TREND ANALYSIS FOR SPARSE DATA

E. Nurminski and N. Vorontsov

September 1979
WP-79-89

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

ABSTRACT

The major theme of this paper is to present some means for an analysis of changes in characteristics of complex systems. Such systems are characterized by a number of parameters which are interdependent. The data available on such systems are sparse in the sense that only a few such systems exist in the real world. Any attempt to choose uniform population with respect to some characteristics will decrease the number of data even more until a statistics approach becomes completely unreasonable.

The alternative approach is based on pattern recognition ideas and uses an idea of separation of different classes of a complex system by multidimensional surfaces. The position of these surfaces demonstrates the trends in the system's development.

The analysis of coal mines with respect to different criteria has been performed as an example.

TREND ANALYSIS FOR SPARSE DATA

E. Nurminski and N. Vorontsov

There are many examples of complicated systems of which adequate representations require innumerable indexes. This is typical for industrial plants, power generating stations, agricultural enterprises, etc. The partial interdependency of the system's characteristics which is unknown in detail is also typical for these systems. The great variety in sizes and values of parameters is observed among operating systems and/or systems under construction. Under such circumstances, it is extremely difficult to draw conclusions about the actual trends in system characteristics, the estimate of the essential and nonessential indexes, and the comparison of the degree of advancement of new technologies and industrial projects.

A possible way to overcome many difficulties connected with the analysis of multidimensional sparse data is based on the geometrical analysis of the sets of points representing different systems in corresponding coordinate spaces. This leads to a type of pattern recognition technique used with different groups of technologies, plants, etc., and is looked upon as a different approach to the analysis of this data.

We are more interested in representing the differences between various classes of objects rather than the traditional

approaches of pattern recognition (Andrews 1972) or cluster analysis (Duran and Odell 1974) which are essentially oriented toward partitioning the data into a number of classes. This study gives the primary partitioning of the systems, namely coal mines. The aim is to determine and interpret the differences between the mines belonging to the different categories.

PROBLEM AND METHOD

The data on coal mines which are used for this analysis are a collection of numerical values and the most important qualitative characterizations of the coal enterprises. These data were collected in 1977-1978 and put into a specialized Information Management System (IMS) (Grenon and Lapillone 1976) which gives extended possibilities for information which is stored and retrieved.

So far as we are interested in the analysis of numerical characteristics, a particular coal mine is represented as a point in multidimensional space of the system's characteristics. The total number of characteristics handled by IMS is approximately 200 and the total number of mines in the data base is approximately 60. The correspondence between these numbers is such that even for a limited choice of parameters, the ratio of the number of data to the number of parameters is very low. This may justify the belief that the statistical approach for analysis of the data under consideration is inefficient and necessitates the search for an alternative approach.

The general idea of the proposed method in a sample case is as follows: There are two sets of coal mines which are singled out with respect to a certain classification rule. For instance, these two sets might be sets consisting of operating mines and mines under construction; mines built before and after a specific date; mines before and after reconstruction; mines in different countries, etc. We will not discuss the classification rule itself as we are mainly interested in the analysis and representation of the differences between two classes. These differences are represented by the aid of the

separating hyperplane. This plane must be defined in such a way that allows for a maximum separation of points of one set from points of another set. Thus, the position of this hyperplane will, in fact, show the direction of changes in the total sum of parameters.

There are many contrasting opinions as to the best possible separation of the given sets. The particular choice made in this paper is not in any way a unique possibility but is most probably the simplest one.

Let us introduce some notations. We will take as an example, two sets of coal mines as two finite subsets (A and B) of a n-dimensional euclidian space. The elements of the set A and B are denoted by $\{a^k\}$ and $\{b^i\}$, respectively,

$$A = \{a^k, i \varepsilon I_A\} \quad ,$$

$$B = \{b^k, j \varepsilon J_B\} \quad .$$

I_A and J_B are finite sets. The unknown coefficients of the separated hyperplane will be denoted as vector $x = (x_1, x_2, \ldots, x_n)$. The equation describing the hyperplane in the parameter space is

$$px - x_o = 0 \quad ,$$

where p is a paremeter vector and x_o a scalar value.

We are seeking a hyperplane which would allow the following expression to receive minimal value:

$$f(x) = \sum_{i \varepsilon I_A} (a^i x - x_o)^+ + \sum_{j \varepsilon J_B} (b^j x - x_o)^- \quad , \qquad (1)$$

where

$$r^+ = \max \{0, r\}$$

$$r^- = \max \{0, -r\} \quad .$$

This problem can be converted into a linear programming problem as follows:

$$\min \sum_{i\in I_A} u_i + \sum_{j\in J_B} v_j$$

$$a^i x - x_0 \leq u_i, i\in I_A$$

$$b^j x - x_0 \leq v_u, j\in J_B \qquad (2)$$

$$0 \leq u_i, i\in I_A$$

$$0 \leq v_j, j\in J_B \quad .$$

It is necessary to impose constraints on vector x to get a bounded solution. The natural boundaries for x would be

$$-1 \leq x_i \leq 1 \quad ,$$

and the boundness of x_0 will regularly follow from that.

IMPLEMENTATION AND RESULTS

The approach described above has been implemented as an experimental code on IIASA's home computer PDP11/70. This implementation uses the linear programming subroutine authored by W. Orchard-Hays. Insignificant changes were made in this routine to simplify the handling of problems with different numbers of data and parameters.

Up to the present time, an interface does not exist between IMS in charge of data storing and this code. However, this is justified by the experimental nature of the development. The programmed interlink between these codes might be provided after a detailed study of the methodological usefulness of this technique. This work is at an early stage and the preliminary results include test examples and the analysis of various real data as well.

The test examples consider the separation of two sets of two-dimensional space. The points of both sets are presented as the first example in Table 1. It is assumed that every point of these sets represents a system which is characterized by two parameters x and y; the numerical values of these parameters are different.

This data gives rise to the linear programming problem with 13 variables and 10 constraints. As a result a separating hyperplane is obtained which is characterized by normal vector p = (0.25,1). The data and the separating hyperplane are shown in Figure 1. The position of the plane and the numerical values of the coordinate of the normal vector show that the changes in both parameters are directed toward the increase of their value; the increase of the second variable is essentially bigger than in the first one. This agrees with common sense and shows that the proposed approach gives reasonable results.

The second test example is a modification of the first one. Now an additional point with the coordinates x = 0.25, y = 1.0 has been added to set A.

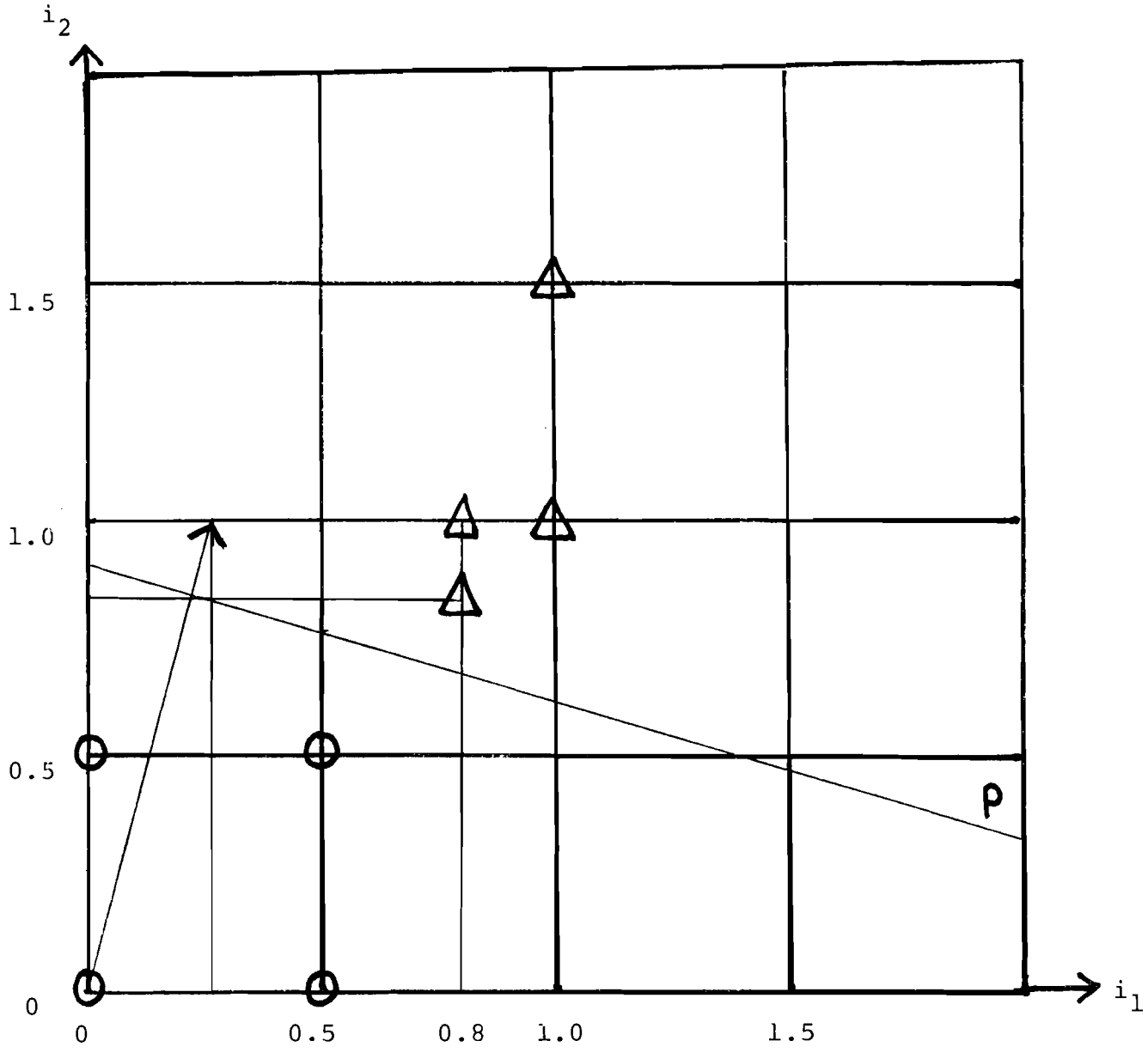
The solution of the correspondent extremum problem gives a slightly different result: p = (0.5, 0.7). The values of the coordinates of the vector p give a relative proportion of the differences in characteristics of the points in sets A and B. The data and the separating hyperplane is illustrated in Figure 2.

The experiments with real data have been made in the following manner. From the resource data base 48, both deep and surface coal mines were chosen. Their characteristics, such as coal reserves (CRESERVES), annual capacity (ANCAP), period of construction (CONSTRUR), estimated life time (LIFET), cost of construction (COST), water consumption (WATER), land occupied (LAND), and manpower (MEN), are given in Table 2.

Every entry in the last column shows the type of mine being considered: a conditional underground coal mine (UNDCON), an underground hydraulic mine (UNDHYD) and a surface coal mine (SURF). This table includes operational mines and mines under construction (projected). Therefore, the decision has been

Table 1. Data for the test examples 1 and 2.

No.	Class A		Class B	
	x	y	x	y
1	0.0	0.0	0.8	0.8
2	0.5	0.0	0.8	1.0
3	0.5	0.5	1.0	1.0
4	0.0	0.5	1.0	1.5

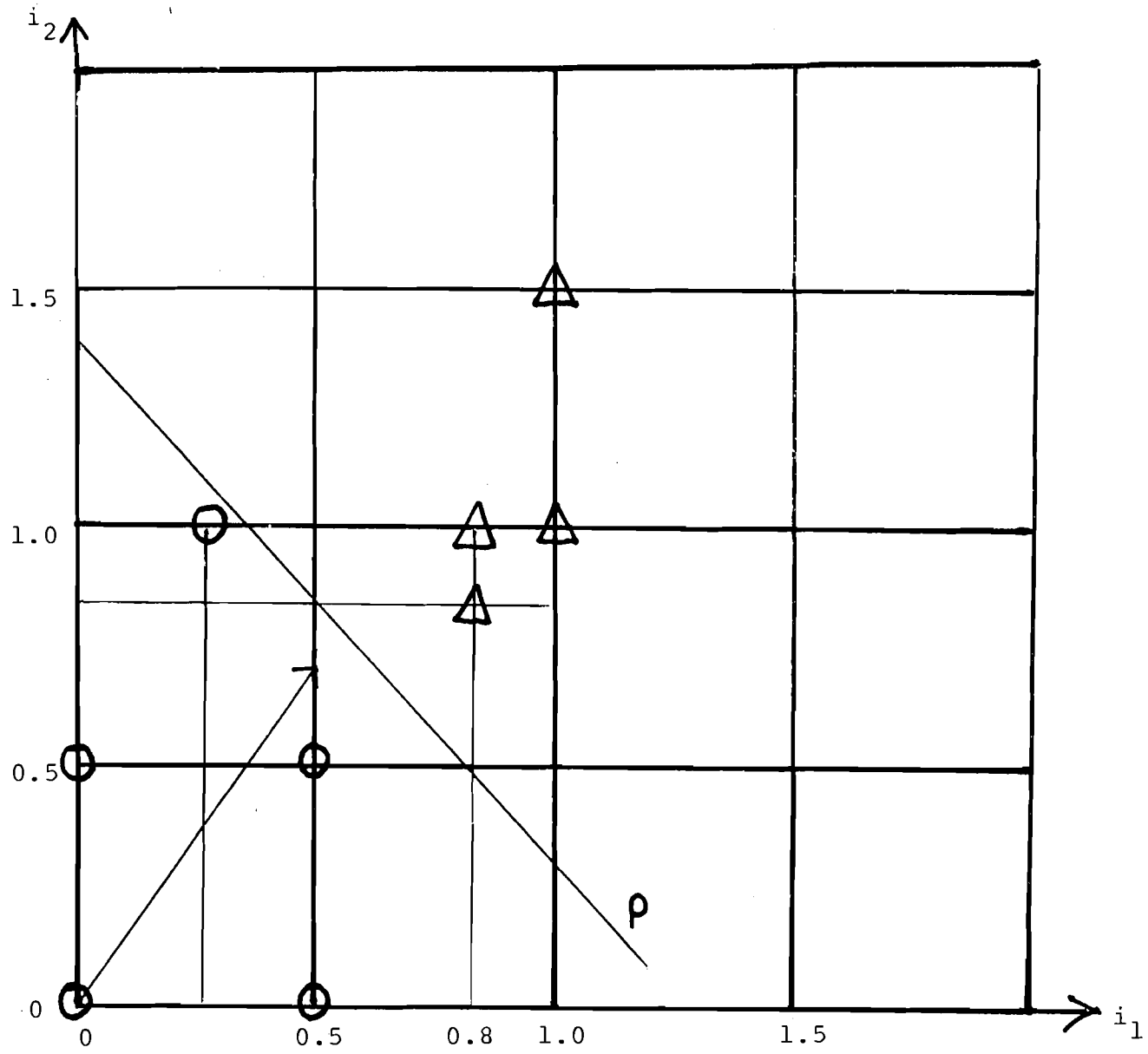


Class A: ○

Class B: △

Separation hyperplane P: $0.25i_1 + i_2 \leq a$

Figure 1. Test example 1.



Class A:

Class B:

Separation hyperplane P: $0.5i_1 + 0.7i_2 \leq a$

Figure 2. Test example 2.

Table 2. Characteristics of surface and underground coal mines.

CODE	BEYIAN	RESERVES	ANCAP	DORSTOUR	LIFET	COST	WATER	LAND	MEN	TECHNO
MINE010	-1	15.300	0.600	4.000	15	40.300	0.010	0.000	586,000	UNDCON
MINE036	-1	-1.000	0.750	-1.000	20	-1.000	-1.000	-1.000	-1,000	UNDCON
MINE213	-1	54.000	1.810	2.500	30	-1.000	-1.000	-1.000	-1,000	UNDCON
MINE055	-1	225.300	3.600	7.000	65	25.250	-1.000	-1.000	-1,000	UNDCON
MINE211	-1	175.000	3.630	2.500	30	-1.000	-1.000	0.400	-1,000	SURF
MINE012	-1	164.000	5.440	2.400	30	-1.000	-1.000	0.080	-1,000	SURF
MINE042	1973	11.700	0.950	-1.000	20	7.100	0.020	-1.000	254,000	UNDCON
MINE041	1973	23.500	1.860	-1.000	20	6.000	0.020	-1.000	461,000	UNDCON
MINE048	1973	143.000	2.720	-1.000	20	6.800	0.017	-1.000	153,000	SURF
MINE042	1973	35.100	2.015	-1.000	20	5.700	0.041	-1.000	674,000	UNDCON
MINE009	1974	-1.000	50.000	-1.000	-1	-1.000	2.000	-1.000	-1,000	SURF
MINE043	1975	21.200	0.960	-1.000	20	9.600	0.036	-1.000	246,000	UNDCON
MINE044	1975	43.600	1.850	-1.000	20	7.800	0.037	-1.000	428,000	UNDCON
MINE045	1975	63.500	2.920	-1.000	20	7.100	0.037	-1.000	633,000	UNDCON
MINE037	1975	87.000	4.350	-1.000	20	10.400	0.017	-1.000	195,000	SURF
MINE046	1975	160.000	4.540	-1.000	20	6.700	0.037	-1.000	997,000	UNDCON
MINE047	1975	58.600	4.550	-1.000	20	10.100	0.063	-1.000	620,000	UNDCON
MINE038	1975	124.000	6.100	-1.000	20	9.200	0.017	-1.000	202,000	SURF
MINE039	1975	167.000	8.340	-1.000	20	3.500	0.017	-1.000	234,000	SURF
MINE035	1976	27.000	0.450	-1.000	40	5.590	-1.000	-1.000	515,000	UNDHYD
MINE034	1976	96.000	1.380	-1.000	80	12.020	-1.000	-1.000	1015,000	UNDCON
MINE052	1976	64.000	1.440	-1.000	55	10.790	-1.000	-1.000	1230,000	UNDCON
MINE051	1976	72.000	3.030	-1.000	35	6.350	-1.000	-1.000	1185,000	UNDHYD
MINE027	1977	55.000	0.900	-1.000	77	30.300	-1.000	0.460	2411,000	UNDCON
MINE050	1977	300.000	1.450	-1.000	32	8.160	0.870	0.000	1360,000	UNDCON
MINE020	1977	90.000	1.550	-1.000	97	26.600	0.070	0.000	3300,000	UNDCON
MINE023	1977	28.000	1.000	-1.000	23	12.000	0.800	0.000	-1,000	UNDCON
MINE024	1977	67.900	1.600	-1.000	52	15.050	0.540	0.000	1970,000	UNDCON
MINE029	1977	93.000	2.300	-1.000	75	14.750	0.290	0.000	2680,000	UNDCON
MINE021	1977	177.000	2.500	-1.000	70	42.000	-1.000	0.000	4193,000	UNDCON
MINE025	1977	112.000	2.650	-1.000	54	18.080	-1.000	0.110	2512,000	UNDCON
MINE028	1977	80.200	2.800	-1.000	36	17.640	0.100	0.290	2443,000	UNDCON
MINE027	1977	57.000	3.150	-1.000	41	16.920	0.020	0.000	2648,000	UNDCON
MINE016	1980	130.000	1.800	6.700	78	25.660	0.310	0.000	1783,000	UNDCON
MINE019	1980	1160.000	20.000	6.000	40	7.450	-1.000	0.050	2657,000	SURF
MINE026	1980	2230.000	30.000	7.000	75	5.670	-1.000	0.037	2130,000	SURF
MINE018	1980	1460.000	30.000	8.000	45	6.720	0.004	0.046	4600,000	SURF
MINE017	1980	3200.000	60.000	7.000	58	-1.000	0.007	0.016	2816,000	SURF
MINE012	1982	370.000	17.000	2.000	25	-1.000	0.093	0.220	796,000	SURF
MINE002	1985	25.500	0.790	5.000	30	-1.000	0.082	0.770	739,000	SURF
MINE003	1985	26.100	0.960	6.000	30	-1.000	0.298	0.260	1605,000	SURF
MINE005	1985	46.500	1.530	6.000	30	-1.000	0.177	0.570	1457,000	SURF
MINE001	1985	49.700	1.550	6.000	30	25.900	0.010	0.700	605,000	SURF
MINE004	1985	65.900	2.190	6.000	30	-1.000	0.031	0.250	490,000	SURF
MINE006	1985	83.500	2.800	6.000	30	-1.000	0.022	0.250	474,000	SURF
MINE024	1985	80.000	2.820	6.000	30	-1.000	0.024	0.300	508,000	SURF
MINE008	1985	111.000	3.650	6.000	30	-1.000	0.079	-1.000	-1,000	SURF
MINE215	1984	-1.000	10.000	7.000	35	-1.000	-1.000	-1.000	-1,000	UNDCON

taken to study the differences between operational and projected units. In accordance with that, all data have been divided into two classes A and B: operational coal mines and those under construction. As the possibility of the experimental code is rather limited, the number of characteristics under consideration are decreased and experiments are made with such characteristics as CRESERVES, ANCAP, WATER, and MEN.

Surface mines have only been studied at the first stage and Tables 3 and 4 give the correspondent data, respectively, for operational mines, projected mines, and mines under construction.

After solving the correspondent linear programming problem, the following changes in indexes are observed:

CRESERVES - -0.12%

ANCAP - -0.10%

WATER - -100.00%

MEN - +0.16% .

These results must be considered as representing a particular pattern of the future development of the coal industry. It shows namely that the most significant changes will occur in water consumption whenever coal reserves and annual capability drop a small amount and manpower increases slightly.

The hypothesis is submitted that this substantial difference is caused by the different scale factors for indexes under consideration. To examine this influence, experiments with normalized data have been performed. In this case, each column of Tables 3 and 4 have been divided by the average value of the correspondent index calculated only from Table 3. The results of this normalization are shown in Tables 5 and 6 and this solution of th linear programming problem gives the percentage of such changes:

CRESERVES - -3.63%

ANCAP - +3.63%

WATER - -100.00%

MEN - +6.56% .

Table 3. Operational surface coal mines.

CRESERVS	ANCAP	WATER	MEN
109.000	3.630	0.540	196.000
164.000	5.440	0.540	196.000
143.000	2.720	0.017	153.000
132.000	50.000	2.000	196.000
87.000	4.350	0.017	195.000
122.000	6.100	0.017	202.000
167.000	8.340	0.017	234.000

Table 4. Projected surface coal mines.

CRESERVS	ANCAP	WATER	MEN
1140.000	20.000	0.080	2657.000
2230.000	30.000	0.080	2130.000
1460.000	36.000	0.004	4800.000
3200.000	60.000	0.007	2818.000
370.000	17.000	0.093	796.000
23.500	0.790	0.082	739.000
28.100	0.960	0.298	1805.000
46.500	1.530	0.177	1457.000
49.700	1.550	0.010	605.000
65.900	2.190	0.031	490.000
83.500	2.800	0.022	474.000
84.000	2.820	0.024	508.000
111.000	3.650	0.079	1606.600

Table 5. Operational surface coal mines normalized data.

CRESERVS	ANCAP	WATER	MEN
0.826	0.318	1.000	1.000
1.240	0.477	0.032	1.000
1.080	0.239	3.700	0.781
1.000	4.390	2.000	1.000
0.659	0.382	0.032	0.995
0.924	0.535	0.032	1.030
1.270	0.732	0.032	1.190

Table 6. Projected surface coal mines normalized data.

CRESERVS	ANCAP	WATER	MEN
8.640	1.750	0.148	13.560
16.900	2.630	0.148	10.870
11.000	3.160	0.007	24.490
24.200	5.260	0.013	14.380
2.800	1.490	0.172	4.060
0.180	0.069	0.152	3.770
0.210	0.084	0.552	9.209
0.352	0.134	0.328	7.434
0.377	0.136	0.018	3.087
0.499	0.192	0.057	2.500
0.633	0.246	0.041	2.418
0.636	0.247	0.044	2.592
0.841	0.320	0.146	8.197

It is clear that normalization does not change the final result: it remains that the changes in water consumption are still the most essential feature of the development. It is interesting to note that normalization has resulted in a different forecast for annual capacity. Instead of the very small decrease predicted on the basis of unnormalized data, a slight growth in ANCAP is estimated. In our opinion, this divergency doesn't change the general picture of the trend of the coal industry revealed by both computer analyses, namely, big drops in water consumption as compared to small changes in other indexes.

This particular pattern of changes might be interpreted as a radical change in surface mining technology with respect to water consumption. On the other hand, this analysis might reveal as well an unreliable character of data on projected mines with respect to this characteristic. Which of these possibilities is in fact true is a matter for specialists in coal mining to judge. Indubitably the approach described above provides a useful assistance in focusing their attention on the key factors as they become apparent throughout formal analysis.

Both computer runs also show a good separability of operating and projected surface coal mines. The values of the objective function in a linear programming problem demonstrates that these two sets have almost no intersection.

In the next experiment performed, water consumption was excluded from the indexes under consideration. After solving the correspondent linear programming problem, the following relative changes in coal reserves, annual capacity and manpower were observed:

CRESERVES	-	-36.16%
ANCAP	-	+12.01%
MEN	-	+66.21% .

This result can be interpreted as a confirmation of the first two. We can see that changes in all indexes have approximately the same order or magnitude. Moreover, the manpower index demonstrates a greater change than the other two. In addition, the parameter coal resources/per mine shows a tendency to decrease, as revealed by the two previous computer runs.

This work can be continued in a number of ways. One which presents an essential practical interest is to increase an analytical power of this approach by changing the algorithmic basis of it. Transformation of the original problem of minimizing the function (1) into a linear programming problem (2) requires an introduction of a large number of artificial variables which increases the size of the problem and correspondingly decreases the number of parameters this approach can handle taking into account the rather modest size of IIASA's home computer PDP 11/70.

The natural alternative to an LP algorithm would be non-differentiable optimization techniques which directly minimize function (1) and work with an essentially smaller number of variables. This work is currently under way and future results will be reported.

CONCLUSION

Complex systems are described by a large number of interdependent characteristics and only a small amount of empirical information exists relating to these systems. This makes a mathematical statistics approach very difficult when analyzing the changes occuring within these systems.

An approach based on a pattern recognition technique gives significant and useful results and a better insight may be obtained into the trend pattern of changes which occur in system characteristics.

A number of test examples have been analyzed and the approach based on a separation technique has been applied to the analysis of changes in the characteristics of surface coal mines.

REFERENCES

Andrews, H.C. (1972) Introduction to Mathematical Technique in Pattern Recognition. New York: Wiley-Interscience.

Duran, B.S. and P.C. Odell (1974) A Survey Lecture Notes in Economics and Mathematical Systems. Vol.100. Berlin: Springer-Verlag.

Grenon, M. and B. Lapillonne (1976) The WELMM Approach to Energy Strategies and Options. RR-76-19. Laxenburg, Austria: International Institute for Applied Systems Analysis.