Research on the raw data processing method of the hydropower construction project

Zhichao Tian
Power China Chengdu Engineering Corporation Limited, Chengdu 610072, Sichuan, China

Abstract. In this paper, based on the characteristics of the fixed data, this paper compares the various mathematical statistics analysis methods and chooses the improved Grabs criterion to analyze the data, and through the analysis of the data processing, the data processing method is not suitable. It is proved that this method can be applied to the processing of fixed raw data. This paper provides a reference for reasonably determining the effective quota analysis data.

1. Introduction
The so-called set, that is, the amount; the amount is the amount or limit. From the broad sense, the quota is the prescribed amount or limit. China's current construction management system, the "fixed" word has its special meaning, that is, "set" is the statutory set, "the amount" is the consumption of labor, materials, the amount of the amount of machinery, the specific can be understood as: In the normal construction conditions, in order to complete a certain unit of measurement in line with national technical standards, technical specifications and quality assessment standards of qualified products must be consumed in the manual, material, mechanical classes standard, reflecting a certain period of productivity and economic level The

Therefore, the quota must be determined from the actual situation, according to the production conditions, quality standards and the existing technical level of workers, such as through the calculation, statistics, analysis and development, and with the above conditions to supplement and revision to meet the needs of production development. Therefore, the quota level is the core of the entire quota work.

2. Basic characteristics of fixed raw data
2.1. The characteristics of the original data
2.1.1. Data Sources. According to the statistics, the data of the major enterprises based on the data and construction enterprises are mainly derived from the direct observation statistics or the experimental statistics. There are two main channels: the first channel is the quotient of the compilers, through the investigation, observation and field technology Determination of income, this data is more accurate. The technical level, manpower and material resources of the observer have a great influence on the fixed field measurement. The second channel is the internal cost of personnel, according to the characteristics of enterprises, collecting historical data of construction projects over the years, dig out from the statistical data. The second type of quota data will be constrained by the level of statistical analysis. By
analyzing the massive database data in the historical database, it can provide the data to support the actual situation of the enterprise.

2.1.2. Data characteristics. According to the characteristics of the construction itself, the original data of the statistical engineering construction has obvious characteristics of less (less) sample, diversification, data incompleteness, and data anomaly.

a). Less sample
Hydropower engineering construction products usually have a wide area, complex conditions, large size and other characteristics, and hydropower project tooling machine demand, a wide range of complex construction complex, changeable, hydropower construction project construction period and other characteristics of the current fixed technology are based on the use of methods in large sample conditions.

b). Massive and diversified
Hydropower construction projects are usually related to the complex environment, the project will be put into a lot of manpower and material resources before the geological survey, surveying and mapping, etc., the construction unit in the construction should also invest a lot of manpower, material development by the construction schedule Construction, machinery and equipment deployment, labor organization, personnel management.

c). Data instability
Hydropower construction process by the staff, the impact of the environment is relatively large. In the construction process has been in the process of dynamic changes, even in the same tenders, the middle and afternoon construction data is not the same, the statistics out of the quota consumption data is also unstable. With the construction of enterprises to complete the project more, the accumulation of historical data will be more mass, so the greater the difficulty of statistical analysis to determine and analyze the instability of the data on the level of quota preparation a great impact.

d). Data anomaly
Hydroelectric engineering construction enterprises in the statistical construction data, usually subject to objective factors and subjective factors in the dual impact, under the dual role of subjective and objective will lead to abnormal consumption of statistical data anomalies. Subjective factors are due to different measured people, the record will affect the accuracy of the data, different observers used in different observation methods, the choice of mathematical theoretical model is also different.

2.2. Quota the acquisition of raw data
The determination of quota consumption is inseparable from the acquisition of raw data, and the original data collection requires a lot of field research. In the data acquisition, it is very important to adopt a scientific and effective method of collecting and analyzing, such as field investigation method, time measurement method and interview questionnaire. The field survey method is a detailed on-site investigation, summarize and analyze the practical application of the construction technology and technical means; timing measurement method is the new technology, new materials, new machinery on-site measurement; interview questionnaire is through the construction of all parties To conduct interviews and surveys to improve the possible disqualification of the target, to make up for deficiencies in the field investigation. Data collection method according to the actual situation, can be a single method, it can be a variety of ways to cross.

3. Research on Quota Raw Data Processing Method

3.1. The Choice and Application of Suspicious Data
It is necessary to carry out a large number of observations and measurements, which are records of quantitative and qualitative determinations, which are the most important basis for the quantitative consumption of scientific research. Therefore, the quality and accuracy of these data are directly related to the final outcome of the quota Correctness and scientifcity. It is inevitable that some of the anomaly data appear in these observations, which may be due to the inherent characteristics of the data generation
mechanism, and may be data entry and transmission errors; observers are confused, false, and concealed Statistical data distortion, loss of data and other factors. The occurrence of these abnormal data is bound to lead to the statistical analysis of the data results of the mutation, and ultimately lead to the results of the quota and the reality of a large deviation, so how to identify the abnormal data, the measured data to fake true, so that the results closer to the actual situation A realistic question of the quota.

There are five main ways to remove the abnormal data: $3\sigma$ criterion, Grubbs criterion, Dixon criterion, Chauvenet criterion, Romanovsky criterion (t test).

The Rhine guidelines do not need to be look-up and easy to use, but when the number of observations $n <10$ or $n = 10$, the criterion is invalid, which is generally applied when the request is not high or the number of observations is large.

The Schweller criterion improves the Rhine criterion, but does not have a certain probability of meaning, early use is more, and when the observed data $n \rightarrow \infty$, the criterion fails. When there is only one outlier in the observed data, the Dixon criterion method can get a better effect, but often the observed data in the abnormal value may exist more than one, and sometimes also appear in the same side. At this time the effect of the method is poor, therefore, in the fixed observation of abnormal data removal is not commonly used.

The Romanovsky Basis method is computationally complex and not often used.

In the above five criteria, the Grabs criterion and the Dixon criterion give strict results, but there are also defects in the Dixon criterion, and the domestic experts replace the average by using the median of the observed data. Effectively eliminating the anomaly data in the same side of the problems, such improvements to make processing more robust, the international community often recommend the use of Grabs criterion. In this study, we mainly introduce the improved Grabus criterion method and how to simplify the calculation method by means of excel tool.

3.2. Improved Grabus Criterion to Determine Abnormal Values

The improved Grabus criterion distinguishes the basic principles of anomaly data, and here only describes its basic steps. as follows:

a). Calculate the quasi-simulated anomaly. Assuming that the repeated observation data are $(X_1, X_2, \ldots, X_n)$, the number of data is $n$, the maximum value is $X_{\text{max}}$, the minimum value is $X_{\text{min}}$, the median is $M_e$, the standard deviation is $s$, The maximum outliers $G_{\text{max}}$ and the quasi-minimum minimum anomalies $G_{\text{min}}$ are:

$$G_{\text{max}} = \frac{X_{\text{max}} - M_e}{s}; \quad G_{\text{min}} = \frac{M_e - X_{\text{min}}}{s}$$

where, $s = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n-1}}$

b). Find the Grabius threshold $G (\alpha, n)$. When the significance level is $\alpha$ (constant $\alpha$ is 0.05 or 0.01, equivalent to 95% or 99% confidence), the data is detected by the Grubbs threshold $G (\alpha, n)$ The Grabius critical value $G (\alpha, n)$ is $n$.

c). Abnormal data identification and elimination. Comparing $G_{\text{max}}$ and $G_{\text{min}}$ with $G (\alpha, n)$, if $G_{\text{max}} > G (\alpha, n)$ or $G_{\text{min}} > G (\alpha, n)$, the corresponding $X_{\text{max}}$ or $X_{\text{min}}$ is abnormal data and should be removed.

d). Duplicate recognition and elimination of exception data. Repeat the above steps for removing the data after the exception data until the maximum or minimum value of the remaining data is less than the Grabus threshold.

3.3. An example of using the improved glabs criterion to eliminate suspicious data with the excel tool.

Because the observation data of the maintenance budget quota are determined according to the different engineering breakdowns, most of the data are accurate to 2 to 3 digits after the decimal point. Therefore, when the calculation result is large, the number of digits after the decimal point is large there will be
errors, in order to eliminate such a human error, reduce the amount of man-made calculations, and you can use the computer excel tool to operate, the following example.

The fixed data of the fixed data of the concrete of 6m3 dump truck for a project are shown in the following table. To determine whether the group of data outliers and be removed.

| Number | Queuing time | Charge waiting time | Loading time | Charge the time of shipment | Discharge waiting time | Discharge time | Empty back time | Sum | Total 1 (min) |
|--------|--------------|---------------------|--------------|----------------------------|------------------------|---------------|----------------|-----|---------------|
| 1      | 2 : 45.50    | 0 : 52.00           | 0 : 27.90    | 0 : 39.92                  | 0 : 12.56              | 0 : 27.90     | 0 : 03.00      | 1   | 35.11         |
| 2      | 2 : 42.76    | 0 : 52.00           | 0 : 49.02    | 0 : 07.11                  | 0 : 00.00              | 0 : 33.22     | 0 : 03.00      | 1   | 18.52         |
| 3      | 5 : 00.00    | 0 : 43.00           | 0 : 32.90    | 0 : 05.00                  | 0 : 00.00              | 0 : 39.30     | 0 : 03.00      | 1   | 21.84         |
| 4      | 3 : 04.70    | 0 : 45.00           | 0 : 22.00    | 0 : 00.00                  | 0 : 00.00              | 0 : 41.70     | 0 : 03.00      | 1   | 24.85         |
| 5      | 1 : 31.26    | 0 : 49.00           | 0 : 18.42    | 0 : 06.98                  | 0 : 00.00              | 0 : 23.18     | 0 : 03.00      | 1   | 39.51         |
| 6      | 2 : 24.00    | 0 : 00.00           | 0 : 21.36    | 0 : 32.38                  | 0 : 00.00              | 0 : 33.09     | 0 : 03.00      | 1   | 59.18         |
| 7      | 3 : 37.04    | 0 : 69.00           | 0 : 10.78    | 0 : 41.00                  | 0 : 00.00              | 0 : 43.56     | 0 : 03.00      | 1   | 18.52         |
| 8      | 0 : 54.06    | 0 : 08.00           | 0 : 15.01    | 0 : 44.00                  | 0 : 00.00              | 0 : 53.82     | 0 : 03.00      | 1   | 19.82         |
| 9      | 0 : 16.95    | 0 : 44.00           | 0 : 14.86    | 0 : 24.00                  | 0 : 00.00              | 0 : 45.10     | 0 : 03.00      | 1   | 17.45         |
| 10     | 0 : 13.37    | 0 : 52.00           | 0 : 46.84    | 0 : 35.05                  | 0 : 00.00              | 0 : 35.42     | 0 : 03.00      | 1   | 58.71         |

Table 1. 6m³ dump truck transportation concrete observation table

EXCEL application steps are as follows:

a). Open the excel, according to the following table input data and the establishment of worksheets, of which the number of data n = 50.

b). Through the excel provided MAX function, MIN function, MEDIAN function, STDEV function, find the data of the maximum, minimum, median, standard deviation. The specific method is as follows:

1. Using the function MAX, the maximum value is Xmax of 25.92;
2. Using the function MIN, the minimum value is Xmin is 13.91
3. Using the function MEDIAN, the median M is 18.33;
4. Using the function STDEV, the standard deviation s is 2.09.

c). Find the approximate maximum outliers Gmax and the minimum outliers Gmin. The specific steps are as follows:

1. The maximum apparent value Gmax is 3.63;
2. And the minimum apparent value Gmin is 2.11.
d). Find the Grabius threshold $G(\alpha, n)$. (The greater the significance level), the more narrow the data recognition interval, the more stringent identification requirements), because the number of data $n$ is 50, then by looking up the table, the Grabus critical value $G(0.05, 50)$ was 2.956. The results are as follows:

- Max $X_{\text{max}} = 25.92$; minimum $X_{\text{min}} = 13.91$; median $M_{\text{e}} = 18.33$; standard deviation $s = 2.09$;
- The maximum apparent value $G_{\text{max}} = 3.63$; the quasi-minimum value $G_{\text{min}} = 2.11$;
- Grabus threshold $G(0.05, 49) = 2.956$.

e). Abnormal data recognition and elimination. Because $G_{\text{max}}$ is 3.63 greater than the critical value $G(0.05, 50) = 2.956$, and $G_{\text{min}}$ is 2.11 less than the critical value of 2.956, when the significant level of $\alpha$ is 0.05 (confidence 95%), $G_{\text{max}}$ corresponds to the maximum value of 25.92 data, From the statistical point of view should be regarded as abnormal data, should be removed.

f). Duplicate recognition and rejecting constant data. After removing 25.92, the remaining data is re-identified by the above steps. The results are:

- Max $X_{\text{max}} = 21.58$; minimum $X_{\text{min}} = 13.91$; median $M_{\text{e}} = 18.32$; standard deviation $s = 1.81$;
- The maximum apparent value $G_{\text{max}} = 1.79$; the quasi-minimum value $G_{\text{min}} = 2.44$;
- Grabus threshold $G(0.05, 49) = 2.948$.

Through the second calculation, the $G_{\text{max}}$ is 1.79; $G_{\text{min}}$ is 2.44, which is less than $G(0.05, 49)$ value 2.948. No abnormal data exists in the data, and the abnormal data is identified and eliminated.

4. Summary
In this paper, the characteristics of the original data are analyzed, the data processing methods are analyzed, the improved Grabus criterion is used to determine the outliers, and the data is processed. Through the actual data example, it has good effect and can be used as quota Data processing analysis reference use.

References
[1] QIAO Peng. Study on the Method of Quantitative Data Determination and Multivariate Data Difference in Highway Engineering.
[2] Wang Shouxu, QIAO Peng, ZHANG Zheng-zhen, et al. Adaptive selection of field construction quota method based on AHP method [J]. Journal of China & Foreign Highway, 2010, 4: 72-76.
[3] Wang Shouxu, Qiao Peng, Zhang Zhengqi, etc. Based on linear regression analysis of quantitative calculation of consumption calculation [J]. Journal of Transportation Science and Engineering, 2011, 4 (2): 36-41.
[4] WANG Jie-xian, QIAO Peng. Study on Establishment and Application of Raw Data Difference Processing Model for Engineering Quota [J]. Value Engineering, 2012, 31