Hua-Xin.; Sch. J. Phys. Math. Stat., 2015; Vol-2; Issue-3A (Jun-Aug); pp-335-340

Scholars Journal of Physics, Mathematics and Statistics **ISSN 2393-8056 (Print)**

Sch. J. Phys. Math. Stat. 2015; 2(3A):335-340 **ISSN 2393-8064 (Online)**

(An International Publisher for Academic and Scientific Resources)

Statistical Model in Data Merging

Hua-Xin

Northeast Petroleum University, Daqing City of Heilongjiang Province, 163318, China

***Corresponding Author:**

Hua-Xin

Email: xinhuayatou@126.com

Abstract: Geologists through two different kinds of instruments and equipment distribution respectively of two particle size of molecules. Based on statistical theory, this paper from three perspectives, by using the methods of regression analysis, parameter estimation and probability. At the same time, considering the particularity of the experimental data, several statistical methods of merging two sets of data are given. Specific solution including the search for the optimal function and least square method, probability calculation, etc. And deduces the corresponding results, and the advantages and disadvantages of the three methods are described in detail, provides the certain model reference for related data analysis personnel.

Keywords: Data Consolidation; Regression,; Parameter Estimation.

INTRODUCTION

Geologists by nitrogen adsorption method, can measure the distribution of small particles of minerals molecule radius of less than 63 nm, large particles can be measured distribution of molecules larger than 6.3 nm, the radius of minerals by mercury method.

Chart 1 and Table 2 are small molecule distribution measured by a nitrogen adsorption method particles, Table 3 is obtained by the mercury distribution of molecules of large particles, wherein the apertures in Table 2 and Table 3 in front of the six identical particles aperture, That radius is located 6.3 nm to 63 nm particles of the two instruments have measurement results. Target merge data from two instruments obtained through certain statistical methods into line on a map. For data obtained by different methods, often you need to merge together during data analysis, but because of the different data sources so that in most cases can not be directly merged. For different problems, many documents are given a number of ways. Obviously, different data characteristics, methods are not the same the same. So not only consider the issue, but also from the data characteristics of view, to find the most effective approaches and methods to achieve the most effective result of the merger.

Table-1: Small particle data distribution frequency

No.	1	2	3	4	5	6	7	8	9	10	11	12	13
Aperture	0.015	0.025	0.04	0.065	0.1	0.16	0.25	0.4	0.63	1	1.6	2.5	4
Frequency	0.002	0.003	0.006	0.008	0.014	0.02	0.034	0.052	0.874	3.067	3.004	3.771	5.684

Table-2: Small particle data distribution frequency

No.	14	15	16	17	18	19
Aperture	6.3	10	16	25	40	63
Frequency	8.322	11.967	9.913	16.516	25.333	11.408

Table-3: The frequency distribution of large particle data

No.	1	2	3	4	5	6	7	8	9	10	11
Aperture	6.3	10	16	25	40	63	100	160	250	400	630
Frequency	3.219	5.926	6.659	8.34	8.852	10.531	11.958	12.958	16.006	11.691	3.854

PREFERENCES

A total of $p=19$ kinds of small particles of molecular content, set the content of each is $\alpha_i, i=1,2,\cdots 19$, $\sum_{i}^{19}\alpha_i=1$, a total of $q=11$ kinds of large particles of molecular content, the amount of each is $\beta_j, j=1,2,\cdots 11$, $\sum_{j}^{11}\beta_j=1$.The second set of data in which small particles and former six kinds of particles larger particles molecules are measured in the two instruments, that were obtained in both instruments in their population share some frequency.

SOLUTIONS

- The optimal function using cross measurement data

The information histograms (Fig-1) can be obtained through 6 group of overlapping data, seen from the graph, two coincident data is linear distribution. We can know that the quadratic fit variance is minimal by finding the best curve fitting, , but taking into account the actual measurement, edge data error is large, and based on past experience test, we used a simple regression as a fitting function [5,6].

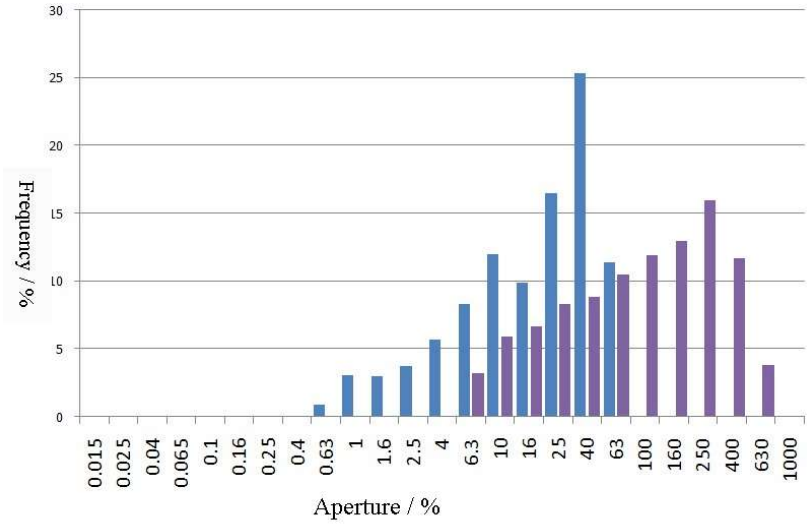

Fig-1: The particle distribution of two kinds

Considering the instrument edge measurement error is large, take the middle four readings, set the regression equation is $y=kx+\varepsilon$ .You can use the least squares method to find the estimator k .Set $y_i=kx_j+\varepsilon_j$,let $i=15, j=2, i=16, j=3, i=17, j=4, i=18, j=5$;

Table-4: Regression parameter table

Coefficient a,b								
Model		Unstandardized Coefficients		Standardized Coefficients	T	Significance	B ^1 Confidence Interval	
		B	Standard error	Beta			Lower limit	Upper limit
1	VAR00005	2.194	.300	.973	7.322	.005	1.241	3.148
a.Response variable\: VAR00001								
b. Through the origin of linear regression								

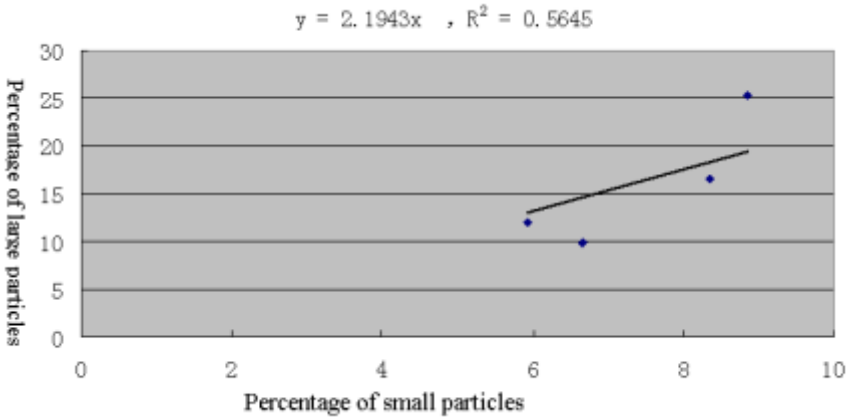

Fig-2: Simple regression best fit function diagram

From the regression results and maps can be obtained parameter estimates is： $\hat{k}=2.19$ ；

We put large particle data β completely mapped to a small particle data α , $\hat{y}=k\hat{x}$.

It should be noted, $\hat{y}_i \neq \alpha_i$, so we can take the average of $\hat{y}_i, \alpha_i$ asthe final reading of the cross data block .

Table-5: Cross data expectations

$\hat{y}_j$	7.050	12.978	14.583	18.265	19.386	23.063
α_j	8.322	11.967	9.913	16.516	25.333	11.408
$\hat{\alpha}_j=(\hat{y}_j+\alpha_j)/2$	7.686	12.472	12.248	17.391	22.359	17.235

Table-6: Small particles frequency distribution $(\hat{\alpha}_i, i=1,2,\ldots,13)$

No.	1	2	3	4	5	6	7	8	9	10	11	12	13
Aperture	0.015	0.025	0.04	0.065	0.1	0.16	0.25	0.4	0.63	1	1.6	2.5	4
Frequency	0.002	0.003	0.006	0.008	0.014	0.02	0.034	0.052	0.874	3.067	3.004	3.771	5.684

Table-7: Frequency distribution of large particles $(\hat{\alpha}_i, i=14,15,\ldots,24)$

No.	14	15	16	17	18	19	20	21	22	23	24
Aperture	6.3	10	16	25	40	63	100	160	250	400	630
Frequency	7.686	12.472	12.248	17.391	22.359	17.235	26.188	28.378	35.053	25.603	8.440

Based on Table-5, Table-6, Table-7 units of either obtain a uniform distribution namely:

$$\hat{\eta}_i=\frac{\hat{\alpha}_i}{\sum_{i=1}^{21}\hat{\alpha}_i}, i=1,2,\cdots 21, \quad (1)$$

The calculation results to omit.

- The overall ratio obtained by the two sets of measurement data

Set the overall ratio of the prior probability of large particles and the small particles is λ ,let ϖ_1 represents a total volume of small particles, let ϖ_2 represents the total volume of large particles. There are $p=19$ kind of small particles

of molecular content, each content accounted for $\alpha_i, i=1,2,\cdots 19$, then $\frac{x_i}{\varpi_1}=\alpha_i$, x_i represents the i - th component of the total volume. Large particles molecule includes a total of 11 kinds of contents, each content representing $\beta_j, j=1,2,\cdots,11$, then $\frac{y_j}{\varpi_2}=\beta_j$, y_j represents the j - the component of the total volume, which has set up six kinds of mineral particles data duplication while belong to two particle component.

Then $x_i=y_j$,when $i=14, j=1, i=15, j=2, i=16, j=3, i=17, j=4, i=18, j=5, i=19, j=6$, that is $\varpi_1\times\alpha_i\%=\varpi_2\times\beta_j\%$,set up $\frac{\varpi_2}{\varpi_1}=\lambda$,on behalf of the six sets of data to obtain $\hat{\lambda}_1=2.585$, $\hat{\lambda}_2=2.019$, $\hat{\lambda}_3=1.488$, $\hat{\lambda}_4=1.980$, $\hat{\lambda}_5=2.862$, $\hat{\lambda}_6=1.083$,then $\hat{\lambda}=\sum_{i=1}^{6}\hat{\lambda}_i$;Consider the actual measurement, the edge reliability of the instrument measured relatively weak, so the first set and the sixth group estimates removed, and finally we take the middle value of four groups seeking to expect $\hat{\lambda}=\sum_{i=2}^{4}\hat{\lambda}_i=2.003$.

Table-8: Cross data expectations

$\hat{\alpha}_j$	6.438	11.852	13.318	16.68	17.704	21.062
α_j	8.322	11.967	9.913	16.516	25.333	11.408
$(\hat{\alpha}_j+\alpha_j)/2$	7.380	11.910	11.616	16.598	21.519	16.235

Processing units the 24 data's,

$$\hat{\eta}_i=\frac{\hat{\alpha}_i}{\sum_{i=1}^{21}\hat{\alpha}_i}, i=1,2,\cdots 21,$$

The results may be calculated as follows:

Table-9: Small particles of molecular frequency distribution

Aperture	0.015	0.025	0.04	0.065	0.1	0.16	0.25
Frequency	0.00093	0.00140	0.00279	0.00373	0.00652	0.00931	0.01583
Aperture	0.4	0.63	1	1.6	2.5	4	
Frequency	0.02422	0.40702	1.42831	1.39897	1.75616	2.64705	

Table-10: Cross data frequency distribution

Aperture	6.3	10	16	25	40	63
Frequency	3.437	5.547	5.410	7.730	10.021	7.561

Table-11: Large particles molecular frequency distribution

Aperture	100	160	250	400	630
Frequency	11.138	12.069	14.908	10.889	3.590

Demand ratio based on three sets of data expected

Set up the 13 kinds of small particles in front of the small particles in general as A, the total proportion of small particles $P(w_1)$ is:

$$w_1\% = \sum_{i=1}^{13} \alpha_i = 16.539\%$$

Set up the 14th to the 19th of small particles in the particles generally is B, the proportion of small particles in the range of $P(w_2)$ is :

$$w_2\% = \sum_{14}^{19} \alpha_i = 83.461\%,$$

Set up overall pre-six kinds of large particles is B', a large proportion of particles $P(w_2)'$ is:

$$w_2'\% = \sum_{j=1}^{6} \beta_j = 43.527\%$$

Set up large particles on the 6th to the 11th particles overall is C, the content is:

$$w_c\% = \sum_{j=7}^{11} \beta_i = 56.467\%$$

We set up total volume $\omega_1, \omega_2, \omega_3$ respectively for the three-part can be obtained as the volume ratio of three parts are:

$$E(\omega_1) : E(\omega_2) : E(\omega_3) = 0.19816 : 1 : 1.29729$$

Then, the three parts are added together, the first part of ω_1 total content of 7.941%, the second part of ω_2 total content of 40.073%,the third part of ω_3 total content of 51.986%,

$$\hat{\alpha}_i = \alpha_i \times (7.941\% + 40.073\%), i = 1, 2, \square, 13 \qquad (2)$$

$$\hat{\beta}_j = \beta_j \times (51.986\% + 40.073\%), j = 1, 2, \square, 11 \qquad (3)$$

When $i = 14, 15, \ldots 19$, $j = 1, 2, \square, 6$ use （2）, （3）calculating average is:

$$\hat{\lambda}_i = \frac{1}{2}\left(\hat{\alpha}_i + \hat{\beta}_i\right), i = 14, 15, \square, 19, j = 1, 2, \ldots 6 \qquad (4)$$

In formula (1) the calculation results are as follows:

Table-12: Frequency distribution of small particles molecule

Aperture	0.015	0.025	0.04	0.065	0.1	0.16	0.25
Frequency	0.00096	0.00144	0.00288	0.00384	0.00672	0.00960	0.01633
Aperture	0.4	0.63	1	1.6	2.5	4	
Frequency	0.02497	0.41964	1.47259	1.44234	1.81061	2.72912	

Table-13:The cross section of the frequency distribution after averaging

Aperture	6.3	10	16	25	40	63
Frequency	3.480	5.601	5.445	7.804	10.156	7.586

Table- 14: Large particles molecular frequency distribution

Aperture	100	160	250	400	630
Frequency	10.254	11.111	13.724	10.024	3.305

CONCLUSION

This selection of the three methods from three different sides to analyze the data, although starting from different angles, but the result is similar to the overall distribution of frequencies listed roughly the same, the overall ratio of large particles and small particles of molecular molecule relations are basically 2 times.

Among them, the first method uses the measurement data obtained cross, finding the optimal function of two readings, the higher the accuracy of this method, but the reading more cases for use, if the data is less than the error is relatively large.

The second method "The overall ratio obtained by the two sets of measurement data" and the third way "Demand ratio based on three sets of data expected "are used to calculate the desired probability by evaluating the various components. The second method is not strictly divided large particles and small particles, irrespective the situation of cross data exists to calculate the ratio of the two population parameters, and then seek a desired of ratio parameters. Make full use of information provided in each reading. The third method is the strict molecular particles into three states, three in the particulate matter directly to obtain a desired ratio. Obviously, if the test data accuracy is high, we are more willing to choose the second scenario, if each set of measurement accuracy is not particularly good, but the overall measurement error is not large, we will use the third method. Therefore, when do the actual data processing problem, should not arrested in a way, and to consider the characteristics of the experimental data, characteristics, in order to find the optimal solution.

Acknowledgement: Youth Foundation of Northeast Petroleum (Data mining technique and algorithm research) No:NEPUQN2014-24

REFERENCE

1. Dempster A P; Upper and Lower probabilities induced by a multi valued mapping. Annals of Mathematical Statistics, 1967; 138:325-339.
2. Demribas K; Distributed sensor data fusion with binary decision tree. IEEE Trans Areosp Syst Electron, 1989; 25(5):643-649.
3. Xiaoyong-Zheng, Jingshun-Yao; A data fusion algorithm is put forward. Mathematics in Practice and Theory, 2003; 152-158.
4. Pan Q, Wang ZF, Liang Y, Yang F, Liu ZG; Basic methods and progress of information fusion(II). Control Theory & Applications, 2012; 29(10): 1233-1244.
5. Jianping-Zhu, Ruifei-Duan; The application of SPSS in statistical analysis. The first edition. Beijing: tsinghua university press, 2007; 30-32.
6. Huixuan-Gao; Applied Multivariate Statistical Analysis. The first edition. Beijing: Peking University press, 2005; 5-118.