University of Glasgow | Department of Chemistry

Visualisation of Unknown Dataset

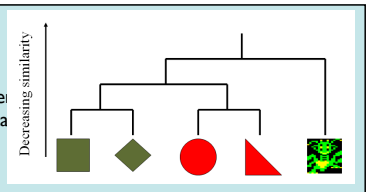
DataMix Workshop, August 2009

Gordon Barr

'Large' numeric dataset
269 countries
Each country has entries for each of 100 years 1950 – 2050
Countries are the 'samples'
Each sample has 100 year variables associated with it.

	A	B	C	D	E	F	G	H	I	J
1		1950	1951	1952	1953	1954	1955	1956	1957	1958
2	Afghanista	8151	8271	8396	8526	8662	8805	8954	9109	9271
3	Albania	1215	1240	1269	1303	1339	1379	1421	1466	1513
4	Algeria	8753	8953	9141	9326	9516	9715	9924	10142	10365
5	American_	19	19	19	20	20	20	20	20	20
6	Andorra	6	7	7	8	9	9	10	11	12
7	Angola	4148	4212	4282	4360	4443	4531	4623	4718	4816
8	Anguilla	5	5	5	5	5	5	6	6	6
9	Antigua_ar	46	48	50	51	52	53	53	53	54
10	Argentina	17150	17518	17887	18254	18616	18974	19325	19670	20011
11	Armenia	1354	1383	1420	1463	1511	1564	1619	1678	1739

Cluster Analysis - Methods for grouping objects of similar kind into separate categories. Numerous differe algorithms; Cluster analysis discovers structures in data without explaining why they exist



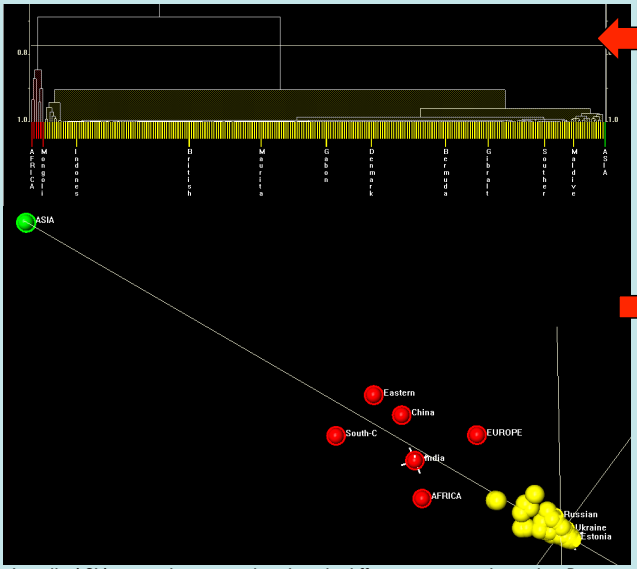
Calculate the 'distance' between each sample – larger distances = more different, small distances = similar
Get a Distance Matrix **d** & correlation matrix **p**
From **d** & **p** can do all sorts of fun statistical analyses
-Cluster Analysis
-PCA
-MMDS 3D Plots
-Minimum Spanning Trees
-Parallel Coord Plots

$$d_{ij} = \left[\sum_{k=1}^{n} w_k \left(x_{ik} - x_{jk} \right)^{\lambda} \right]^{\frac{1}{\lambda}}$$

$\Rightarrow \mathbf{d}\boldsymbol{\rho}\Rightarrow$

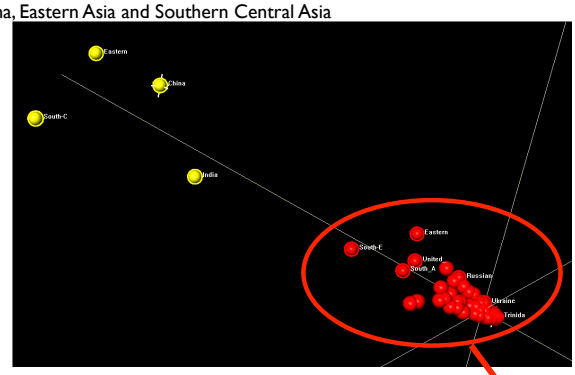
	Albania	American_Sa	Andorra	Anguilla	Antigua_and_	Armenia	Aruba	Bahamas	Bahr
Albania	1.000	0.640	0.641	0.634	0.643	0.962	0.643	0.668	0.70
American_Samoa	0.640	1.000	0.998	0.994	0.996	0.622	0.996	0.971	0.52
Andorra	0.641	0.998	1.000	0.993	0.996	0.623	0.996	0.972	0.52
Anguilla	0.634	0.994	0.993	1.000	0.991	0.617	0.991	0.965	0.52
Antigua_and_Ba	0.643	0.996	0.996	0.991	1.000	0.626	0.998	0.974	0.52
Armenia	0.962	0.622	0.623	0.617	0.626	1.000	0.626	0.659	0.68
Aruba	0.643	0.996	0.996	0.991	0.998	0.626	1.000	0.974	0.52
Bahamas	0.668	0.971	0.972	0.965	0.974	0.650	0.974	1.000	0.95
Bahrain	0.706	0.926	0.927	0.920	0.928	0.685	0.928	0.953	1.00
Barbados	0.666	0.972	0.972	0.967	0.976	0.649	0.976	0.985	0.94
Belize	0.663	0.974	0.976	0.969	0.977	0.645	0.977	0.994	0.95
Bermuda	0.640	0.997	0.996	0.994	0.997	0.623	0.997	0.971	0.52
Bhutan	0.704	0.933	0.934	0.927	0.936	0.685	0.936	0.962	0.98
Bosnia_and_Her	0.638	0.510	0.511	0.505	0.514	0.867	0.514	0.536	0.56
Botswana	0.814	0.808	0.810	0.803	0.811	0.797	0.811	0.837	0.86
British Virgin Isla	0.636	0.995	0.994	0.999	0.992	0.618	0.992	0.966	0.93

3D Plot of MMDS calculation: plot spheres to represent countries/samples. Positions of spheres relate to distance matrix – the further apart 2 spheres are, the more different the countries are based on data provided. Colours taken from clustering estimates from dendrogram.

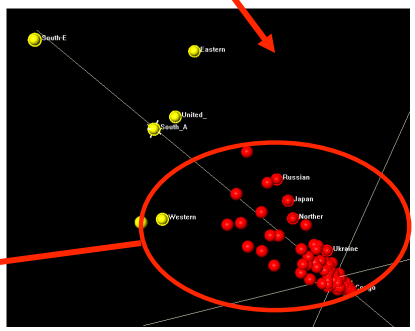


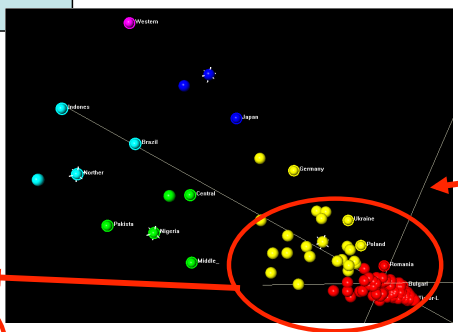
Initially, ASIA is an obvious outlier; hugely different to everything else. Re-run calc without the combined 'Continents' entries (Asia, Europe, etc)

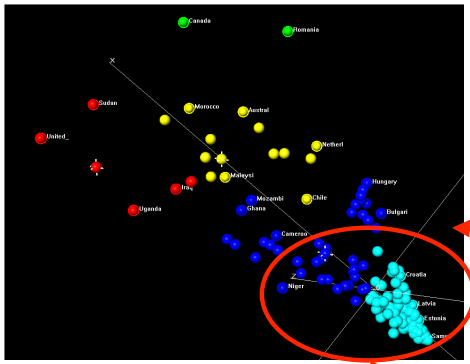
Better discrimination of differences now; major outliers are now: India, China, Eastern Asia and Southern Central Asia



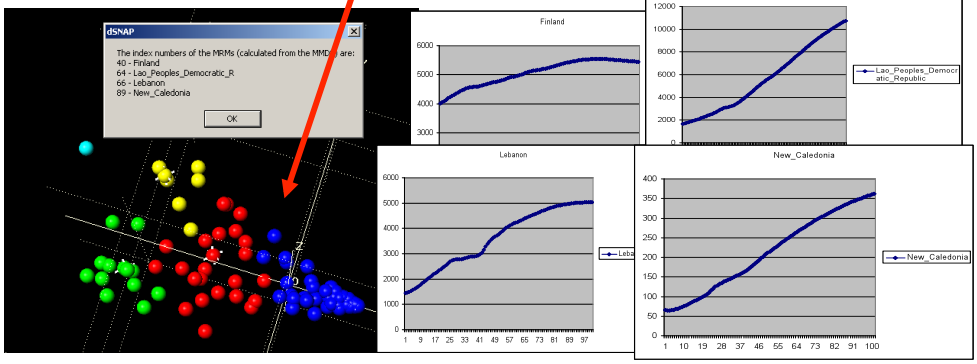
Can now select just the red cluster and repeatedly 'zoom in' to determine smaller, more detailed differences



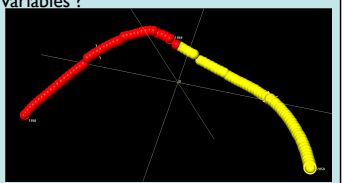




Can identify 'Most Representative Member' of each cluster group to look at in more detail



Alternative: What if we treat the years as the samples & the countries as the variables ?



Too much of a continuum to cluster

Conclusion: for gradually changing data without discrete divisions, cluster analysis is perhaps not best suited. However the ability to 'drill down' into the data set, starting off with an overview and gradually zooming in to expose finer and finer grained detail can be very powerful

[Obviously, in the interactive version, you can zoom rotate and select points of interst; hovering the mouse over any sample shows a tooltip with the full name.]