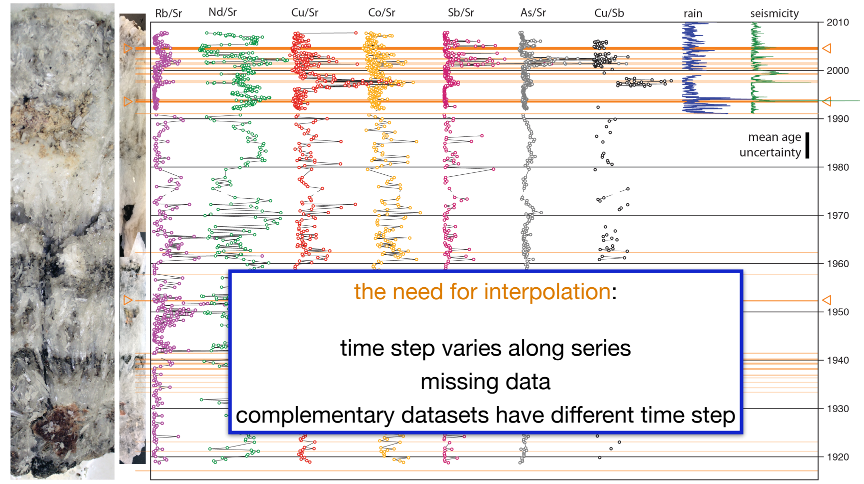
Data analysis and Geostatistics

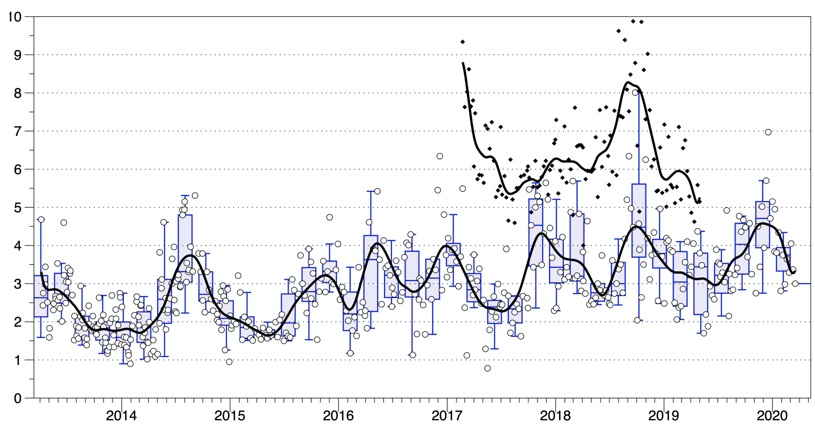
Short Course on the use of statistical techniques in the geosciences

Vincent van Hinsberg • McGill University

Interpolation



Interpolation



Multi-variate techniques

Regression analysis; quantitative description of trends in data - allows for interpolation and extrapolation beyond the input data

Discriminant function analysis; a means to differentiate groups in a data set - used to differentiate and classify

Principal component and factor analysis; determine directions in a data set to reduce the number of variables and/or look for processes in the data

Cluster analysis; group data into homogenous clusters - used to differentiate and to split up multi-modal data sets for use in other stat techniques

Spatial geostatistics; techniques for mining spatially distributed data

Multi-variate techniques: regression

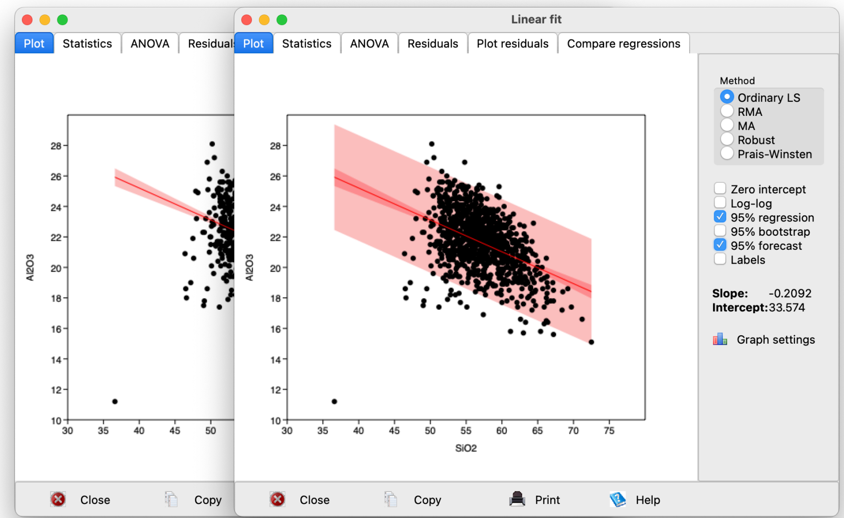
Key aspects of regression analysis

It generates a model of your data; quantitative description of trends in data - allows for interpolation and extrapolation beyond the input data

Strict requirements; normality and no trends or bias in the residuals, no overly influential data points

Significant, meaningful and predictive; need to test that the coefficients and model are significant ($r \neq 0$, $b_i \neq 0$), that the equation chosen is the most appropriate and that the model is predictive (no overfitting)

Multi-variate techniques: regression



Multi-variate techniques: regression

regression analysis versus curve-fitting

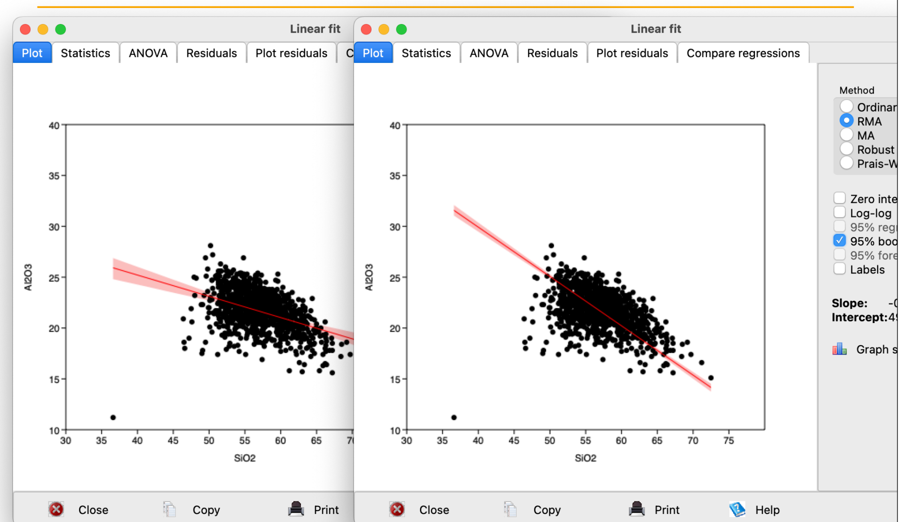
In common use, and in software, these terms have a lot of overlap

the purpose in both cases is to fit a model to data to be used for something

my view:

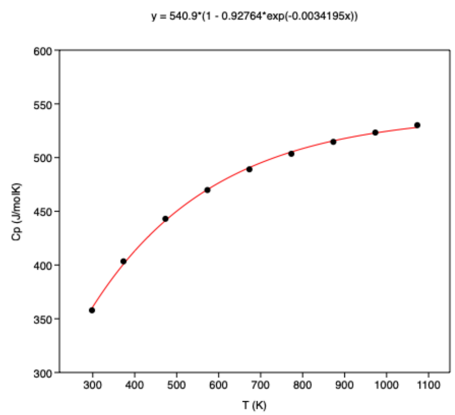
regression:	curve-fitting:
variables not equal	variables can be equal
uncertainty in y	uncertainty in x and y
data define the model	a-priori knowledge of model
generally uses least-squares model search	can use least-squares model search

Multi-variate techniques: regression



Multi-variate techniques: regression

regression analysis versus curve-fitting



Model equation

- Linear (slope and intercept)
- Quadratic (2nd order polynomial)
- Power (allometric equation)
- Exponential (increase or decay)
- von Bertalanffy (growth model)
- Michaelis-Menten
- Logistic (sigmoidal)
- Gompertz (growth model)
- Gaussian (normal distribution)
- Hill's equation (sigmoidal)

Zero constant
95% confidence

a: 540.9
b: 0.92764
c: 0.0034195

Akaike ICc: 35.127 **R^2:** 0.9991

Labels Graph settings

Multi-variate techniques: regression

regression analysis versus curve-fitting

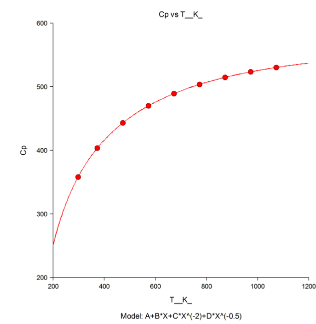
Y Variable: Cp. X Variable: T__K_.
Model Fit: A+B*X+C*X^(-2)+D*X^(-0.5)

Parameter Estimates for All Groups

Groups	Count	Iter's	R2	A	B	C
All	9	29	1.00000	768.98919274225	-0.034861417982246	8-1939102.3069131

Groups	Count	Iter's	R2	D
All	9	29	1.00000	-6542.39510935947

Combined Plot Section

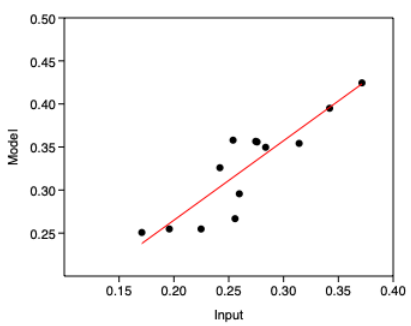


heat capacity equation:

$$C_P = a + b\,T + c\,T^{-2} + d\,T^{-0.5}$$

Multi-variate techniques: regression

regression analysis versus curve-fitting



Ordinary Least Squares Regression: Input-Model

Slope *a*:	0.92301	**Std. error *a*:**	0.14152
***t*:**	6.5221	***p* (slope):**	4.296E-05
Intercept *b*:	0.080364	**Std. error *b*:**	0.038448

95% bootstrapped confidence intervals (*N*=1999):

Slope *a*:	(0.71555, 1.0766)
Intercept *b*:	(0.03107, 0.14159)

Correlation:

***r*:**	0.89137
***r*2:**	0.79454
***t*:**	6.5221
***p* (uncorr.):**	4.296E-05
Permutation *p*:	0.0001

Ordinary Least Squares Regression: Input-Model

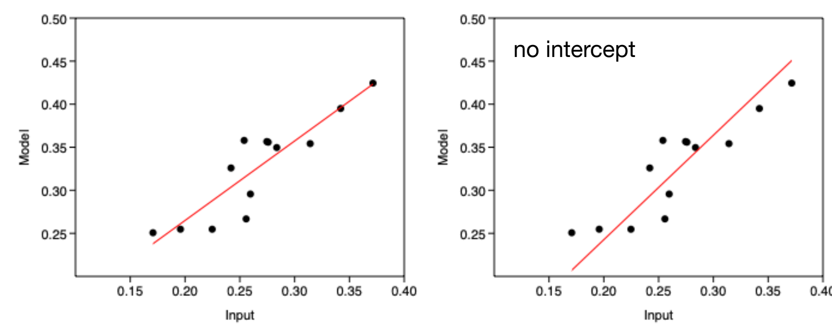
	SS	df	MS	F	p
Regression	0.030908	1	0.030908	42.538	4.296E-05
Residual	0.0079926	11	0.0007266		
Total SS	0.038901				

multivariate regression model:

$$V_{Tur} = X_{Uv}V_{Uv} + X_{Drv}V_{Drv} + X_{Sch}V_{Sch} + \ldots$$

Multi-variate techniques: regression

regression analysis versus curve-fitting

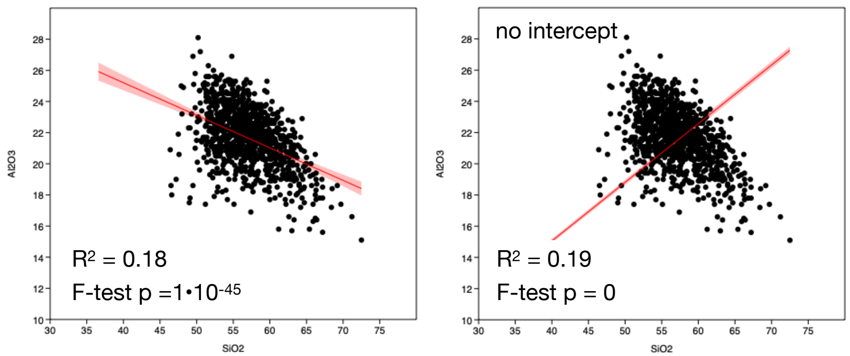


multivariate regression model:

$$V_{Tur} = X_{Uv}V_{Uv} + X_{Drv}V_{Drv} + X_{Sch}V_{Sch} + \ldots$$

Multi-variate techniques: regression

regression analysis versus curve-fitting



Geotop Short Course in Data Analysis and Geostatistics

Sample classification: DFA and clustering

geotop

Multi-variate techniques

Have now finished data description and statistical testing

will now move to more advanced (multi-variate) techniques:

Regression analysis; quantitative description of trends in data - allows for interpolation and extrapolation beyond the input data

Discriminant function analysis; a means to differentiate groups in a data set - used to differentiate and classify

Principal component and factor analysis; determine directions in a data set to reduce the number of variables and/or look for processes in the data

Cluster analysis; group data into homogenous clusters - used to differentiate and to split up multi-modal data sets for use in other stat techniques

Spatial geostatistics; techniques for mining spatially distributed data

Separation and classification of data

Two main statistical techniques used to separate and classify:

Discriminant function analysis - DFA

Cluster analysis

Goals of these techniques:

- to separate
  majority of statistical techniques cannot be applied to multi-modal data sets: have to split them into homogenous groups.

- to classify
  to what group should a sample be assigned. Examples: soil classification, rock classification, etc. Use the combination of a variety of characteristics to link unknowns to specific (pre-defined) groups.

Separation and classification of data

The two techniques have a somewhat different focus:

Discriminant function analysis:
find a function/vector that best separates the groups in your data set

Cluster analysis:
group samples into clusters based on their similarity

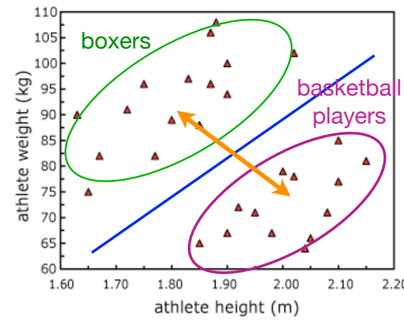
both techniques allow you to quantify the degree of membership to each cluster

Discriminant function analysis

Examples of discriminant function analysis

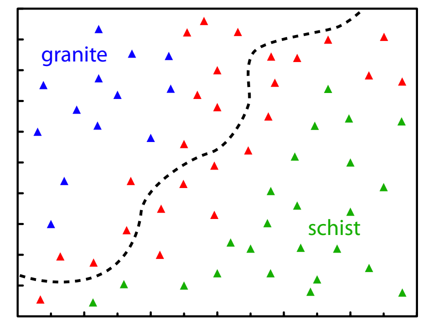
2D case: difference between athletes

can directly visualize the DF



multi-D case: boundary mapping

DF combines multitude of characteristics that are then plotted in space

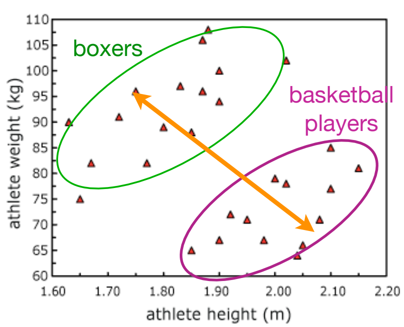


Discriminant function analysis

How do we determine a discriminant function ?

Need a training set that defines the groups: data with known grouping

e.g. a characteristic group of boxers and basketball players



Next: search within this training set for the vector that leads to optimal separation

This function can then be used to classify unknowns

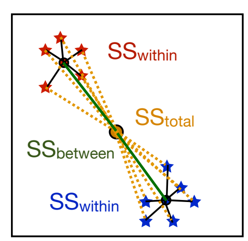
Discriminant function analysis

How do we determine a discriminant function ?

The vector of maximum separation can be obtained by sum of squares methodology

so let's have another look at the sum of squares:

$SS = \sum (x_i - \bar{x})^2$ the cumulative deviation from a mean



SS_{within}: the cumulative deviation of the data from their respective group's mean - within variance

SS_{total}: the cumulative deviation of the data from the overall data mean - total variance

$SS_{between}$: the cumulative deviation of the group means from the overall mean - between variance

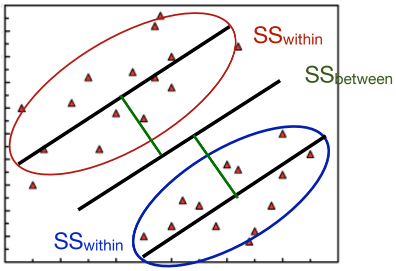
Discriminant function analysis

a good DF is a function where $SS_{between} >> SS_{within}$

Find the best DF by optimizing the function for maximum $SS_{between} / SS_{within}$

$DF = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + ...$

fitting of the b - coefficients is generally done by iteration and is thus best performed by a computer program.



when data strongly correlated: the mean not the best descriptor when calculating the cum. dev.

Instead: use the cumulative deviation from the covariance trend: the mean vector

to work: correlations within groups have to be similar between groups

Discriminant function analysis

Not all variables in the DF are necessarily significant

Have to check if each variable adds something to the separating power of the equation - if not: remove the variable from the DF

$DF = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + ...$

How to check for significance:

include everything and test the significance using F and tolerance tests, then rerun with subset of significant variables

F-tests: does my fit significantly improve by including this variable ?
tolerance: is this var's separation already covered by another var ?

include variables stepwise and determine how the fit (correct assignment of training set) improves as you add variables

Both are affected by the order of inclusion/exclusion of variables .....

Discriminant function analysis

requirements for discriminant function analysis:

data must be derived from multi-variate normal distributions

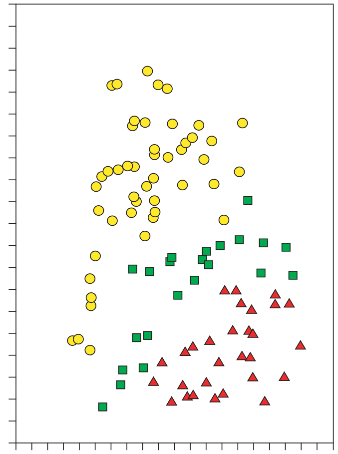
covariance matrices should be same for each group
(the mean vectors should be parallel)

if not:

can still apply discriminant function analysis, but the resulting functions will not be linear, and significance and goodness-of-fit are much more difficult to assess

Discriminant function analysis

DFA to determine the location of a geological boundary



the contact between a granite and a schist:

two sets for training and a set of unknowns

26 major and trace elements have been determined on river sediments in this area

river sediment compositions are a mixture of the drainage area, so boundaries are diffuse

use these to derive a discriminating function with which to assign the unknowns and thereby pinpoint the location of the boundary

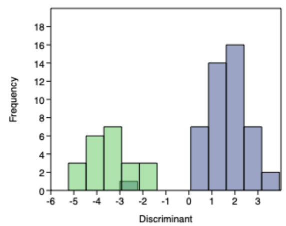
Discriminant function analysis with PAST

Untitled

Show: Row attributes · Column attributes | Click mode: Select · Drag rows/columns | Edit: Cut · Copy · Paste · Select all | View: Bands · Binary · Recover windows · Decimals: -

	geolcode	SiO2	Al2O3	Fe2O3					
45	7	60.4	20.9	6.7					
46	7	52.7	22.7	7.9					
47	7	52.8	23.2	8.4					
48	7	46.4	20.9	11.7					
49	?	58.8	22.6	5.9					
50	?	62.3	21.6	3.1					
51	?	62	22.2	5.3					
52	?	61.9	21.6	4					
53	?	63.1	22.2	3.1					
54	?	60.6	22.6	3.3					
55	?	61.8	21.7	2.8					
56	?	60	21.8	3.3					
57	?	62.2	20.8	2.7					
58	?	64.2	21.7	2.3					
59	?	63.4	21.7	4.1					
60	?	57.7	23.9	5.7					
61	?	61.6	22.6	4.9					
62	?	57.4	19.8	4.7	1.3	1.3	4.6	0.1	0.78
63	?	60.3	24	5.3	1.3	2.1	5.6	0.1	1.05
64	?	56.7	22.4	5.1	1.5	1.8	4.9	0.2	1.03
65	?	57.9	24.5	5.8	1.2	2.5	4.5	0.1	1.29
66	?	56.6	20.6	3.5	1.9	1.6	7.1	0.1	0.83
67	?	58.2	23	5.1	1.3	2.3	5.7	0.1	1.21

Plot | Scores | Loadings | Classifier | Confusion matrix



X axis: Axis 1

Bins: 12

Group labels

Axis	Eigval	Percent
1	6.1144	100

Graph settings

Discriminant function analysis with PAST

Plot | Scores | Loadings

	Axis 1
2	1.7346
3	1.7325
4	-2.2507
5	0.58595
6	0.90136
7	1.35
8	1.1304
9	2.5792
10	1.3678
11	2.3333
12	2.5206
13	1.3713
14	1.3343
15	2.1572
16	1.6589
17	1.14

Plot | Scores | Loadings

	Axis 1
SiO2	0.074055
Al2O3	0.17649
Fe2O3	1.0443
CaO	-0.57169
MgO	1.3128
K2O	-0.027256
MnO	-5.3349
TiO2	-0.0263
P2O5	0.00055494
Li	0.0027225
Be	0.0028415
B	-0.0090147
V	0.0059463
Cr	-0.071659
Co	0.061826
Ni	0.03194
Cu	-0.033276
Zn	-8.8537E-05
As	-0.0052967
Sr	0.017747
Y	-0.15769
Nb	-0.033292
Mo	0.30053

Plot | Scores | Loadings | Classifier | Confusion matrix

	7	10	Total
7	46	1	47
10	0	22	22
Total	46	23	69

Rows: Given groups
Columns: Predicted grps

Jackknifed

% correctly classified: 98.55

Discriminant function analysis with PAST

Plot | Scores | Loadings | Classifier | Confusion matrix

Point	Given group	Classification	Jackknifed
46	7	7	7
47	7	7	7
48	7	7	7
49	?	10	
50	?	10	
51	?	10	
52	?	10	
53	?	10	
54	?	10	
55	?	10	
56	?	10	
57	?	10	
58	?	10	
59	?	10	
60	?	10	
61	?	10	
62	?	10	
63	?	7	
64	?	10	
65	?	10	

Discriminant function analysis

Discriminant function analysis with NCSS: check the tutorial

Variable Influence Section

Variable	Removed Lambda	Removed F-Value	Removed F-Prob	Alone Lambda	Alone F-Value	Alone F-Prob	R-Squared Other X's
SiO2	0.929825	3.40	0.071940	0.532149	65.06	0.000000	0.897321
Al2O3	0.957967	1.97	0.166838	0.918458	6.57	0.012403	0.841157
Fe2O3	0.997191	0.13	0.723460	0.469680	83.55	0.000000	0.977096
CaO	0.866789	6.92	0.011653	0.996300	0.27	0.601687	0.969604
MgO	0.913752	4.25	0.045116	0.959101	3.16	0.079778	0.960142
K2O	0.764824	13.84	0.000551	0.761159	23.22	0.000007	0.874850
MnO	0.996012	0.18	0.673246	0.912984	7.05	0.009686	0.843421
TiO2	0.812242	10.40	0.002347	0.615201	46.29	0.000000	0.959324
P2O5	0.892082	5.44	0.024167	0.708030	30.52	0.000000	0.785085
Li	0.919225	3.95	0.052854	0.227281	251.59	0.000000	0.946061
Be	0.966909	1.54	0.221043	0.241559	232.34	0.000000	0.962685
B	0.972051	1.29	0.261353	0.389654	115.91	0.000000	0.888401
V	0.958531	1.95	0.169777	0.533296	64.76	0.000000	0.968710
Cr	0.910433	4.43	0.040994	0.985079	1.12	0.293171	0.974521
Co	0.973536	1.22	0.274603	0.630067	43.45	0.000000	0.960442
Ni	0.975333	1.14	0.291743	0.781352	20.71	0.000021	0.975646
Cu	0.960009	1.87	0.177750	0.412601	105.35	0.000000	0.899286
Zn	0.996957	0.14	0.712653	0.940833	4.65	0.034234	0.826045
As	0.978912	0.97	0.330094	0.938756	4.83	0.031136	0.761502
Sr	0.975353	1.14	0.291945	0.991157	0.66	0.419100	0.934216
Y	0.996994	0.14	0.714331	0.928123	5.73	0.019204	0.915249
Nb	0.830807	9.16	0.004074	0.999987	0.00	0.975186	0.855943
Mo	0.997030	0.13	0.715968	0.625719	44.26	0.000000	0.781241
Sn	0.933963	3.18	0.081209	0.199167	297.55	0.000000	0.960175
Sb	0.986739	0.60	0.440835	0.956418	3.37	0.070328	0.633268
Ba	0.997823	0.10	0.755466	0.565783	56.79	0.000000	0.739602
La	0.970122	1.39	0.245282	0.961321	2.98	0.088608	0.988756
Ce	0.986945	0.60	0.444426	0.965349	2.66	0.107397	0.989428
Pb	0.986018	0.64	0.428595	0.612586	46.80	0.000000	0.738825
Zr	0.988784	0.51	0.478639	0.769770	22.13	0.000012	0.818772

tutorial tells you what all input and output means + requirements

check for

significance of the variables with F-tests:

- removed F-prob should be < α
- alone F-prob should be < α

check for tolerance issues with R^2 :

if $1-R^2$ is low, the var doesn't add diff

Discriminant function analysis

Variable Influence Section

Variable	Removed Lambda	Removed F-Value	Removed F-Prob	Alone Lambda	Alone F-Value	Alone F-Prob	R-Squared Other X's
K2O	0.743857	24.10	0.000006	0.761159	23.22	0.000007	0.355752
P2O5	0.894571	8.25	0.005390	0.708030	30.52	0.000000	0.347597
B	0.715202	27.87	0.000001	0.389654	115.91	0.000000	0.669410
Cu	0.900233	7.76	0.006876	0.412601	105.35	0.000000	0.642587
Sn	0.588693	48.91	0.000000	0.199167	297.55	0.000000	0.669063

Linear Discriminant Functions

	geolcode	
Variable	7	9
Constant	-35.90635	-42.90045
K2O	8.120147	3.320065
P2O5	2.697954E-03	6.058516E-03
B	-3.353538E-02	0.1448863
Cu	0.3907292	0.1507453
Sn	0.1338246	1.0996

Classification Count Table for geolcode

	Predicted		
Actual	7	9	Total
7	46	1	47
9	0	29	29
Total	46	30	76

Reduction in classification error due to X's = 97.4%
Reduction in classification error due to X's = K2O, P2O5, B, Cu, Sn

Canonical Variate Analysis Section

Fn	Inv(W)B Eigenvalue	Ind'l Pcnt	Total Pcnt	Canon Corr	Canon Corr2	F-Value	Numer DF	Denom DF	Prob Level	Wilks' Lambda
1	10.974471	100.0	100.0	0.9573	0.9165	153.6	5.0	70.0	0.0000	0.083511

all variables are now significant in the DF

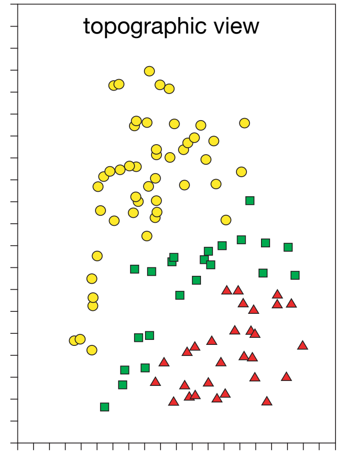
no tolerance issues

NCSS gives you two discriminant vectors based on these vars

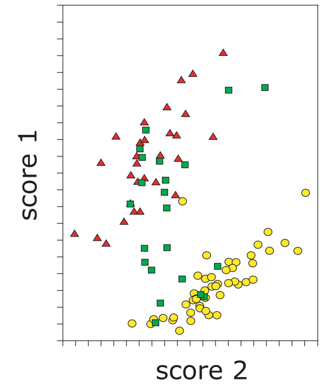
only one sample is assigned incorrectly:
outlier

DF is highly significant

Discriminant function analysis

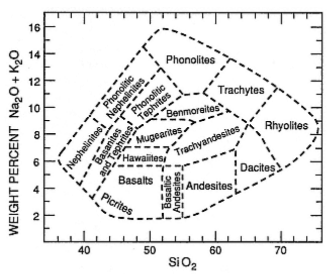


schist - granite - unknown

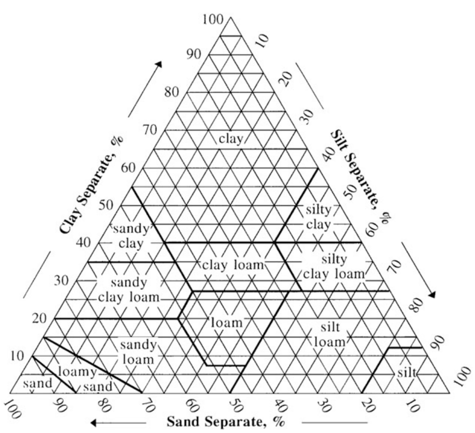


score 1 and 2 are two combinations of the 5 vars that lead to maximum separation of the groups - two vectors in multi-D space

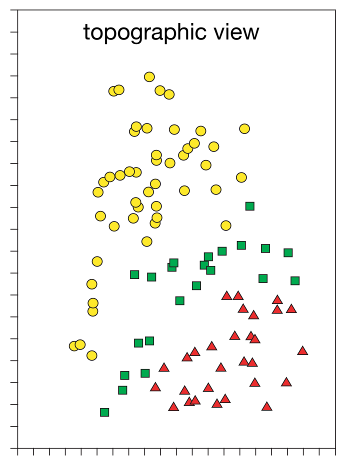
Why separating vectors instead of boundaries ?



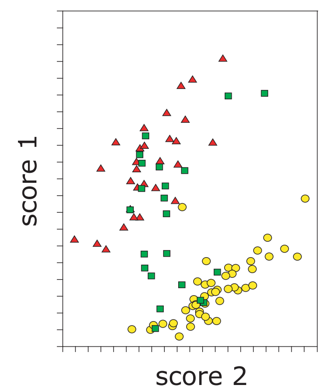
boundaries are rules in the space defined by the separating vectors



Discriminant function analysis



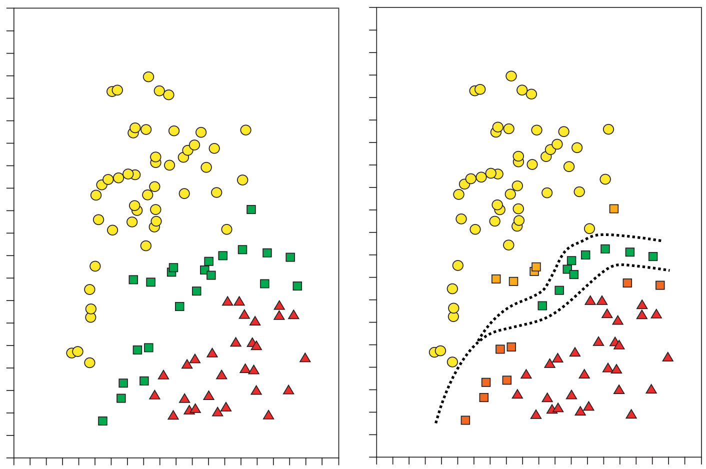
schist - granite - unknown



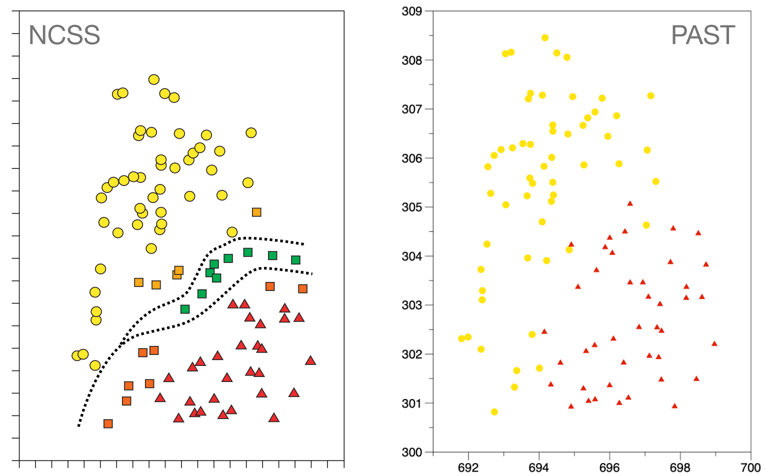
score 1 and 2 are two combinations of the 5 vars that lead to maximum separation of the groups - two vectors in multi-D space

Discriminant function analysis

Use the vectors to assign the unknowns - not all fit with these groups



Comparison of LDA results in NCSS and PAST



Other discriminating approaches

Given how important classification is, there are many more techniques that have been devised for this;

QDA - quadratic discriminant function

PCA-LDA - discriminant analysis on transformed coordinate axes (principal components)

PLS-DA - discriminant analysis on transformed coordinate axes with axis directions optimized for discrimination

mapping (hypercube logic, random forest, etc) - mapping “routes” in multivariate space to the desired outcome

Cluster analysis

Group samples into clusters based on similarity

Cluster analysis requires substantial user input

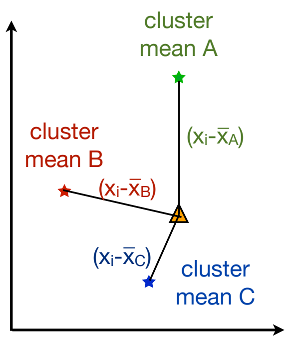
(selection of number of clusters, clustering routine, similarity criteria, etc)

and results can therefore be ambiguous:

always give detailed information on how your cluster analysis was performed

Cluster analysis

Group samples into clusters based on similarity



whichever deviation between sample and cluster mean is smallest:

assigned to that cluster

Cluster analysis is again controlled by the sum of squares:

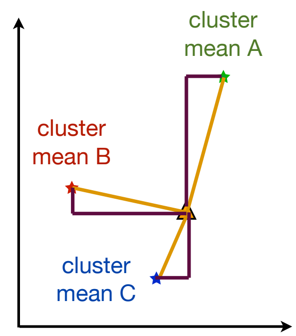
SS_{within}	$SS_{between\ A\text{-}B}$
SS_{within}	$SS_{between\ B\text{-}C}$
SS_{within}	$SS_{between\ A\text{-}C}$

small variance within: tight clusters
large variance between: good separation

increasing the number of clusters will decrease the within variance, until all samples are their own cluster. That result is however meaningless....

Cluster analysis - sample assignment criteria

range of techniques that can be used to determine similarity



Wide range of techniques - see book for details

- Euclidian distance - r or r²
- city block of Manhattan distance - this is useful when the two variables are separate characteristics (fossil length and width, the diagonal is not of interest)
- correlation similarity - sample with the same correlation are grouped together: deals with dilution effects
- association values - especially useful when you have only presence/absence data - specialized

Cluster analysis - two types

Ward's method: groups are linked to minimize within variance

UPGMA: linked based on average dissimilarity of each group

Two varieties of clustering: hierarchical and partitioning methods

hierarchical techniques: represent similarity in a tree or dendrogram

the method:

1. all samples are a separate cluster
2. link the two most similar samples
3. link two other samples to form a new cluster or add a third sample to the first cluster depending on similarities
4. continue until only one cluster remains

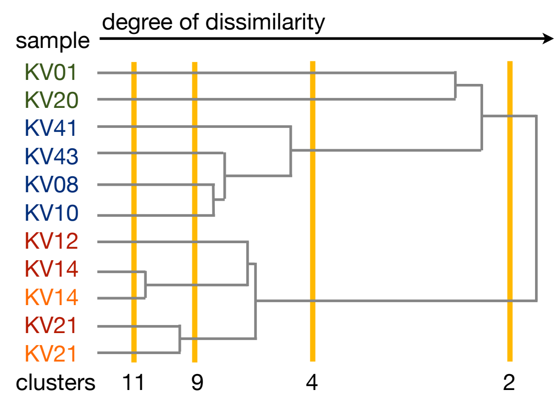
in this technique all intermediate steps and cluster associations are immediately available - depends on the user to select an appropriate "pruning" level in the tree

there are many ways to link samples and these do result in different trees (see book for details)

Hierarchical cluster analysis

An example of hierarchical clustering:

the composition of a number of lava samples from Kawah Ijen volcano:



dissimilarity based upon nearest neighbour criterium

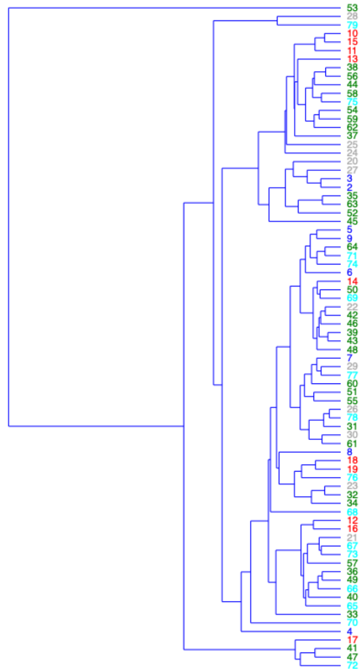
the resulting tree can be "pruned" at any level: up to the user to select

should test if difference between groups is significant (which test?)

basalt dacite
andesite duplicate

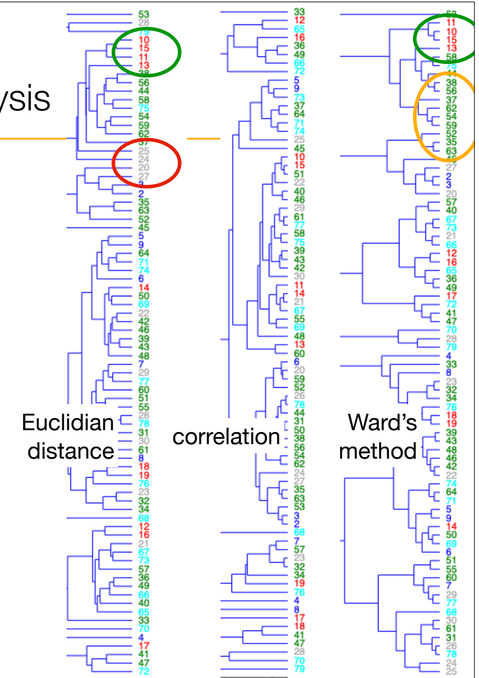
Hierarchical cluster analysis

Compositional data (major elements and trace elements) with colour coding for geological unit

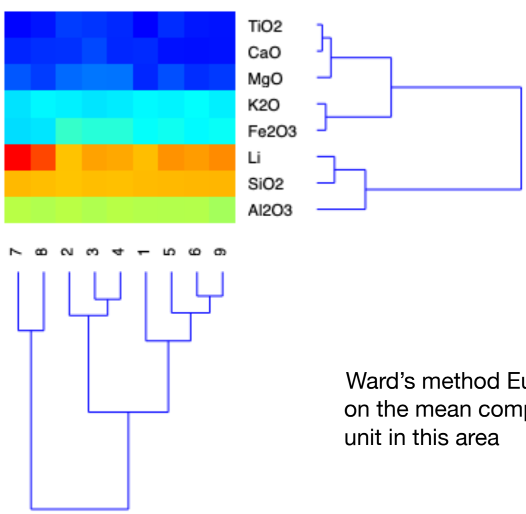


Hierarchical cluster analysis

Compositional data (major elements and trace elements) with colour coding for geological unit



Hierarchical cluster analysis: two-way



Ward's method Euclidian distance clustering on the mean composition of each geological unit in this area

Clustering - partitioning techniques

Two varieties of clustering: hierarchical and partitioning methods

partitioning techniques: assigns samples to a known number of clusters based upon similarity criteria

the method:

1. samples are assigned to the cluster they are most similar to in multi-dimensional space
2. each assignment results in a shift in the characteristics of the cluster centre (means + variance or only variance)
3. samples are re-assigned where necessary and this routine is iterated until the system stabilizes

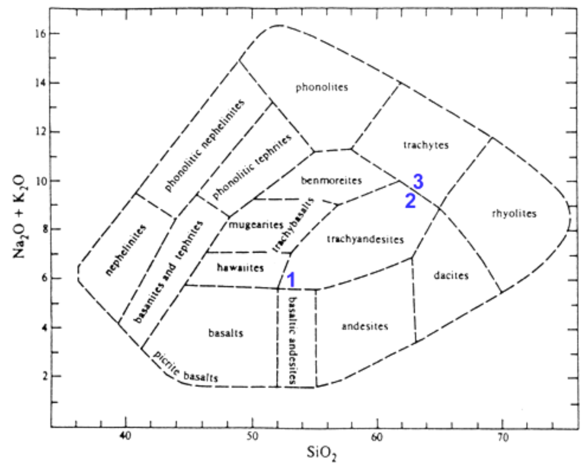
There are two main approaches:
clustering with specified cluster means (i.e. known groups) and
clustering where the means are obtained during clustering

both have their pros and cons:

Partitioning techniques

	advantages	disadvantages
specified/ fixed	‣ you always get the same answer during classification ‣ groups can relate to real dividing phenomena ‣ unknowns are (generally) easily classified	‣ boundaries commonly based on consensus (artificial) ‣ 2 samples close together can be in different clusters ‣ 2 very different samples can be in same cluster
assigned/ sought	‣ data groups not split up over different clusters ‣ boundaries always in regions of low data density ‣ easy to apply to data sets with many variables	‣ instability issues: more data will result in shift in cluster means and sample assignment ‣ no fixed boundaries so unsuitable for classification schemes

Clustering with hard boundaries



2 samples close together can be in different clusters

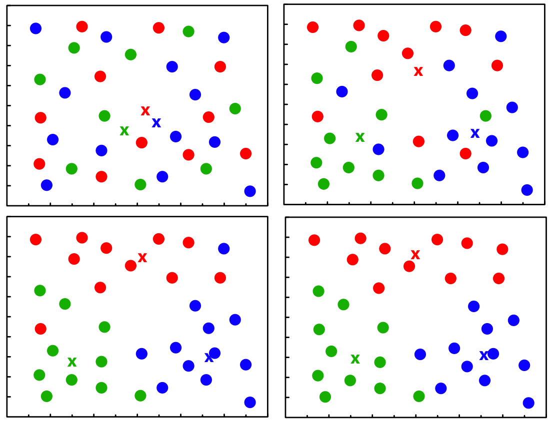
2 very different samples can be in same cluster

Partitioning techniques

	advantages	disadvantages
specified/ fixed	‣ you always get the same answer during classification ‣ groups can relate to real dividing phenomena ‣ unknowns are (generally) easily classified	‣ boundaries commonly based on consensus (artificial) ‣ 2 samples close together can be in different clusters ‣ 2 very different samples can be in same cluster
assigned/ sought	‣ data groups not split up over different clusters ‣ boundaries always in regions of low data density ‣ easy to apply to data sets with many variables	‣ instability issues: more data will result in shift in cluster means and sample assignment ‣ no fixed boundaries so unsuitable for classification schemes

Cluster means assigned during clustering:

when cluster means are specified: use minimum distance to mean to assign
if not: randomly assign each sample to a cluster and iterate to stable solution



both cluster means and cluster assignment change during the iteration

process stops when samples no longer change their assignment

● cluster A
● cluster B
● cluster C
x center

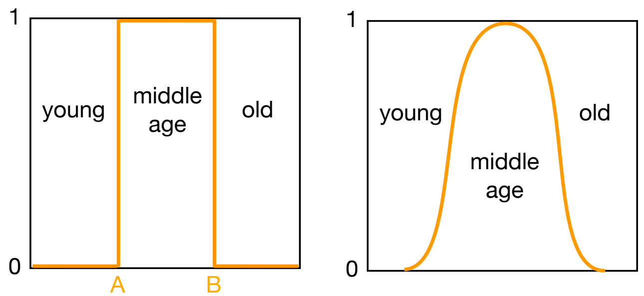
Cluster analysis - method of assignment

Samples are normally assigned to a cluster in a "hard" way:

samples are unambiguously attributed to a specific cluster - 0 or 1 assignment

However, mother nature is rarely so black and white....

"middle age" cluster depends very much on percon/country/continent



if age is between A and B: middle age

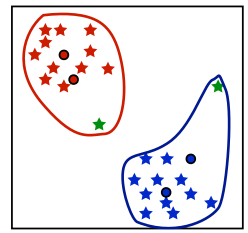
fuzzy approach: samples have cluster memberships between 0 and 1

Fuzzy clustering

fuzzy clustering has a number of distinct benefits:

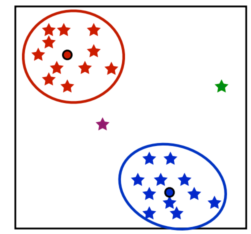
can deal with intermediate cases - not force-assigned
samples have share multiple clusters - extra information:
(0.7 young + 0.3 middle age versus 0.5 young + 0.5 middle age)
ensures that single samples do not overly control individual clusters
can have a separate outlier assignment

most flexible and powerful: fuzzy clustering with seeking of cluster means



hard clustering

strained assignment due to outlier and intermediate value



fuzzy clustering

outlier not a problem and intermediate shown

Clustering in NCSS - the eating habits of Europe

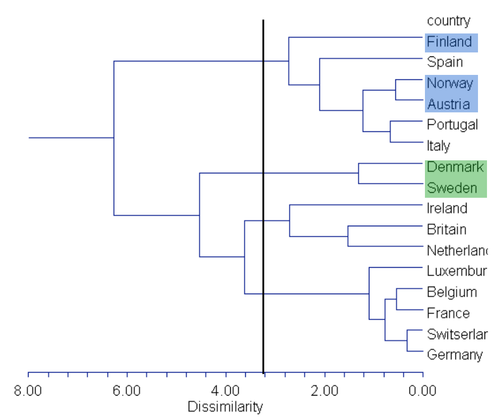
can we distinguish the Europeans by their eating habits?

the data (missing value = -999):

country	Coffee	Nescaf	Tea	Sweete	Biscuit	Pack_s	Tin_so	Frozen	Frozen	Fresh_	Tin_fru	Jam	Garlic	Butter	Marger	Olive_	Yoghur
Germany	90	49	88	19	57	51	19	27	21	81	44	71	22	91	85	74	30
Italy	82	10	60	2	55	41	3	4	2	67	9	46	80	66	24	94	5
France	88	42	63	4	76	53	11	11	5	87	40	45	88	94	47	36	57
Netherlands	96	62	98	32	62	67	43	14	14	83	61	81	15	31	97	13	53
Belgium	94	38	48	11	74	37	25	13	12	76	42	57	29	84	80	83	20
Luxemburg	97	61	86	28	79	73	12	26	23	85	83	20	91	94	94	84	31
Britain	27	86	99	22	91	55	76	20	24	76	89	91	11	95	94	57	11
Portugal	72	26	77	2	22	34	1	20	3	22	8	16	89	65	78	92	6
Austria	55	31	61	15	29	33	1	15	11	49	14	41	51	51	72	28	13
Switserland	73	72	85	25	31	69	10	19	15	79	46	61	64	82	48	61	48
Sweden	97	13	93	31	-999	43	43	54	45	56	53	75	9	68	32	48	2
Denmark	96	17	92	35	66	32	17	51	42	81	50	64	11	92	91	30	11
Norway	92	17	83	13	62	51	4	30	15	61	34	51	11	63	94	28	2
Finland	98	12	84	20	64	27	10	18	12	50	22	37	15	96	51	17	-999
Spain	70	40	40	-999	62	43	2	23	7	59	30	38	86	44	25	91	16
Ireland	13	52	99	11	80	75	18	5	3	57	46	89	5	97	-999	31	3
	Real coffee	Nescafe	Tea	Sweetener	Biscuits	Pack. soup	Tinned soup	Frozen fish	Frozen veg.	Apples	Tinned fruit	Jam	Garlic	Butter	Margerine	Olive oil	Yoghurt

Clustering in NCSS - the eating habits of Europe

hierarchical clustering of this data set: clear clustering



lots of options available:

use parametric and non-parametric data and even mix these (length + color)

variety of linkage types: nearest neighbour, furthest neighbour, Ward's method

distance: Euclidian or Manhattan city block

see the NCSS hierarchical clustering tutorial for more information

Clustering in NCSS - the eating habits of Europe

hard and fuzzy clustering of this data set:

	K-means - hard	fuzzy-prob 1	fuzzy-prob 2	fuzzy-prob 3	fuzzy-prob 4
Germany	2	0.04	0.02	0.83	0.11
Italy	3	0.01	0.93	0.04	0.02
France	2	0.05	0.12	0.77	0.06
Netherlands	2	0.23	0.07	0.53	0.16
Belgium	2	0.08	0.25	0.56	0.11
Luxembourg	2	0.09	0.06	0.75	0.1
Britain	4	0.92	0.01	0.04	0.03
Portugal	3	0.02	0.92	0.04	0.03
Austria	3	0.03	0.87	0.06	0.05
Switzerland	2	0.05	0.05	0.86	0.05
Sweden	1	0.05	0.04	0.08	0.82
Denmark	1	0.03	0.02	0.07	0.88
Norway	1	0.03	0.07	0.1	0.8
Finland	1	0.06	0.22	0.16	0.57
Spain	3	0.02	0.83	0.11	0.04
Ireland	4	0.88	0.04	0.06	0.03

Plotting clusters on maps - Massif Central dataset

Will look at an example from the Massif Central in France. A dataset of the chemical composition of stream sediments collected in an area with a diverse geology and old, now abandoned, mining for Sb, As, Pb, Au, Ba & F

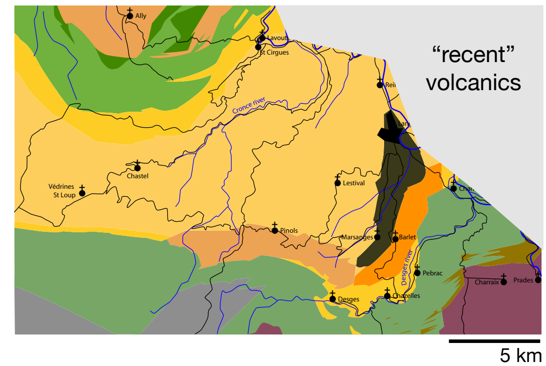
Geology consists of:

felsic gneisses

mafic gneisses and schists

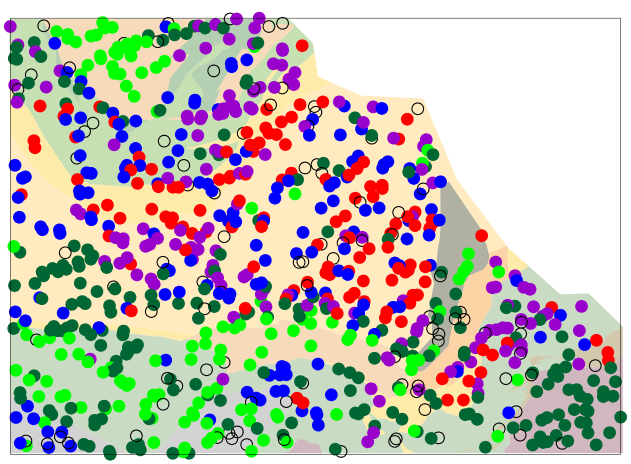
(meta) - granite

sediment (incl coal)



Clustering - groups in Massif Central dataset

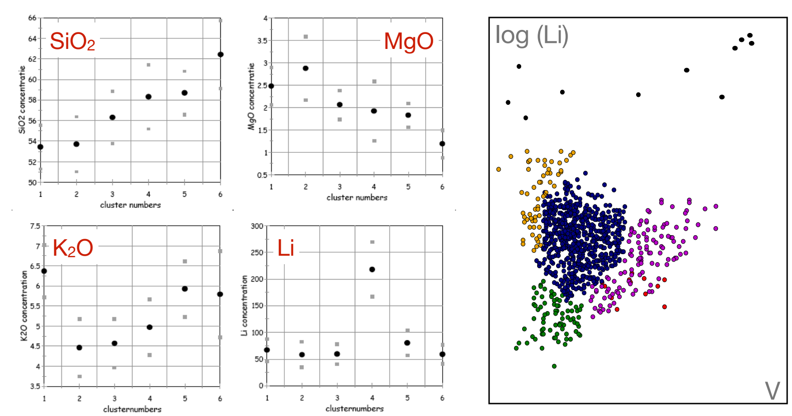
The dataset is best described when split up into six clusters



clear link to the bedrock geology, but not 1 to 1

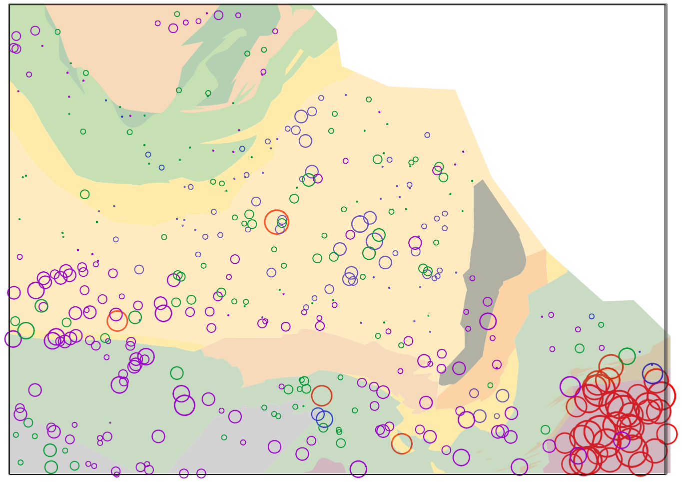
Clustering - properties per cluster

when the data have been clustered: can look at the characteristics of each cluster (mean + stdev) and correlations within this



Clustering - groups in Massif Central dataset

Can plot clusters individually to look at spatial distribution and contents



cluster 1
cluster 2
cluster 3
cluster 4

cluster separation isn't perfect

only cluster 4 is distinct in Li: multi-element separation

Clustering - number of clusters

the main difficulty in cluster analysis is choosing the no. of clusters

NCSS, PAST and other clustering packages will calculate assignments for a cluster number range

the residual variance will decrease with every additional cluster so this is not a good indicator of optimal no. of clusters

instead:
choose no. of clusters where variance no longer strongly decreases

use the averaged silhouette value: comparison between a value's dissimilarity with its cluster and the dissimilarity with its nearest neighbour: ranges from 1 to -1: > 0.75: good model < 0.25: poor model

Use the fuzziness of the model (0; completely fuzzy to 1; hard) Fc(U) and Dc(U) parameters: max Fc(U) + min Dc(U) = best model

DFA and cluster analysis summarized

why:

need data to be in homogenous groups
group and classify as a data analysis tool

how:

discriminant function analysis
derive separating vectors from training set

cluster analysis
fixed/specified cluster means/medians or obtained in clustering
hierarchical, hard or fuzzy

requires:

lots of normally distributed variables