Indicators Valuation using FCA and Moebius Inversion Function

Petr GAJDOŠ[1], Daniela ĎURÁKOVÁ[1], Lubor TVRDÝ[2], George LEV[3]

[1]*Department of Computer Science, VŠB-Technical University of Ostrava*
tř. 17. listopadu 15, 708 33 Ostrava-Poruba
Petr.Gajdos@vsb.cz, Daniela.Durakova@vsb.cz

[2]*Deparment of Regional Economics, VŠB-Technical University of Ostrava*
Sokolská 33, 701 21 Ostrava 1
Lubor.Tvrdy@vsb.cz

[2]*Interconcult A.S.A.*
glev@redshift.bc.ca

Abstract. The article shows a new method of selection of important key indicators used to best describe a specific geographical area. An indicator is defined as a characteristic number or a time series representing in a unique manner certain important feature of the area in question. Using a method of expert estimate, weights are assigned to each indicator in different categories. Using FCA, interdependencies among the indicators are analysed and the indicators are sorted according to their importance and uniqueness in the area description. A new method based on Moebius Inversion Function was used to evaluate the set of indicators deemed representative for a pilot area. The output of the method described in this article is a sequence of objects / indicators ordered according to their weight. Main goal of this work was to compare the results of this approach with subjective evaluation carried out by the method of expert estimate. The method described in the article appears to be a valued contribution to evaluation of regional competitiveness in a 5th Framework Programme RTD project "Iron Curtain".

Key words: Formal concept analysis, Moebius function, diversity, weight, evaluation, indicators

1 Introduction

The data used for practical application of the new approach originate in a 5th FP RTD project "Iron Curtain" QLRT-2000-01401 ("Innovative models of critical key indicators as planning and decision support for sustainable rural development and integrated cross border regional management in former Iron Curtain areas based on north to south European reference studies.")

Karel Ježek (ed.), DATAKON 2004, Brno, 23.-26. 10. 2004, pp. 1-10.

The project is focused on 6 reference areas along the former “Iron Curtain”.

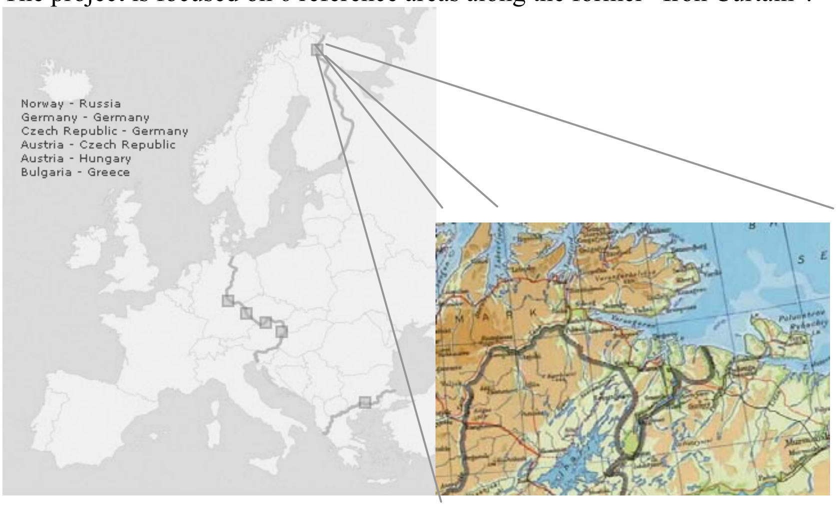

Figure 1. Pilot areas - http://www.ironcurtainproject.com/reference.php

Main goal of the project is to find key indicators to evaluate a regional competitive level and territorial capital of selected pilot areas. The four principal dimensions of territorial competitiveness as defined by the LEADER methodology concept are:

- Economic competitiveness
- Social competitiveness
- Environmental competitiveness
- Positioning in the global context - (relation with the outside world)

From the 6 reference areas analysed in the “Iron Curtain” project, we have chosen for testing and evaluation of indicators the RA #1 located on the border between Norway and Russia, between administrative units of Finnmark County, Sor-Varanger municipality on the Norwegian side and Murmansk Oblast (Region), Pechenga District. It is situated approximately 400 km north of the Arctic Circle and approximately delimited by coordinates 67-69N and 26-29 E or the Norwegian – Russian part of catchment of Pasvik river. Due to its geographic position, historical development, and its dependence on natural resources the area has played a crucial role in history in connection with the development of contacts in the Northern Calotte Area and contact development between Norway and North-west Russia. It plays a unique role in international co-operation in general and in particular in the Barents Euro-Arctic Region.

For detailed description of the area, 44 indicators were selected and were assigned weights on scale of 0-10 in each dimension of the Leader Concept by method of expert evaluation.

A simple degree scale was used.

0	1	2	3	4	5	6	7	8	9	10
❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
No Importance						Very Important				

2 Background

This section shows some definitions and theoretical background important for our later valuating. We define basic notions of formal concept analysis (FCA) and the basic features of the multi-attribute model developed in theory of diversity (TD).

2.1 Formal Context and Concept

FCA is defined by R. Wille [2] and it can be used for hierarchical order of objects based on object's features. The basic terms are formal context and formal concept.

Definition 1. *A formal context C := (G,M,I) consists of two sets G and M and relation I between G and M. The elements of G are called the objects and the elements of M are called the attributes1 of the context. In order to express that an object g is in a relation I with an attribute m, we write gIm or* $(g,m) \in I$ *and read it as "the object g has the attribute m". The relation I is also called the incidence relation of the context.*

Definition 2. *For a set* $A \subset G$ *of objects we define*

$$A^{|} = \{m \in M \mid gIm \text{ for all } g \in A\}$$

(the set of attributes common to the objects in A). Correspondingly, for a set B of attributes we define

$$B^{|} = \{g \in G \mid gIm \text{ for all } m \in B\}$$

(the set of objects which have all attributes in B).

Definition 3. *A formal concept of the context* (G,M,I) *is a pair* (A,B) *with* $A \subseteq G$*,* $B \subseteq M$*,* $A^{|} = B$ *and* $B^{|} = A$*. We call A the extent and B the intent of concept* (A,B)*. Power set* $\mathbf{B}(G,M,I)$ *denotes the set of all concepts of context.*

Definice 4. *Let F be the totality of all features deemed relevant in the specific context, and denote by* $R \subset X \times F$ *the "incidence" relation that describes the features possessed by each object, i.e.* $(x,f) \in R$ *whenever object* $x \in X$ *possesses feature* $f \in F$*. For each relevant feature* $f \in F$*, let* $\lambda_f \geq 0$ *quantify the value of realization of f. Upon normalization,* λ_f *can thus be thought of as the relevant importance, or weight of feature f. The diversity value of a set S is defined as*

$$v(S) = \sum_{f \in F:(x,f) \in R,\ for\ some\ x \in S} \lambda_f \ . \tag{1}$$

The diversity value of a set is given by the total weight of all different features possessed by some objects in S [3]. Note especially that each feature occurs at most once at sum. In particular, each single object contributes to diversity the value of all those features that are not possessed by any already existing objects.

For any subset $A \subseteq X$ of objects denote by F_A the set of features that are possessed exactly the objects in A. Each feature in F_A is possessed by all elements of and not possessed by any element of $X \setminus A$. Then we can write

$$v(S) = \sum_{A \cap S \neq 0} \sum_{f \in F_A} \lambda_A \ . \tag{2}$$

Then, for each subset $A \subseteq X$ denote by $\lambda_A := \sum_{A \cap S \neq \emptyset} \lambda_A$ the total weight of all features with extension A, with the convention that $\lambda_A = 0$ whenever $F_A = \emptyset$. With this notation we write

$$v(S) = \sum_{A \cap S \neq 0} \lambda_A \ . \tag{3}$$

2.2 Conjugate Moebius Inverse

Theorem 1. *For any function* $v : 2^X \to R$ *with* $v := 0$ *there exists unique function* $\lambda : 2^X \to R$*, the Conjugate Moebius Inverse, such that* $\lambda_\emptyset = 0$ *and, for all S,*

$$v(S) = \sum_{A : A \cap S \neq 0} \lambda_A \tag{4}$$

Furthermore, the Conjugate Moebius Inverse λ *is given by the following formula. For all* $A \neq \emptyset$,

$$\lambda_A = \sum_{A : A \cap S \neq \emptyset} (-1)^{|A|-|S|+1} v(S^c) \ , \tag{5}$$

where S^c *denotes the complement of S in X.* [4]

3 Description of our method

The application of FCA renders relations (dependencies) between objects, in this particular case relations between indicators describing the pilot area. First, the input data is converted. By method of expert estimate, a table/matrix of weights for each indicator according to importance of indicator in each of the four principal competitiveness dimensions is obtained. A scaling method [1], [2] is used to make an input incidence matrix. Then the concepts are computed.

index	values from experts				description of indicator
	E	G	N	S	
1	8	5	2	2	Accommodation capacity
2	8	5	5	4	Number of Overnight Stays (per inhabitant)
3	8	3	4	3	Average Overnight Stays
4	4	5	3	4	Number of border crossings per 100km of the border
5	5	3	0	4	Number of cell phones per 100 inhabitants
6	5	5	2	7	Commuting distance to work
7	8	3	3	5	Share of employment in Agriculture, forest and fishing
8	9	5	3	6	Share of employment in Industry
9	9	4	0	7	Share of employment in Services
10	0	3	8	3	Percent of days per year rated as having good air (imissions under limits)
11	5	0	4	3	Electricity consumption per capita per year
12	5	0	6	2	Share of consumption of renewable energy resources
13	7	4	2	6	Employees in small and medium enterprises
14	2	4	9	1	spatial share of forest
15	5	0	8	0	Use of fertilizers (kg/ha of arable land)
16	5	7	0	5	Percentage of population speaking at least one foreign language
17	10	6	0	7	Average salary (ratio CZ/AU)
18	10	6	0	6	GDP per capita
19	10	9	0	7	wage /avarge income
20	8	6	0	10	population age structure
21	3	3	0	9	Share of elementary educated
22	7	7	1	10	Share of University graduated
23	5	4	6	5	Human and economic loss due to natural disasters
24	7	5	2	8	net migration rate
25	6	5	2	9	population change in 10 years (1990-2000)
26	4	4	4	8	Population density
27	5	3	0	4	price for broadband access (128kbps) Euro per month (relative to avg. Wage)
28	10	6	0	4	Foreign direct investments
29	1	5	0	2	Number of regional information centres (per district)
30	2	4	0	0	Number of multilingual regional informational webpages
31	3	2	0	6	Level of crime (crimes per 10000 inhabitants per year)
32	3	4	9	4	Land used by settlements
33	6	1	6	1	Total area of quarries
34	5	1	7	1	Number of mining licences
35	0	4	0	7	Number of inhabitants per one medical practitioner
36	2	5	10	0	Share of protected areas in RA
37	0	0	8	0	Abundance of Selected Key Species
38	0	0	6	6	Wastewater treatment coverage (share of the region's population connected)
39	0	3	8	0	Share of threatened/extinct species as a percent of total native species
40	8	5	0	6	number of unemployed per 1 vacancy
41	4	3	0	6	Percent of population in urban areas
42	9	6	0	10	Unemployment rate
43	0	0	9	0	Concentration of Nitrates in the surface water in mg/l
44	0	0	8	0	Concentration of Phosphates in the surface water in mg/l

Table 1. Input data – weighted values of indicators of Norway/Russia area.

Legend: The Leader competitiveness dimensions: E-Economical, S-Social, N-Environmental, G-Position in the Global Context

The (presented) concept represents a relation between objects. The weight of this relation is given by weight of common (shared..?) features. The diversity of all objects of every concept is the next important variable. The diversity is an output of $v : 2^X \to R$ function mentioned in Theorem 1.

The weights and diversities for each concept are computed, according to (4),(5) and the final results table is generated. All following steps will be demonstrated on a selected concept.

sign	Variable description
v(S)	diversity of the set S. The elements of S are all objects in selected concept
λ_A	weight of selected concept - Möebius function
do_g	diversity of one specific object g
sdo	sum of diversities of objects
wca	weight of common attributes of object in selected concept
$impo_g$	the importance of object

Table 2. List of variables

Important formulas needed for the concept computation:

The diversity of object g $\boldsymbol{do_g}$ is the sum of all weights of all features in relation with object in incidence matrix. It conveys information about partial importance of object but doesn't clearly display next dependences.

$$do_g = \sum_{f:f\in F \text{ and } (g\, I\, f)\in R} \lambda_f \tag{6}$$

The sum of diversities of objects **sdo** means significant "overlapping" of selected concept over the set of attributes according their weights,

$$sdo = \sum_{i=1}^{\text{size of extent}} do_i \ , \tag{7}$$

and next we can compute *the sum of weights of attributes* for each concept

$$wca = \sum_{\text{For all attributes of concept}} \lambda_f \tag{8}$$

The final result of our method is *the computed importance of the object g*.

$$impo_g = \sum_{C:g\in C} \frac{sdo}{v(S)} \lambda_A do \tag{9}$$

This value represents the importance of each object from these aspects:

Uniqueness – Is there any similar object?

Range of description – What type of dimension does the object describe?

Weight of description – What is the weight of object in each dimension?

4 Method application

We have already processed incidence matrix and computed all possible concepts. Indicators are represented by objects in incidence matrix. Features are represented by scaled dimensions.

For simplification we want to demonstrate the method on one selected concept, an identical approach will be applied for other concepts.

For instance, we selected the concept called C_{10} = ({ 2 4 5 27 28 32 },{g4}).

We read next values from incidence matrix. In the following Table 3, only a part of the incidence matrix, is presented due to size of the matrix and limited space.

	weight of feature	3	4	5	8	10	3	4	5	6	3	5	9	4
	dimension after scaling	**e3**	**e4**	**e5**	**e8**	**e10**	**g3**	**g4**	**g5**	**g6**	**n3**	**n5**	**n9**	**g4**
Objects of incidence matrix	**2**	0	0	0	1	0	0	0	1	0	0	1	0	1
	4	0	1	0	0	0	0	0	1	0	1	0	0	1
	5	0	0	1	0	0	1	0	0	0	0	0	0	1
	27	0	0	1	0	0	1	0	0	0	0	0	0	1
	28	0	0	0	0	1	0	0	0	1	0	0	0	1
	32	1	0	0	0	0	0	1	0	0	0	0	1	1

Table 3. The incidence matrix for selected concept (selected section)

The concept diversity is then computed using the diversity function (4). The weight of this concept is 4 according to (5), but it can also be determined from the Table 3.

In that case we get following results.

According to (7) we compute sum of diversities of objects.

$$sdo = 22 + 16 + 12 + 12 + 20 + 20 = 102$$

According to (8) we get weight of common attributes

$$wca = 4,$$

as only one attribute is owned by all objects in concept C_{10} and the weight of this attribute is 4.

The following table shows the values of our variables for each concept of concept lattice.

C	λ_A	v(S)	sdo	wca	C	λ_A	v(S)	sdo	wca	C	λ_A	v(S)	sdo	wca
1	0	202	672	0	30	9	34	43	9	59	0	32	49	17
2	8	66	119	8	31	6	71	106	6	60	0	23	23	23
3	5	103	185	5	32	0	31	42	11	61	0	22	22	22
4	0	32	58	13	33	0	23	23	23	62	9	26	26	26
5	2	69	99	2	34	4	86	132	4	63	10	39	49	10
6	0	59	80	7	35	0	20	31	11	64	0	24	24	24
7	2	29	38	2	36	0	20	20	20	65	9	28	37	9
8	0	18	25	7	37	8	19	46	8	66	3	40	46	3
9	0	17	17	17	38	0	14	25	11	67	0	15	15	15
10	4	69	102	4	39	0	14	14	14	68	1	25	25	25
11	0	29	38	9	40	0	12	12	12	69	0	20	20	20
12	5	22	22	22	41	6	36	59	6	70	8	34	42	8
13	3	55	101	3	42	0	22	33	11	71	0	22	22	22
14	0	26	37	11	43	0	13	13	13	72	6	30	36	6
15	3	31	44	3	44	7	50	66	7	73	0	22	22	22
16	0	26	32	6	45	0	32	41	9	74	0	20	20	20
17	4	39	50	4	46	0	19	19	19	75	0	20	20	20
18	0	23	30	7	47	2	31	39	2	76	1	8	8	8
19	0	18	18	18	48	0	16	22	6	77	2	11	11	11
20	3	47	58	3	49	9	23	36	13	78	0	20	20	20
21	0	31	39	8	50	1	41	44	1	79	1	26	28	2
22	4	32	36	4	51	0	16	16	16	80	0	14	14	14
23	0	16	16	16	52	0	13	13	13	81	7	14	14	14
24	5	74	132	5	53	7	35	42	7	82	10	17	17	17
25	0	12	24	12	54	0	27	37	10	83	0	12	12	12
26	7	57	99	7	55	0	17	17	17	84	0	19	19	19
27	0	19	19	19	56	10	42	91	10	85	0	0	0	0
28	5	41	56	5	57	6	51	89	6					
29	0	19	19	19	58	0	33	65	16					

Table 4. Main values for each concept (we need to truncate the values because of no space)

In the next step, the value of importance for each object needs to be computed. Formula (9) is used to obtain the following table of ordered indicators:

order	index	impo	order	index	impo	order	index	impo	order	index	impo
1	19	1113,439	12	32	662,2255	23	14	437,4005	34	5	241,122
2	17	1040,165	13	23	634,4211	24	15	411,5249	35	27	241,122
3	20	989,7818	14	6	625,6123	25	16	410,5192	36	38	227,8462
4	42	976,4167	15	40	618,6148	26	4	374,0199	37	11	219,6623
5	18	928,8513	16	9	610,7486	27	36	365,4441	38	35	201,2717
6	22	851,4167	17	7	574,1264	28	21	309,0659	39	37	181,8947
7	8	783,7608	18	3	523,2969	29	39	307,9822	40	44	181,8947
8	28	779,5942	19	13	520,656	30	12	277,8482	41	31	160,6423
9	2	754,9931	20	26	518,6116	31	33	268,568	42	43	158,4783
10	24	681,395	21	1	491,2164	32	34	252,9662	43	29	100,8102
11	25	680,7461	22	10	451,5902	33	41	249,5463	44	30	51,93398

Table 5. Weights of indicators

5 Comparison of results of our methods with the opinion of Experts

The order of indicators doesn't depend only on sum of values in each dimension. It can be said that a more meaningful relations between indicators have been found.

index	impo	values from experts				description of indicator
		E	G	N	S	
19	1113,4386	10	9	0	7	wage /avarge income
17	1040,16491	10	6	0	7	Average salary (ratio CZ/AU)
20	989,781818	8	6	0	10	population age structure
42	976,416667	9	6	0	10	Unemployment rate
18	928,851282	10	6	0	6	GDP per capita
22	851,416667	7	7	1	10	Share of University graduated
8	783,760796	9	5	3	6	Share of employment in Industry
28	779,594203	10	6	0	4	Foreign direct investments
2	754,993105	8	5	5	4	Number of Overnight Stays (per inhabitant)
24	681,395015	7	5	2	8	net migration rate
25	680,746107	6	5	2	9	population change in 10 years (1990-2000)
32	662,22548	3	4	9	4	Land used by settlements
23	634,421109	5	4	6	5	Human and economic loss due to natural disasters
6	625,612267	5	5	2	7	Commuting distance to work
40	618,614751	8	5	0	6	number of unemployed per 1 vacancy
9	610,748592	9	4	0	7	Share of employment in Services
7	574,12636	8	3	3	5	Share of employment in Agriculture, forest and fishing
3	523,29691	8	3	4	3	Average Overnight Stays
13	520,655979	7	4	2	6	Employees in small and medium enterprises
26	518,611615	4	4	4	8	Population density
1	491,216357	8	5	2	2	Accommodation capacity
10	451,590231	0	3	8	3	Percent of days per year rated as having good air
14	437,400485	2	4	9	1	spatial share of forest
15	411,524893	5	0	8	0	Use of fertilizers (kg/ha of arable land)
16	410,519183	5	7	0	5	Percentage of population speaking at least one foreign l.
4	374,019863	4	5	3	4	Number of border crossings per 100km of the border
36	365,444096	2	5	10	0	Share of protected areas in RA
21	309,065934	3	3	0	9	Share of elementary educated
39	307,982186	0	3	8	0	Share of threatened species as a percent of total native species
12	277,848245	5	0	6	2	Share of consumption of renewable energy resources
33	268,56798	6	1	6	1	Total area of quarries
34	252,966178	5	1	7	1	Number of mining licences
41	249,546341	4	3	0	6	Percent of population in urban areas
5	241,12201	5	3	0	4	Number of cell phones per 100 inhabitants
27	241,12201	5	3	0	4	price for broadband access (128kbps) Euro per month
38	227,846154	0	0	6	6	Wastewater treatment coverage
11	219,662263	5	0	4	3	Electricity consumption per capita per year
35	201,271726	0	4	0	7	Number of inhabitants per one medical practitioner
37	181,894737	0	0	8	0	Abundance of Selected Key Species
44	181,894737	0	0	8	0	Concentration of Phosphates in the surface water
31	160,642308	3	2	0	6	Level of crime (crimes per 10000 inhabitants per year)
43	158,478261	0	0	9	0	Concentration of Nitrates in the surface water in mg/l
29	100,810177	1	5	0	2	Number of regional information centres (per district)
30	51,9339835	2	4	0	0	Number of multilingual reg. informational webpages

Table 6. Comparison of results.

Consulting of the results of the described method with the experts familiar with the area who were also the authors of weights for individual indicators it has been proven that this method of approach delivers qualitatively new information to the indicator analysis.

6 Conclusion

The demonstrated method of approach has shown that in comparison with subjective evaluation by expert estimates, the Moebius function and FCA offer a well-ordered, quantified hierarchy of indicators with a value indicating the 'uniqueness" or the ability of an indicator to be replaced or substituted by another indicator from the selected set. This hierarchy offers the decision-makers an easy solution to a rather complex and convoluted problem of subjective indicator evaluation and selection of a subset of critical key indicators from the available data.

It is apparent that the output depends primarily on the values/weights assigned to each indicator in the four main dimensions of the Leader Project [5] and a possible subjective error of judgement of an expert evaluator may propagate through the FCA approach, but the structure of the output suggest a possibility of further analysis and evaluation of relative position of the indicators, thus highlighting discrepancies and enabling an experienced analyst to determine the subjective validity of the expert evaluation.

The parameter $impo_g$ (the importance of object g) produced by the method delivers an aggregated information containing the 3 important descriptors (Uniqueness, Range of Description, Weight of Description), combining the most important features of each indicator and allowing the analyst/evaluator to select the most critical indicators/data fields for monitoring, further analysis and as an input for predictive modelling tools used in advanced decision-support systems.

References

1. Ďuráková, D. and Snášel, V.: *Hierarchické uspořádání výsledků prostorového dotazu. Datakon 2001,* Bieliková M. (Ed.) Vydavatelstvo STU, Bratislava (2001), 231-241.
2. Ganter, B. and Wille, R.: *Formal Concept Analysis*. Springer-Verlag, Berlin, Heidelberg, 1999.
3. Nehring, K.: A Theory of Diversity. *Ecometrica 70* (2002) 1155-1198.
4. Nehring, K. and Puppe, C.: Modelling phylogenetic diversity. *Resource and Energy Economics* (2002).
5. Project web site, http://www.ironcurtainproject.com/