



Definition of Region for Clinical Trials

Mathematical Optimization Approach

Alex Zolot (Zolotovitski), PhD

Alex.Zolot@StatVis.com

alexzol@microsoft.com

Yoko Tanaka
Eli Lilly and
Company





Alex Zolot (Zolotovitski), PhD

Senior Researcher at Microsoft (Bing).

Formerly - a Senior Statistician / Sr. Engineer at Sun Microsystems, an Analytic Science Manager at FICO, executed successful data mining projects for Kraft Foods, Visa, Discover Financial, Cox Communications.

Fields of expertise include statistical analysis and modeling, and data mining, mainly predictive analytics.

Ph.D in Theoretical and Mathematical Physics, and a Ph.D. in Economics.

Certified Advanced SAS Programmer, MCP.

www.zolot.us

- **InnoCentive** - www.Innocentive.com - provides connection services between "Seekers" and "Solvers." Seekers are the companies searching for solutions to critical challenges. Solvers are the 185,000 registered members of the InnoCentive crowd who volunteer their solutions to the Seekers. Solvers whose solutions are selected by the Seekers are compensated for their ideas by InnoCentive, which acts as broker of the process.
- **Eli Lilly** posted the challenge, "to identify proposals for new regions for clinical trials which are supported by information which is currently publicly available (publication, clinicaltrials.gov, medline, etc.)."

- The task of definition of regions is universal for different actions – Clinical Trials or Soccer Cup.
- No universal solutions for clinical trials (CTReg)
 - Optimal regions depends on many parameters, so in our solution we describe not definition of regions, but **procedure** how to get the definition if we have required parameters.
- We also attach small program in R that generates regions given a set of parameters (“weights”) provided by experts.

Stage 1 – Experimental Design (DoE)

Typically clinical trials may be designed to do assess the safety and effectiveness of some medications or devices on a specific kind of patient and could be formalized as analysis of response

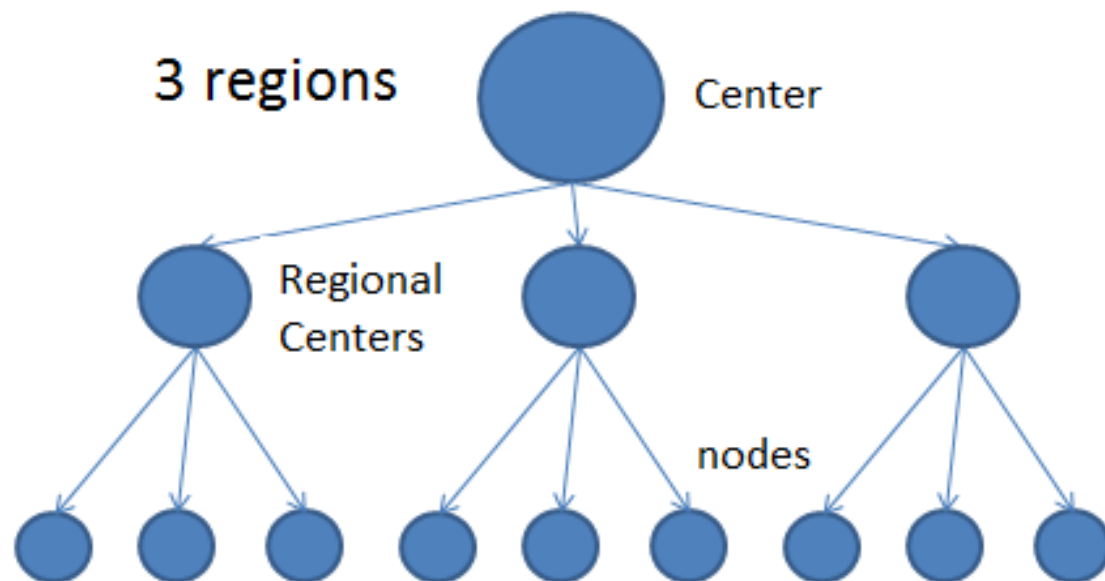
$$Y_i = f(p_j, x_k)$$

health criteria Y on treatment variables x . The response depend on parameters p of patients, including mentioned by Seeker: culture, ethnicity, language, medical practice, patient/disease characteristic, regulatory filing system demographic and so on. The objective of clinical trials is reveal (find a good approximation) for function f , that is a typical task of regression that could be non-linear and non-parametric. The regression procedure is out of scope of this work. For simplicity let us suppose linear regression

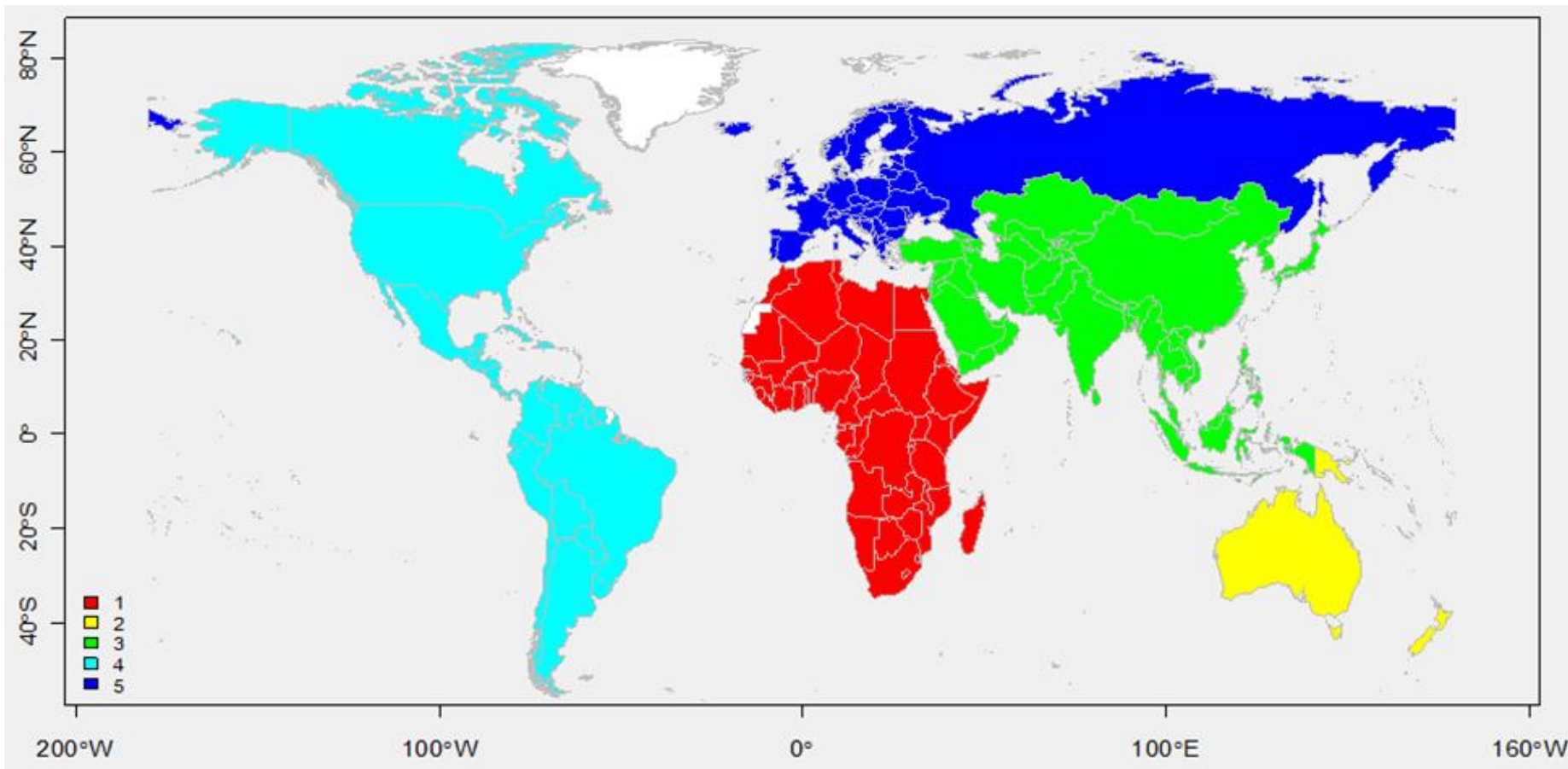
$$Y_i = \sum a(p_j) * x_k,$$

then objective of clinical trials is to estimate $a(p_j)$. For definition of CTRegs is convenient to aggregate parameters p_j into one aggregated variable p (“aggregated parameter”, AP).

Result of DoE – sample/block size - supposed to be done – out of the scope.

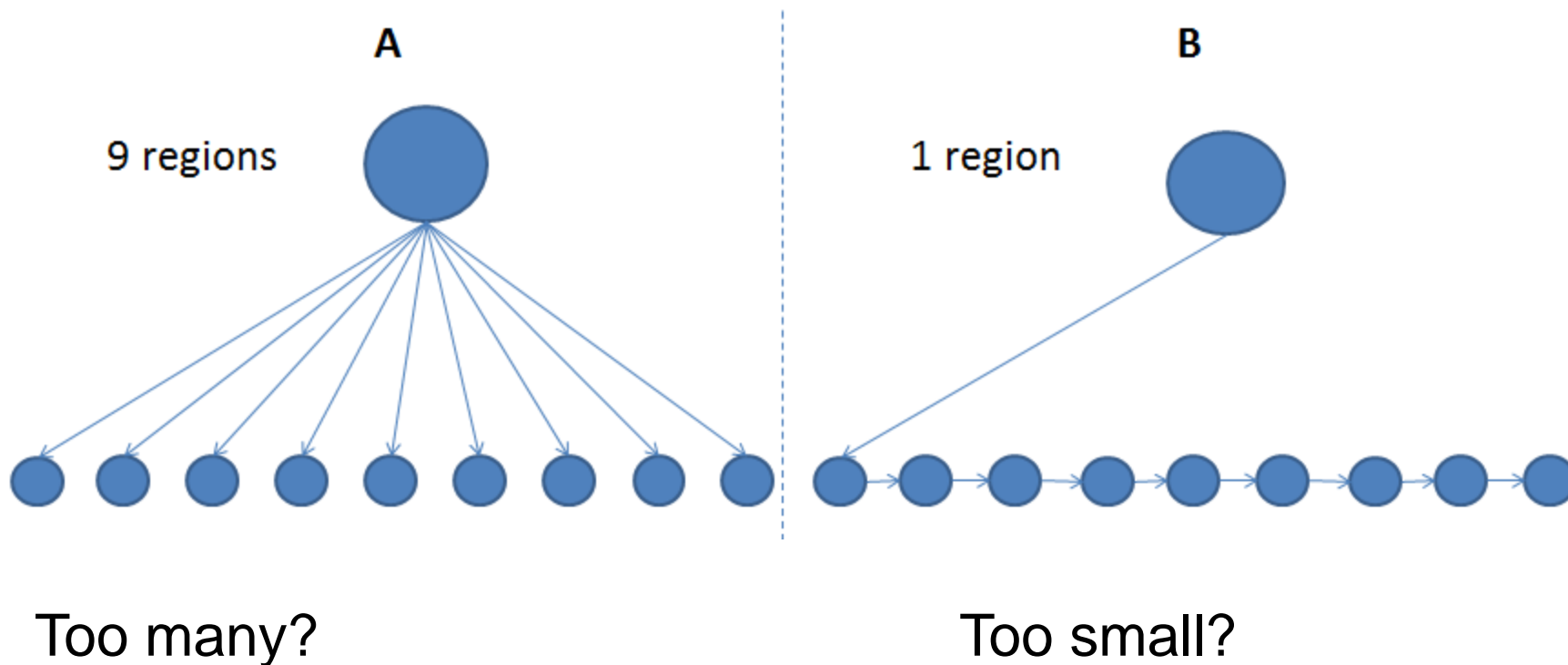


UN Regions – 5 regions/23 subregions, ~ 200 nodes (countries)



Why 5 or 23 regions? Optimal Number of Regions -?

Optimal Number of Regions -?



Optimal Number of Regions n -?

- Minimization of

$$\begin{aligned} \text{Cost} = & \text{AdminCost}(n_+) \\ & + \text{SizeCost}(R_+) \\ & + \text{NonUnifCost}(R_+) \end{aligned}$$

“+“ indicates increasing function,

R is “average” size of region

Optimal Number of Regions n -?

$$\text{Cost} = \sum_{u \in \text{CTR}} (A_u + B_u n_u + C_u n_u \langle (R - R_{cu})^2 \rangle) \quad (2)$$

$$= \sum_{u \in \text{CTR}} (A_u + B_u n_u + C_u n_u \text{Var}(R)_u) \quad (3)$$

where Σ means summation by Regions u ,

n_u – number of nodes in Region u ,

A_u, B_u, C_u, D_u are cost parameters that depends on Regions u and

R_u are known constants - average size of Region u that could be estimated as $\text{AREA}_u^{1/2}$

Optimal Number of Regions n -?

Minimization of

$$\text{Cost} = \sum_{u \in CTR} (A_u + B_u n_u + C_u n_u \text{Var}(R)_u) \quad (4)$$

if coefficients do not depend on regions

$$= A n_{\text{reg}} + \cancel{B} n_{\text{nodes}} + C n_{\text{nodes}} \text{mean}(\text{Var}(R)) \quad (5)$$

R (geo) → General set of parameters $\Sigma w_j p_j$

$$\text{Cost} = A n_{\text{reg}} + n_{\text{countries}} \text{mean}(\Sigma w_j^2 \text{Var}(p_j)) \quad (6)$$

$$= A n_{\text{reg}} + \Sigma w_j^2 (p_j - p_{c_j})^2 \quad (7)$$

A – cost to create one Reg.Center

w_j^2 – cost of intra-region variance of p_j

Out of scope –
task dependent -
defined by experts

p - raw variables or

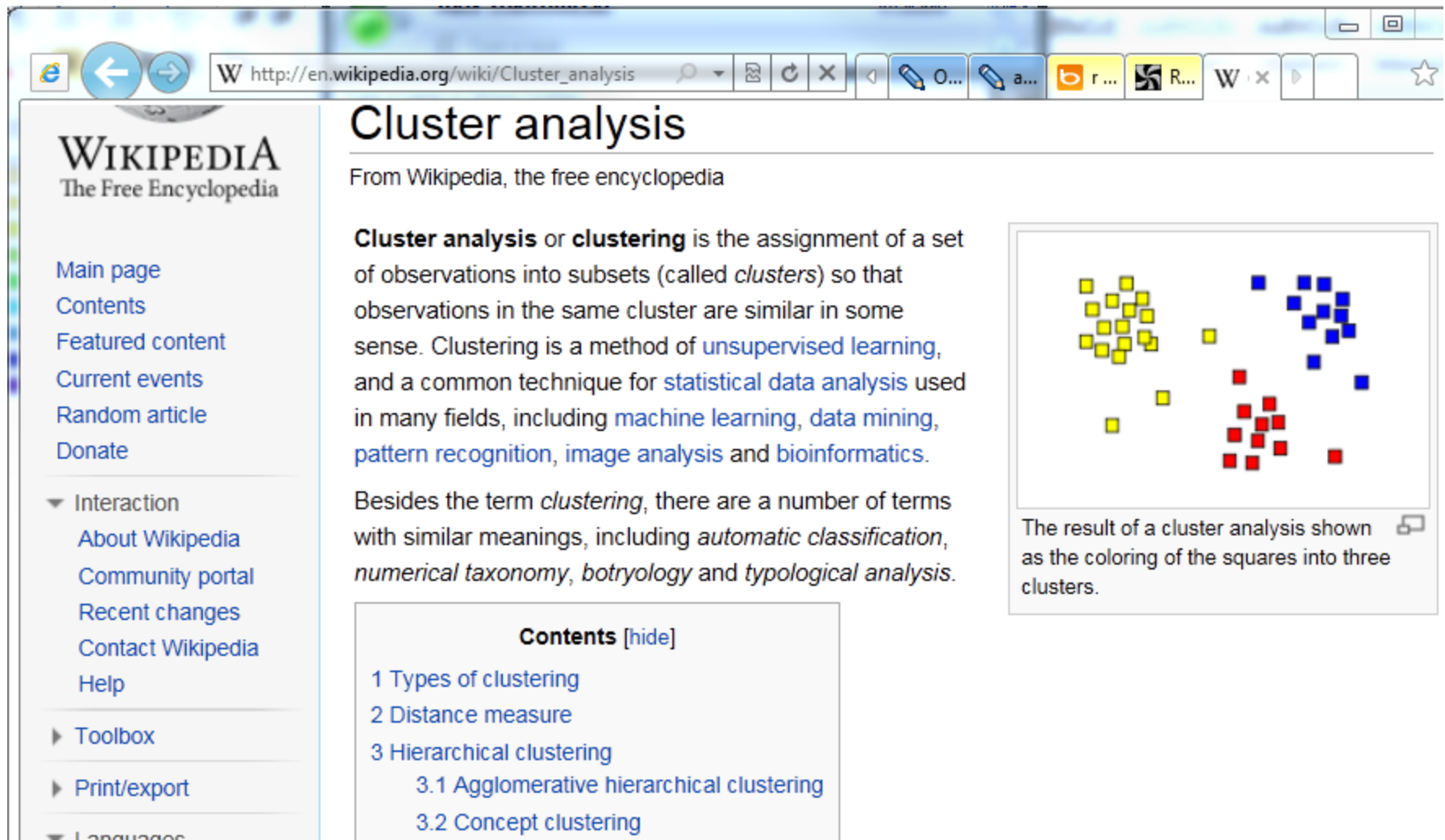
- Indexes of raw variables or

- Factors build on raw variables

Categorical variables → dummy numerical variables

$$\text{Min(Cost)} = \min(A n_{\text{reg}} + \sum w_j^2 (p_j - p_{c_j})^2) \quad (8)$$

= typical task of clustering



The screenshot shows a web browser window displaying the Wikipedia page for "Cluster analysis". The browser's address bar shows the URL "http://en.wikipedia.org/wiki/Cluster_analysis". The page content includes the Wikipedia logo, a navigation sidebar on the left, and the main article text. The article text defines cluster analysis as the assignment of observations into subsets (clusters) and mentions its application in unsupervised learning, statistical data analysis, machine learning, data mining, pattern recognition, image analysis, and bioinformatics. A diagram on the right illustrates the result of a cluster analysis, showing a set of squares colored into three distinct clusters: yellow, blue, and red. Below the diagram, a caption reads: "The result of a cluster analysis shown as the coloring of the squares into three clusters."

WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate](#)

Interaction

- [About Wikipedia](#)
- [Community portal](#)
- [Recent changes](#)
- [Contact Wikipedia](#)
- [Help](#)

Toolbox

Print/export

Languages

Cluster analysis

From Wikipedia, the free encyclopedia

Cluster analysis or **clustering** is the assignment of a set of observations into subsets (called *clusters*) so that observations in the same cluster are similar in some sense. Clustering is a method of [unsupervised learning](#), and a common technique for [statistical data analysis](#) used in many fields, including [machine learning](#), [data mining](#), [pattern recognition](#), [image analysis](#) and [bioinformatics](#).

Besides the term *clustering*, there are a number of terms with similar meanings, including *automatic classification*, *numerical taxonomy*, *botryology* and *typological analysis*.

Contents [hide]

- [1 Types of clustering](#)
- [2 Distance measure](#)
- [3 Hierarchical clustering](#)
 - [3.1 Agglomerative hierarchical clustering](#)
 - [3.2 Concept clustering](#)

The result of a cluster analysis shown as the coloring of the squares into three clusters.

Minimization = Clustering

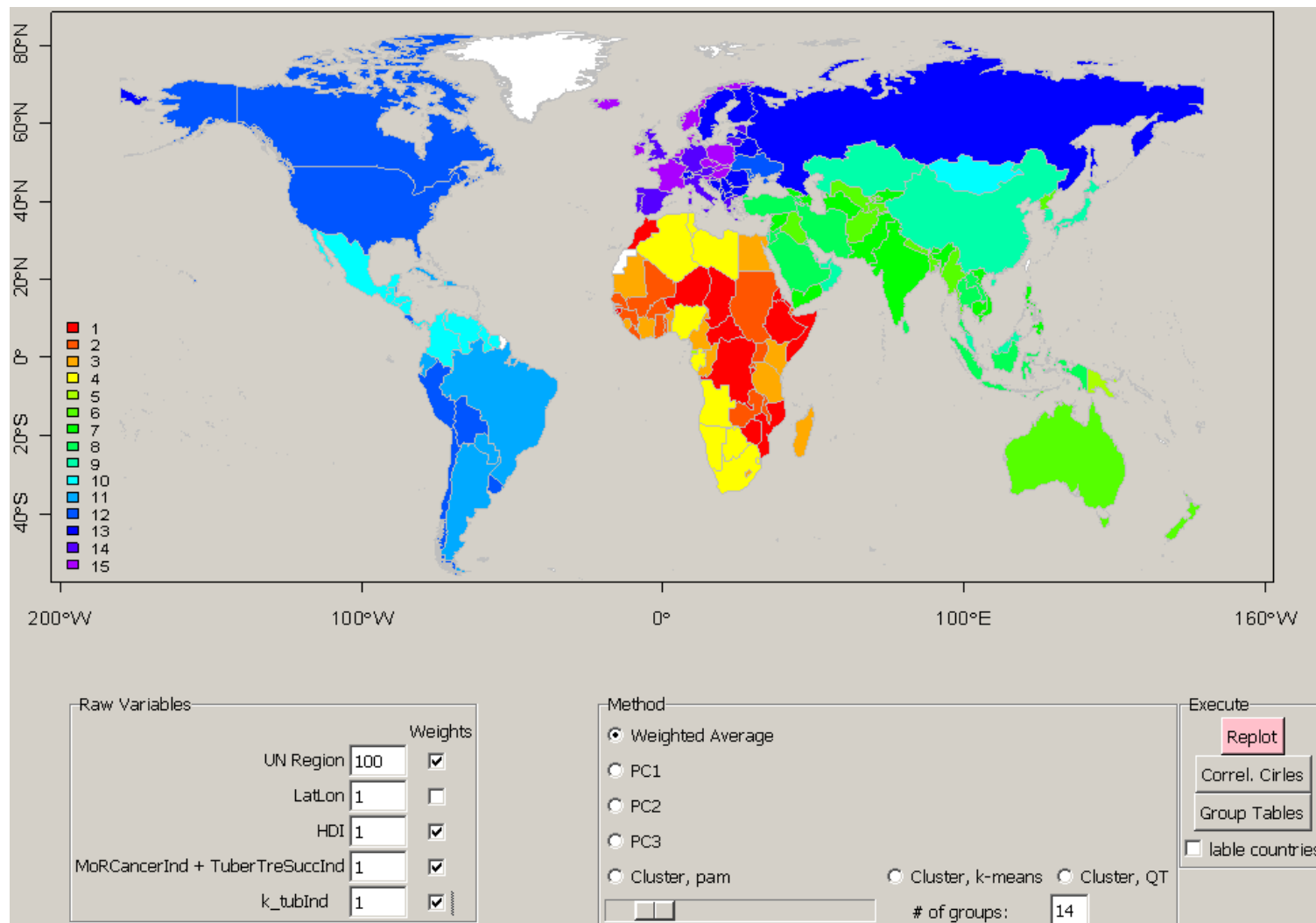
$$\text{Min(Cost)} = \min(A n_{\text{reg}} + \sum w_j^2 (p_j - p_{c_j})^2) \quad (8)$$

= typical task of clustering in data mining

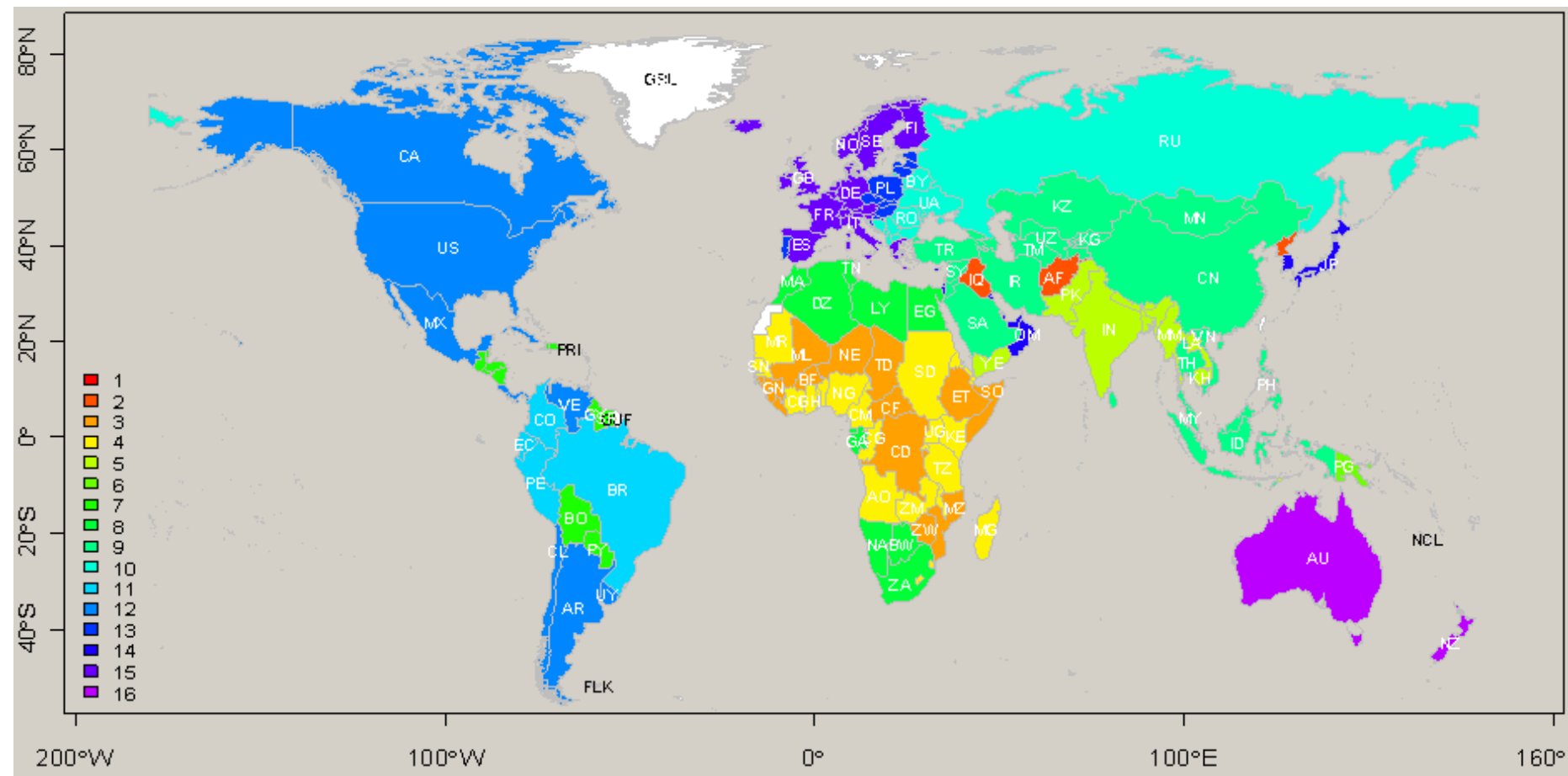
We can start from wide set (hundreds or thousands) parameters, characterizing countries (nodes).

Then dimension reduction - e.g. via factor (PCA) analysis.

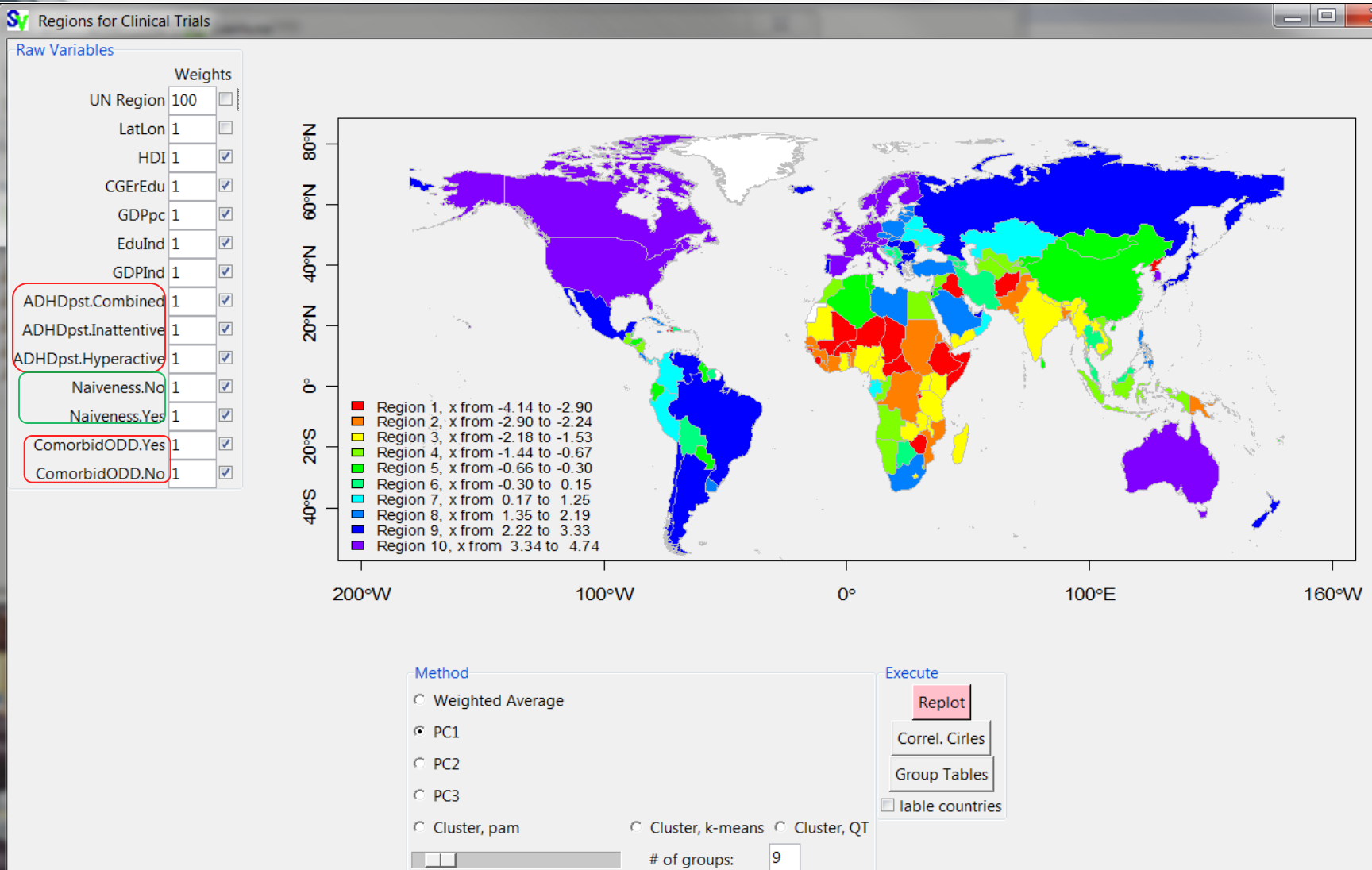
In our simple mock-up we use GDP, Human Development Index, Mortality Rate of Cancer and Tuberculosis Treatment Success.



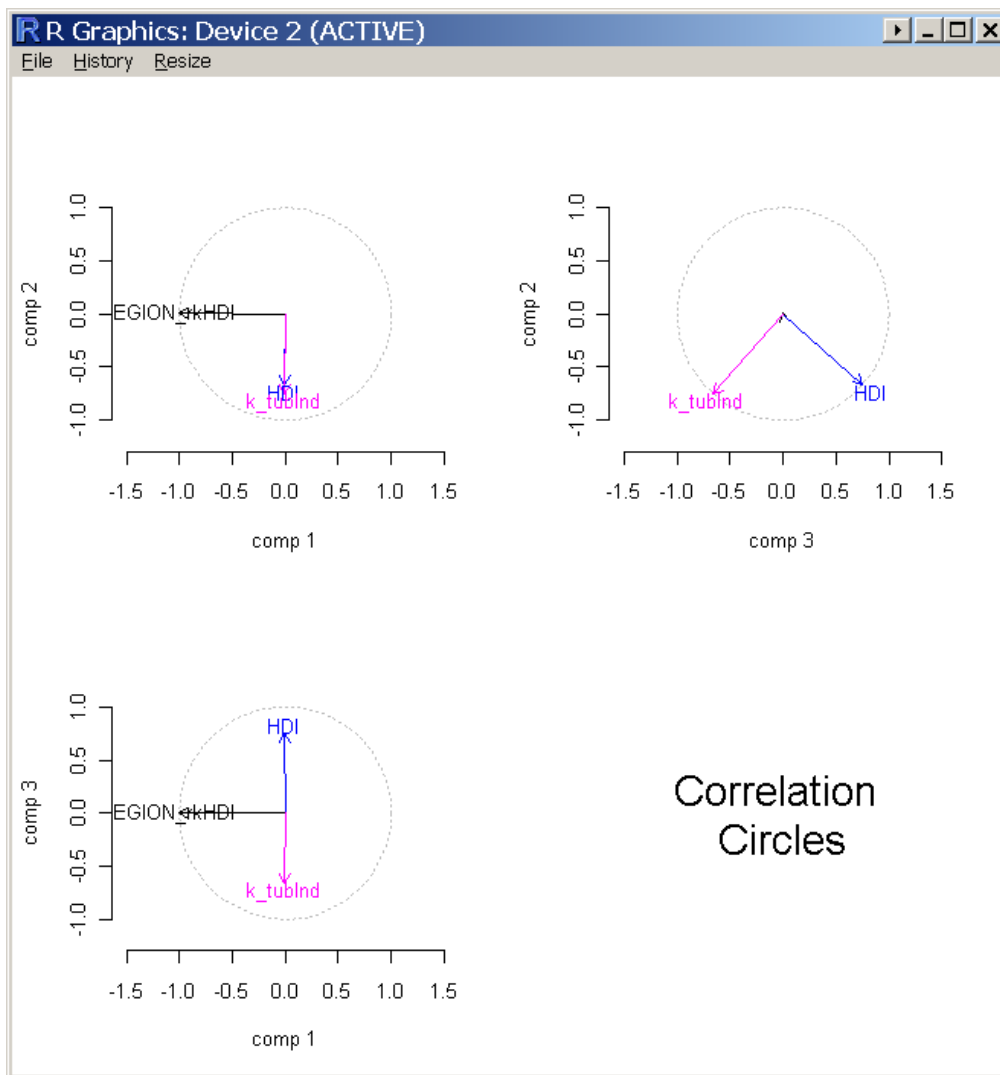
14 CTRegs, equal weights, Weighted Average



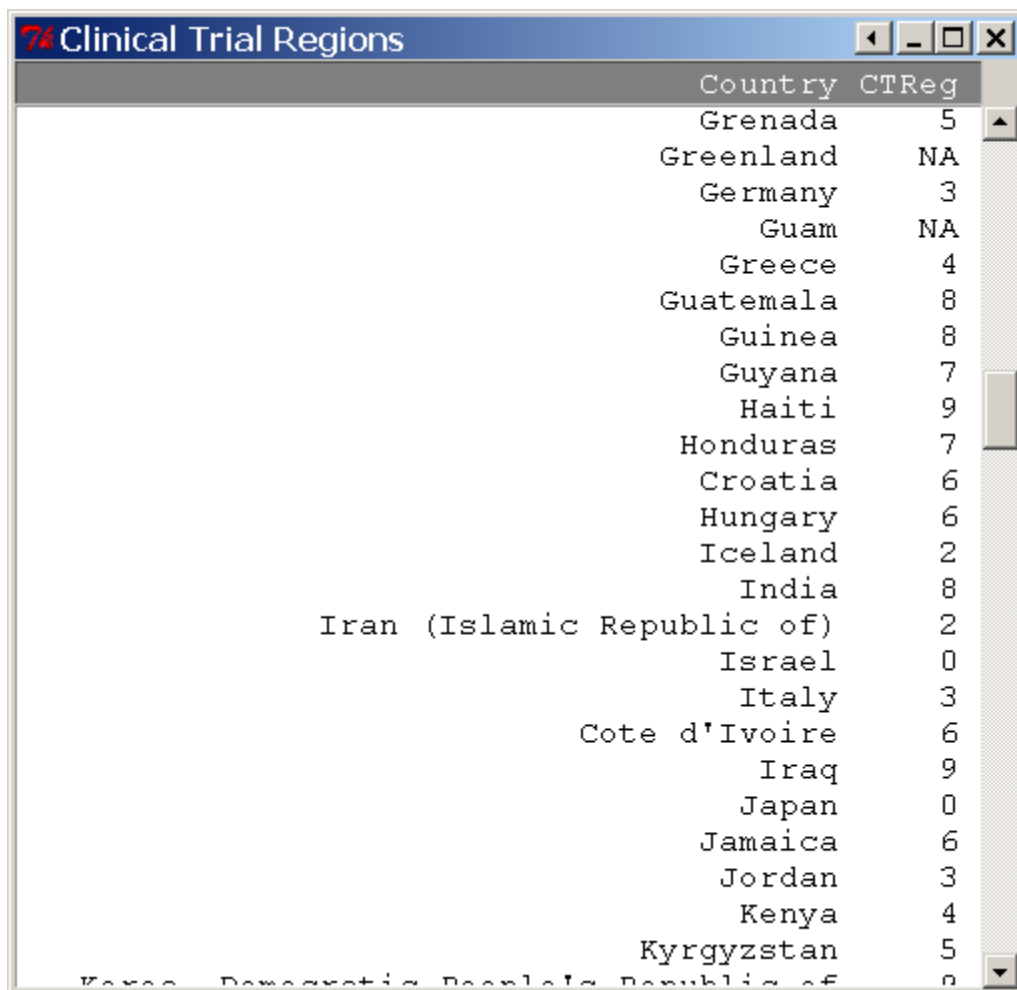
14 CTRegs, HDI only, Clusters (pma), labels on



No geo, 3 Categorical variables-> 7 dummy, 1st factor, 10 levels



Output for PCA Analysis: Correlation Circles.

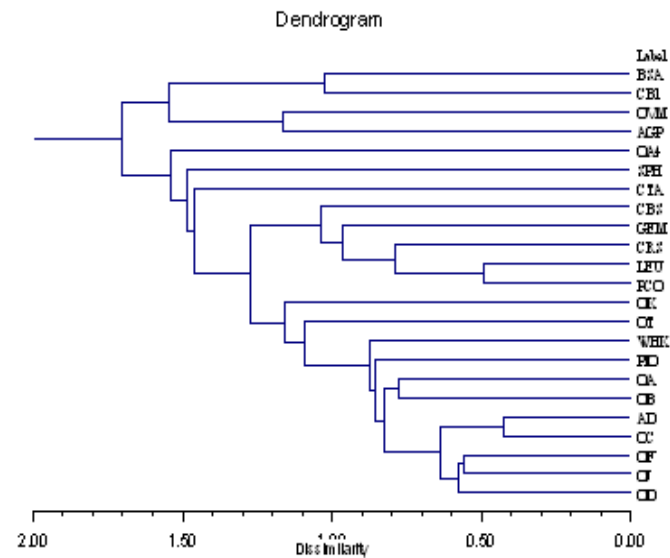


The screenshot shows a window titled "Clinical Trial Regions" with a table containing the following data:

Country	CTReg
Grenada	5
Greenland	NA
Germany	3
Guam	NA
Greece	4
Guatemala	8
Guinea	8
Guyana	7
Haiti	9
Honduras	7
Croatia	6
Hungary	6
Iceland	2
India	8
Iran (Islamic Republic of)	2
Israel	0
Italy	3
Cote d'Ivoire	6
Iraq	9
Japan	0
Jamaica	6
Jordan	3
Kenya	4
Kyrgyzstan	5
Korea, Democratic People's Republic of	0

Output: Table of CTRegs.

- Why we need PCA?
- Generalization: Hierarchy – hierarchical clustering.



Conclusion

We developed

- General optimization approach to create optimized CT Regions using standard clustering technics and expert defined weights (or cost coefficients) for arbitrary set of important input variables
- Mock-up R application implementing the technic and visualizing results of clustering, levels of factors (or main principal components) or raw variables



Thank you

????????

www.StatVis.com

alex@zolot.us