How to make Integral Indicators using Data and Expert Estimations

Vadim Strijov
Computing Center of the Russian Academy of Sciences

1-Sept-2010, OR2010Munich
10-Sept-2010, Universität Siegen

Russian Academy of Sciences

joins the network of scientific research institutes from across the Russian Federation as well as scientific and social units.

- **Founded in 1724 by decree of Emperior Peter I the Great**
- **Now**
  - 470 institutions
  - 55,000 researchers
  - 16 Nobel laureates
- Section of Applied Mathematics and Informatics,
  - **Computing Center**

Computing Center of RAS

- **Founded in 1955**
- **Fields of the scientific research**
  - computational methods
  - mathematical modeling
  - mathematical methods of pattern recognition
- **276 researchers**
  - **8** academicians and corresponded members of RAS
  - **75** researchers have DSc degree
  - **136** researchers have PhD degree

Data mining

Machine learning
Multivariate statistics

is a collection of methods for extracting

- unexplored,
- nontrivial,
- useful,
- and interpretable

patterns, models and facts from the data.

Data mining is important to support decisions in various fields of science, economics and finance.

Non-supervised learning

- Clustering
- Principal Component Analysis
- Visualizing

our decisions are based only on

mathematical models

we have no

expert opinions / historical data

Supervised learning

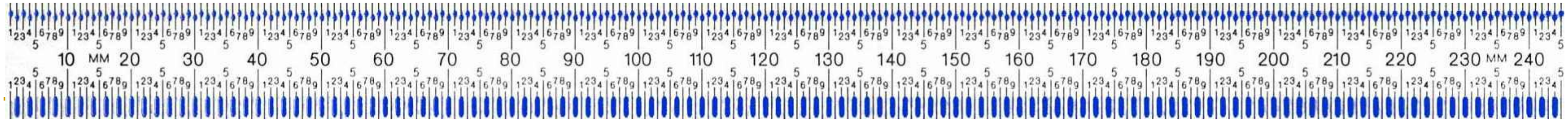
- Regression / Forecasting
- Classification / Scoring
- Model parameter estimation

we have

1) **mathematical models**,
2) **expert opinions** / historical data

What is the Integral Indicator?

- The **integral indicator** is a **measure** of object's quality.
  - It is a scalar, corresponded to an object.
- The **integral indicator** is an **aggregation** of object's features that describe various components of the term "quality".
  - Expert estimation of object's quality could be an integral indicator, too.



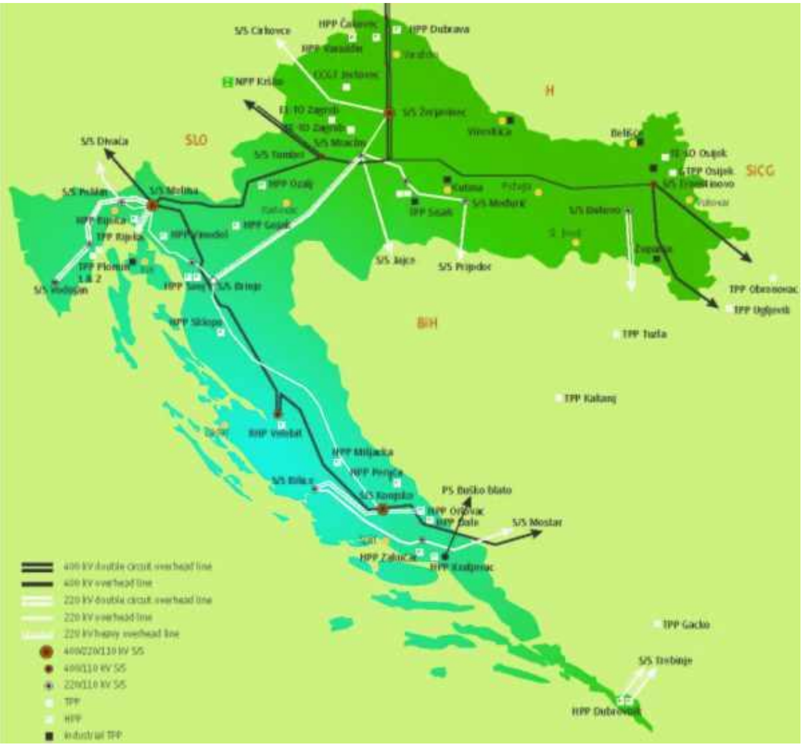
Examples

Integral Indicator	**Objects**	**Features**	**Model**
TOEFL exams	Students	Tests	Sum of scores
Eurovision	Singers	Televotes, Jury votes	Linear (weighted sum)
S&P500, NASDAQ	Time-ticks	Shares (prices, volumes)	Non-linear
Bank ratings	Banks	Requirements	By an expert commission
Integral Indicator of Thermal PP's	**Thermal Power Plants**	**Waste measurements**	**Linear**

There is a set of objects

- Croatian Thermal Power Plants and Combined Heat and Power Plants

1. Plomin 1 TPP
2. Plomin 2 TPP
3. Rijeka TPP
4. Sisak TPP
5. TE-TO Zagreb CHP
6. EL-TO Zagreb CHP
7. TE-TO Osijek CHP
8. *Jetrovac TPP*



There is a set of features

- Outcomes and Waste measurements

1. Electricity (GWh)
2. Heat (TJ)
3. Available net capacity (MW)
4. SO_2 (t)
5. NO_X (t)
6. Particles (t)
7. CO_2 (kt)
8. Coal (kt)
9. Sulphur content in coal (%)
10. Liquid fuel (kt)
11. Sulphur content in liquid fuel (%)
12. Natural gas (10^6 m^3)

How to construct an Integral Indicator?

1. Assign a comparison criterion
   Ecological footprint of the Croatian Power Plants
2. Gather a set of comparable objects
   TPP and CHP (Jertovec TPP excluded)
3. Gather features of the objects
   Waste measurements
4. Make a data table: objects/features
   See 7 objects and 10 features in the table below
5. Select a model
   Linear model (with most informative coefficients)

Data table and feature optimums

The criterion is: **the Ecological Footprint of a Power Plant**

N	Power Plant	Electricity (GWh)	Heat (TJ)	Available net capacity (MW)	SO_2 (t)	NO_x (t)	Particles (t)	CO_2 (kt)	Coal (kt)	Sulphur content in coal (%)	Liquid fuel (kt)	Sulphur content in liquid fuel (%)	Natural gas (10^6 m^3)
1	Plomin 1 TPP	**452**	**0**	98	1950	1378	140	454	198	0.54	0.43	0.2	0
2	Plomin 2 TPP	**1576**	**0**	192	581	1434	60	1458	637	0.54	0.37	0.2	0
3	Rijeka TPP	**825**	**0**	303	6392	1240	171	616	0	0	200	2.2	0
4	Sisak TPP	**741**	**0**	396	3592	1049	255	573	0	0	112	1.79	121
5	TE-TO Zagreb CHP	**1374**	**481**	337	2829	705	25	825	0	0	80	1.83	309
6	EL-TO Zagreb CHP	**333**	**332**	90	1259	900	19	355	0	0	39	2.1	126
7	TE-TO Osijek CHP	**114**	**115**	42	1062	320	35	160	0	0	37	1.1	24
				max	**min**	**min**	**min**	**min**	**min**	**min**	**min**	**min**	**min**

Each feature has its own optimal value (min, max)

Notations

$A=\{a_{ij}\}$ – $(n \times m)$ real matrix, **data set**,

$\mathbf{q} = [q_1, \ldots, q_m]^T$ – vector of integral indicators,

$\mathbf{w} = [w_1, \ldots, w_n]^T$ – vector of feature importance weights,

$\mathbf{q}_0$, $\mathbf{w}_0$ – **expert estimations of indicators and weights**.

		w=			
		w_1	w_2	…	w_n
	q_1	a_{11}	a_{12}	…	a_{1n}
q=	q_2	a_{21}	a_{22}	…	a_{2n}
	…	…	…	…	
	q_m	a_{m1}	a_{m2}	…	a_{mn}

Data preparation

Convert data to the comparable scales,

$$a_{ij} \mapsto (-1)^{s_j} \frac{a_{ij} - \min\limits_i(a_{ij})}{\max\limits_i(a_{ij}) - \min\limits_i(a_{ij})} + s_j.$$

And put it to the principle "*the bigger the better*":

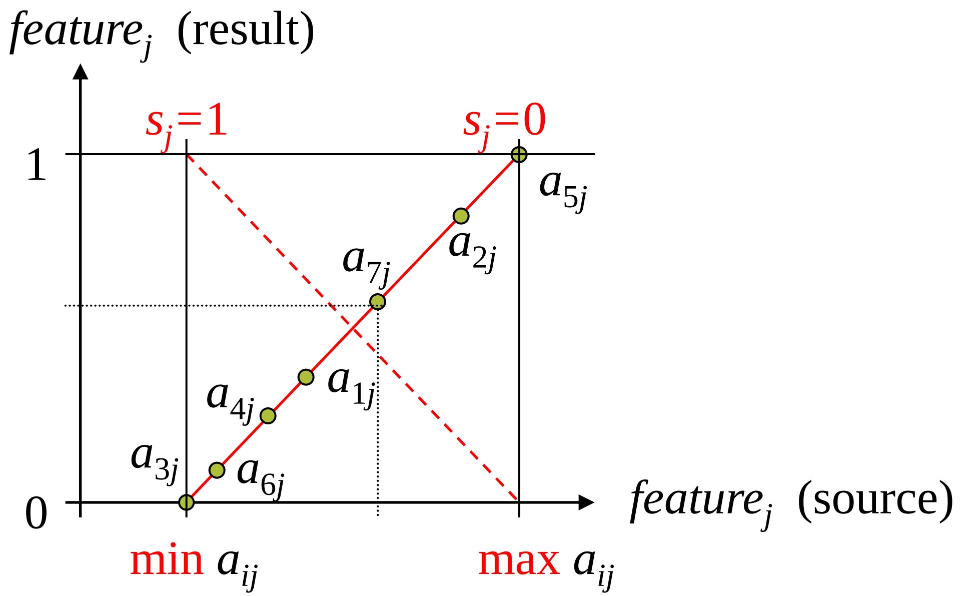
$s_j = 0$, if the desired value of j-th feature is **max**;

$s_j = 1$, if the desired value is **min**.

Usually, data prepared so that

1. the minimum of each feature equals 0, while the maximum equals 1;
2. the bigger value of each implies better quality of the integral indicator.

Data preparation, explanation



"*The bigger the better*" principle:
greater value of *i*-th object, given feature, involves greater value of the integral indicator for this object.

The algorithms

1. Pareto-Slicing
2. Metric Algorithms
3. Weighted Sum*
4. Principal Components Analysis
5. Expert-Statistical Technique*
6. Linear/Ordinal Specification*

---

* Expert estimations required

Integral indicators and expert estimations

There are lot of ways to construct integral indicators. However, when algorithms are chosen and some results obtained, the following question arises:

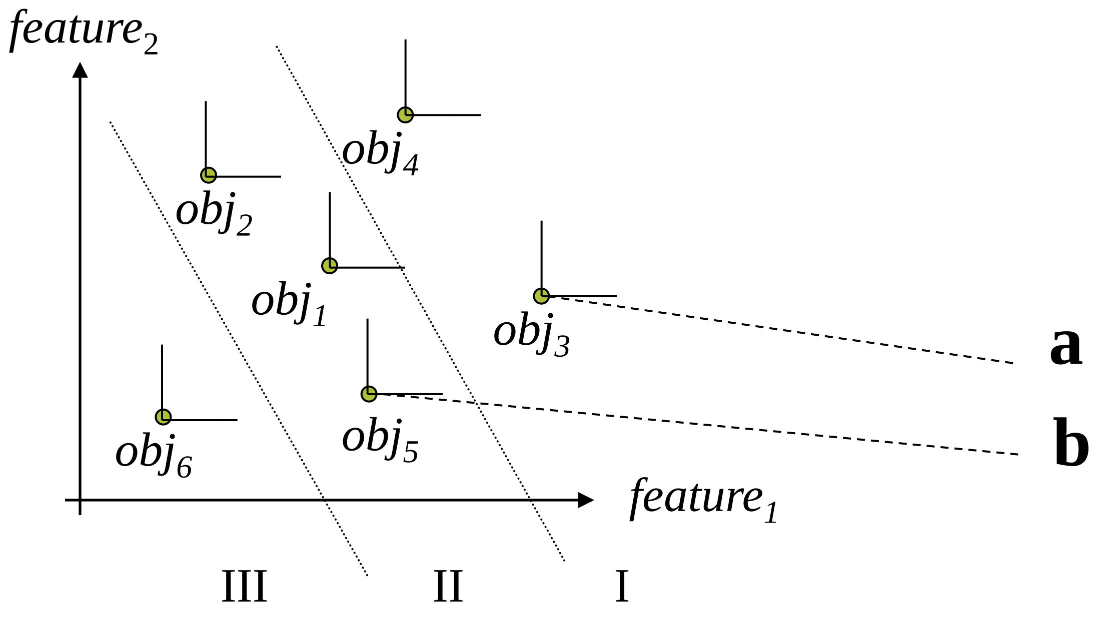
- **How to show adequacy of the calculated integral indicators?**

To answer the question analysts invite experts. The experts express their opinion and then the second question arises:

- **How to show that expert estimations are valid?**

The first method, Pareto slicing

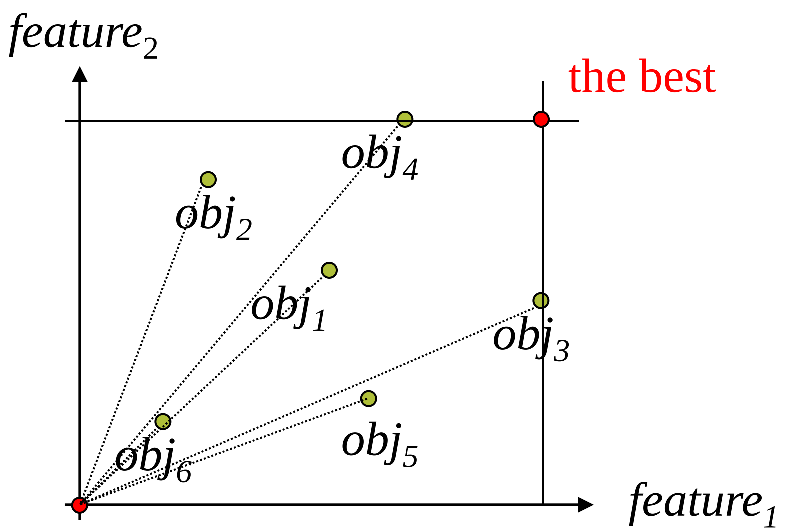
Find non-dominated objects at each slicing level.



The object **a** is non-dominated if there is no $\mathbf{b}_i$ such that $b_{ij} \geq a_i$ for all features j.

The second method, Metric algorithm

The worst (best) object is an object that contains the minimal (maximal) values of the features.



$$q_i = \sqrt[r]{\sum_{j=1}^{n} (a_{ij} - a_j^{worst})^r}$$

For $r = 1$, this algorithm coincides the weighted sum with equal weighs.

Supervised way-1,

the Weighted sum

$$\mathbf{q}_1 = A\,\mathbf{w}_0,$$

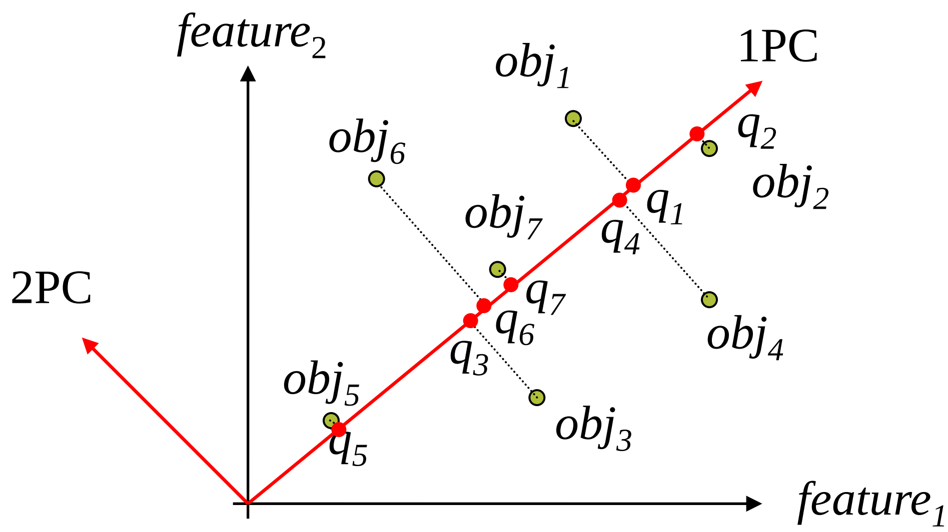
$$\begin{pmatrix} q_1 \\ \vdots \\ q_m \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}.$$

Unsupervised way,

Principal Components Analysis

$Q=AW$, where W—rotation matrix of the principal components.

$\mathbf{q}_{\mathrm{PCA}}=A\mathbf{w}_{\mathrm{1PC}}$, where $\mathbf{w}_{\mathrm{1PC}}$ is the 1st column vector of W.



PCA gives minimal mean square error between objects and their projections.

useful tool for PCA

$$A = ULW^T$$

$$A^T A = WLU^T ULW^T$$

$$A^T AW = WL^2$$

Supervised way-2,

the Expert-Statistical Technique

$$\mathbf{w}_1 = \arg\min \|\mathbf{q}_0 - A\,\mathbf{w}\|^2,$$

least squares, $\mathbf{w}_1 = (A^T A)^{-1} A^T \mathbf{q}_0.$

The problem of specification

- **We have**

the data table A,

expert estimations $\mathbf{q}_0$, $\mathbf{w}_0$,

calculated weights and indices $\mathbf{q}_1$, $\mathbf{w}_1$.

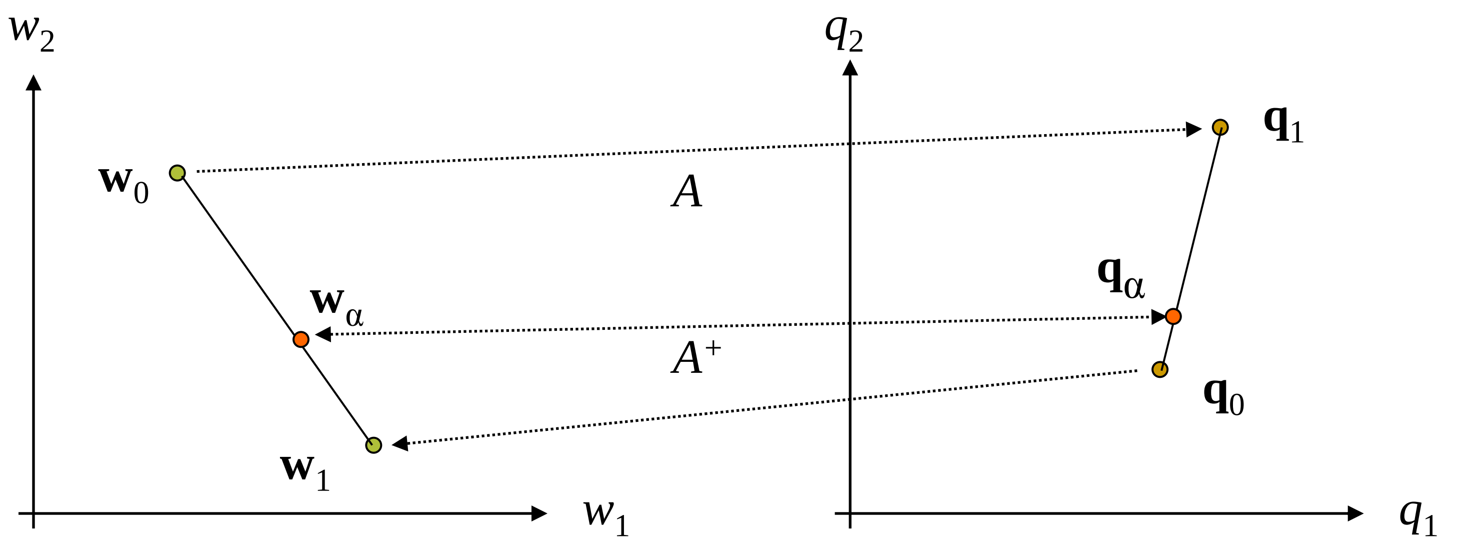
- **Contradiction**

$$\textbf{Neither } \mathbf{q}_0 \neq A\mathbf{w}_0, \textbf{ nor } \mathbf{w}_0 \neq A^{+}\mathbf{q}_0.$$

Calculated indices are not the same as the expert estimations for the indices;

as well, calculated weights are not the same as the expert estimations of the weights.

Linear specification



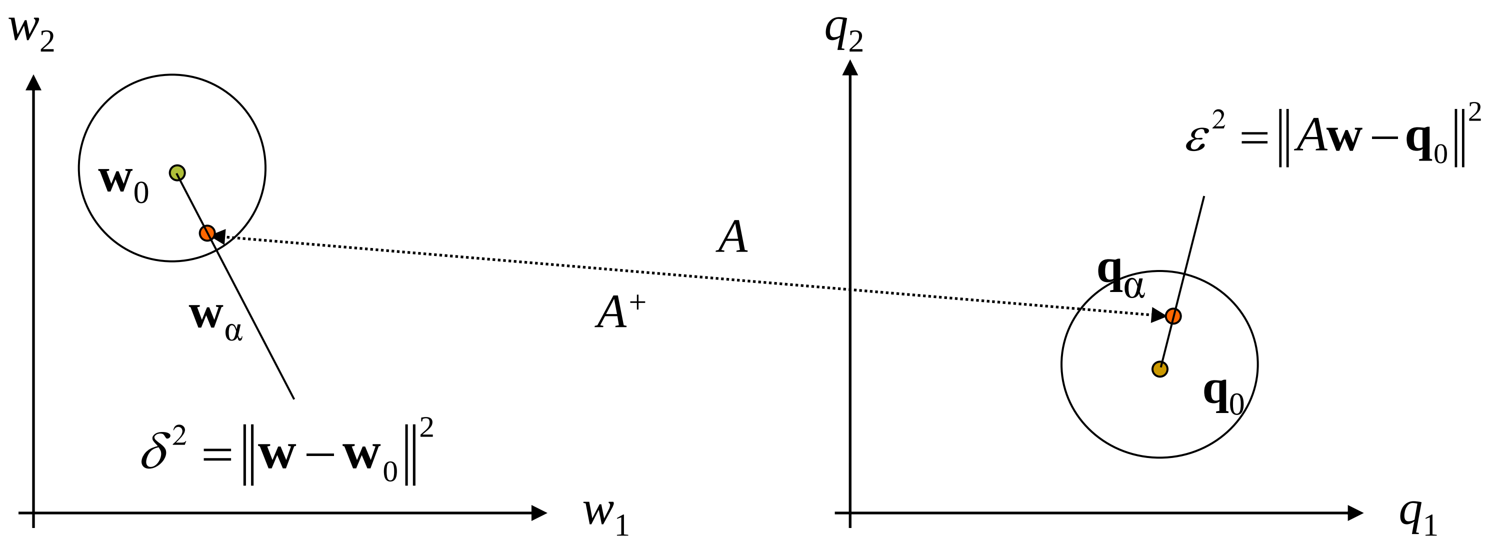
$$\mathbf{w}_\alpha = \alpha A^+\mathbf{q}_0 + (\mathbf{1}\text{-}\alpha)\mathbf{w}_0, \qquad \mathbf{q}_\alpha = (1\text{-}\alpha)A\mathbf{w}_0 + \alpha\mathbf{q}_0.$$

Parameter α is in [0,1].

α = 0, we trust expert estimations of the weights,

α = 1, we trust expert estimations of the indices.

Quadratic specification

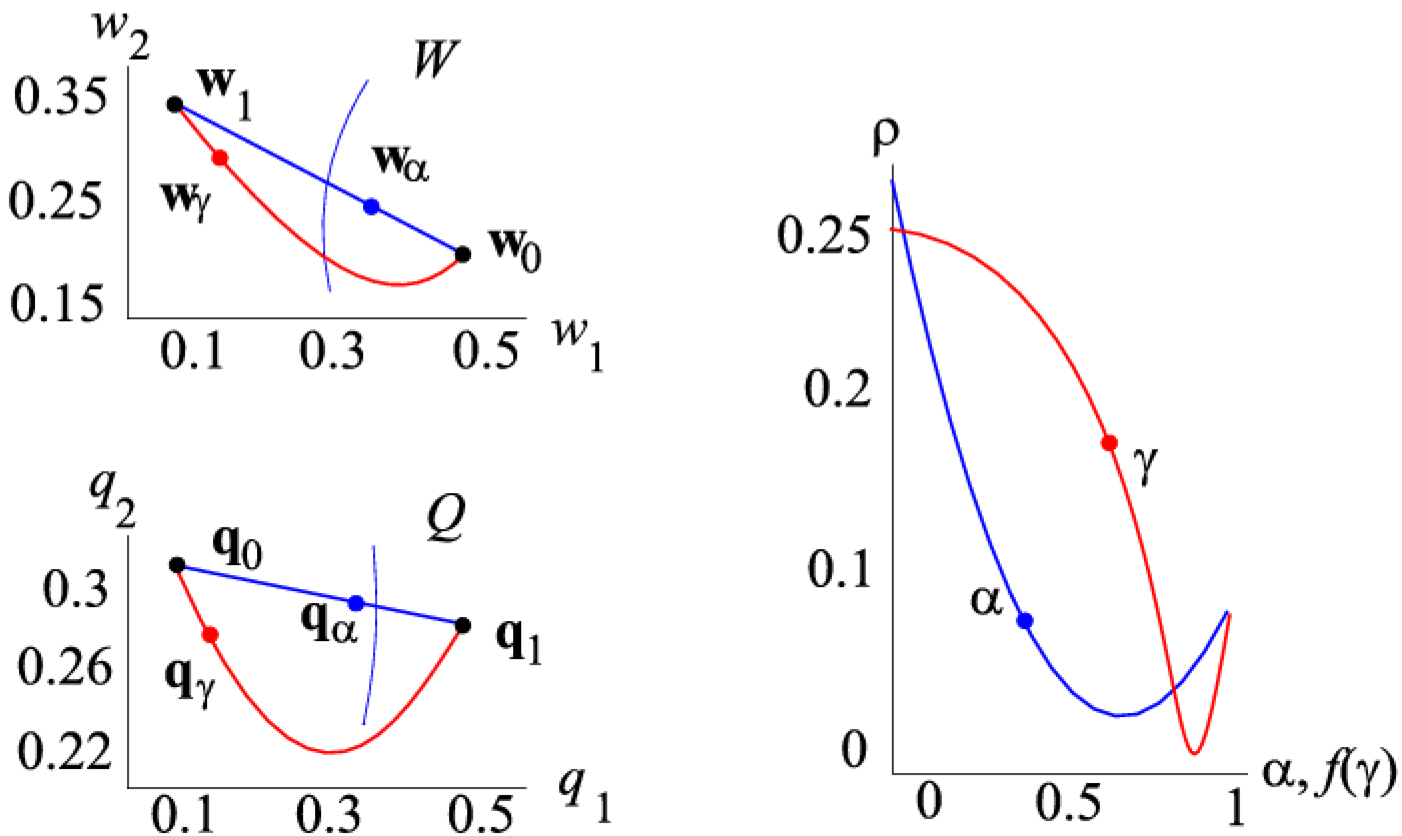


$$\mathbf{w}_\gamma = \arg\min_{\mathbf{w}\in W}(\varepsilon^2 - \gamma^2\delta^2), \quad \mathbf{w}_\gamma = (A^T A + \gamma^2 I)^{-1}(A^T\mathbf{q}_0 + \gamma^2\mathbf{w}_0).$$

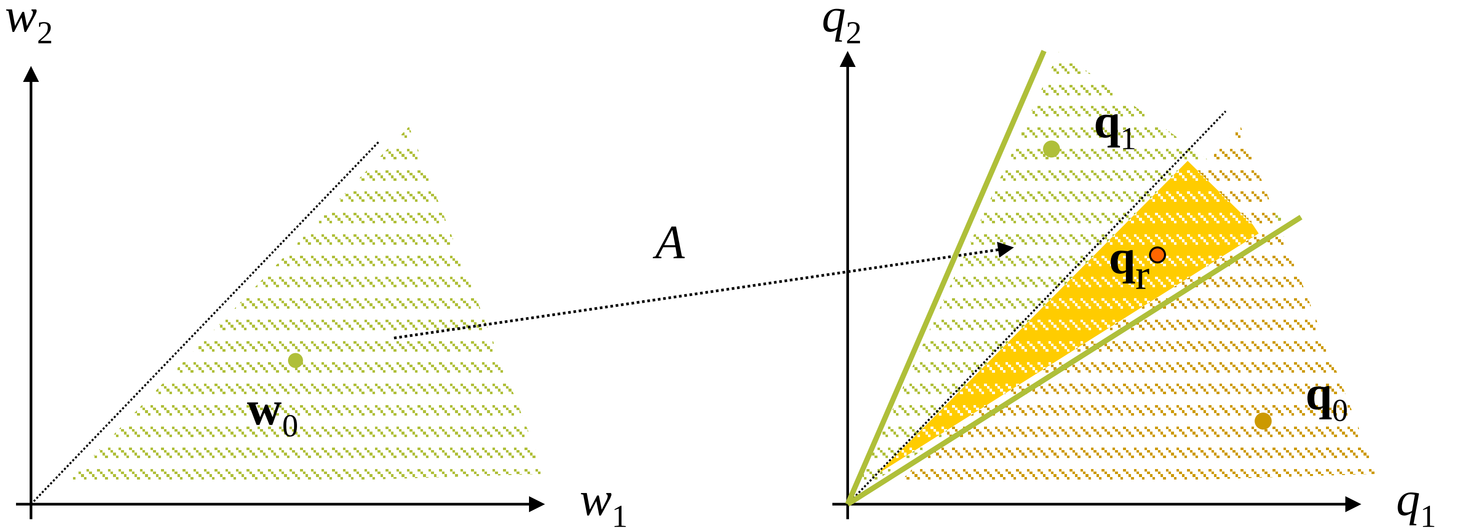
If parameter γ^2 is 0, then we trust expert estimations of the indices.

Comparison of the methods,

what is the difference?



Ordinal specification



$$\mathbf{w}_0 = [w_1 \geq w_2 \geq \ldots \geq w_n \geq 0]^T, \mathbf{q}_0 = [q_1 \geq q_2 \geq \ldots \geq q_m \geq 0]^T.$$

Rank-scaled expert estimations

$$\mathbf{w}_0 = [w_1 \geq w_2 \geq \ldots \geq w_n \geq 0]^T, \mathbf{q}_0 = [q_1 \geq q_2 \geq \ldots \geq q_m \geq 0]^T.$$

$$Q_0 = \{\mathbf{q}_0 \mid J_m \mathbf{q}_0 \geq \mathbf{0}\},$$

$$W_0 = \{\mathbf{w}_0 \mid J_n \mathbf{w}_0 \geq \mathbf{0}\}.$$

$$J = \begin{pmatrix} 1 & -1 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

The cones intersection exists

$$\mathbf{q}_1 \in AW_0 \bigcap Q_0,$$

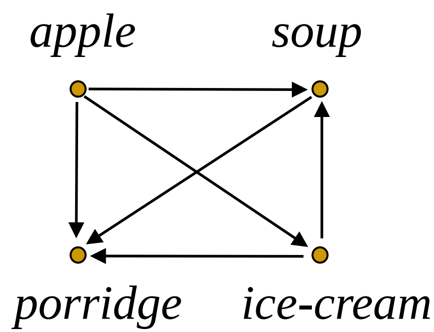
or not, then specify

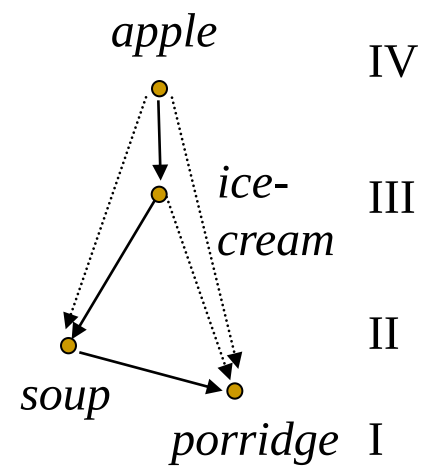
$$\mathbf{q}_\alpha = (1-\alpha)A\mathbf{w}' + \alpha\mathbf{q}', \quad \text{where}$$

$$\mathbf{w}', \mathbf{q}' = \arg \min_{\substack{\mathbf{w}\in W_0, \|\mathbf{w}\|^2=1 \\ \mathbf{q}\in Q_0, \|\mathbf{q}\|^2=1}} \left\| A\mathbf{w} - \mathbf{q} \right\|^2 .$$

Pair-wise comparison

	a	*s*	*p*	*i-c*
apple	●	+	+	+
soup		●	+	–
porridge			●	–
ice-cream				●





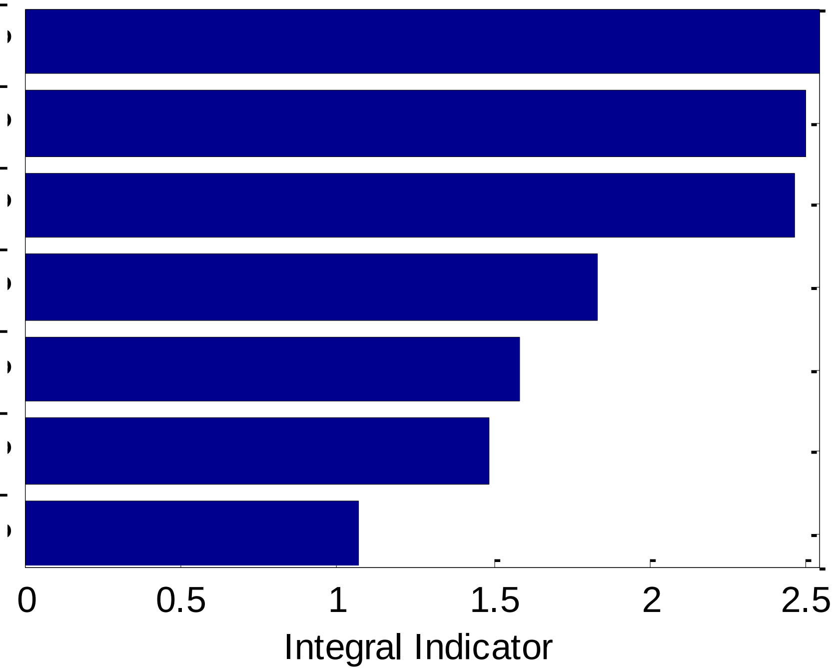
If an object in a row is better than the other one in a column then put "+", otherwise "-".

Make a graph, *row* + *column* means *row* ●——▶● *column.*

Find the top and remove extra nodes.

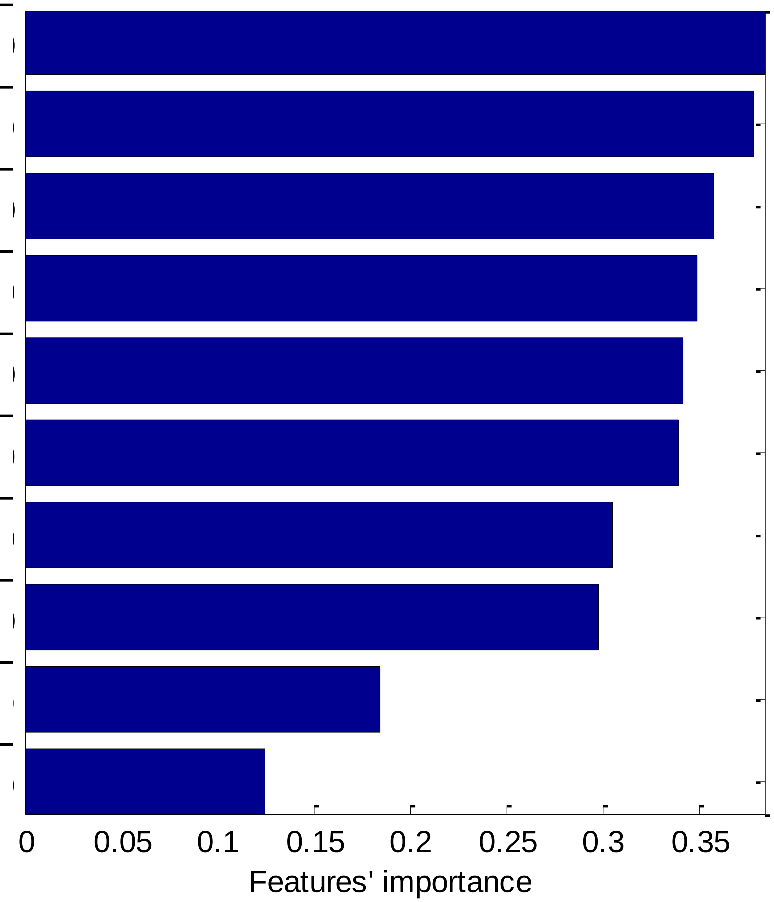
The Integral Indicator of Ecological Footprint for the Croatian Thermal Power Plants

Power Plant	Integral Indicator
TE-TO Zagreb CHP	2.53
EL-TO Zagreb CHP	2.49
TE-TO Osijek CHP	2.46
Plomin 2 TPP	1.83
Rijeka TPP	1.57
Sisak TPP	1.48
Plomin 1 TPP	1.07

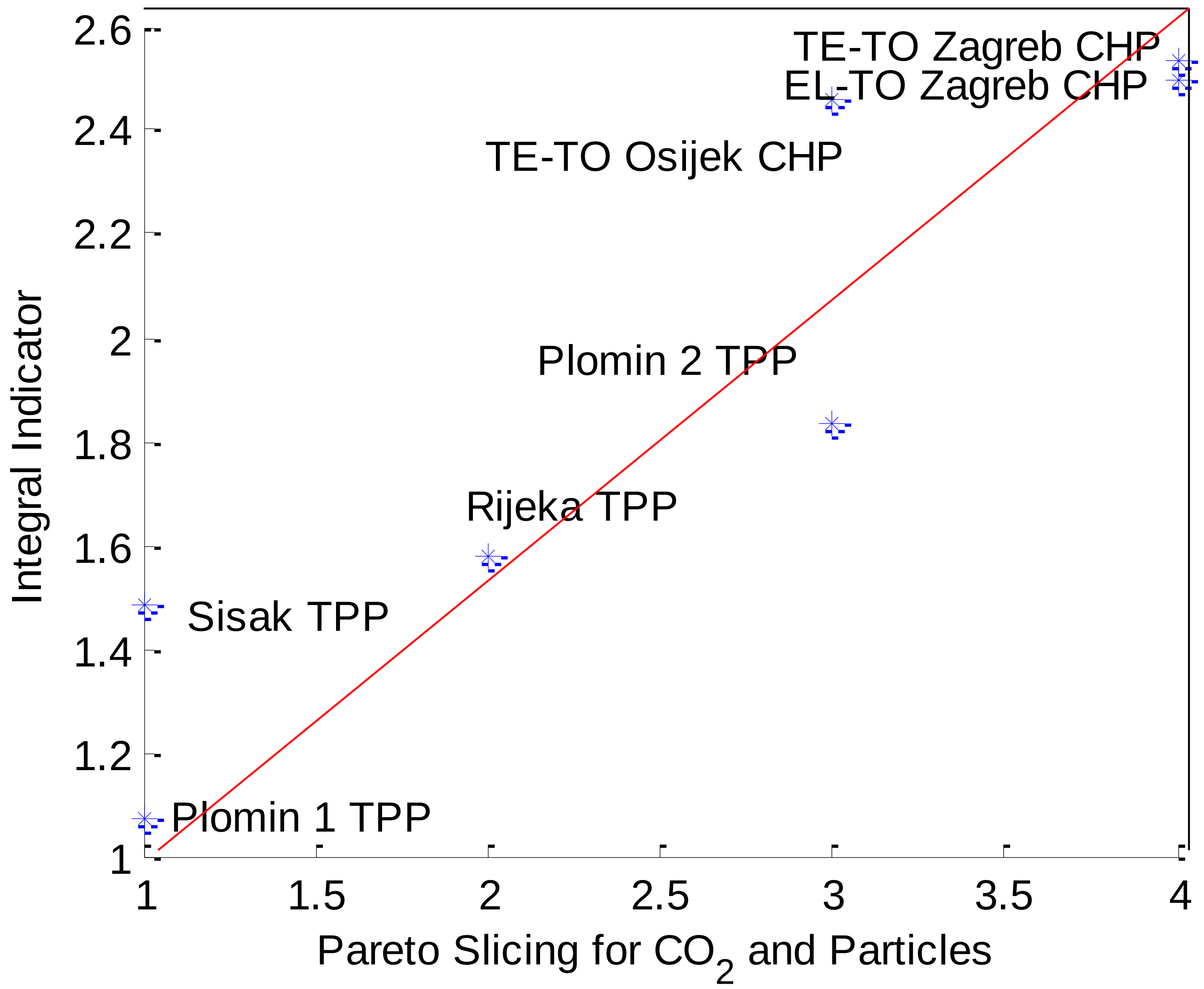


The Importance weights of the Features

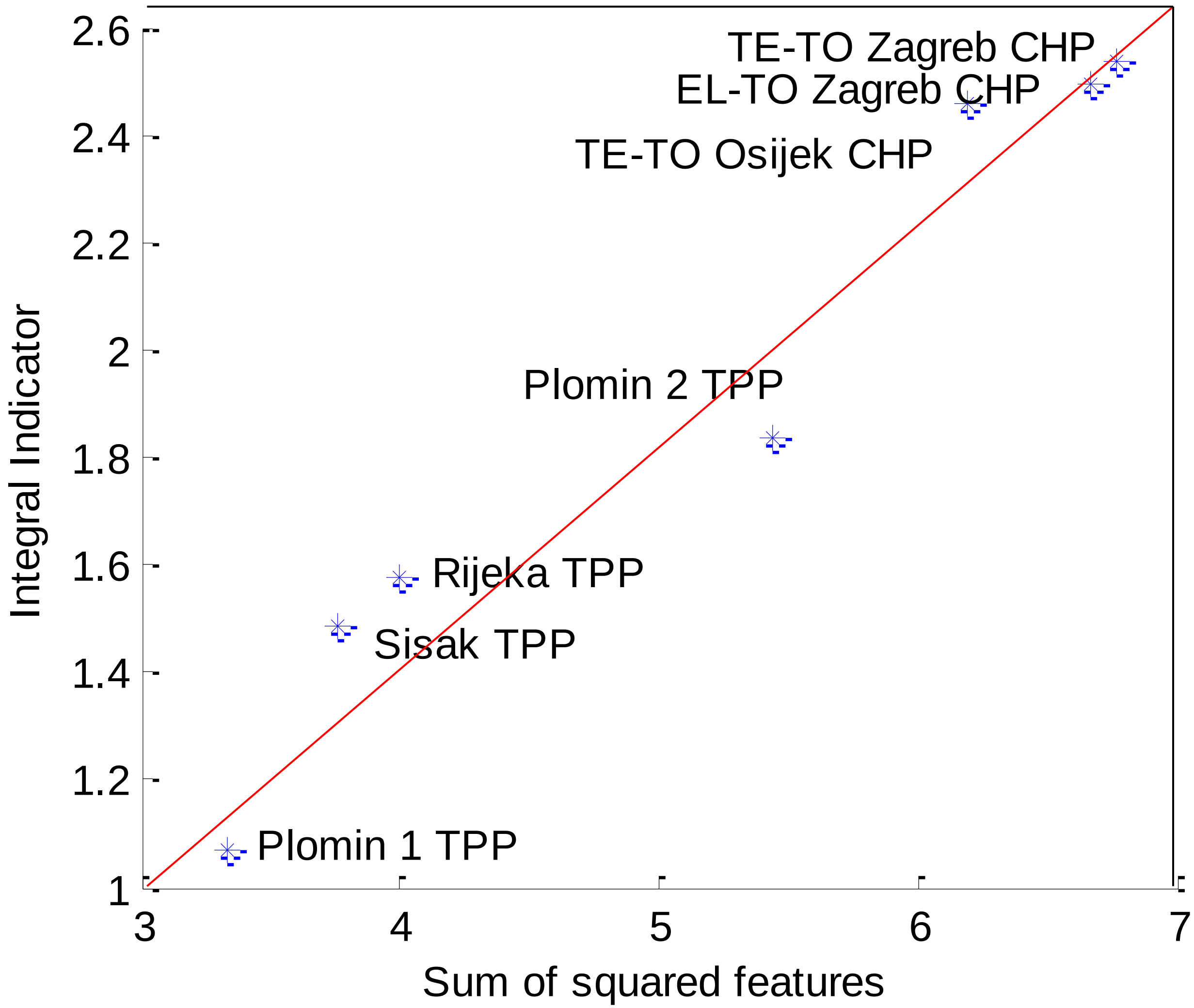
Feature	Weight
Coal (t)	0.38
Sulphur content in coal (%)	0.37
NO_x (t)	0.35
Liquid fuel (t)	0.34
SO_2 (t)	0.34
Particles (t)	0.33
Natural gas (10^3 m^3)	0.30
CO_2 (kt)	0.29
Sulphur content in l.fuel (%)	0.18
Available net capacity (MW)	0.12



The Integral Indicator versus Pareto Slicing



The Integral Indicator versus Metric Algorithm

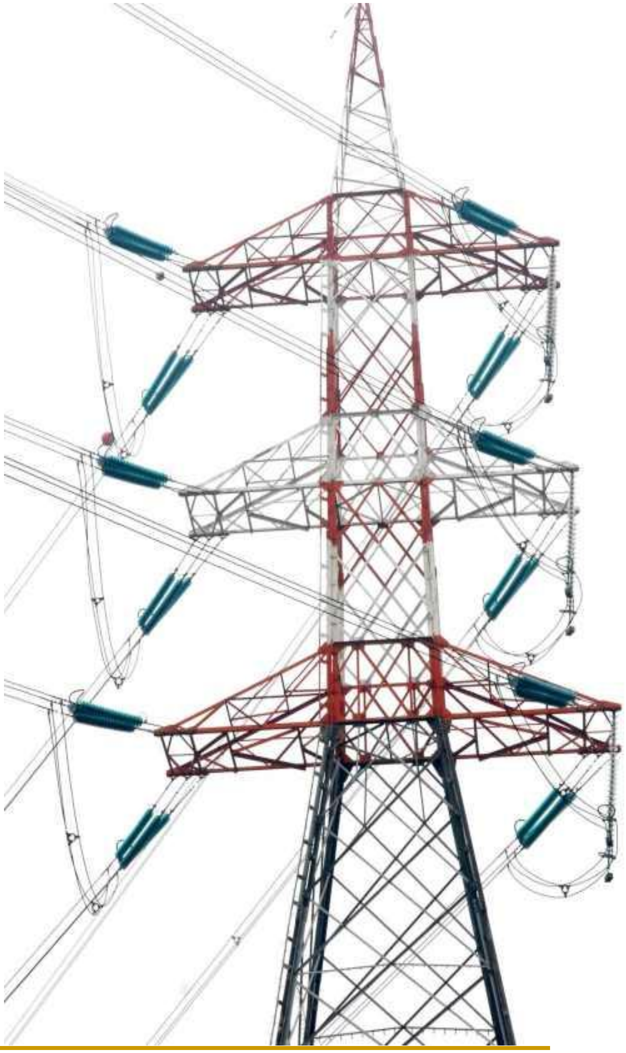


The results of the specification are

- adequate indices,
- explained expert estimations.

We know why our expert valued each object

and what contribution each feature makes to the indicators.



Strong sides of the methodology

- The Integral Indicator usually is based on the open-source data
- The model of the Integral Indicator and the methodology of construction are published

→ Anybody can check the results

- The Integral Indicator could include expert estimations
- The methodology of the expert estimations specification is suggested

→ Experts are welcome to show opinions

List of the constructed indices

1. Integral indicators of the quality of life in the Russian regions
2. Human development index in Russia
3. Kyoto-index: power plant ecological footprints in the USA, Ohio
4. Protected area management effectiveness in Russia
5. Index of rare and Red List species in Russia
6. Econometrical index of the Russian economy state
7. The high school science effectiveness for the Ministry of Education
8. Croatian power plant ecological footprints