SMALL SIZE SAMPLE MATHEMATICAL MODELING

V. Popukaylo
T.G. Shevchenko Dniester State University, Tiraspol, Moldova

INTRODUCTION

In the modern industry there are such productions which because of technological limitations cannot provide a sufficiently large size sample, in accordance with the laws of experiment planning theory to get adequate mathematical model suitable for managing complex control object. This state of things exists at many enterprises with small-scale production, as well as enterprises producing high-tech and expensive products.

Similar examples can be found in medicine, biology, economy and other branches of human activity. In this paper we propose a method of multidimensional point distribution allowing to obtain adequate mathematical models of complex object-based multidimensional small samples.

To eliminate the loss of information when processing small samples is necessary to abandon groups of observations and to go to the methods of considering each individual realization as a distribution center of a virtual sample with the appropriate parameters.

1. RESEARCH METHODS

The aim of this work is to compare mathematical models obtained after the analysis of the basic data and data obtained after application of multidimensional pointed distributions method.

In the work "Small size samples" (by D.V. Gaskarov, V. Shapovalov) the specific methods principles of statistical small samples processing are most clearly articulated and substantiated. Development of this work led to the definition of the small size samples upper range limit $n = 15$ [1], and later to create a point distributions method (PDM) [2].

To eliminate the loss of information when processing small samples is necessary to abandon groups of observations and to go to the methods of considering each individual realization as a distribution center of a virtual sample with the appropriate parameters [3]. These methods include PDM, using which each measurement is considered as a distribution center with the known law. The usage of PDM allows to obtain the accuracy of calculations corresponding to size sample 3-5 times larger than the initial.

However, in real production a lot of factors affect the target function and required regression equation to be multidimensional. There are various methods for passive experiment tables processing, among which there is the method of least squares with pre-orthogonalization factors (MLSO) and the modified random balance method (MRBM) [4].

One of the oldest and most developed methods for passive data modeling is method of least squares (MLS) which is based on selection of equation of regression for the sum of squares of a difference between the equation and experimental data was the smallest of all possible. However, there is a problem when the recognition of any factor is insignificant, it is necessary to exclude it from consideration and to do all computing procedure from the very beginning. MLSO, which proposes to choose special system of linearly independent functions for each regression task, so that the normal equations matrix is single, became the solution of this problem [4]. In this case, there is no need to look for the inverse matrix, and it is possible to reject insignificant coefficients of regression without the others. The choice of function system is carried out with use of orthogonal polynoms of Chebyshev so that the *Y(X)* curve decayed on the chosen system of functions in a row, *Xkj* which is quickly meeting in each point. Thus the system of functions has to be defined on that interval of values of the *Xkj* variable on which experimental points are located. However, MLS is sensitive to the order of sequence factors in order of importance, as well as increasing the number of factors and decrease the number of lines is much more complicated and increases the processing error.

Also one of the most known and most convenient methods of modeling of passive experiments is the random balance method(RBM). The essence of RBM is to construct a planning matrix with a random distribution of factor levels in the experiment on the matrix and in specific data processing experiment. Later this method has been developed to a modified random balance method (MRBM), which is complex and cumbersome graph-analytical procedure estimates the coefficients of the model is replaced by easier analytical procedure. This method has a high resolution (the ability to allocate strongly influencing factors), and low sensitivity (i.e., the ability to allocate significant model parameters which characterize the factors that have a relatively weak effect) [4]. However, as the modified random balance method (MRBM) is the eliminating method, so its application to small selections is not possible.

For solving this problem, below is shown a method that combines the ideas of two other methods. The first part of the calculations performed by the method of point distributions, treating each factor by the initial sample point distributions and knowing the nature of the distribution law may artificially increase the sample size in order to be able to use one of the methods for obtaining adequate mathematical models for passive data. Joining individual factor samples in a single multi-dimensional large size sample occurs in the lines with the highest level of non-normalized probability density and with simultaneous cutting off of all incomplete lines.

There was thus developed a fundamentally new multidimensional distributions point method (MSPM) to obtain adequate mathematical models of complex multidimensional object based on the initial samples of small size.

Algorithm:

1. A correlation analysis, the purpose of which is to find highly related factors.
2. By means of MSP for all *Xi* and *Y* to build tables for calculating non-normalized probability densities in the virtual domain.
3. For each line *l* of the initial experimental data table to construct a virtual data table , in which to simultaneously bring in the values of two *Xij* columns from corresponding table of non-normalized probability densities and *Xi*l column. Alignment (joining) pairs of columns *Xij* and *Xil* (and) *Yj* and *Yl* should occur at the maximum probability density level.
4. From all tables found in the preceding paragraph of this algorithm is filled with rows and all columns indicating the non-normalized probability density are not completely erased. The joining of edited tables occurs in numerical order of input data table rows. The received virtual data table is 15-20 times longer than initial data table, it allows to achieve the bigger accuracy and reliability during its processing.
5. According to the table of complete virtual sample we determine coefficients of correlation of all factors and output size by the principle "everyone with everyone", for the detailed analysis we use correlation pleiades method in conjunction with an expert weighting coefficients of importance method.
6. According to the received table we make mathematical model by methods of passive experiment, such as: the modified random balance method, the smallest squares method with pre-orthogonalization of factors, or the combined method.

Thus, we can construct a mathematical model appropriate for small size sample, even if the initial small sample was supersaturated up.

In this article we will show the modeling process for the small-size sample.

We also compare the simulation results before and after increase in the virtual sample size.

2. RESULTS AND DISCUSSION

Let's take as a result from the production $n = 8$ product units (parties) the following numerical values of control parameters (X_i – the parameters controlled during technological process; Y – output quality indicator of a product. All names of dimensions for simplicity are omitted)

Table 1. Table of initial experimental data.

Num. of product	Factors X_i				Output value, Y
	X_{1f}	X_{2f}	X_{3f}	X_{4f}	
1	0,695	89,65	66,71	-27,29	57,18
2	0,644	99,40	68,58	-32,09	75,48
3	0,674	108,50	64,97	-36,08	79,12
4	0,695	92,50	67,71	-28,32	72,03
5	0,711	95,80	66,11	-28,90	76,34
6	0,685	100,90	68,13	-27,45	72,22
7	0,692	102,60	65,78	-30,21	81,90
8	0,697	90,60	66,85	-31,83	55,94

At the beginning, we construct a correlation table in the table of initial data, for this we determine correlation coefficients of all factors and output size by the principle "everyone with everyone". We will use the Pearson's correlation coefficient, which varies from -1 to 1.

Table 2. Table of correlation coefficients.

	X_1	X_2	X_3	X_4	Y
X_1	1	-0,453	-0,402	0,488	-0,271
X_2	-0,453	1	-0,363	-0,603	0,805
X_3	-0,402	-0,363	1	0,451	-0,225
X_4	0,488	-0,603	0,451	1	-0,296
Y	-0,271	0,805	-0,225	-0,296	1

Having analyzed the correlation matrix we conclude that the input factors are independent.

To handle such a table of random balance modified method is not possible because of the small row number, so use the method of least squares with pre-orthogonalization factor that is less sensitive to this factor.

As a result of calculations the adequate model was received:

$$Y = -180.1 + 186.17X_1 + 1.2674X_2$$

The adequacy dispersion of this model = 45.18

The average weighted dispersion = 23.324

Fisher criterion Fr = 1.937

When the tabulated value is Ft = 3.87

Thus the resulting model is adequate, but it has a great adequacy dispersion and calculated value of the Fisher criterion.

We try to apply this multidimensional point distributions method for a better mathematical model of researched process. To do this using the point distributions method for all Xi and Y we construct a table for calculating non-normalized probability densities in the virtual domain. As an example, a calculation for X_2 factor is presented in Table 3.

For every line f of table of initial experimental data we construct the tables of virtual data in which we simultaneously bring in the values of two X_{ij} columns from the corresponding table of unrationed density probabilities(similar to Table 3) and the X_{if} column. Alignment pairs of columns X_{ij} and X_{il}, Y_j and Y_l should occur at the maximum probability density. The joining of edited tables occurs in numerical order table rows of input data. The result is a virtual sample that is presented in Table 4.

According to the experiment planning theory only independent factors are liable to modeling. At the next step according to full virtual sample table we determine the correlation coefficients of all the factors and all the output value according to the principle "everyone with everyone". The results are put in Table 5.

If a detailed analysis of coefficient pair correlation table is needed, it is recommended to use the correlation pleiades method [5] combined with an expert method of weighting importance coefficients [4].

Having analyzed the correlation matrix we conclude that the factor X_4 is strongly associated with factors X_1 and X_3. We combine three of these factors in the pleiad, choose a factor, which characterize the pleiad.

Then we start modeling through one of the methods, which help to receive adequate mathematical model processing passive data: method of least squares with pre-ortogonalization factors or random balance modified method.

Table 3. Table probability densities.

X_{1j}	X_{2f}							
	89,65	*90,6*	*92,5*	*95,8*	*99,4*	*100,9*	*102,6*	*108,5*
82,85	0,23	0,14	0,05					
83,89	0,34	0,24	0,09					
84,94	0,49	0,36	0,16	0,02				
85,99	0,65	0,50	0,26	0,05				
87,03	0,80	0,66	0,38	0,08				
88,08	0,92	0,81	0,53	0,15	0,02			
89,12	0,99	0,93	0,69	0,24	0,03	0,01		
90,17	0,99	0,99	0,84	0,36	0,06	0,02		
91,22	0,92	0,99	0,95	0,51	0,12	0,05	0,02	
92,26	0,80	0,91	1,00	0,67	0,19	0,09	0,03	
93,31	0,65	0,79	0,98	0,82	0,30	0,16	0,06	
94,36	0,49	0,64	0,90	0,94	0,44	0,25	0,11	
95,40	0,35	0,48	0,76	0,99	0,60	0,38	0,19	
96,45	0,23	0,33	0,61	0,99	0,76	0,53	0,30	
97,49	0,14	0,22	0,45	0,91	0,89	0,69	0,43	0,02
98,54	0,08	0,13	0,31	0,79	0,98	0,84	0,59	0,04
99,59	0,04	0,07	0,20	0,63	1,00	0,95	0,75	0,08
100,63	0,02	0,04	0,12	0,47	0,95	1,00	0,88	0,14
101,68		0,02	0,07	0,33	0,85	0,98	0,97	0,22
102,72			0,03	0,21	0,70	0,90	1,00	0,34
103,77			0,02	0,13	0,54	0,77	0,96	0,49
104,82				0,07	0,39	0,61	0,85	0,65
105,86				0,04	0,26	0,45	0,71	0,80
106,91				0,02	0,16	0,31	0,55	0,92
107,96					0,09	0,20	0,40	0,99
109,00					0,05	0,12	0,27	0,99
110,05					0,03	0,07	0,17	0,93
111,09					0,01	0,04	0,10	0,81
112,14						0,02	0,05	0,65
113,19							0,03	0,49

Table 4. Table of virtual sample.

Num of pro-duct	**Factors X_i**				**Output value, Y**
	X_{1f}	X_{2f}	X_{3f}	X_{4f}	
1	0,680	84,939	65,673	-29,795	49,640
2	0,683	85,985	65,870	-29,319	51,186
3	0,687	87,031	66,067	-28,843	52,731
4	0,690	88,078	66,264	-28,366	54,277

5	0,693	89,124	66,461	-27,890	55,822
6	0,696	90,170	66,658	-27,414	57,367
7	0,700	91,216	66,855	-26,938	58,913
8	0,703	92,263	67,052	-26,461	60,458
9	0,706	93,309	67,249	-25,985	62,004
10	0,709	94,355	67,446	-25,509	63,549
11	0,712	95,401	67,643	-25,033	65,095
12	0,716	96,448	67,840	-24,556	66,640
13	0,719	97,494	68,037	-24,080	68,185
14	0,722	98,540	68,234	-23,604	69,731
15	0,725	99,586	68,431	-23,128	71,276
16	0,641	98,540	68,431	-32,652	74,367
17	0,645	99,586	68,628	-32,176	75,913
18	0,648	100,632	68,826	-31,700	77,458
19	0,651	101,679	69,023	-31,224	79,003
20	0,654	102,725	69,220	-30,747	80,549
21	0,658	103,771	69,417	-30,271	82,094
22	0,661	104,817	69,614	-29,795	83,640
23	0,664	105,864	69,811	-29,319	85,185
24	0,667	106,910	64,490	-36,939	75,913
25	0,670	107,956	64,687	-36,462	77,458
26	0,674	109,002	64,885	-35,986	79,003
27	0,677	110,048	65,082	-35,510	80,549
28	0,680	111,095	65,279	-35,034	82,094
29	0,683	112,141	65,476	-34,557	83,640
30	0,687	113,187	65,673	-34,081	85,185
31	0,667	82,847	65,870	-32,652	57,367
32	0,670	83,893	66,067	-32,176	58,913
33	0,674	84,939	66,264	-31,700	60,458
34	0,677	85,985	66,461	-31,224	62,004
35	0,680	87,031	66,658	-30,747	63,549
36	0,683	88,078	66,855	-30,271	65,095
37	0,687	89,124	67,052	-29,795	66,640
38	0,690	90,170	67,249	-29,319	68,185
39	0,693	91,216	67,446	-28,843	69,731
40	0,696	92,263	67,643	-28,366	71,276
41	0,700	93,309	67,840	-27,890	72,822
42	0,703	94,355	68,037	-27,414	74,367
43	0,706	95,401	68,234	-26,938	75,913
44	0,709	96,448	68,431	-26,461	77,458
45	0,712	97,494	68,628	-25,985	79,003
46	0,716	98,540	68,826	-25,509	80,549
47	0,719	99,586	69,023	-25,033	82,094
48	0,722	100,632	69,220	-24,556	83,640
49	0,725	101,679	69,417	-24,080	85,185
50	0,729	102,725	69,614	-23,604	86,731
51	0,680	84,939	64,096	-33,605	60,458
52	0,683	85,985	64,293	-33,129	62,004
53	0,687	87,031	64,490	-32,652	63,549
54	0,690	88,078	64,687	-32,176	65,095
55	0,693	89,124	64,885	-31,700	66,640
56	0,696	90,170	65,082	-31,224	68,185
57	0,700	91,216	65,279	-30,747	69,731
58	0,703	92,263	65,476	-30,271	71,276
59	0,706	93,309	65,673	-29,795	72,822
60	0,709	94,355	65,870	-29,319	74,367
61	0,712	95,401	66,067	-28,843	75,913
62	0,716	96,448	66,264	-28,366	77,458
63	0,719	97,494	66,461	-27,890	79,003
64	0,722	98,540	66,658	-27,414	80,549
65	0,725	99,586	66,855	-26,938	82,094
66	0,729	100,632	67,052	-26,461	83,640
67	0,732	101,679	67,249	-25,985	85,185
68	0,735	102,725	67,446	-25,509	86,731
69	0,651	90,170	66,067	-32,176	57,367
70	0,654	91,216	66,264	-31,700	58,913
71	0,658	92,263	66,461	-31,224	60,458
72	0,661	93,309	66,658	-30,747	62,004
73	0,664	94,355	66,855	-30,271	63,549
74	0,667	95,401	67,052	-29,795	65,095
75	0,670	96,448	67,249	-29,319	66,640
76	0,674	97,494	67,446	-28,843	68,185
77	0,677	98,540	67,643	-28,366	69,731
78	0,680	99,586	67,840	-27,890	71,276
79	0,683	100,632	68,037	-27,414	72,822
80	0,687	101,679	68,234	-26,938	74,367
81	0,690	102,725	68,431	-26,461	75,913
82	0,693	103,771	68,628	-25,985	77,458
83	0,696	104,817	68,826	-25,509	79,003
84	0,700	105,864	69,023	-25,033	80,549
85	0,703	106,910	69,220	-24,556	82,094
86	0,706	107,956	69,417	-24,080	83,640
87	0,709	109,002	69,614	-23,604	85,185
88	0,712	110,048	69,811	-23,128	86,731
89	0,664	93,309	64,096	-34,557	68,185
90	0,667	94,355	64,293	-34,081	69,731
91	0,670	95,401	64,490	-33,605	71,276
92	0,674	96,448	64,687	-33,129	72,822
93	0,677	97,494	64,885	-32,652	74,367
94	0,680	98,540	65,082	-32,176	75,913

95	0,683	99,586	65,279	-31,700	77,458
96	0,687	100,632	65,476	-31,224	79,003
97	0,690	101,679	65,673	-30,747	80,549
98	0,693	102,725	65,870	-30,271	82,094
99	0,696	103,771	66,067	-29,795	83,640
100	0,700	104,817	66,264	-29,319	85,185
101	0,703	105,864	66,461	-28,843	86,731
102	0,706	106,910	66,658	-28,366	88,276
103	0,709	107,956	66,855	-27,890	89,821
104	0,712	109,002	67,052	-27,414	91,367
105	0,716	110,048	67,249	-26,938	92,912
106	0,719	111,095	67,446	-26,461	94,458
107	0,683	85,985	66,067	-33,605	49,640
108	0,687	87,031	66,264	-33,129	51,186
109	0,690	88,078	66,461	-32,652	52,731
110	0,693	89,124	66,658	-32,176	54,277
111	0,696	90,170	66,855	-31,700	55,822
112	0,700	91,216	67,052	-31,224	57,367
113	0,703	92,263	67,249	-30,747	58,913
114	0,706	93,309	67,446	-30,271	60,458
115	0,709	94,355	67,643	-29,795	62,004
116	0,712	95,401	67,840	-29,319	63,549
117	0,716	96,448	68,037	-28,843	65,095
118	0,719	97,494	68,234	-28,366	66,640
119	0,722	98,540	68,431	-27,890	68,185
120	0,725	99,586	68,628	-27,414	69,731
121	0,729	100,632	68,826	-26,938	71,276
122	0,732	101,679	69,023	-26,461	72,822

Table 5. Table of correlation coefficients

	X_1	X_2	X_3	X_4	Y
X_1	1	0,178	0,292	0,694	0,265
X_2	0,178	1	0,344	0,213	0,868
X_3	0,292	0,344	1	0,727	0,308
X_4	0,694	0,213	0,727	1	0,284
Y	0,265	0,868	0,308	0,284	1

Applying the method of least squares with pre-orthogonalization factors we built adequate mathematical models that are presented with their characteristics in Table 6.

As it is seen the received models have a lower dispersion adequacy and best calculated value of the Fisher criterion than the initial, and thus could be considered more operable.

Applying the modified method of random balance we built adequate mathematical models that are presented with their characteristics in Table 7.

Table 6. MLSO mathematical models.

	The adequacy dispersion	**The average weighted dispersion**	**Fisher criterion (*Ft=1,5*)**
$Y = -86,68 +55.148X_1 +1.2368X_2$	26,7164	28,084	0,9513
$Y = -37,97 +1.234X_2 +0.34848X_4$	26,9648	28,084	0,9601

Table 7. MRBM mathematical models.

	The adequacy dispersion	**The average weighted dispersion**	**Fisher criterion (*Ft=1,5*)**
$Y=72,204 +4,00X_1 +11,29X_2$	4,1423	27,5436	0,1504
$Y=71,62 +11,39X_2 +3,63X_3 - 5,55X_2X_3$	19,7125	25,2240	0,7815
$Y=72,262 +10,46X_2 +4,59X_4$	13,7916	25,6464	0,5378

As it is seen the received models also have a lower dispersion adequacy and best calculated value of the Fisher criterion than the initial.

The next task is to select the best model. After analyzing the constructed models we choose the most operable mathematical model.

These data indicate that this model is:

$Y=72,204 +4,00X_1 +11,29X_2$

It is noticeable that the resulting model includes the same factors that enter the model built on the initial data. However, the calculated characteristics, such as adequacy dispersion and Fisher criterion were significantly better.

3. CONCLUSIONS

1. Suggested a fundamentally new method of constructing adequate multidimensional models by small size samples.
2. Possibility of receiving more efficient model at application of a method of multidimensional pointed distributions is proved.
3. It is required the expansion of this method to different character data for solving various problems.

4. It is required to evaluate the impact of blunders on the experiments results.

5. It is required to develop software to facilitate non-normalized density probability and virtual data tabulation.

References

1. *Stolyarenko Y.A.* Control' kristallov integral'nih sxem na osnove statisticheskogo modelirovaniya metodom tochechnyx raspredelenij. Diss. na soisk. uch. stepeni kand. tech. nauk. – M..: GUP NPC «Spurt», 2006. 192 p.

2. *Dolgov A.Y.* Povyshenie effektivnosti statisticheskix metodov kontrolya i upravleniya tehnologicheskimi protzessami izgotovleniya mikrosxem. Diss. na soisk. uch. stepeni kand. tech. nauk. – M..: MGAPI, 2000. 218 p.

3. *Gaskarov D.V., Shapovalov V.I.* Malaya viborka. M.: Statistika, 1978. 248 p.

4. *Dolgov Y.A., Stolyarenko Y.A.* Modelirovanie: Uchebnoe posobie. Tiraspol: Izd-vo Pridnestrov-skogo universiteta, 2006. 96 p.

5. *Druzhinin G.V.* Metodi ocenki I prognoziro-vaniya kachestva. M.: Radio i svyaz'. 1982. 160 p.

Recommended for publication 23.09.2015