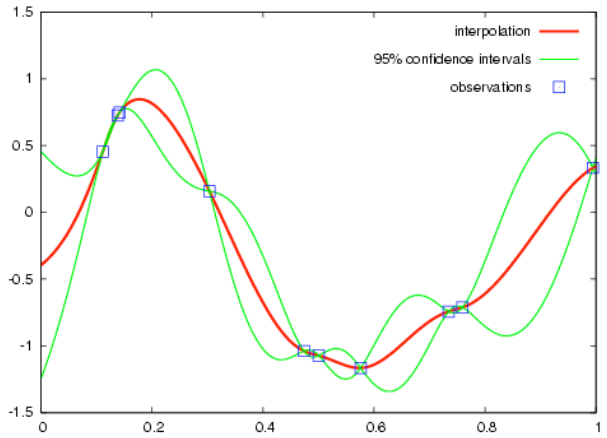
in kriging, only samples that are within the range are used to determine the value at a given intermediate position and the weighing for each sample is derived from its associated semivariance

$$A(x_i,y_i) = wt_1 \cdot A(x_1,y_1) + wt_2 \cdot A(x_2,y_2) + wt_3 \cdot A(x_3,y_3) + \ldots$$

as an added bonus this also gives us the variance associated with each interpolated value (the uncertainty), so we can immediately see where our interpolations are reliable and where they are not

because weights are based on the semivariance, obvious trends in the data should be removed as this leads to a continuous rise in the semivariance: can be done by first subtracting a trend surface

Estimate of uncertainty for each interpolated value

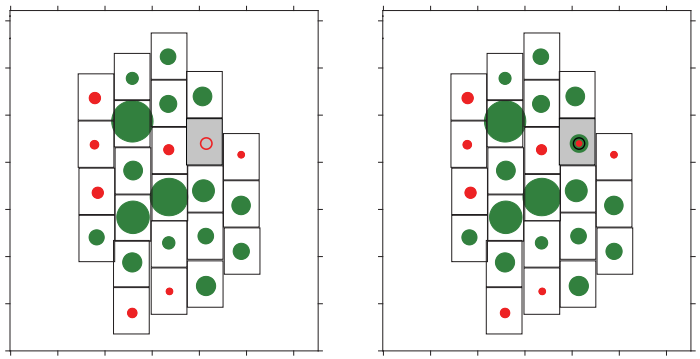


source: wikipedia.org

Uncertainty in block kriging of grades

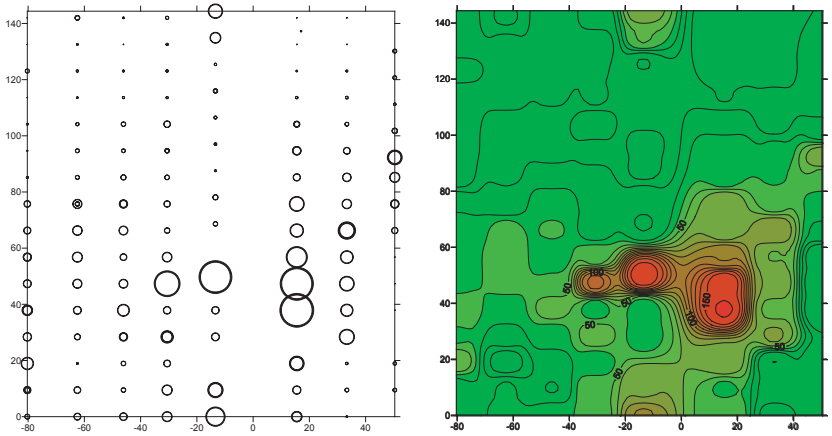
Kriging is commonly applied to estimate the grade of blocks in open pit mining using a sample grid or the grade of adjacent blocks (or both).

In such cases it is invaluable to know the uncertainty on the grade estimate



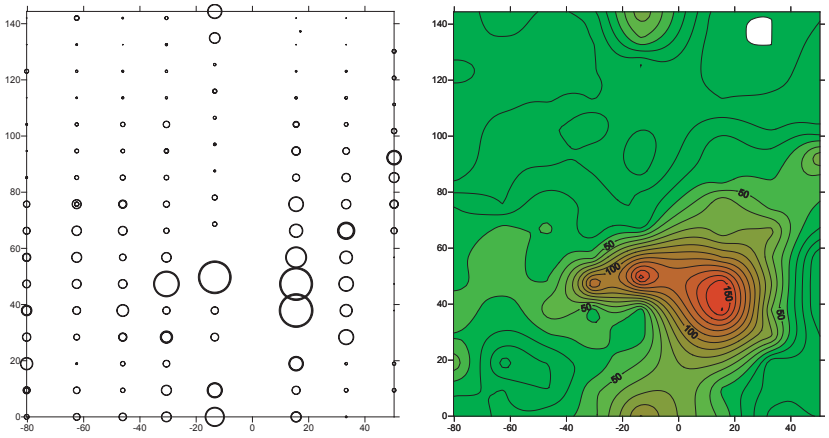
Back to our example

Results of different interpolation techniques:



And now using kriging as the interpolation method

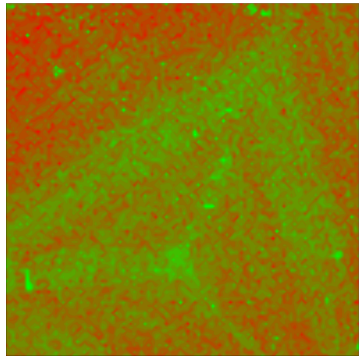
Results of kriging on this data set:



Some data are not suited to interpolation/kriging

There is a strong tendency to directly start with the most complex or fancy technique, such as kriging. However, kriging is not always appropriate !

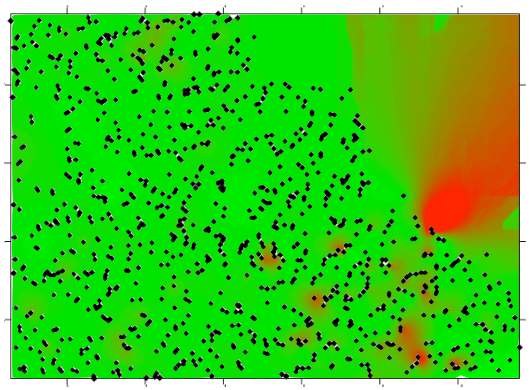
raw concentrations plotted



optimized kriging map

Kriging and sample coverage

Kriging works best when you have a high sample density and a more or less uniform distribution of data over the sample are. If not ➛ get artefacts



Areas without samples need to be blanketed out, not just removed afterwards

Geotop Short Course in Data Analysis and Geostatistics

Short course summary

Eigenvector techniques - highlights

Techniques to locate the principle directions in your dataset

- useful to reduce the dimensionality of a dataset to its principle directions - greatly facilitates the interpretation
- the principle directions generally represent the underlying processes that control the data distribution - process identification

some practicalities:

two main techniques: principle component analysis and factor analysis - very similar, but PCA is a true data transformation (no loss of info) whereas FA retains only a subset of the variance

eigenvector techniques are basically a clustering of the variables based on their correlation/covariance similarity - high cor/cov: same trend, low cor/cov: different trend

have to carefully decide the number of significant factors - use the scree plot. If a more appropriate interpretation can be made using more or less factors than the number suggested by the scree plot - no problem

Spatial analysis - highlights

Spatial analysis of data is a great technique to:

- interact with your data, spot trends, correlations, outliers, clustering, and thereby suggests ways to analyze and interpret your data
- link your data to all kinds of other spatial information, such as position of roads, towns, rivers, ice cover, topography, geology, soil type, vegetation, etc
- disseminate your results to others: easy to understand

some practicalities:

advanced methods need a dedicated sampling design, otherwise stick to the more basic techniques such as bubble plots

when a dense uniform sampling grid is available, best results for Earth Science datasets are generally obtained by using kriging and semivariance

trend surfaces are a further powerful technique to interpret spatial data and de-trending should be performed before kriging

Clustering techniques - highlights

Clustering of data is used to:

- split up multi-modal datasets so they can be analyzed with other statistical techniques, such as t-tests and ANOVA
- look for homogeneous groups in the data, which can tell you something about the main separating processes acting upon the data
- classify samples: assign samples to pre-determined groups

some practicalities:

many varieties of separation techniques: DFA, hierarchical clustering, fixed or sought cluster means, partitioning clustering using hard and fuzzy rules, etc

fuzzy clustering is the most powerful for geochemical datasets as it gives the partial membership to each cluster, thereby being able to cope with intermediate samples

as in eigenvector techniques, the main difficulty is in deciding the number of clusters. A variety of parameters can help you make that decision, but feel free to deviate (e.g. outliers commonly get their own cluster)

Regression analysis - highlights

Regression analysis is a technique:

- that allows you to fit a quantitative model to data that can subsequently be used in mathematical models. Also allows for inter- and extrapolation
- that allows you to determine whether a variable explains a significant part of the variance in the dataset: in other words, whether it belongs in the model
- test what the best model is to describe your data (linear, quadratic, logarithmic, exponential, multiple linear, etc)

some practicalities:

the best regression fit has maximum variance along the regression line and minimal on either side. The ratio of explained over total variance is R^2.

important assumptions in regression analysis that have to be met: always check normality of residuals, multi-collinearity, significance of coefficients, etc

Testing - highlights

Statistical testing:

- test the validity of a hypothesis at a specified confidence interval α
- rejection of the null-hypothesis is the stronger results: choose your hypotheses carefully
- all techniques work in exactly the same way: each test has a probability distribution: exceed the critical probability (α) and the hypothesis is rejected, otherwise: no reason to reject the null hypothesis
- crucial to keep the errors in mind when testing: type I - known, specified as the confidence interval in testing results; type II - unknown
- many statistical tests, optimized for specific hypotheses, data distributions, etc (e.g. t-test, Z-test, F-test, ANOVA, Kolmogorov-Smirnov, χ^2-test)
- most commonly used: t-test/ANOVA - determine whether a number of groups/clusters are significantly different from each other
  χ^2-test - determine whether two data distributions or curves are significantly different

Basic techniques - highlights

data description:

central value: mean, median, mode

measures of spread: range, stdev, IQR, percentile, accuracy vs. precision

normal versus robust techniques

type of distribution: normal, lognormal, multi-modal, outliers

data visualization: histograms, boxplots, scatter diagrams, violin plots, etc

correlation:

correlation between variables expressed by a Pearson or Spearman correlation coefficient. To quickly assess correlations for a complex data matrix: cor matrix

error propagation:

technique to propagate the uncertainty on the measured values to the property you are deriving. Easiest way to do this: split up the equation to its most basic operators: add - subtract - multiply - divide

The end....

If you take nothing else away from this course, remember these:

garbage in = garbage out

most scientists use statistics as the drunkard uses a lamppost; for support rather than illumination