Analysis of Statistical Modeling Methods for Small-Volume Samples

Popukaylo V.
T.G. Shevchenko Dniester State University
Tiraspol, Moldova
vsp.science@gmail.com

***Abstract* — There are offered some statistical modeling methods for small-volume samples on the base of passive observational study which may be used for getting multidimensional adequate mathematical models on samples of small volume.**

***Index Terms* — statistical modeling, small-volume samples, passive observation, modeling of the technological process, point distribution method.**

I. Introduction

In the modern industry there are such productions which because of technological limitations cannot provide a sufficiently large sample volume, 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.

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.

II. Research Methods

In the work "Small-volume samples" (by DV 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 volume samples upper range limit n = 15 [1], and 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 sample volume 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 sample volume occurs in the lines with the highest level of offensive 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 volume.

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 volume sample, even if the initial small sample was supersaturated up.

III. Results And Discussion

Let's take as a result from the production n = 5 produc units (parties) the following numerical values of control parameters (Xi – the parameters controlled during technological process; Y – output quality indicator of a product. All names of dimensions for simplicity are ammited)

TABLE I. Table Of Initial Experimental Data

Number of product	Factors X_i		Output value, *Y*
	X_{1f}	X_{2f}	
1	0,549	2,1682	72,22
2	0,478	2,1371	71,65
3	0,607	1,8629	46,65
4	0,485	2,5204	65,8
5	0,441	2,4838	48,85
6	0,397	2,0652	43,87
7	0,257	2,0801	60,63
8	0,342	2,1557	80,84

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:

$$Y = -2297 + 1193{,}2X_1 + 1839{,}4X_2 - 744{,}38\, X_1 X_2 + \\ + 475{,}15\, X_1^2 - 325{,}63\, X_2^2$$

The adequacy dispersion of this model = 139.8398

The average weighted dispersion = 46.51099

Fisher criterion Fr = 3.007

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, here is presented a calculation for X_1factor.

TABLE II. Table Probability Densities

J	X_{1j}	X_{1f}							
		0,55	***0,48***	***0,61***	***0,49***	***0,44***	***0,40***	***0,26***	***0,34***
1	0,184						0,01	0,56	0,07
2	0,202						0,02	0,72	0,12
3	0,220						0,03	0,86	0,20
4	0,238					0,01	0,06	0,96	0,31
5	0,256					0,02	0,11	1,00	0,45
6	0,274		0,01		0,01	0,05	0,19	0,97	0,60
7	0,292		0,02		0,02	0,09	0,30	0,87	0,76

8	0,310		0,05		0,03	0,15	0,44	0,74	0,89
9	0,328		0,08		0,07	0,25	0,59	0,58	0,98
10	0,346	0,01	0,15		0,12	0,37	0,75	0,42	1,00
11	0,364	0,02	0,24		0,20	0,52	0,89	0,29	0,95
12	0,382	0,05	0,36		0,31	0,68	0,97	0,18	0,84
13	0,400	0,09	0,51	0,01	0,45	0,83	1,00	0,11	0,70
14	0,418	0,15	0,67	0,02	0,61	0,94	0,95	0,06	0,54
15	0,436	0,24	0,82	0,04	0,77	1,00	0,85	0,03	0,38
16	0,453	0,37	0,94	0,08	0,90	0,98	0,71	0,01	0,26
17	0,471	0,52	1,00	0,13	0,98	0,90	0,55	0,01	0,16
18	0,489	0,68	0,99	0,22	1,00	0,77	0,39		0,09
19	0,507	0,83	0,91	0,34	0,95	0,62	0,26		0,05
20	0,525	0,94	0,78	0,48	0,84	0,46	0,17		0,03
21	0,543	1,00	0,63	0,64	0,69	0,32	0,10		0,01
22	0,561	0,98	0,47	0,79	0,53	0,21	0,05		0,01
23	0,579	0,91	0,33	0,92	0,38	0,12	0,03		
24	0,597	0,78	0,21	0,99	0,25	0,07	0,01		
25	0,615	0,62	0,13	0,99	0,16	0,04	0,01		
26	0,633	0,46	0,07	0,93	0,09	0,02			
27	0,651	0,32	0,04	0,81	0,05	0,01			
28	0,669	0,21	0,02	0,66	0,02				
29	0,687	0,13	0,01	0,50	0,01				
30	0,705	0,07		0,35	0,01				

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 II) and the X_{if} column. Alignment (joining) pairs of columns X_{ij} and X_{il} (and) Y_j and Y_l should occur at the maximum probability density level. The joining of edited tables occurs in numerical order table rows of input data. The result is a complete virtual sample that is presented in Table III.

TABLE III. TABLE OF VIRTUAL SAMPLE

Number of product	Factors X_i		Output value, *Y*
	X_{1f}	X_{2f}	
1	0,346	1,782	49,311
2	0,364	1,817	51,493
3	0,382	1,852	53,675
4	0,400	1,887	55,858
5	0,418	1,922	58,040
6	0,436	1,957	60,223
7	0,453	1,992	62,405
8	0,471	2,027	64,587
9	0,489	2,062	66,770
10	0,507	2,097	68,952
11	0,525	2,132	71,134
12	0,543	2,167	73,317
13	0,561	2,202	75,499
14	0,579	2,237	77,682
15	0,597	2,272	79,864
16	0,615	2,307	82,046
17	0,633	2,342	84,229
18	0,651	2,376	86,411
19	0,669	2,411	88,593
20	0,687	2,446	90,776
21	0,274	1,747	47,128
22	0,292	1,782	49,311
23	0,310	1,817	51,493
24	0,328	1,852	53,675
25	0,346	1,887	55,858
26	0,364	1,922	58,040
27	0,382	1,957	60,223
28	0,400	1,992	62,405
29	0,418	2,027	64,587
30	0,436	2,062	66,770
31	0,453	2,097	68,952
32	0,471	2,132	71,134
33	0,489	2,167	73,317
34	0,507	2,202	75,499
35	0,525	2,237	77,682
36	0,543	2,272	79,864
37	0,561	2,307	82,046
38	0,579	2,342	84,229
39	0,597	2,376	86,411
40	0,615	2,411	88,593
41	0,633	2,446	90,776
42	0,651	2,481	92,958
43	0,507	1,677	36,217
44	0,525	1,712	38,399
45	0,543	1,747	40,581
46	0,561	1,782	42,764
47	0,579	1,817	44,946
48	0,597	1,852	47,128
49	0,615	1,887	49,311
50	0,633	1,922	51,493
51	0,651	1,957	53,675
52	0,669	1,992	55,858

53	0,687	2,027	58,040
54	0,705	2,062	60,223
55	0,274	2,097	40,581
56	0,292	2,132	42,764
57	0,310	2,167	44,946
58	0,328	2,202	47,128
59	0,346	2,237	49,311
60	0,364	2,272	51,493
61	0,382	2,307	53,675
62	0,400	2,342	55,858
63	0,418	2,376	58,040
64	0,436	2,411	60,223
65	0,453	2,446	62,405
66	0,471	2,481	64,587
67	0,489	2,516	66,770
68	0,507	2,551	68,952
69	0,525	2,586	71,134
70	0,543	2,621	73,317
71	0,561	2,656	75,499
72	0,579	2,691	77,682
73	0,274	2,167	29,669
74	0,292	2,202	31,852
75	0,310	2,237	34,034
76	0,328	2,272	36,217
77	0,346	2,307	38,399
78	0,364	2,342	40,581
79	0,382	2,376	42,764
80	0,400	2,411	44,946
81	0,418	2,446	47,128
82	0,436	2,481	49,311
83	0,453	2,516	51,493
84	0,471	2,551	53,675
85	0,489	2,586	55,858
86	0,507	2,621	58,040
87	0,525	2,656	60,223
88	0,543	2,691	62,405
89	0,274	1,817	29,669
90	0,292	1,852	31,852
91	0,310	1,887	34,034
92	0,328	1,922	36,217
93	0,346	1,957	38,399
94	0,364	1,992	40,581
95	0,382	2,027	42,764
96	0,400	2,062	44,946
97	0,418	2,097	47,128
98	0,436	2,132	49,311
99	0,453	2,167	51,493
100	0,471	2,202	53,675
101	0,489	2,237	55,858
102	0,507	2,272	58,040
103	0,525	2,307	60,223
104	0,543	2,342	62,405
105	0,561	2,376	64,587
106	0,579	2,411	66,770
107	0,597	2,446	68,952
108	0,184	1,957	51,493
109	0,202	1,992	53,675
110	0,220	2,027	55,858
111	0,238	2,062	58,040
112	0,256	2,097	60,223
113	0,274	2,132	62,405
114	0,292	2,167	64,587
115	0,310	2,202	66,770
116	0,328	2,237	68,952
117	0,346	2,272	71,134
118	0,364	2,307	73,317
119	0,382	2,342	75,499
120	0,400	2,376	77,682
121	0,418	2,411	79,864
122	0,436	2,446	82,046
123	0,453	2,481	84,229
124	0,184	1,852	62,405
125	0,202	1,887	64,587
126	0,220	1,922	66,770
127	0,238	1,957	68,952
128	0,256	1,992	71,134
129	0,274	2,027	73,317
130	0,292	2,062	75,499
131	0,310	2,097	77,682
132	0,328	2,132	79,864
133	0,346	2,167	82,046
134	0,364	2,202	84,229
135	0,382	2,237	86,411
136	0,400	2,272	88,593
137	0,418	2,307	90,776
138	0,436	2,342	92,958

As according to the experiment planning theory only independent factors are liable to modeling, at the next step according to full of 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 IV.

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].

TABLE IV. TABLE OF CORRELATION COEFFICIENTS

	X_1	X_2	Y
X_1	1	0,441	0,448
X_2	0,441	1	0,551
Y	0,448	0,551	1

Having analyzed the correlation matrix we conclude that the factors are independent and we start modeling through one of the methods, which helps to receive adequate mathematical model by passive data.

Applying the method of least squares with pre-orthogonalization factors we were not able to build adequate mathematical model.

Applying the modified method of random balance was formed following adequate mathematical model:

$$Y = 62,005 + 6,23X_1 + 8,26X_2 + 9,17\, X_1\, X_2$$

(with numerical value codes $-1 \leq X_i \leq +1$).

The adequacy dispersion of this model = 44,83844

The average weighted dispersion = 140,9533

Fisher criterion Fr =0,3181085 when the tabulated value is Ft = 1,5

The received model has a lower dispersion adequacy and best calculated value of the Fisher criterion than the initial, and thus can be considered more operable.

IV. CONCLUSIONS

1. Suggested a fundamentally new method of constructing adequate multidimensional models by small volume 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 and for solving various problems.

REFERENCES

[1] Столяренко Ю.А. Контроль кристаллов интегральных схем на основе стати-стического моделирования методом точечных распределений. – Дисс. на соиск. уч. степ. канд. техн. наук. – М..: ГУП НПЦ «Спурт», 2006. – 192 с.

[2] Долгов А.Ю. Повышение эффективности статистических методов контроля и управления технологическими процессами изготовления микросхем. – Дисс. на со-иск. уч. степ. канд. техн. наук. – М..: МГАПИ, 2000. – 218 с.

[3] Гаскаров Д.В., Шаповалов В.И., Малая выборка. – М.: Статистика, 1978. – 248 с.

[4] Долгов Ю.А. Столяренко Ю.А. Моделирование: Учебное пособие – Тирасполь: Изд-во Приднестровского университета, 2006. – 96 с.

[5] Дружинин Г. В. Методы оценки и прогнозирования качества. – М.: Радио и связь, 1982 – 160 с.