The analysis and evaluation method for small samples from the perspective of regression

YIN Boya[1*] LI Chenyi[2]

[1] *Department of Mathematics college of science, Zhejiang University of Technology;*

[2] *JianXing Honors College, Zhejiang Universities of Technology, Hangzhou City, Zhejiang Province, China, 310023*

Received 1 November 2014, www.cmnt.lv

Abstract

It is well-known that regression analysis often suffers from the small sample problem while big data is a necessity to greatly improve the credibility of research. We propose an analysis and evaluation method to judge the quality of small samples accurately through taking full advantage of these representative data. To achieve it, regression analysis was adopted to describe and expand the small amount of data. Meanwhile, principal component analysis would be performed to give the comprehensive evaluation of the forwarded data. We demonstrate the method by taking micro alloying of adding aluminium and rare earth as an example to explain the feasibility and accuracy of this analysis system. It proceeds as follows: stepwise regression →data expansion and forwarding→ principal component analysis. The paper ends with recommendations that adding rare earth benefited the ensemble more to illustrate this analysis method can describe the general trend of data.

Key words: polynomial interpolation, regression analysis, stepwise regression, positive data, principal component analysis

1 Introduction

It's known that the greater amount of data benefits the statistics more. While in the engineering application, a huge-invested experiment usually got only a group of data. Thus to reduce the cost, it's the most direct way to do the experiment by using the most representative data. With these data analysed, a better result would be obtained to do experiment examination again, which can greatly limit waste. [7]

Here we need to make such two appointments. Firstly, we should avoid the likely distortion of data itself when performing the analysis of the small samples. Otherwise, approximate analysis method and data distortion would eventually cause severe deviation. Secondly, sample data must be representative. It's told that the trend of data will not have a large fluctuation in a small range. We can approximately turn the trend into a simple shape of straight line or parabola. This is the basic theory to be used in the small sample analysis through regression.

In the condition of small samples, we can no longer blindly pursue goodness of fit of data. Ensuring the strong ability of trend description for the fitting results counts more. In the case of traditional interpolation, statistics could be completely fitted by seven interpolations with 8 points known. But there will be a serious phenomenon of Runge leading to the distortion.[1]~[3] So regression at about two times is reasonable in small samples. Meanwhile, adding the cross terms in dealing with the multidimensional is more convenient, which can describe the general trend of the data and present good results on the goodness of fit after many experiments.

These papers were analysed with a real problem "Micro alloying of aluminium alloy in changing properties of it". It demonstrated that we used regression to fit and expand the data of small sample with the complete process of principal component analysis for data evaluation. On the premise of the above two appointments, this set of data analysis and evaluation method has a strong universality.

2 The statement of the problem

Micro alloying is an important approach to improving the microstructure and performance of aluminium alloy. In the laboratory, we determined the tensile strength (MPa) and elongation (%) according to respectively adding a trace of vanadium (V) and rare earth (RE) to the base alloy, Al-4.5Zn-1.0Mg-0.8Cu. Hot cracking statistics (HCS) were shown in the following two tables.

TABLE1. Mechanical properties and thermal cracking performance of products under respectively adding vanadium (V) and rare earth (RE)

V(wt %)	Tensile strength (MPa)	Elongation (%)	HCS	RE(wt%)	Tensile strength (MPa)	Elongation (%)	HCS
0	133	4.66	112	0	133	4.66	112
0.05	160	6.60	83	0.05	165	7.58	88
0.1	154	8.21	100	0.08	169	8.26	80
0.2	140	5.10	131	0.1	171	8.99	72
				0.12	186	8.73	64
				0.15	175	7.71	72
				0.2	156	7.00	80
				0.25	152	6.49	116

[*] ***Corresponding author's*** e-mail: 549773255@qq.com

TABLE2. Mechanical properties and thermal cracking performance of products under adding the mixture of vanadium (V) and rare earth (RE)

Test serial number	V (wt%)	RE (wt%)	Tensile strength (MPa)	Elongation (%)	HCS
1	0.05	0	160	6.60	83
2	0.05	0.1	153	7.22	80
3	0.05	0.25	130	5.57	80
4	0	0.12	186	8.73	64
5	0.05	0.12	146	6.66	68
6	0.1	0.12	133	6.00	74

The main emphasis was placed on the problem of how and how much to add the vanadium and rare earth would perform effectively through the analysis of data from these two tables.

3 The adding respectively data's expansion and evaluation (Table1)

For a small amount of data, blindly pursuing goodness of fit wasn't advocated. Taking separately adding RE for example, there were eight points, using seven times polynomial to fit the data could completely guarantee that residual error was 0. But the generated Runge phenomenon would cause serious distortions of pictures, ultimately made accurately analysing the data be impossible.

It was necessary to assurance the rough description of this category of data. On this basis, appropriate methods could be adopted to improve the goodness of fit. According to the authors' experiment, this kind of data fitted in 1 ~ 3 times of unilabiate or multivariate regression could achieve the ideal effect.

3.1 THE FITTING RESULTS OF ADDING VANADIUM

Define ensile strength as pv_1, elongation as pv_2,hot cracking statistics as pv_3,the content of vanadium as x.

Due to only four data of separately adding vanadium presented, Runge phenomenon wouldn't appear. In this condition, polynomial fitting could be operated directly to acquire result:

$$pv_1(x) = 3.233\times10^{10}x^3 - 1.145\times10^{8}x^2 + 1.032\times10^{5}x + 133$$

$$pv_2(x) = -1.78\times10^{9}x^3 + 2.01\times10^{6}x^2 + 3320x + 4.66$$

$$pv_3(x) = -4.7\times10^{10}x^3 + 1.625\times10^{8}x^2 - 1.275\times10^{5}x + 112$$

3.2 THE REGRESSION RESULTS OF ADDING RARE EARTH

Define ensile strength as pre_1, elongation as pre_2,hot cracking statistics as pre_3,the content of rare earth as x.

In consideration of eight groups of statistics of separately adding rare earth presented, if interpolation was used, it would cause serious Runge phenomenon. Therefore, regression analysis would be considered to describe the directions of data roughly without losing numerous precisions. It was more appropriate to take 2 ~ 3 times.

$$pre_1(x) = -2.27\times10^{7}x^2 + 6.10\times10^{4}x + 136$$

Decision coefficient: 0.846 F statistic: 13.69

$$pre_2(x) = 1.27\times10^{9}x^3 - 6.60\times10^{6}x^2 + 9344x + 4.59$$

Decision coefficient: 0.970 F statistic: 43.45

$$pre_3(x) = 2.88\times10^{7}x^2 - 7.26\times10^{4}x + 114.8$$

Decision coefficient: 0.962 F statistic: 63.20

3.3 THE COMPREHENSIVE EVALUATION

Carrying on the comprehensive evaluation on the data of tensile strength, elongation and hot cracking statistics helped us choose the best program to add vanadium and rare earth with the best influence.

Principal component analysis was responsible to sort and process data here.

3.4 THE EXPANSION OF DATA

We considered using the pre-existing regression equation to expand data for lack of original data. It implemented through mat lab selecting small intervals in a reasonable range. This article took 0.001% as a interval, and adding no more than 0.25% of vanadium and rare earth.

3.5 THE FORWARDING OF DATA

Referring to the experience and reference [3], the implementation of principal component analysis required forwarding the data, or it would fail to obtain reasonable results. Specifically, the higher the comprehensive score got, the better data offered.

Methods of data forwarding: [5]

1).If the data is required as big as possible, data forwarding wouldn't be necessary.

2).If the data is required as small as possible, $1/x_{ij}$ would be used instead of the original data.

3.6 PRINCIPAL COMPONENT ANALYSIS

Principal component analysis was carried out on the expanded and forwarded data.

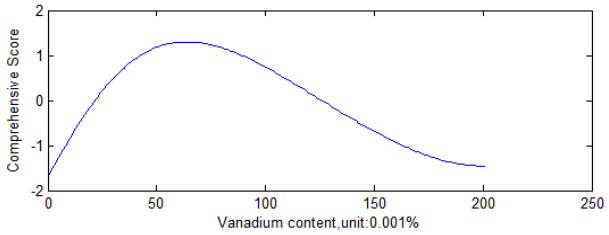

FIGURE 1 Composite Scores of Adding Vanadium Alone

Around 60, namely adding 0.06% of vanadium, the overall situation was good. Through calculation, when adding 0.062% of vanadium, the overall situation reached peak. Three properties were as follows:

tensile strength: 160MPa	elongation: 6.99%	HCS: 83.848

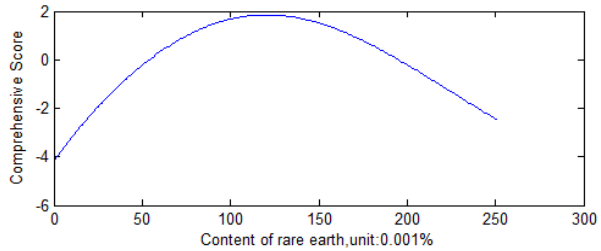

FIGURE 2 Composite Scores of Adding Rare Earth Alone

Through calculation, when adding 0.12% of rare earth, the overall situation reached peak.

tensile strength: 176.4MPa	elongation: 8.48%	HCS: 69.1

4 The adding mixture data's expansion and evaluation (Table2 and Table1)

Define ensile strength, elongation and hot cracking statistics as $pvre_i$, i=1, 2, 3, the content of vanadium as v, the content of rare earth as re.

Here the theory of stepwise regression was mainly applied. From the above results, it accessed rough-trend function which had returned or fitted the data less than three times to describe the data. So the following regression models were given.

$$pvre_i(v,re) = c_1 v + c_2 v^2 + c_3 v^3 + c_4 re + c_5 re^2 + c_6 re^3 + c7$$

Besides the existing adding content of vanadium and rare earth, those without being added were defined as 0 and added to the data of mixed additives to do regression of the increased data.

Nevertheless, the consequence of regression didn't confront with previous intention. In terms of tensile strength, its decision coefficient was only 0.4445 while F statistic was 1.4672. Although what we need was the trend function, the low coefficient of decision would lead to serious distortion of the data, which suggested the improvement of model.

4.1 THE REGRESSION RESULTS OF ADDING MIXTURE

According to the experience, it was guessed that the interaction between the contents of vanadium and rare earth in the model might affect these three properties. Naturally, the cross terms including $v \times re$, $v^2 \times re$, $v \times re^2$ were chose to represent the interaction and improve the model.

$$pvre_i(v,re) = c_1 v + c_2 v^2 + c_3 v^3 + c_4 re + c_5 re^2 + c_6 v \times re + c_7 v^2 \times re + c_8 v \times re^2 + c_9 re^3 + c_{10}$$

i=1,2,3

It aroused the greatly increase on the determination coefficient which were all over 0.9. F statistics also fully

Embodied the significance of the model. But some coefficients weren't significant enough, which propelled these coefficients to be rejected. Therefore, we took the stepwise regression into consideration. Eventually, the following results were produced.

$$pvre_1(v,re) = 42712.41v - 2.06\times10^7 v^2 + 59570.60re - 2.24\times10^7 re^2 - 9.33\times10^7 v\times re + 2.70\times10^{10} v\times re^2 + 137.436$$

$$pvre_2(v,re) = 4575.44v - 1.08\times10^9 v^3 + 9478.94re - 6.67\times10^6 re^2 - 1.21\times10^7 v\times re + 3.19\times10^9 v^2\times re + 3.22\times10^9 v\times re^2 + 1.27\times10^9 re^3 + 4.57$$

$$pvre_3(v,re) = -1.26\times10^5 v + 1.60\times10^8 v^2 - 4.62\times10^{10} v^3 - 4.82\times10^4 re + 1.53\times10^8 v\times re - 8.05\times10^{10} v^2\times re - 4.77\times10^{10} v\times re^2 + 8.03\times10^9 re^3 + 111.835$$

4.2 THE ANALYSIS OF MIXED ADDITIVE'S RESULT

Dualistic equation given inspired us of the contour map to describe mixed additive's result. It was easy to promote the conclusion that 0.1 ~ 0.15% of rare earth and 0.01 ~ 0.02% of vanadium performed best by observing the tensile strength in the Figure 3. Elongation and hot cracking statistics could be similarly available without details here.

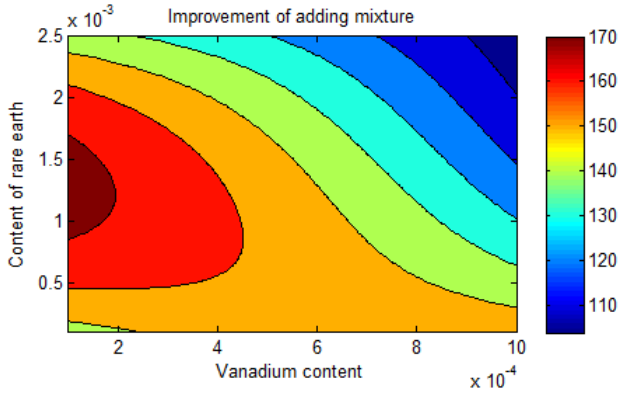

FIGURE3. Ascending Goodness of Tensile Strength When Adding Mixtures

According to the existed original data, it was concluded that adding rare earth benefited the ensemble than adding vanadium. Our suggested content for rare Earth is 0.12% while for vanadium is 0.062%. Meanwhile, when mixed adding, vanadium tended to affect the function of rare earth by decreasing the quality of the it. In consequence, mixed additive of vanadium and rare earth is not recommended.

Surely, error is inevitable, for this conclusion of analysis was based on small sample. This analysis method still can describe the general trend of data, and the procedure of method operates as regression (generally choose 2 ~ 3 regression equation as well)→ data expansion→ data forwarding→ principal component analysis and evaluation. The method can deduce splendidly for similar problems

5 Conclusions

Paper has mainly carried on the demonstration of the idea of using regression and principal component to analyse small samples. The result is accurate through the experimental test of micro alloy. Researchers also use other data to test the method. The relative error between the results and the real value is generally not more than 5%. Therefore, premise of meeting the two agreements in introduction, the method can be used in all of the quantitative relationship.

References

[1] Klaus Backhaus, Bender Ekrison, Wulf Plinke, Wang Xiyi and Julv Weber. Multivariate Statistical Analysis-Using the tool of SPSS. [M] . Shanghai, China, 2009.

[2] Jiang Qiyuan, Xie Jinxing, and Ye Jun. Mathematical Model (3rd ed.). [M]. China Higher Education Press, 2007.

[3] Chen Zhiqiang. Vanadium micro-alloying and the influence of the cold rolling process on 5182 strips [D]. Chongqing University, 2007.

[4] Liu Xinhua. Necessity and Software Operation of Positive Management in Factor Analysis. [J]. Journal of Chongqing Institute of Technology, 2009, 23(9).

[5] Ye Zongyu. The choice of index forwarding and dimensionless method in the multi-index comprehensive evaluation[J]. Zhejiang statistics, 2003, (4).

[6] HURVICH, CM (HURVICH, CM); TSAI, CL (TSAI, CL) REGRESSION AND TIME-SERIES MODEL SELECTION IN SMALL SAMPLES. JUN 1989

[7] Chen, LF (Chen, LF); Liao, HYM (Liao, HYM); Ko, MT (Ko, MT); Lin, JC (Lin, JC); Yu, GJ (Yu, GJ), A new LDA-based face recognition system which can solve the small sample size problem, PATTERN RECOGNITION OCT 2000

[8] Lu Jianfang, Xie Congcong, Lian Xiaopeng. Numerical calculation of foundation. Science Pr

[9] TAN, ZC (TAN, ZC); SUN, GY (SUN, GY); SUN, Y (SUN, Y); YIN, AX (YIN, AX); WANG, WB (WANG, WB); YE, JC (YE, JC); ZHOU, LX (ZHOU, LX), AN ADIABATIC LOW-TEMPERATURE CALORIMETER FOR HEAT-CAPACITY MEASUREMENT OF SMALL SAMPLES . JOURNAL OF THERMAL ANALYSIS ,JUL-AUG 1995

[10] Zhang hengxi, Guo jilian, Zhu jiayuan, Northwestern Polytechnical University press. 2002-9

[11] Brazzale, AR (Brazzale, AR); Davison, AC (Davison, AC); Reid, N (Reid, N). Applied Asymptotics: Case Studies in Small-Sample Statistics. APPLIED ASYMPTOTICS: CASE STUDIES IN SMALL-SAMPLE STATISTICS 2007

[12] Robinson, MD (Robinson, Mark D.); Smyth, GK (Smyth, Gordon K.) . Small-sample estimation of negative binomial dispersion, with applications to SAGE data. BIOSTATISTICSAPR 2008sis.\

Author

<Yin Boya >, <1993.11>,< Hangzhou, Zhejiang Province, P.R. China>

Current position, grades: Department of Mathematics college of science, Zhejiang University of Technology, China.
University studies: Ongoing undergraduate study and research.
Scientific interest: His research interest fields include the data analysis and complex network.
Publications: more than 3 papers published in various journals.
Experience: He has won a second prize in a national contest and some one prize at the provincial level

<Li Chenyi>, <1994.9>,< Shaoxing, Zhejiang Province, P.R. China>

Current position, grades: JianXing Honors College, Zhejiang University of Technology, China.
University studies: Ongoing undergraduate study and research.
Scientific interest: Her research interest fields are Financial data processing.
Publications: She published several articles in the regular newspapers.
Experience: She has won some one and second prize at the provincial level