Supported by

Co-funded by the European Union

DIM ESEE-2 innovative workshop

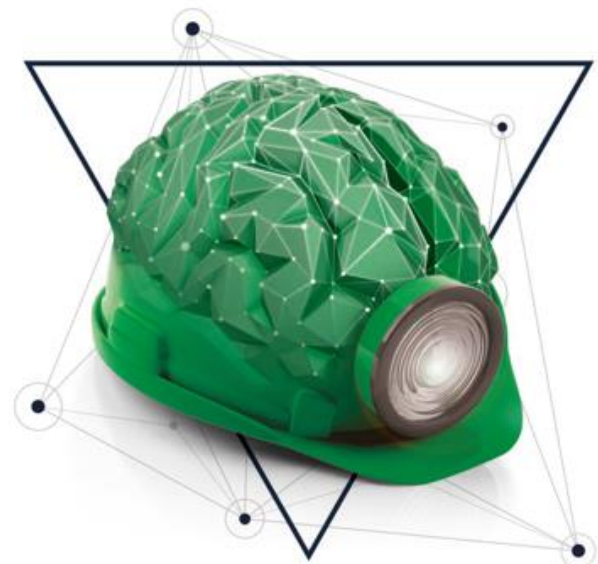
DIM ESEE 2021: Innovation in exploration

Prof. Dr. Norbert Péter Szabó

University of Miskolc, Department of Geophysics

22 October 2021

Advanced statistical analysis of multivariate (big) datasets



October 20th – 22nd, 2021
IUC Dubrovnik, Croatia / online

Supported by

Co-funded by the European Union

Introduction

- He obtained his M.Sc. degree in geophysical engineering in 1999 from Faculty of Mining Engineering, University of Miskolc. He has been continuously working from graduating at the University of Miskolc. He obtained his Ph.D. in 2005. Since 2019, he has been a full professor at the Department of Geophysics. He is currently the head of Geophysical Department and vice-dean for scientific affairs at the Faculty of Earth Science and Engineering. In addition, he is senior research fellow at the MTA-ME Geoengineering Research Group. In 2020, he defended his D.Sc. dissertation at the Hungarian Academy of Sciences.
- He conduct researches on geophysical inversion and exploratory (multivariate) statistical methods and their applications in earth sciences (mainly water and hydrocarbon prospecting). He delivers lectures on well logging, gravitational and magnetic exploration methods, engineering and environmental geophysics and geostatistics in the framework of BSc, MSc and PhD training programs.

Supported by
RawMaterials ACADEMY
Co-funded by the European Union

Course Description

- Introduction to basic univariate and multivariate statistical methods. The advantage of using robust statistical methods
- The Most Frequent Value (MFV) method as robust statistical estimatior
- Exploratory factor analysis of geospatial variables
- Evolutionary computation-based factor analysis and its applications for improved lithological analysis and quantitative estimation of petrophysical properties
- Cluster analysis of multidimensional data objects and its applications for improved lithological analysis and quantitative estimation of petrophysical properties
- Machine learning tools as an aid for a more relaible interpretation of geophysical data. Well logging applications and examples

Supported by
RawMaterials ACADEMY
Co-funded by the European Union

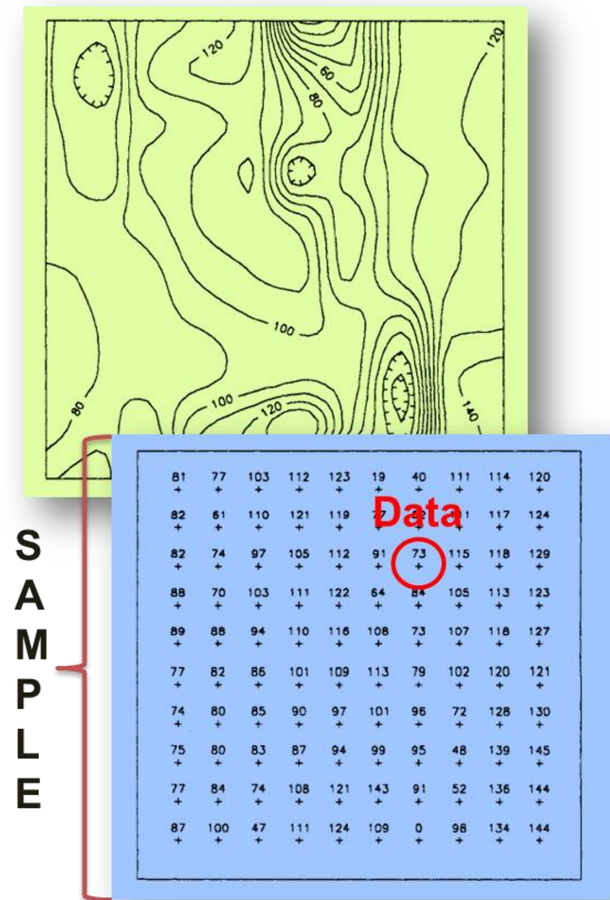
Selected Bibliography

- Edward H. Isaacs, R. Mohan Srivastava, 1989. An Introduction to Applied Geostatistics. Oxford University Press
- Martin H. Trauth, 2006. MATLAB Recipes for Earth Sciences. Springer
- Ferenc Steiner, 1991. The Most frequent value: introduction to a modern conception of statistics. Academic Press Budapest
- William Menke, 2012. Geophysical Data Analysis: Discrete Inverse Theory. MATLAB Edition. Elsevier
- Michalewicz Z., 1992. Genetic Algorithms + Data Structures = Evolution Programs. Springer
- Papers of mine: https://www.researchgate.net/profile/Norbert-Szabo-8/research

Supported by
RawMaterials ACADEMY
Co-funded by the European Union
DIM ESEE 2021

Why Geostatistics?

- How often does a specific value of data occur in the data set?
- How many data occur below a specific value?
- How can data frequency modelled mathematically?
- What is the most characteristic vale in an area?
- What is the standard deviation of a data set?
- How to handle incorrect data?
- How can we estimate measurements at points which have not been measured based on other measurements?
- What kind of relation a data type has with other data?



Isaaks and Srivastava, 1989

Supported by
RawMaterials ACADEMY
Co-funded by the European Union

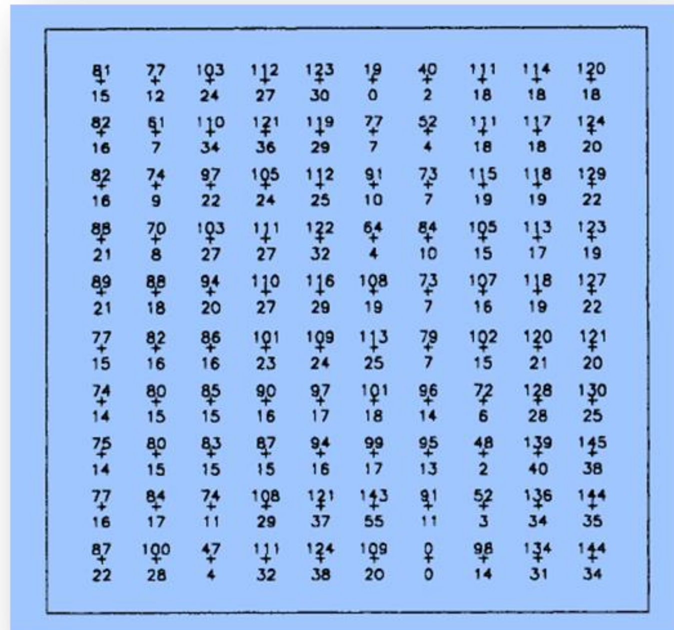
Supported by

Co-funded by the European Union

Multivariate Statistics

- What is the probability of joint occurrence of data?
- Is there a relation between data sets or are they independent?
- How strong is the relation between data sets, positive or negative?
- How do we describe this function relation mathematically and use it to interpolate the result to unmeasured locations?
- How do we estimate the model parameters from the data?
- What is the error of estimation?
- How do we sort data in case of a big dataset?

Isaaks and Srivastava, 1989



Statistical sample of two variables

Supported by

RawMaterials ACADEMY

Co-funded by the European Union

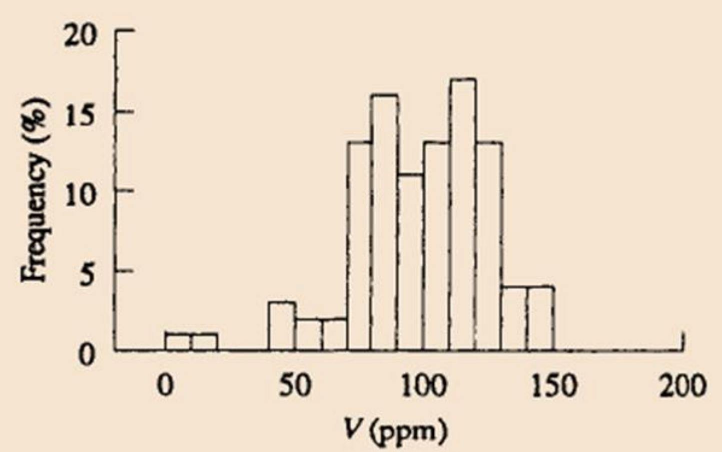
2021 ESEE

Frequency of Data

81	77	103	112	123	19	40	111	114	120
82	61	110	121	119	77	52	111	117	124
82	74	97	105	112	91	73	115	118	129
88	70	103	111	122	64	84	105	113	123
89	88	94	110	116	108	73	107	118	127
77	82	86	101	109	113	79	102	120	121
74	80	85	90	97	101	96	72	128	130
75	80	83	87	94	99	95	48	139	145
77	84	74	108	121	143	91	52	136	144
87	100	47	111	124	109	0	98	134	144

S
A
M
P
L
E

Class	Number	Percentage
$0 \leq V < 10$	1	1
$10 \leq V < 20$	1	1
$20 \leq V < 30$	0	0
$30 \leq V < 40$	0	0
$40 \leq V < 50$	3	3
$50 \leq V < 60$	2	2
$60 \leq V < 70$	2	2
$70 \leq V < 80$	13	13
$80 \leq V < 90$	16	16
$90 \leq V < 100$	11	11
$100 \leq V < 110$	13	13
$110 \leq V < 120$	17	17
$120 \leq V < 130$	13	13
$130 \leq V < 140$	4	4
$140 \leq V < \infty$	4	4



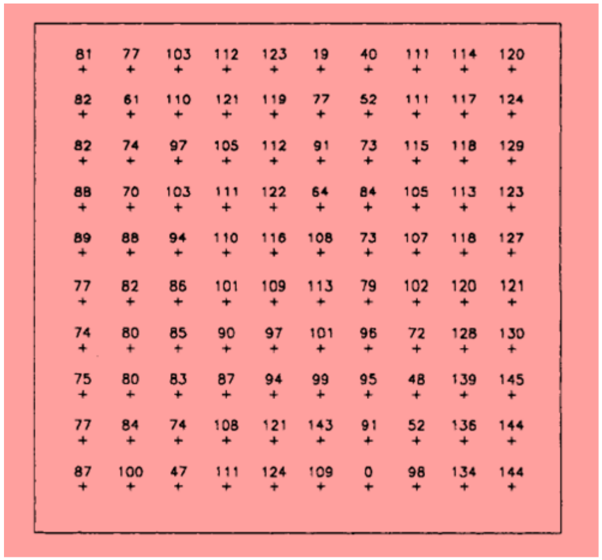
- Empirical probability density function (histogram)
- Walker Lake data set, Nevada (Isaaks and Srivastava, 1989)

Co-funded by the European Union

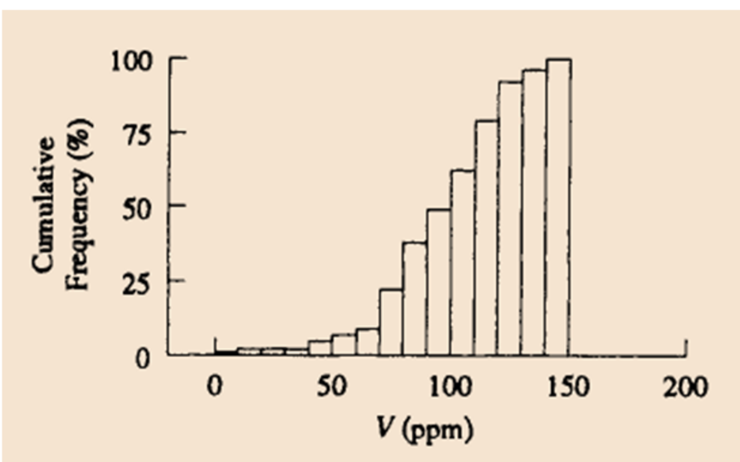
Supported by

Co-funded by the European Union

Frequency of Data



Class	Number	Percentage
$V < 10$	1	1
$V < 20$	2	2
$V < 30$	2	2
$V < 40$	2	2
$V < 50$	5	5
$V < 60$	7	7
$V < 70$	9	9
$V < 80$	22	22
$V < 90$	38	38
$V < 100$	49	49
$V < 110$	62	62
$V < 120$	79	79
$V < 130$	92	92
$V < 140$	96	96
$V < \infty$	100	100



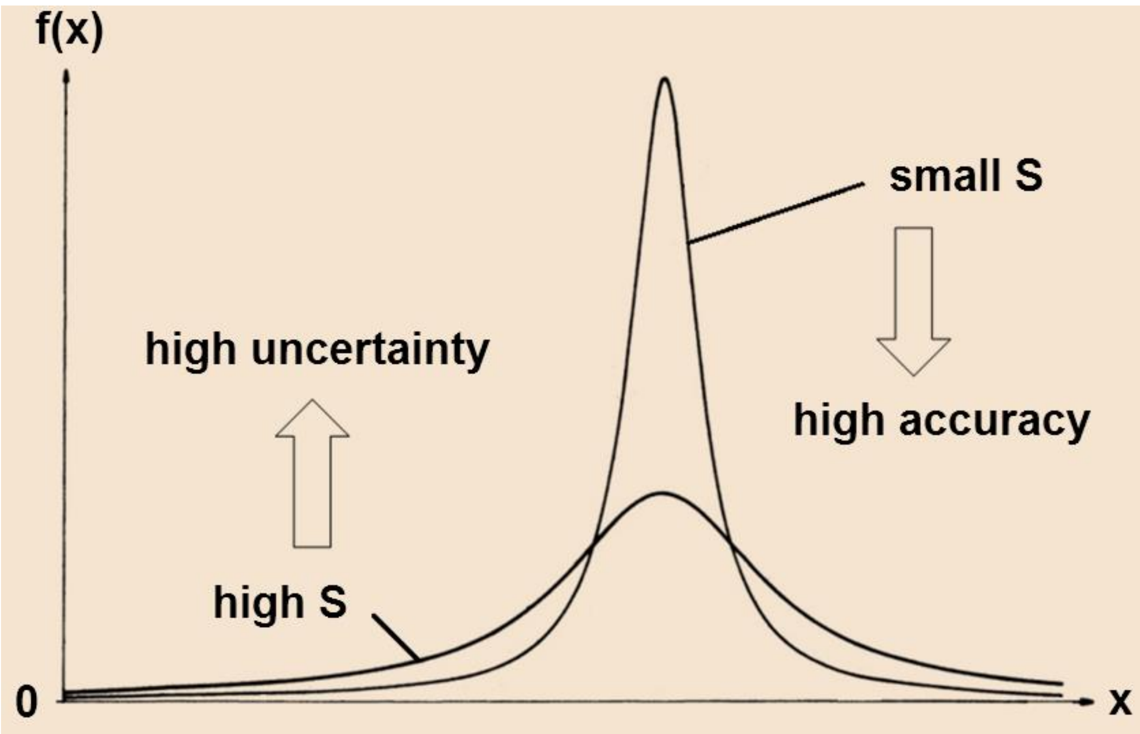
- Empirical probability distribution function
- Walker Lake data set, Nevada (Isaaks and Srivastava, 1989)

Co-funded by the European Union

Supported by

Co-funded by the European Union

Gaussian Distributed Data



- General formula of p.d.f.

$$f_G(x) = \frac{1}{S \cdot \sqrt{2\pi}} e^{-\frac{(x-T)^2}{2S^2}}$$

- Standardized form of p.d.f. (T=0, S=1)

$$f_G(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

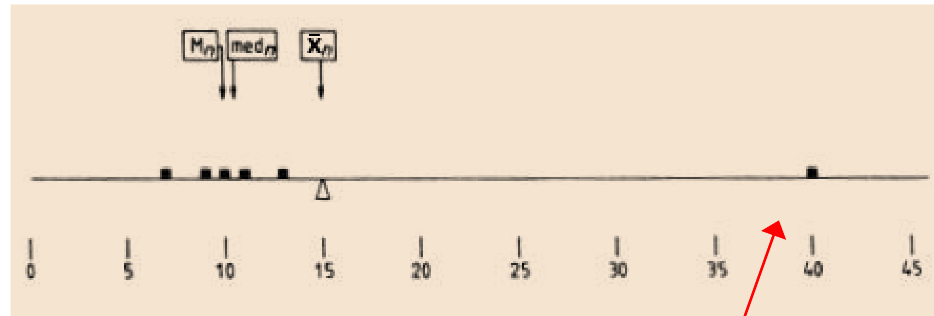
Supported by

Co-funded by the European Union

Characteristic Values of Sample

- Arithmetic mean of sample

$$x_n = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{x_1 + x_2 + \cdots + x_n}{n}$$



Outlier

- Weighted mean of sample

$$\bar{x}_{n,w} = \frac{\sum_{k=1}^{n} w_k \cdot x_k}{\sum_{k=1}^{n} w_k} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n}{w_1 + w_2 + \cdots + w_n}$$

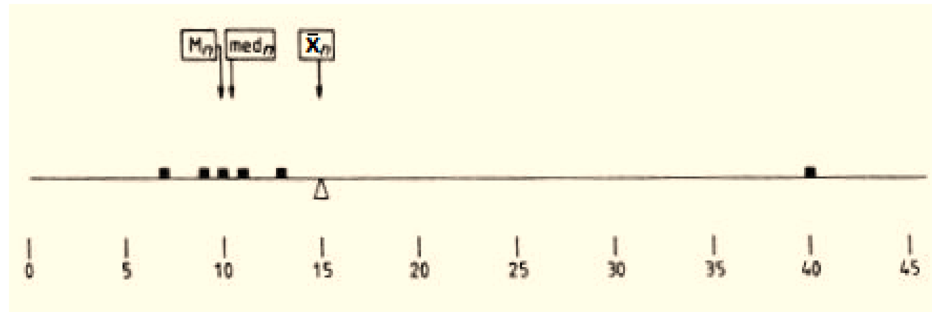
- Median of sample (more robust estimation)

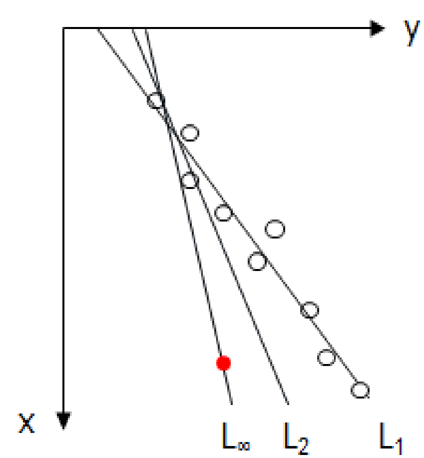
$$med_n = \begin{cases} x_{(n+1)/2}, & n \text{ is odd} \\ \dfrac{x_{n/2} + x_{(n+2)/2}}{2}, & n \text{ is even} \end{cases}$$

Supported by
RawMaterials ACADEMY
Co-funded by the European Union

Robust Estimation

- Balance example: consider the following data set including six data and one of them is an outlier. The source of outliers can be a defective instrument, wrong measurement, data transfer or recording etc.
- It can be seen that the sample mean is very sensitive to the presence of the outlier, the median and the most frequent value given more realistic estimations
- **Resistance**: the estimator is almost entirely insensitive to the presence of outliers
- **Robustness**: this kind of estimation procedure gives reliable results for a wide variety of data distributions





$$L_p = \left[\frac{1}{n}\sum_{i=1}^{n}\left|y_i^{(m)} - f(x_i)\right|^p\right]^{1/p}$$

$$p = 1: \; L_1 = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{(m)} - f(x_i)\right|$$

$$p = 2: \; L_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(m)} - f(x_i)\right)^2}$$

$$p = \infty: \; L_\infty = \max_{i=1}^{n}\left|y_i^{(m)} - f(x_i)\right|$$

Supported by

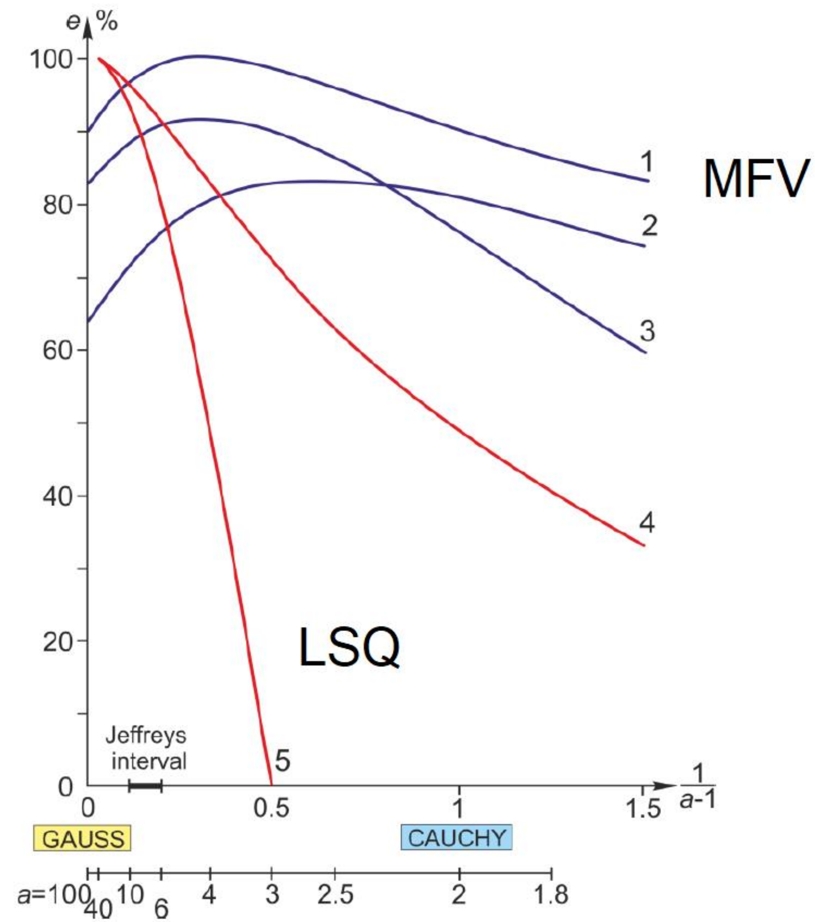
Co-funded by the European Union

Most Frequent Value

- Weighted average - data far from the most of the data get small weights, data at near the MFV get higher weights

$$M = \frac{\sum_{i=1}^{n} x_i \varphi_i}{\sum_{i=1}^{n} \varphi_i}, \qquad \varphi_i = \frac{\varepsilon^2}{\varepsilon^2 + (x_i - M)^2}$$

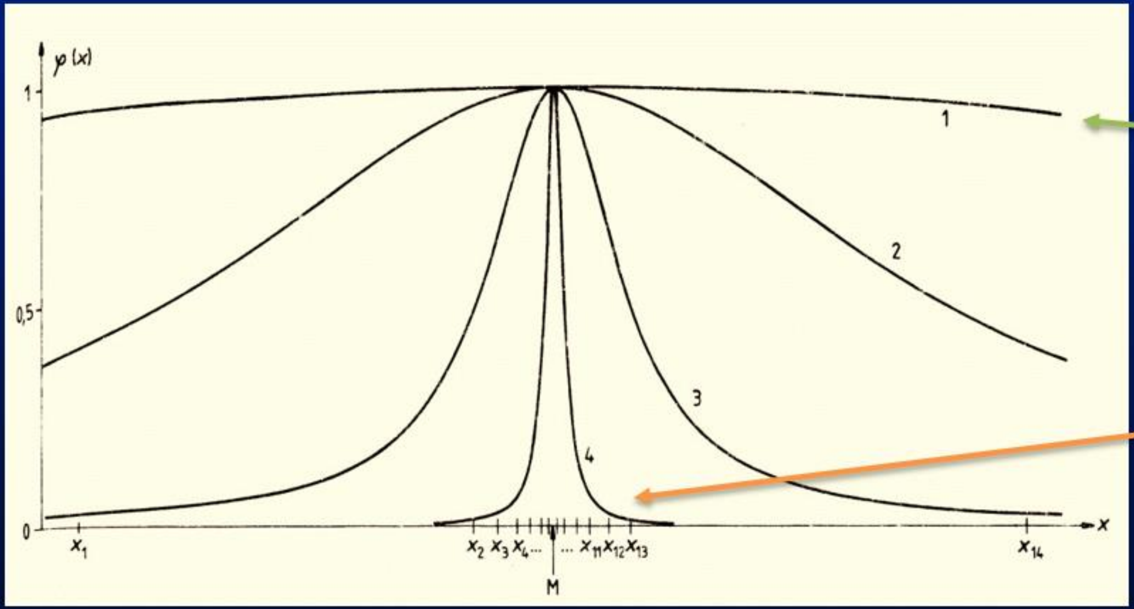
- Automated iterative process - in general the values of M and ε are calculated simultaneously by a recursion formula. Optimal weights are automatically estimated to the given dataset



Supported by
eit RawMaterials ACADEMY
Co-funded by the European Union

Most Frequent Value

- M is the most frequent value - location parameter
- ε is dihesion - scale parameter



Large dihesion =
Big weights go to each data

Small dihesion =
Small or zero weights to outlying data

eit RawMaterials ACADEMY
Co-funded by the European Union

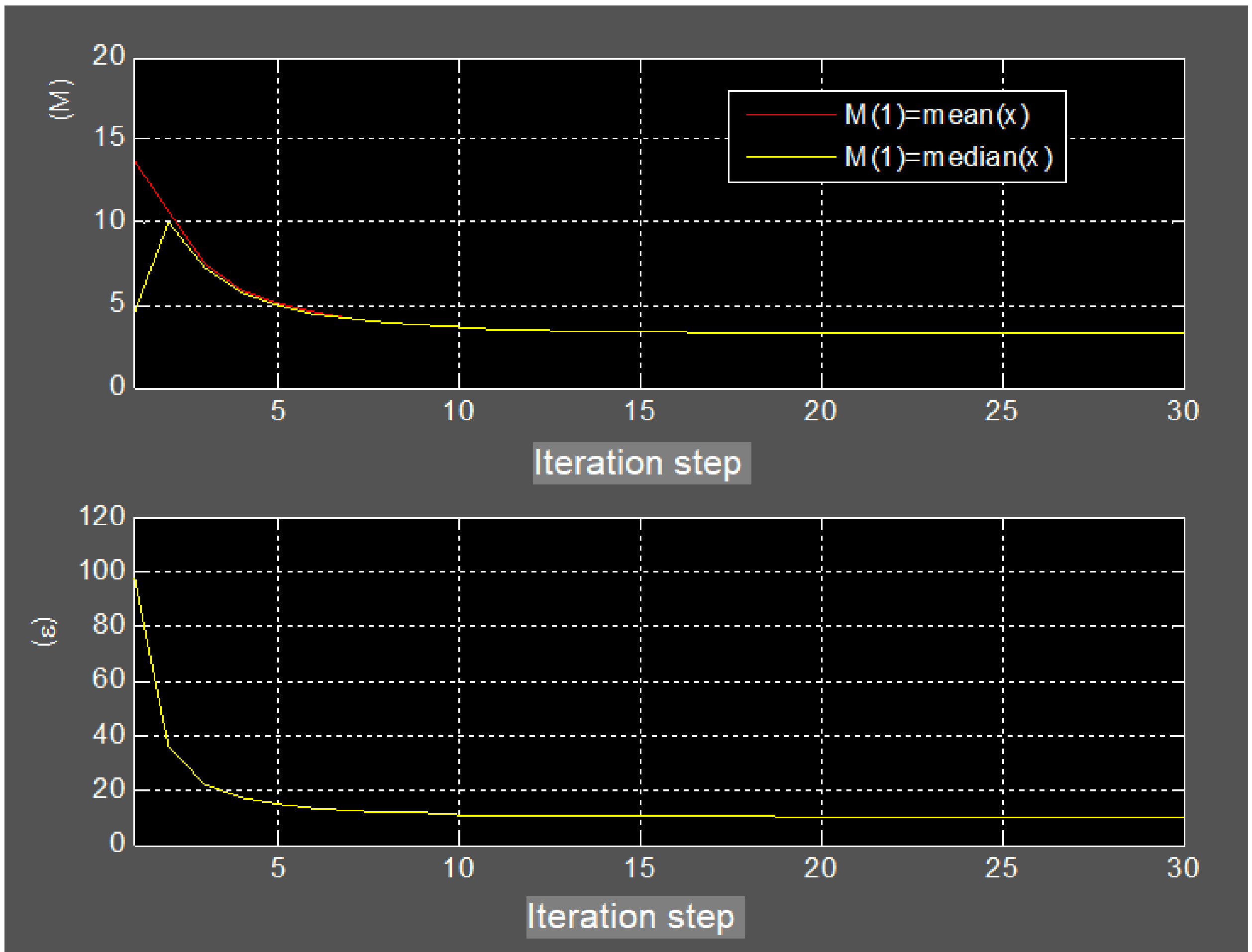
Supported by

Co-funded by the European Union

MFV estimation

Outlier

`x=[-12.5 -6.7 -2 -1.5 0.1 2.4 6.8 9.8 15 23.5 30 100]`



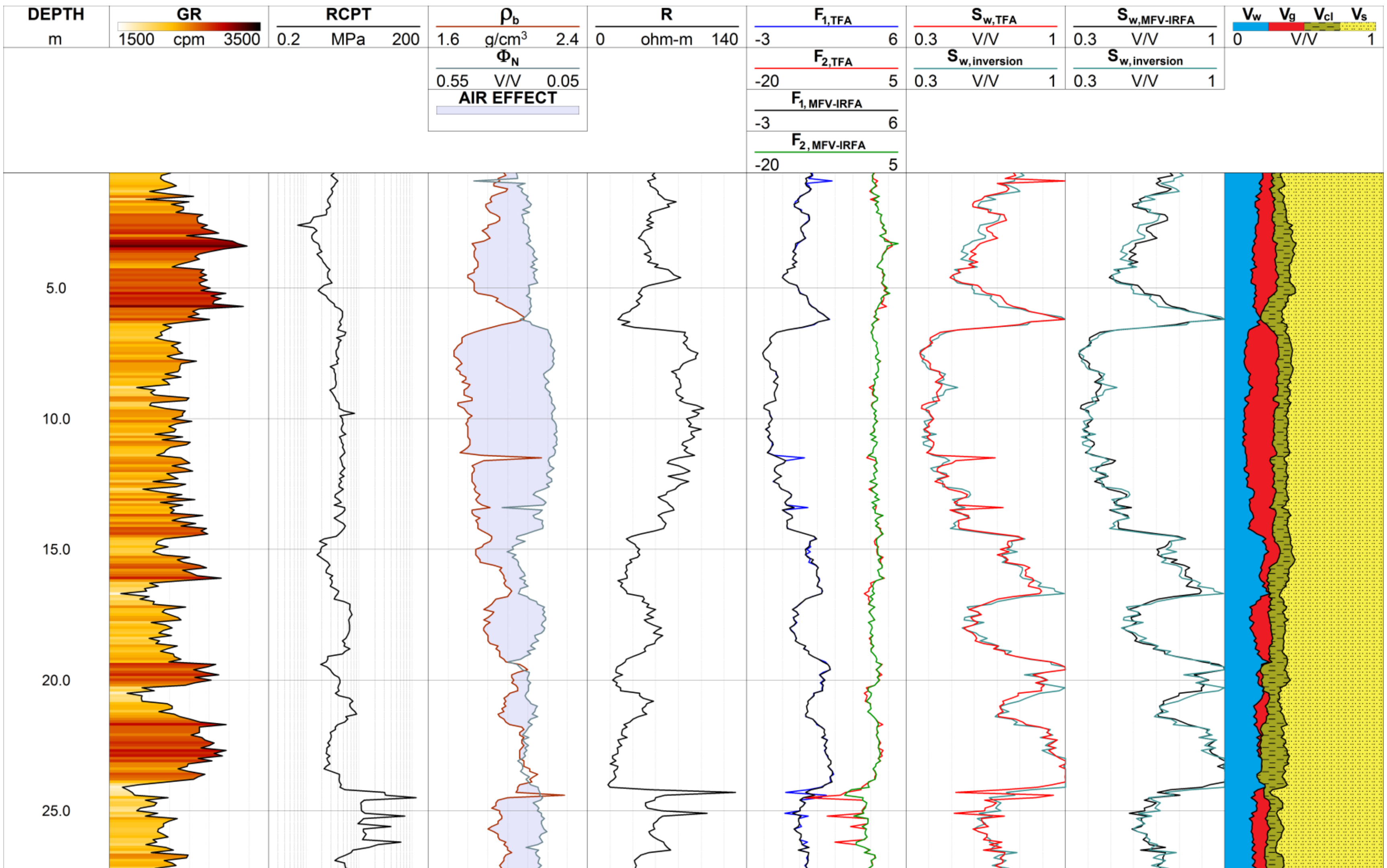
Supported by

EIT RawMaterials ACADEMY

Co-funded by the European Union

DIM ESEE 2021

Well Logging Example



Supported by

Co-funded by the European Union

Model of Factor Analysis

- Standardized well-logging data are stored in N-by-K matrix

$$\mathbf{D}=\begin{pmatrix} GR_1 & SP_1 & RD_1 & RS_1 & DEN_1 & PHIN_1 & AT_1 & CAL_1 & TE_1 \\ GR_2 & SP_2 & RD_2 & RS_2 & DEN_2 & PHIN_2 & AT_2 & CAL_2 & TE_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ GR_k & SP_k & RD_k & RS_k & DEN_k & PHIN_k & AT_k & CAL_k & TE_k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ GR_N & SP_N & RD_N & RS_N & DEN_N & PHIN_N & AT_N & CAL_N & TE_N \end{pmatrix}$$

- Decomposition of data matrix

$$\mathbf{D}=\mathbf{F}\mathbf{L}^T+\mathbf{E}$$

F: N-by-a matrix of factor scores
L: K-by-a matrix of factor loadings
E: N-by-K matrix of residuals
M: number of factors

$$\begin{pmatrix} GR_1 & DEN_1 & NPHI_1 & RES_1 \\ GR_2 & DEN_2 & NPHI_2 & RES_2 \\ GR_3 & DEN_3 & NPHI_3 & RES_3 \\ GR_4 & DEN_4 & NPHI_4 & RES_4 \\ GR_5 & DEN_5 & NPHI_5 & RES_5 \\ GR_6 & DEN_6 & NPHI_6 & RES_6 \\ GR_7 & DEN_7 & NPHI_7 & RES_7 \\ GR_8 & DEN_8 & NPHI_8 & RES_8 \\ GR_9 & DEN_9 & NPHI_9 & RES_9 \\ GR_{10} & DEN_{10} & NPHI_{10} & RES_{10} \end{pmatrix} = \begin{pmatrix} F_1^{(1)} & F_1^{(2)} \\ F_2^{(1)} & F_2^{(2)} \\ F_3^{(1)} & F_3^{(2)} \\ F_4^{(1)} & F_4^{(2)} \\ F_5^{(1)} & F_5^{(2)} \\ F_6^{(1)} & F_6^{(2)} \\ F_7^{(1)} & F_7^{(2)} \\ F_8^{(1)} & F_8^{(2)} \\ F_9^{(1)} & F_9^{(2)} \\ F_{10}^{(1)} & F_{10}^{(2)} \end{pmatrix} \begin{pmatrix} L_{11} & L_{12} & L_{13} & L_{14} \\ L_{21} & L_{22} & L_{23} & L_{24} \end{pmatrix}$$

Supported by

Co-funded by the European Union

Quick (Non-Iterative) Solution

- Factors are linearly independent, matrices $\mathbf{FL}^T$ and $\mathbf{E}$ are uncorrelated, correlation matrix of observed data ($\mathbf{\Psi}$ is matrix of specific variances)

$$\mathbf{R} = \mathrm{N}^{-1}\mathbf{D}^{\mathbf{T}}\mathbf{D} = \mathrm{N}^{-1}\left(\mathbf{FL}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{FL}^{\mathrm{T}}\right) + \mathrm{N}^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{E} = \mathbf{LL}^{\mathrm{T}} + \mathbf{\Psi}$$

- Jöreskog's non-iterative approximate algorithm

$$\mathbf{L} = \left(\mathrm{diag}\,\mathbf{S}^{-1}\right)^{-1/2}\mathbf{\Omega}\left(\mathbf{\Gamma} - \theta\mathbf{I}\right)^{1/2}\mathbf{U}$$

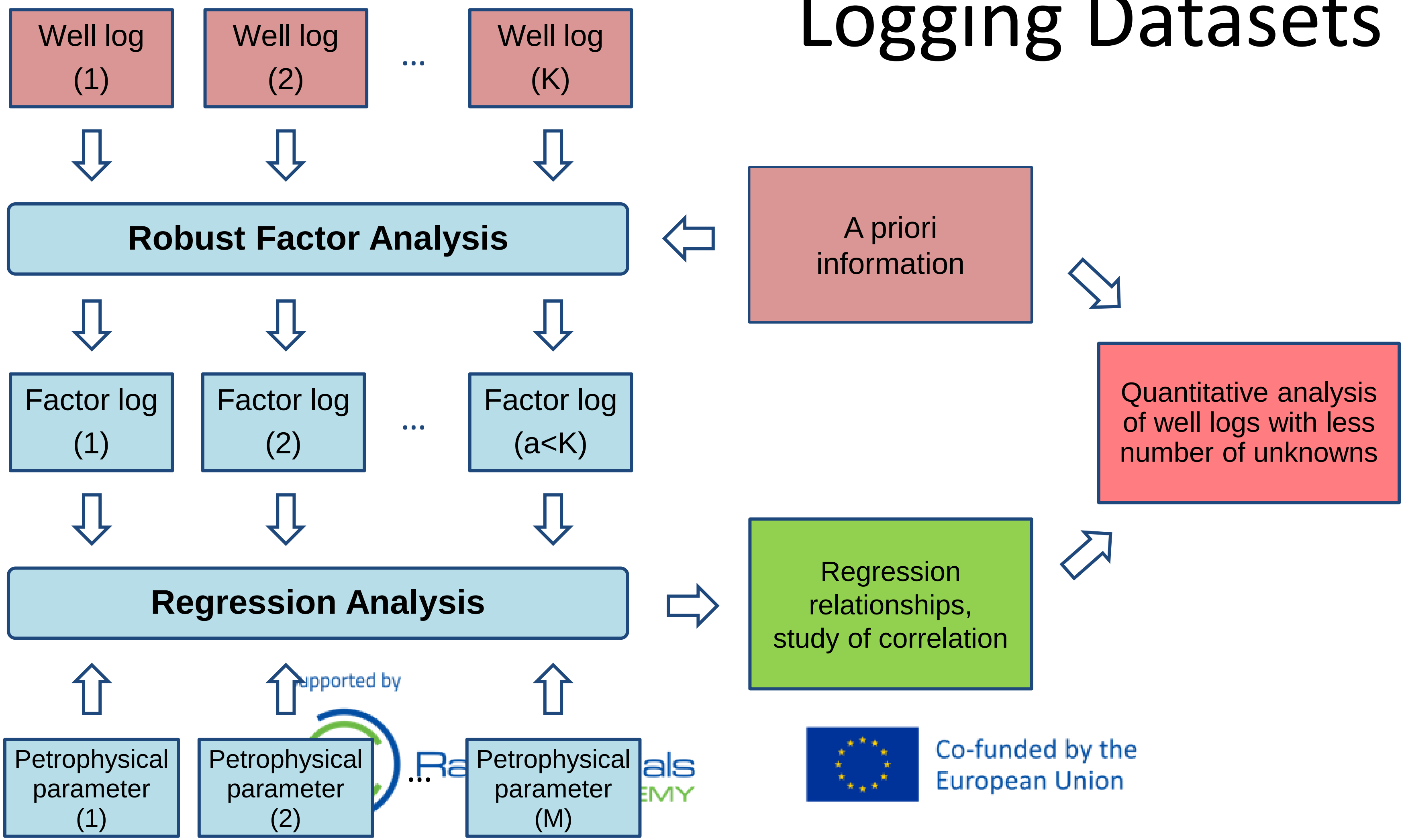
S: sample covariance matrix
Ω: matrix of eigenvectors
Γ: matrix of eigenvalues
U: arbitrary orthogonal matrix,
θ: constant for specifying factors

- Bartlett's hypothesis of linearity leads to an unbiased solution

$$P = -\left(\mathbf{D} - \mathbf{FL}^{\mathrm{T}}\right)^{\mathrm{T}}\mathbf{\Psi}^{-1}\left(\mathbf{D} - \mathbf{FL}^{\mathrm{T}}\right) = \max \quad \rightarrow \quad \mathbf{F} = \left(\mathbf{L}^{\mathrm{T}}\mathbf{\Psi}^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{\Psi}^{-1}\mathbf{D}$$

Supported by
RawMaterials ACADEMY
Co-funded by the European Union

Exploratory Factor Analysis of Well Logging Datasets



Supported by
Co-funded by the European Union

Supported by
RawMaterials ACADEMY
Co-funded by the European Union
DIM ESEE 2021

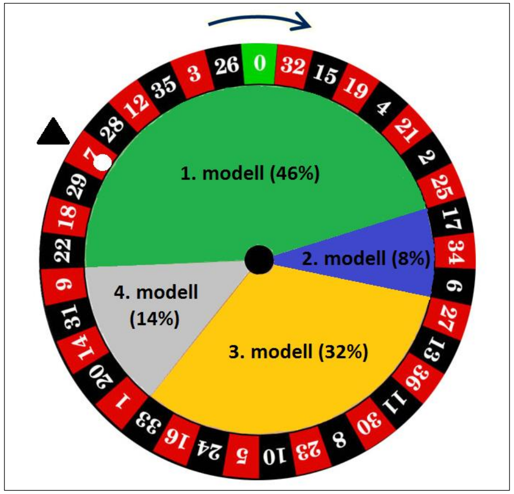
Genetic Algorithm (Machine Learning) Assisted Factor Analysis

- Fitness function is related to the data deviation vector

$$F = -\left\| \mathbf{d} - \tilde{\mathbf{L}}\mathbf{f} \right\|_2^2 = \max$$

- Probability of selecting the *i*-th factor score vector by geometric ranking selection

$$P\left(\mathbf{f}^{(i)}\right) = \frac{q}{1-(1-q)^S}(1-q)^{r_i - 1}$$



- Heuristic crossover gives an extrapolation of two individuals

$$\mathbf{f}^{(new,1)} = \mathbf{f}^{(old,1)} + \gamma\left(\mathbf{f}^{(old,1)} - \mathbf{f}^{(old,2)}\right)$$

$$\mathbf{f}^{(new,2)} = \mathbf{f}^{(old,1)}$$

- The ν-th factor score is randomly changed by uniform mutation

$$\mathbf{f}^{(new)} = \begin{cases} \eta, & \text{if} \quad \nu = h \\ f_\nu^{(old)}, & \text{otherwise} \end{cases}$$

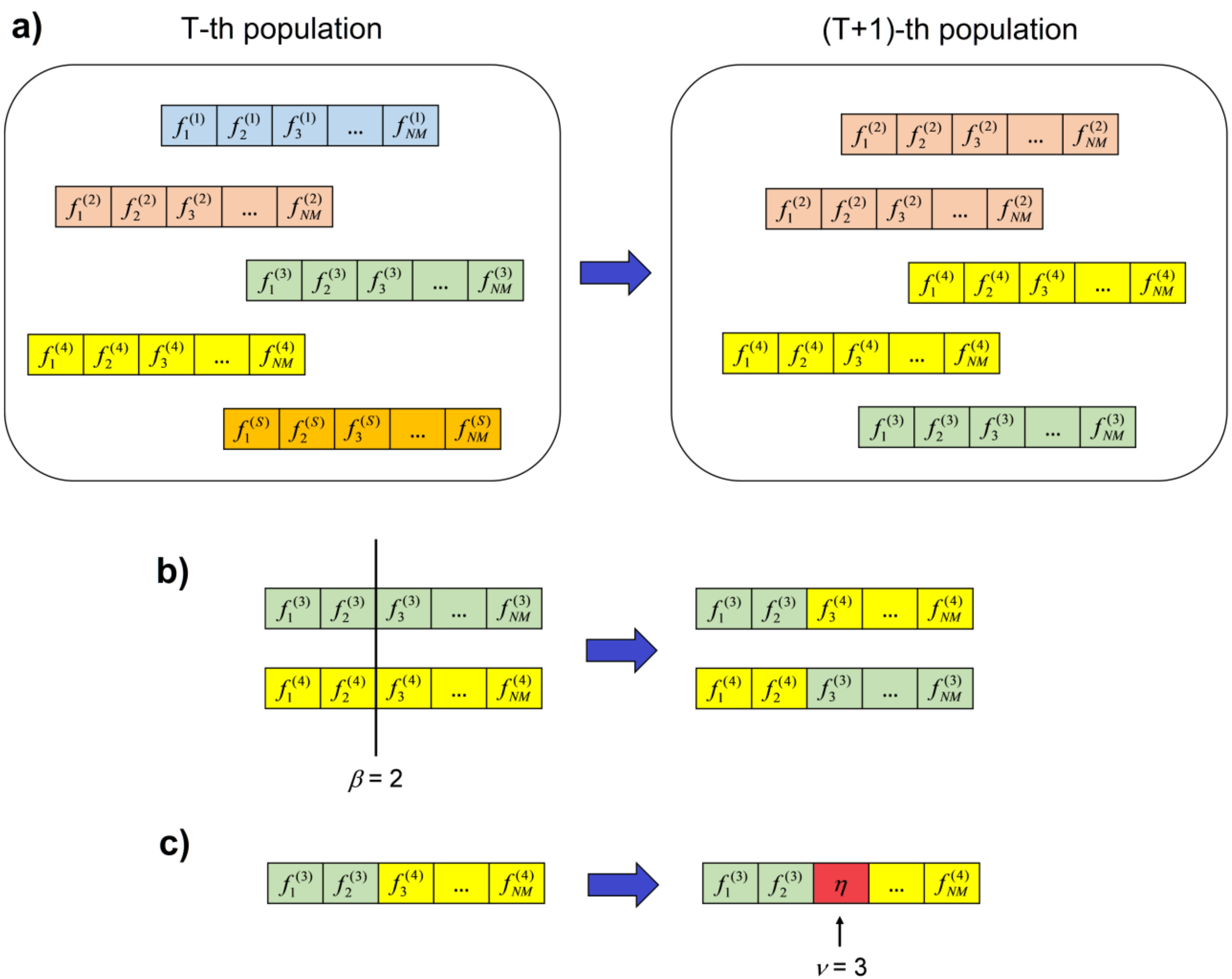
Supported by
RawMaterials ACADEMY
Co-funded by the European Union

Supported by

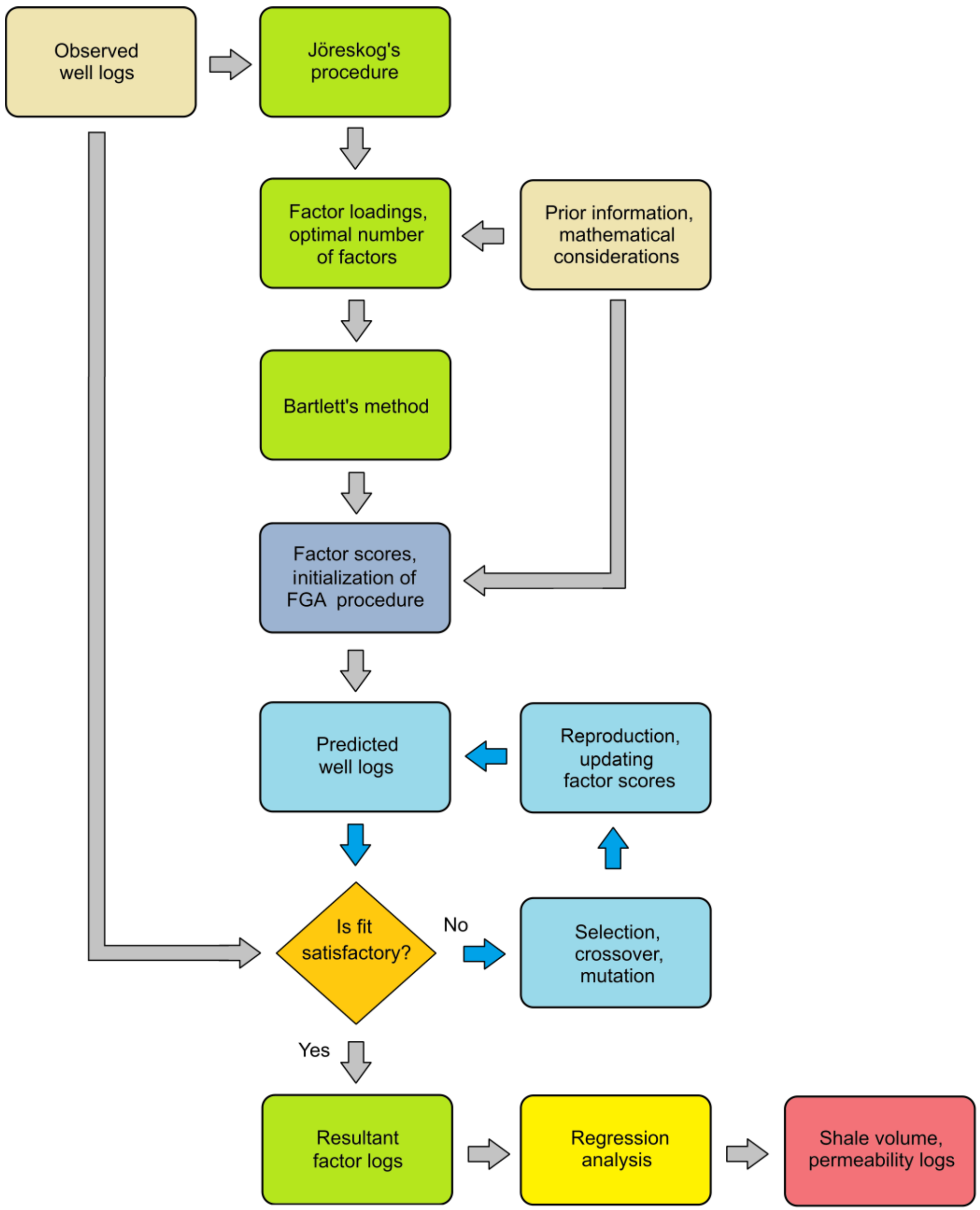
EIT RawMaterials Academy

Co-funded by the European Union

Genetic Algorithm (Machine Learning) Assisted Factor Analysis



Supported by



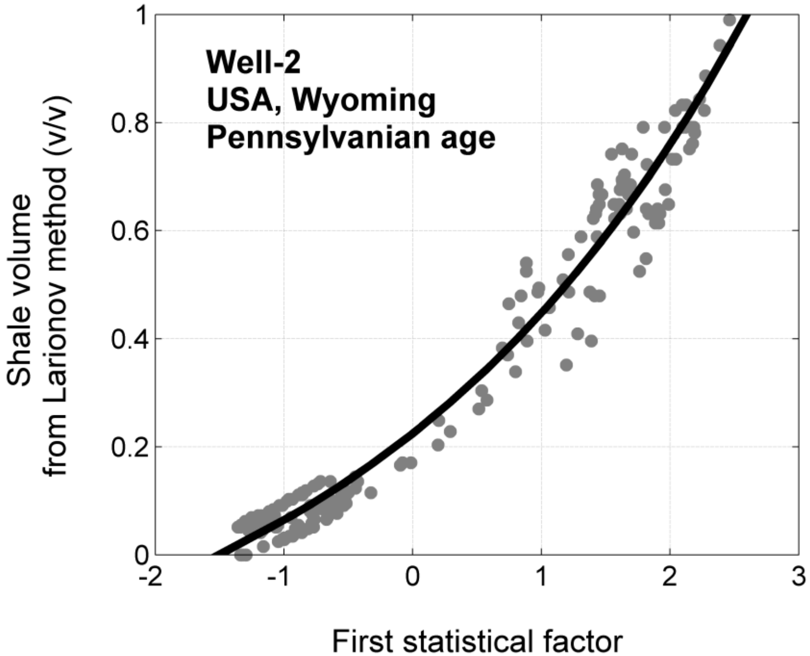
Genetic Algorithm (Machine Learning) Assisted Factor Analysis

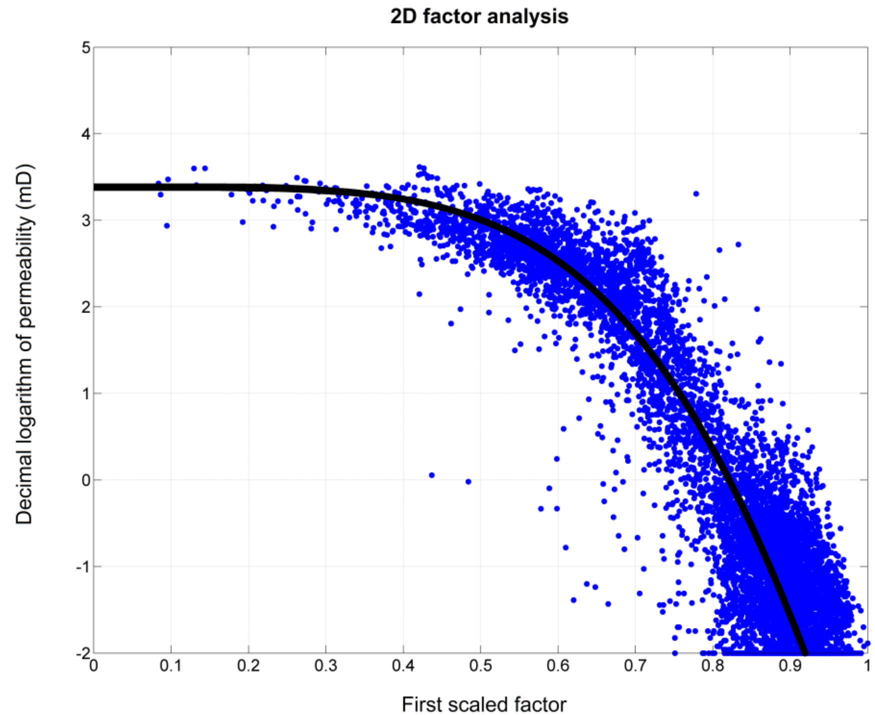
Co-funded by the European Union

Supported by

Co-funded by the European Union

Quantitative Estimation of Petrophysical Properties

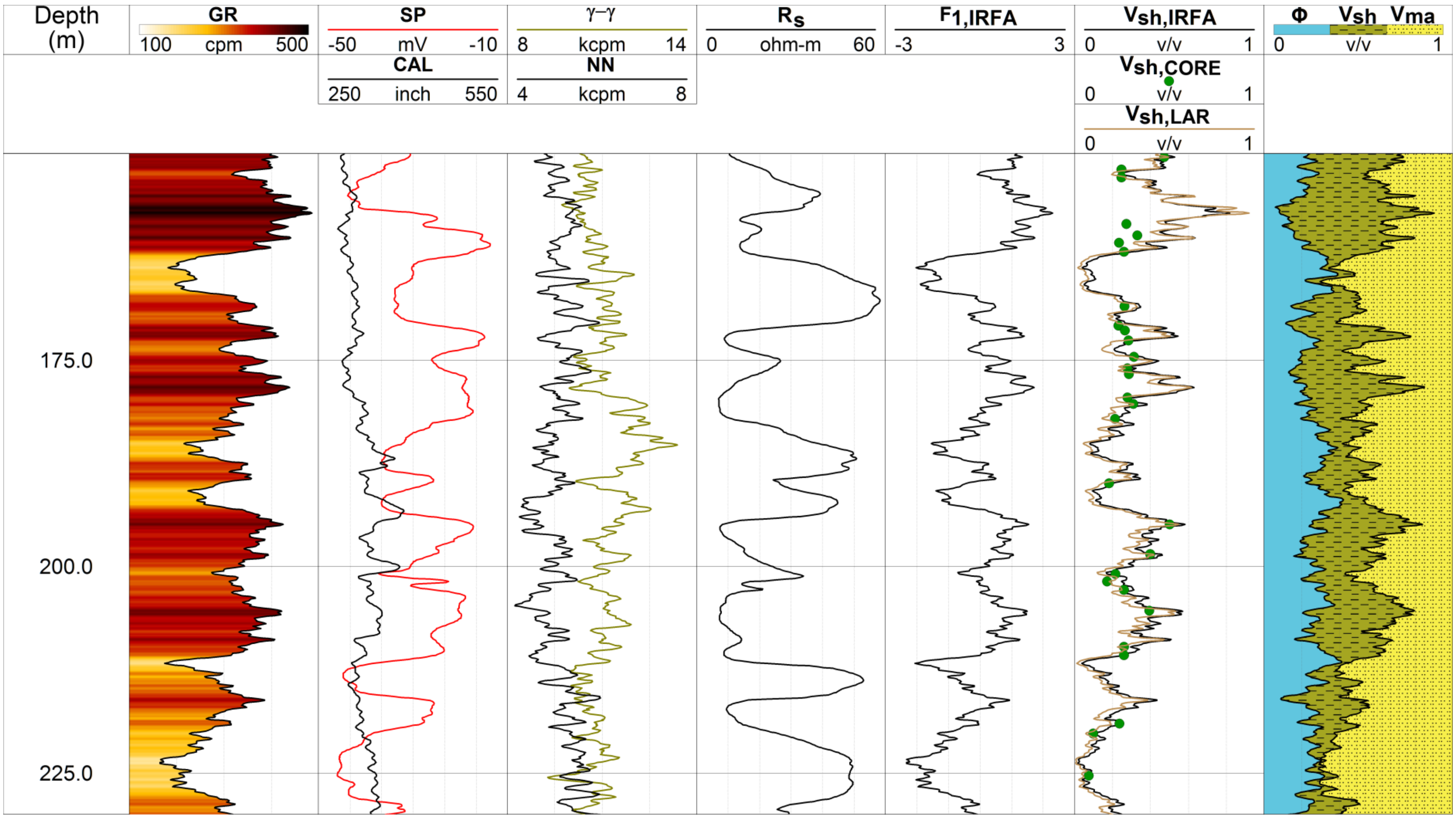




Supported by

Co-funded by the European Union

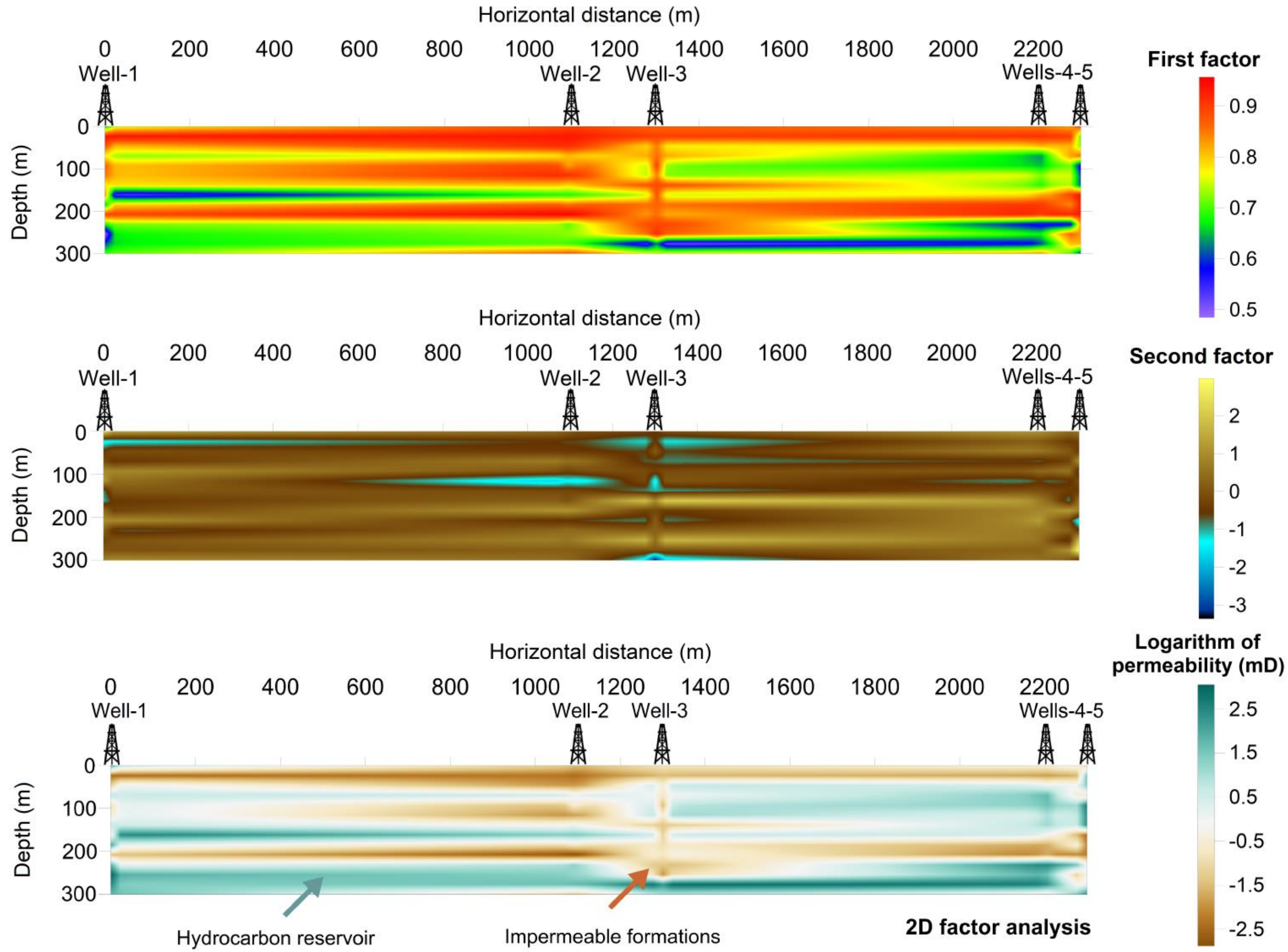
Shale Volume Estimation



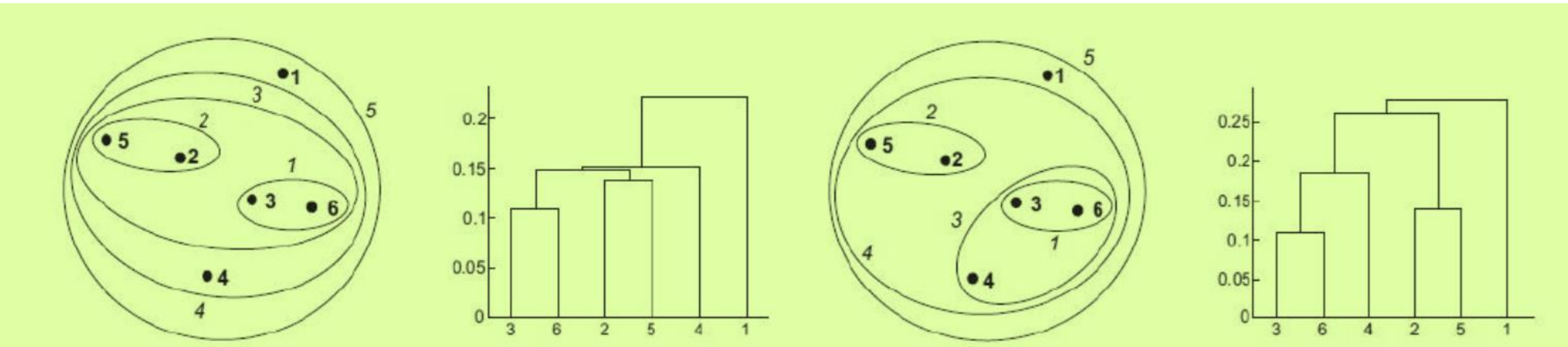
Supported by

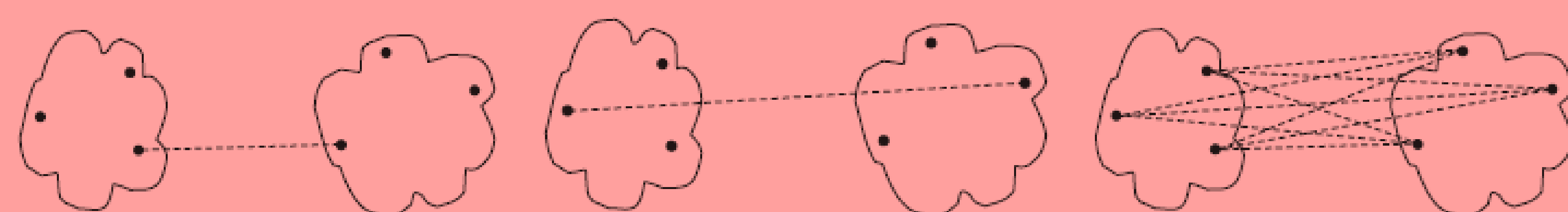
Co-funded by the European Union

Multidimensional Factor Analysis



Hierarchical Cluster Analysis





MIN (single link.) MAX (complete link.) Group average.

Supported by
RawMaterials ACADEMY
Co-funded by the European Union
DIM 2021 ESEE

Measure of Similarity

Let the vectors $\boldsymbol{x}^{(i)}$ *and* $\boldsymbol{x}^{(j)}$ denote two multivariate observations from a population with *p* random variables X_1, … X_p. In a more detailed form, the *i*-th and *j*-th observations are $\boldsymbol{x}^{(i)} = \left\{x_1^{(i)}, \dots, x_p^{(i)}\right\}^T$ and $\boldsymbol{x}^{(j)} = \left\{x_1^{(j)}, \dots, x_p^{(j)}\right\}^T$, which represent two so-called objects in the data space, respectively. In order to group the objects (or more objects) into clusters a measure for the similarity of elements needs to be defined. To determine the similarity between two objects, distance measures can be used. The TCA uses the Euclidean distance:

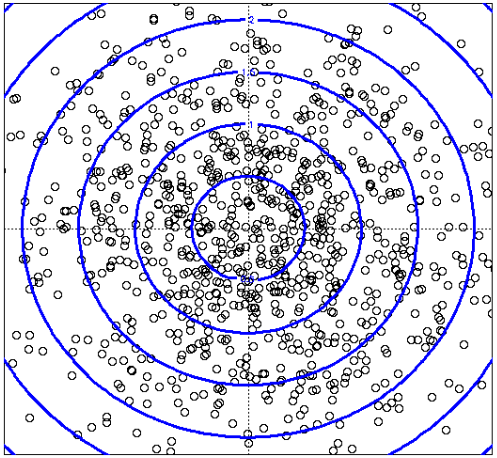
$$D\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \sqrt{\{(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})^T(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(i)})\}}.$$

By weighting it with the covariance matrix, we get the Mahalanobis distance:

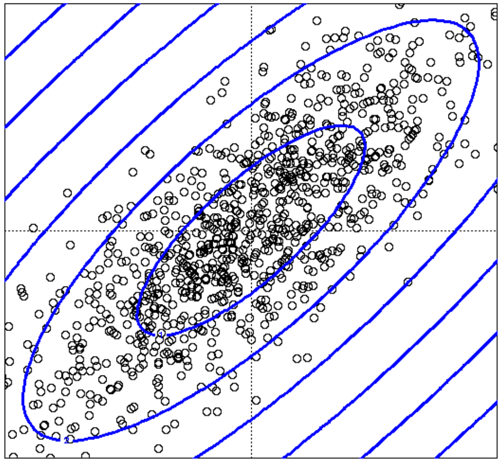
$$D\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \sqrt{\{(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})^T S^{-1}(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(i)}\}},$$

where $\mathbf{S} = \mathbf{C}^{\mathbf{T}}\mathbf{C}/(n-1)$ is the covariance matrix derived from the standardized data matrix **C**.

Contour plot of the Euclidian distance to the origin



Contour plot of the Mahalanobis distance to the origin



Supported by
RawMaterials ACADEMY

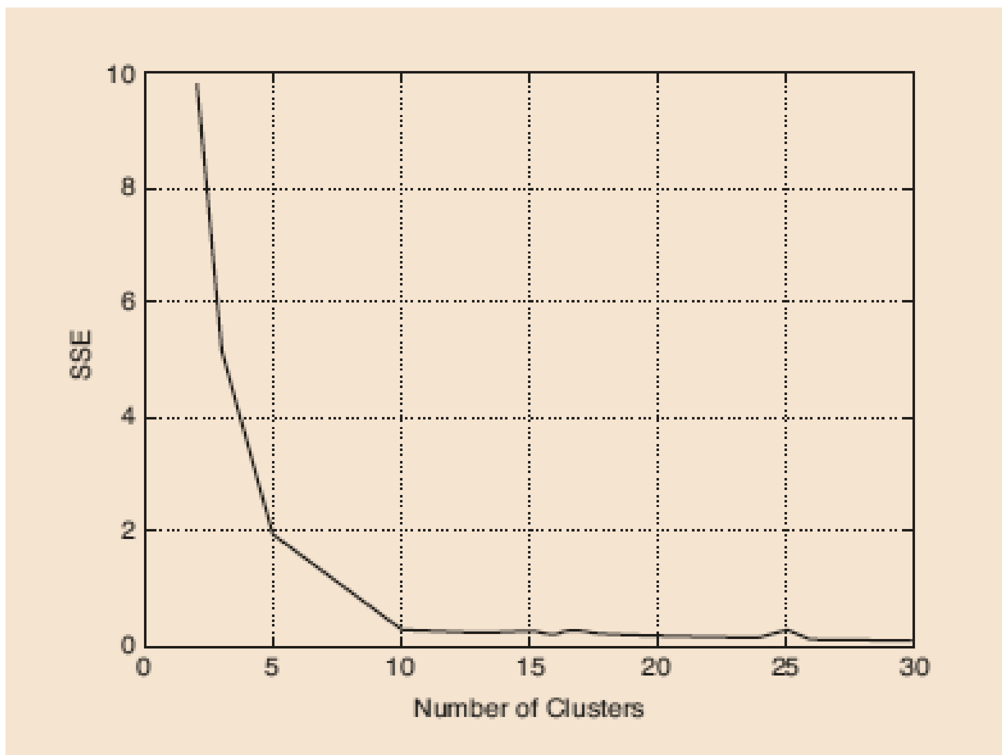
Supported by

Co-funded by the European Union

K-Means Clustering

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{n_i} d^2(c_i, x_j)$$

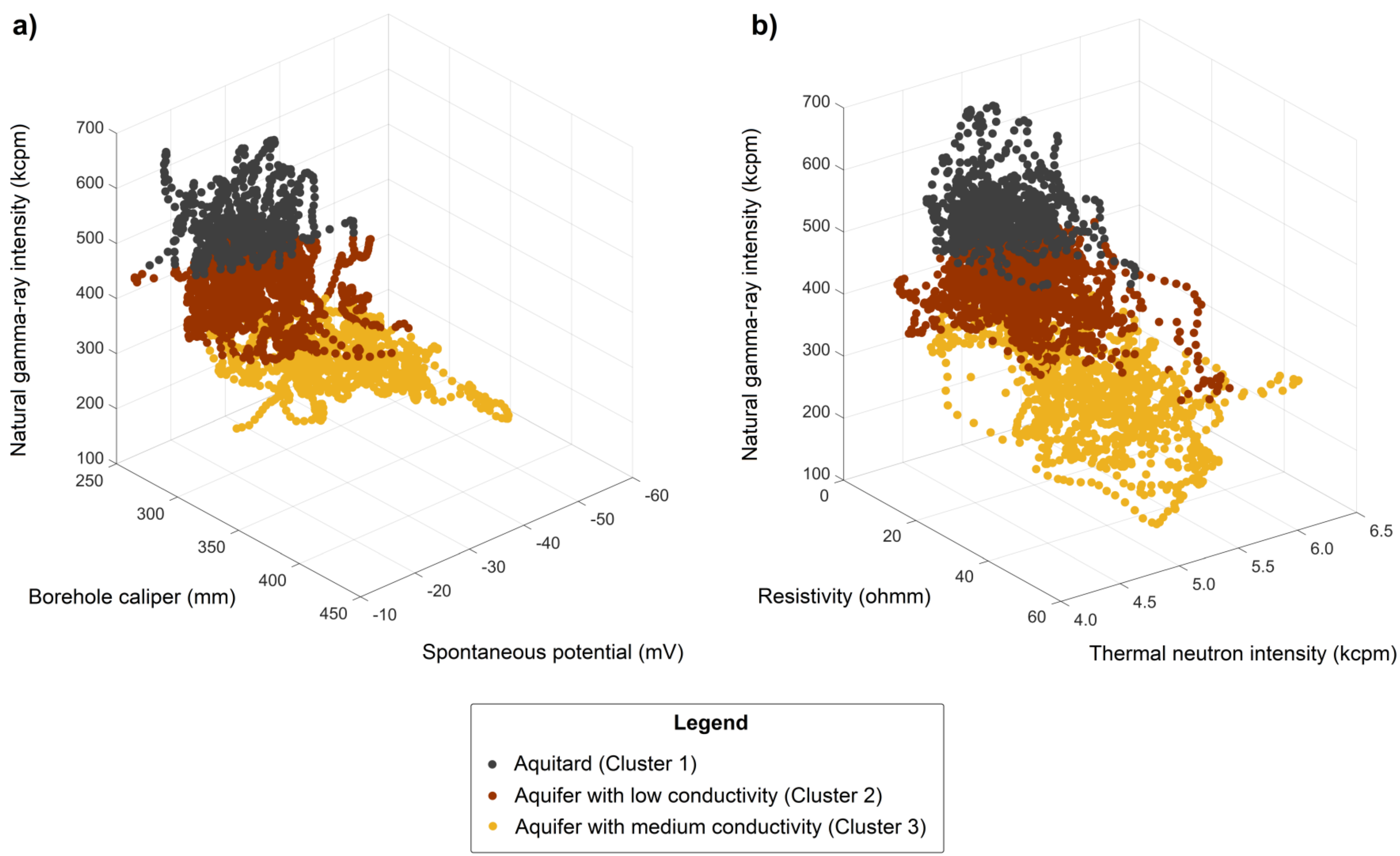
K is predefined number of clusters



Supported by

Co-funded by the European Union

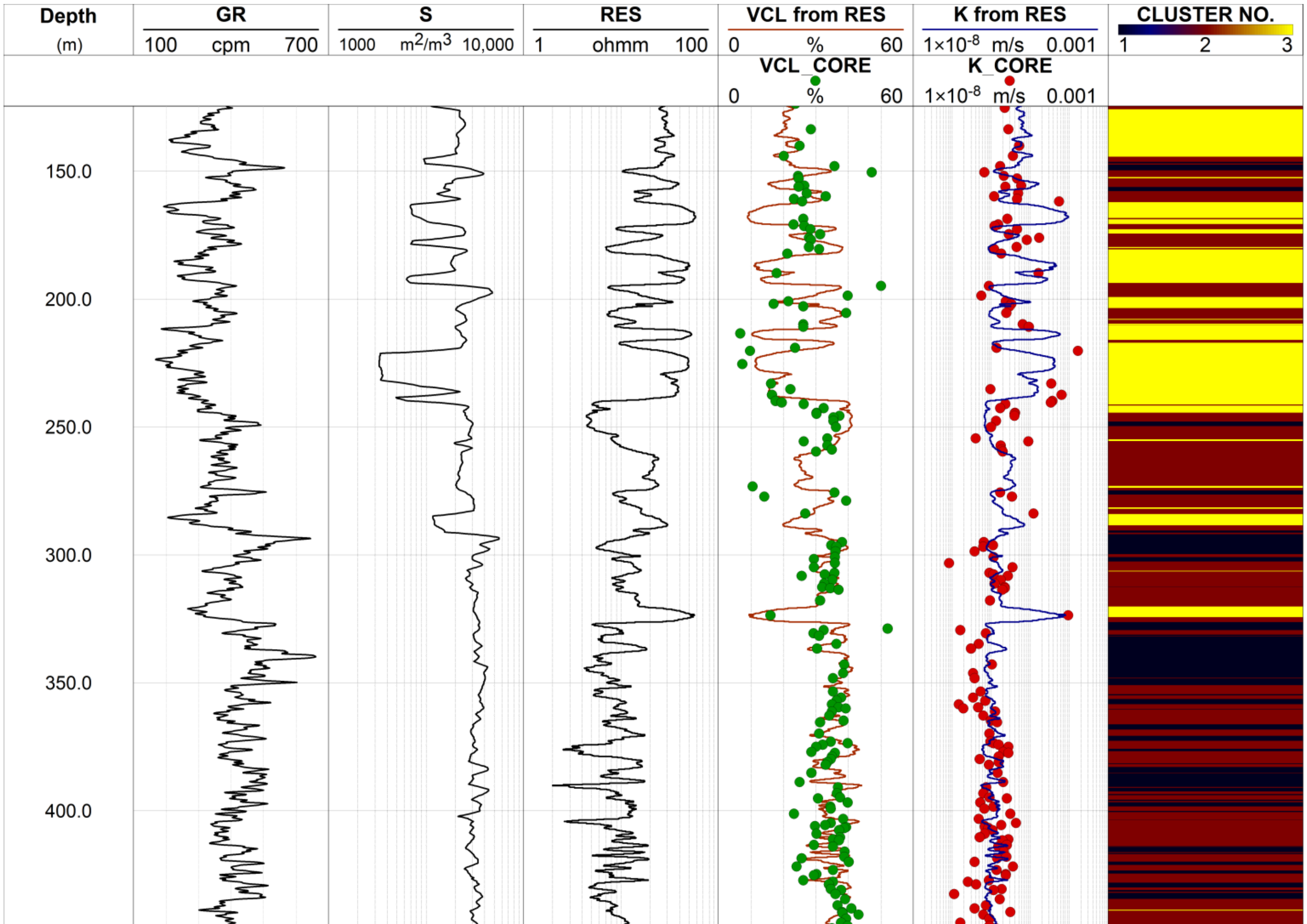
Hydrogeophysical Logging



Supported by

Co-funded by the European Union

Hydrogeophysical Logging



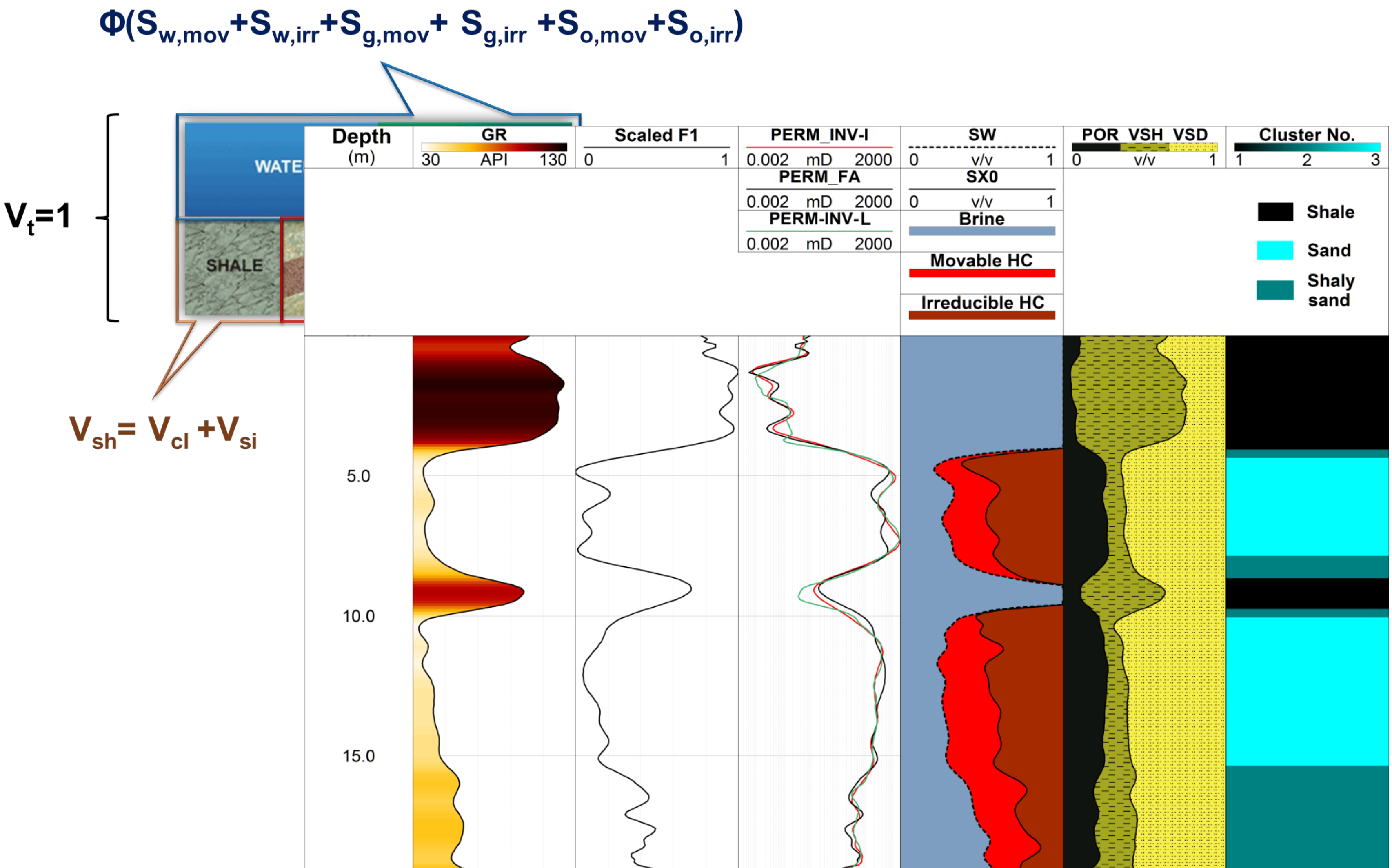
Supported by

RawMaterials ACADEMY

Co-funded by the European Union

2021 ESEE

Petrophysical Example



Supported by

Co-funded by the European Union

Replacement of Missing Data

a)

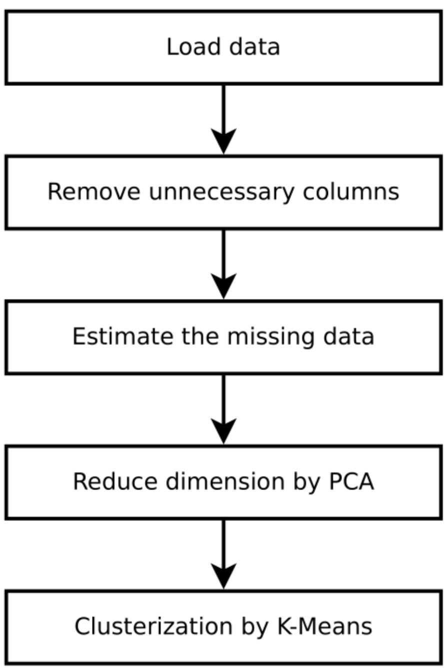
	x	y
0		
1	5	
2	2	7
3		3
4	1	3
5		
6	3	6
7		9

b)

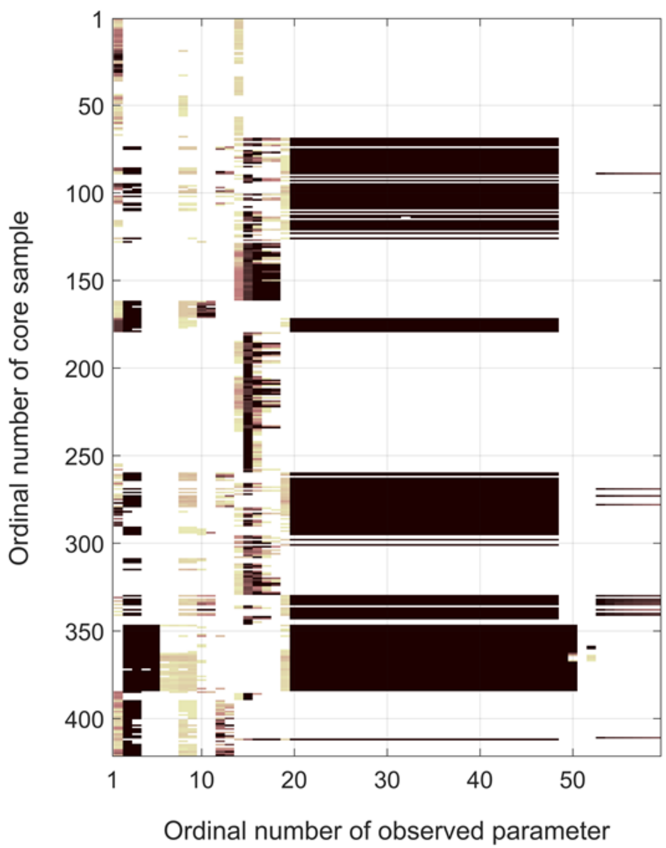
	x	y
0		
1	5	
2	**2**	**7**
3		3
4	**1**	**3**
5		
6	**3**	**6**
7		9

c)

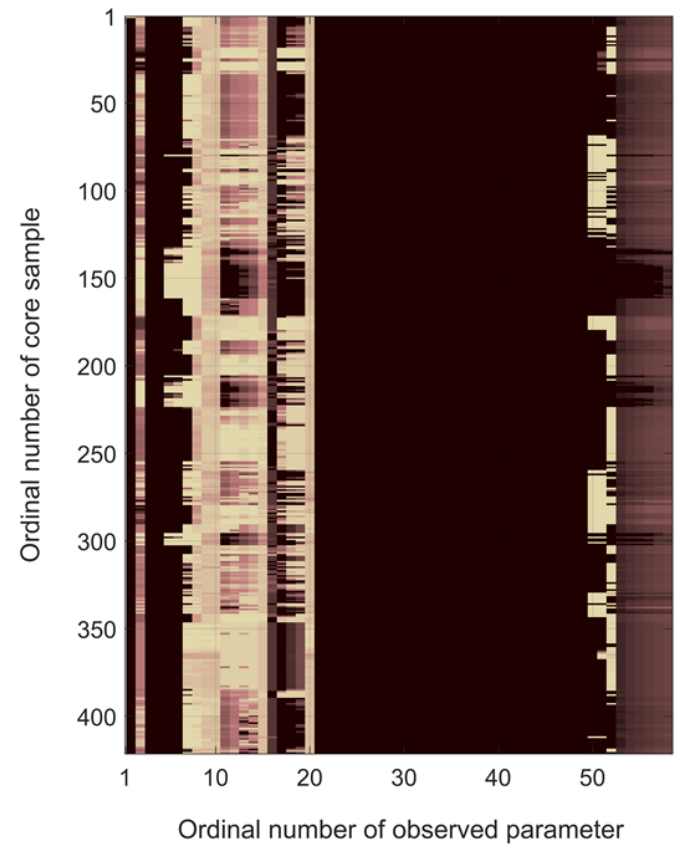
	x	y
0		
1	5	**9.83**
2	**2**	**7**
3	**1.19**	3
4	**1**	**3**
5		
6	**3**	**6**
7	**3.27**	9



a)



b)

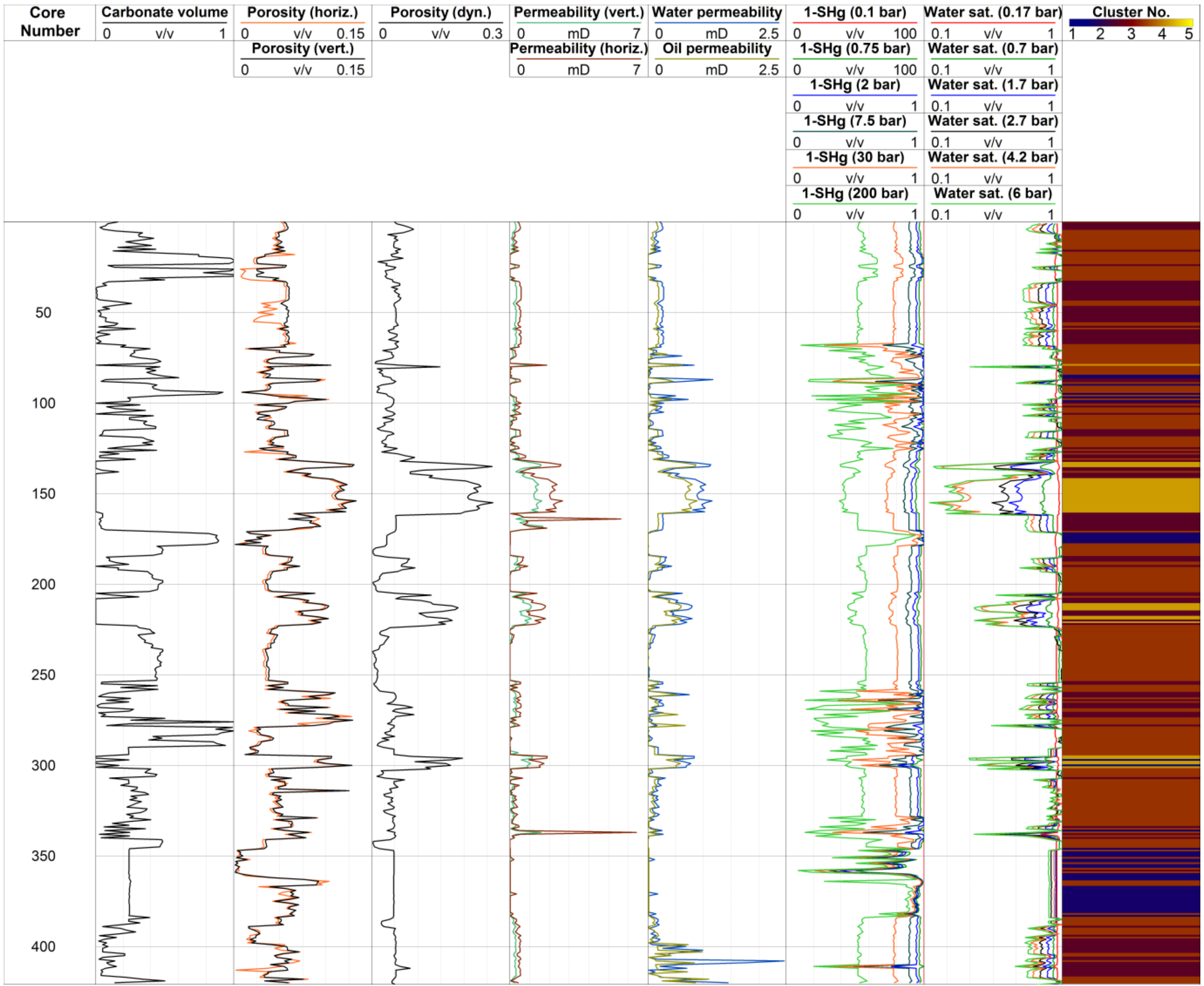


Supported by

Co-funded by the European Union

2021 ESEE

Core Data Based Rock Typing



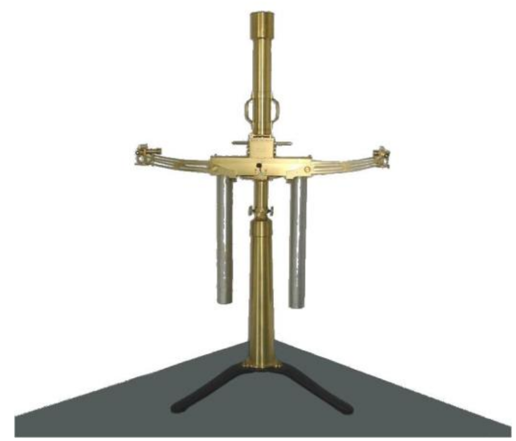
Supported by

Co-funded by the European Union

Thank you for your attention.

Prof. Dr. Norbert Péter Szabó
Institute of Geophysics and Geoinformation Science,
University of Miskolc

- info@dim-esee.eu
- gfnmail@uni-miskolc.hu



Supported by

Co-funded by the European Union