Research on the Quality Evaluation Strategy of Multi-source Heterogeneous Aggregation Data

Gaofeng Zhan¹, Yong Jiang¹, Xiaofeng Wang¹, Yuyang Zhao² and Tong Zhao³,4*
¹ The first Research Institute of the Ministry of Public Security, Beijing, 100048, P.R. China
² Beijing Langrun Yihua Technology Co., Ltd, Beijing, 100429, P.R. China
³ University of Chinese Academy of Sciences, Beijing, 100049, P.R. China
⁴ Key Laboratory of big data mining and knowledge management, Chinese Academy of Sciences, Beijing, 100049, P.R. China
*Corresponding author’s e-mail: zhaotong@ucas.ac.cn

Abstract. In the era of big data, the application of multi-source heterogeneous aggregation data is more and more extensive. If the quality of aggregation data is uneven, it will bring a lot of troubles to the subsequent data mining, and then lead to inaccurate decision-making. A comprehensive quality evaluation method for aggregation data is proposed in this paper, based on factor analysis and multivariate variance analysis which is from the perspective of multivariate statistical inference. The case study shows that the method proposed in this paper is feasible and adaptive for the long-term evaluation of the quality of multi-source heterogeneous aggregation data.

1. Introduction
The convergence of information technology with the economic society has led to the rapid growth of data, which has become a strategic resource for countries and enterprises in the 21st century. No matter from the current e-commerce shopping, online government affairs or the positioning and tracking of the close contacts of the COVID-19, all of them are needed the close cooperation of many data suppliers (e-commerce sites, telecommunications networks, governments, Center for Disease Prevention and Control) which are connected with the produced data in various real-time applications such as identification, secure payment, mobile phone location tracking. The aggregation, integration and analysis of multi-source heterogeneous data play an important role.

Data quality is not only the basis to ensure correct and effective analysis of data in the database, but also the most important premise and guarantee for data mining, decision makings. With the development of information technology, data quality and its relevant researches have been widely concerned for a long time. Huang [1] believes that data quality is “fit for use”, which is widely recognized by the industry. Wand et al. [2] proposed five data quality dimensions based on the information system model: accuracy, integrity, consistency, timeliness and reliability. In literature [3], by analyzing many characteristics of data quality one by one, 15 commonly used quality evaluation indexes are defined. Redeman[4] concentrates data quality into three categories: concept view data, data value and data format.
Data quality evaluation is an important part of data processing. The ability of evaluation is directly related to the whole information system construction and the data sorting. Wang et al. [5] believe that data quality can be divided into objective quality indicators and subjective quality parameters, and users should evaluate data quality according to their own choices. Among them, the subjective quality can be spot checked; some objective and quantifiable quality problems can be defined according to the rules and spot checked every day, and the found quality problems should be recorded, traced and reprocessed. Such quality problems need to be corrected by re-define the rules in the Extraction-Transform-Load (ETL) process. The user extracts the required data from the data source and loads the processed data into the database after data conversion. The workload of ETL accounts for about 60% - 70% of the total.

For multi-source heterogeneous data, if the data source is inconsistent or inaccurate, it is likely to lead to unreliable data in the database, which will bring a lot of troubles to the data analysis and decision-making on the database level. According to Huang [1], the most common data quality problems in the field of data analysis are generally caused by the following five reasons: (1) data integrity. Data comes from different data sources, and each independent data source has different rules, quality and interpretation methods. In addition, it will be not the same in the storage format of each data source or the field is empty or not; (2) data consistency. Different data sources may have different definitions of fields with the same meaning; (3) uniqueness. Some data fields from different data sources may be duplicate; (4) data accuracy. Whether there are errors in the data, some errors may be caused by the measurement unit error; (5) data validity. Whether the data format, type and business logic are correct, etc.

The traditional data aggregation work is mainly completed by various ETL tools. The built-in data quality early warning functions do not meet the needs of large-scale data aggregation work. Therefore, this paper intends to propose a novel idea of quality evaluation for multi-source heterogeneous aggregation data from the perspective of multivariate statistical inference. The evaluation of data quality can be summarized from the following aspects: (1) The evaluation method of the quality of multi-source heterogeneous aggregation data; (2) How to find out that the data quality of data source has changed obviously with time and whether the evaluation method needs to be adjusted; (3) How to adjust the quality evaluation strategy adaptively in the long-term data aggregation.

### 2. Multiple Index Evaluation for the Aggregation Data’s Quality based on Factor Analysis

The quality evaluation of multi-source heterogeneous aggregation data can be regard as a comprehensive evaluation process with the above-mentioned five data elements. As these evaluation elements do not correspond to some specific statistical indexes one by one, several statistical indexes are summarized as follows, combined with our understanding, which may be applied to describe the multi-source data aggregation problem:

| Category   | Statistical Indexes |
|------------|---------------------|
| Integrity  | $A_i$: Number of records with null value in the field |
| Consistency| $A_s$: Number of records with inconsistent field meanings |
| Uniqueness | $A_u$: Number of records with different data in the same field |
| Accuracy   | $A_f$: Number of records with different common sense in field content |
| Validity   | $A_v$: Number of records with illegal fields |

The quality evaluation of multi-source heterogeneous aggregation data can be seen mathematically as the mapping of high-dimensional space to one-dimensional coordinate system and the following questions should be considered carefully: (1) Whether these statistical indexes are representative for
the quality evaluation; (2) Whether the five indexes are completely independent or there are some correlations between the indicators, which may be affected by certain common factors. (3) How to integrate these indicators for evaluation, considering that the above multi-dimensional evaluation criteria are needed to be nondimensionalized meanwhile the weight of each dimension is not clear; (3) How to obtain a general method for mapping high-dimensional indicators into one-dimensional index.

In multivariate statistical methods, factor analysis is a kind of statistical analysis technology to extract common factors from many variables. It was first proposed by Charles Spearman, a British psychologist. After studying the scores of 33 students in classical, French and English, he believes that there is a certain correlation between these subjects. Students with good scores in one subject often have good scores in other subjects, so he speculates whether there are some potential common factors that affect students' learning results. Factor analysis technology can find out the hidden representative factors in many variables, and put the same essential variables into one factor. In the problems studied in this paper, factor analysis technology can find the potential internal structure among the indicators, decompose the external observable indicators from the table and the internal and comb out different types of structures, each structure represents a common factor.

Although the common factors in factor analysis exist objectively, they are usually unobservable and are generally called latent variables. We may set the expected value \( \mu \) for every vectors \( \mathbf{y} = (Y_1, \cdots, Y_p)' \), which is from a large number of samples with the statistical indicators collected in the processing of data aggregation. Where \( p \) is the dimension of the multivariate variables, which is corresponding to \( p \) calculation indexes to be investigated in the quality evaluation, such as \( A_1 \) to \( A_5 \). The orthogonal factor model assumes that \( \mathbf{y} \) is linearly dependent on \( m \) common factor vectors \( \mathbf{f} = (F_1, \cdots, F_m)' \) with \( p \) specific factors \( \mathbf{\epsilon} = (\epsilon_1, \cdots, \epsilon_p)' \) which are not observable, and usually \( m < p \).

\[
\begin{align*}
Y_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \epsilon_1 \\
Y_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \epsilon_2 \\
&\quad \vdots \\
Y_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \epsilon_p
\end{align*}
\] (1)

Among them, the coefficient \( l_{jk} \) is called the loading of the \( j \) variable on the \( k \) factor, which reflects the explanatory power of the common factor to this variable. The expression can be simplified as: \( \mathbf{y}_{\text{vec}} - \mu_{\text{vec}} = \mathbf{L}_{\text{pm}}\mathbf{f}_{\text{mvec}} + \mathbf{\epsilon}_{\text{vec}} \), where \( E(\mathbf{f}) = \mathbf{0}_{\text{mvec}}, \text{COV}(\mathbf{f}) = I_{\text{mvec}} \) meets at the same time: \( E(\mathbf{\epsilon}) = \mathbf{0}_{\text{pvec}}, \text{COV}(\mathbf{\epsilon}, \mathbf{f}) = \mathbf{0}_{\text{pmvec}} \).

For the above factor analysis, the most critical is to obtain the load matrix. The coefficients in the matrix represent the linear transformation relation between the current index system and the potential internal structure of the indicators. The matrix \( \mathbf{L}_{\text{pmvec}} \) can be obtained by handling the following steps one by one, and then the contribution of each index to the data quality can be achieved and which is the most important parameters to count out the comprehensive evaluation score:

**Step 1.** Make clear the indexes for statistical analysis. In the process of multi-source data aggregation, it is necessary to make clear which representative calculation indicators are needed to be counted and recorded. For example, the common five indicators of \( A_1 \) to \( A_5 \).

**Step 2.** Find the correlation matrix \( \mathbf{R} \). Find out the degree of correlation between each variable.

**Step 3.** Obtain factor load matrix \( \mathbf{L}_{\text{pmvec}} \) by principal component analysis, and extract common factor and contribution rate \( T_i \) of common factor.
Step 4. Factor rotation. Make clear of the factor’s meaning based on the explanatory power of the extracted common factors to each variable.

Step 5. Count out evaluation score according to the result of factor scores. After obtaining the load matrix $L_{pm}$, we can calculate the score $Q_i$ of each factor $F_i$, and make $W_i = \frac{T_i}{\sum_{i=1}^{m} T_i}$ the score weight of each common factor. Considering that in the index $A_i$ to $A_s$, the logarithm function is used to adjust the final evaluation score: $S = \ln \left( 1 + \frac{1}{\sum_{i=1}^{m} Q_i \cdot W_i} \right)$.

3. Long-term Variability Assessment of Data’s Quality based on Multivariate Analysis of Variance

The quality of data source may be deeply affected due to many factors such as system upgrade, adjustment of structure or secondary development. Such influences may inevitably affect the quality evaluation in data aggregation processing. How to perceive such changes accurately and adjust the data quality evaluation method in time are crucial to the multi-source heterogeneous aggregation data quality evaluations.

Generally speaking, there are some random fluctuations may be observed in the values of data quality evaluation indexes with time dimension. The fluctuations of these indexes are naturally caused by some random factors, such as the random input error of the operator, the occasional failure of the system software or hardware system, which can also be considered as random errors. For a certain evaluation index, when the volatility is large, it may be regarded as the systematic error, which can be confirmed by the idea of hypothesis test: whether there is a systematic change in the time period before and after the index, with significant differences. Considering that the multi-dimensional indexes are not completely independent, therefore it seems that the possibility of making the first kind of statistical errors will significantly increase by using the univariate t-test method for each evaluation index.

The method of multivariate analysis of variance (MANOVA) is introduced in this paper, and the multivariate analysis mainly investigates whether there are significant differences among multiple populations. Its statistical idea is derived from one-way ANOVA, and it is the extension of one-way ANOVA. Different from univariate analysis of variance, multivariate analysis of variance needs to generalize the sum of the squares of the differences between groups and the sum of the squares of the differences within groups into the matrix forms. Then, the hypothesis test is constructed based on the ratio between the inter group dispersion matrix and the intra group dispersion matrix. Wilks test or Hotelling test are applied in MANOVA to test whether there is significant difference in variance.

Focus on the basic idea of multivariate analysis of variance, there are $K$ consecutive time periods are set in the problem. meanwhile, each collected data of quality assessment have their own independent population, which are presented as: $Y^1, Y^2, \cdots, Y^K$. The samples in such populations satisfies the following normal distribution:

$$
\begin{align*}
&y^1, \cdots, y^1_{n_1} \sim N_p(\mu_1, \Sigma) \\
&\quad \quad \quad \cdots \quad \quad \quad \quad \cdots \\
&y^K, \cdots, y^K_{n_K} \sim N_p(\mu_K, \Sigma)
\end{align*}
$$

(2)

Subsequently, they are independent of each other in $\{y^i, j = 1, \cdots, K; i = 1, \cdots, n_K\}$ and are suitable for being carry out statistical test with type (2) to find out whether the numerical difference between the quality evaluation index data collected in each time period has changed significantly.
Within the statistical test model, the original hypothesis $H_0$ can be set: at least one set of $(i \neq j)$ multivariate data quality indicators satisfies that their expected vector is $\mu_i \neq \mu_j$. That is to say, it is found that the comprehensive expectations of data quality evaluation indicators are different in at least two time periods. In the same time, the alternative hypothesis $H_1$ satisfies: $\mu_1 = \mu_2 = \cdots = \mu_K$, which means that the comprehensive expectations of data quality evaluation indexes have not changed significantly in the above $K$ timeslots.

Through the above statistical test, we can quickly figure out whether the data quality of the data source has significant changes in several consecutive time periods, and then a decision can be made whether the next stage should continue to use the existing statistical indicators and weights to calculate the data quality according to the results of factor analysis, or ready to complete the quality evaluation analysis processing again.

4. Quality Evaluation Process of Multi-source Heterogeneous Aggregation Data
Based on the above two sections of statistical inference method, the quality evaluation strategy can be divided into two parts: (1) factor analysis which is to solve the problem of the optimal quality evaluation of the aggregated data in single time segment; (2) the method of variance analysis which is focus on dynamic adjustment in a long time period. The adaptive analysis process of the whole data quality evaluation is shown in Figure 1:

![Figure 1. Adaptive Evaluation Process of Multi-source Heterogeneous Aggregation Data Quality](image)

5. Case Analysis
The data resources related to public security are more and more abundant, with the progresses in Digital City Project. The prominent problem in these information systems is the lack of the ability of resource integration, analysis and decision-making and infeasible to form an effective mechanism to integrate all kinds of data resources and comprehensive analysis. In this paper, aims to be embedded in the ETL tools, a set of data quality aggregation and long-term adaptive adjustment process are designed in the process of data cleaning and conversion for multiple data sources according to different data types, combined with the basic requirements of public security business for data quality.

| Table 2. KMO and Bartlett’s Test. |
|-----------------------------------|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy | 0.523 |
| Bartlett's Test of Approx. Chi-Square | 113.501 |
| df | 10 |
In the case study, it is found that the information in the five indexes $A_1$ to $A_5$ are highly extracted by the factor analysis based on the principal component (Table 3), with the interval from 71.3% to 95.8%. This means that the internal information of the current indexes can be well explained by the factor analysis method. Statistical calculation shows that there are two indexes with eigenvalues greater than 1 and cumulative variance contribution rate of 82.89% (Table 4). The information contented of these two main components can reflect 82.89% of the information of the index and can basically explain most of the information of the original data. Therefore, these two common factors are selected as indicators of the indicator system.

### Table 3. Communalities

| Initial Extraction | 
|--------------------|------|
| A1                 | 1.0  | .953 |
| A2                 | 1.0  | .813 |
| A3                 | 1.0  | .898 |
| A4                 | 1.0  | .768 |
| A5                 | 1.0  | .713 |

### Table 4. Total Variance Explained (Interception part) a.

| Component | Initial Eigenvalues | Rotation Sums of Squared Loadings |
|-----------|---------------------|----------------------------------|
|           | Total               | % of Variance                    | Total | % of Variance |
| 1         | 2.637               | 52.737                           | 2.296 | 45.916        |
| 2         | 1.508               | 30.156                           | 1.849 | 36.977        |
| 3         | .508                | 10.164                           | .626  | 6.524         |
| 4         | .326                | 6.524                            | .326  | 6.524         |
| 5         | .021                | .419                             | .021  | .419          |

*a Extraction Method: Principal Component Analysis.*

A component score table (Table 5) may be achieved with regression analysis method and there are two main components scored with $Q_1$ and $Q_2$ in this table respectively. The corresponding calculation formula is obtained from Table 5, and for the convenience of calculation, the five indicators had been standardized from $A_1 \cdots A_5$ to $X_1 \cdots X_5$.

\[
\begin{align*}
Q_1 &= 0.405X_1 - 0.089X_2 + 0.333X_3 + 0.423X_4 + 0.062X_5 \\
Q_2 &= 0.025X_1 + 0.512X_2 + 0.144X_3 - 0.270X_4 - 0.471X_5
\end{align*}
\]  \(3\)

### Table 5. Component Score Coefficient Matrix

|        | 1       | 2       |
|--------|---------|---------|
| A1     | .405    | .025    |
| A2     | -.089   | .512    |
| A3     | .333    | .144    |
| A4     | .423    | -.270   |
Therefore, the weights of the final evaluation of multi-source heterogeneous aggregation data is $W_1 = \frac{45.916}{82.893}$ and $W_2 = \frac{36.977}{82.893}$. To sum up, we can get $S = 0.554Q_1 + 0.446Q_2$.

After four time periods, it is checkpoint to judge whether the quality evaluation criteria need to be adjusted.

In multivariate analysis of variance, the results are partly shown in Table 6, from which we can finally confirm whether the current quality evaluation strategy needs to be adjusted.

Table 6. Tests of Between-Subjects Effects (Interception part).

| Source | Variable | Sum of Square | df | Mean Square | F    | Sig. |
|--------|----------|---------------|----|-------------|------|------|
| Group  | A1       | 6233418.390   | 3  | 2077806.130 | .951 | .418 |
|        | A2       | 5148550.332   | 3  | 1716183.444 | 2.302| .081 |
|        | A3       | 441911.269    | 3  | 147303.756  | .904 | .442 |
|        | A4       | 749436.078    | 3  | 249812.026  | 1.176| .322 |
|        | A5       | 22300.783     | 3  | 7433.594    | 2054.907 | .000 |

6. Conclusion
In this paper, multivariate statistical analysis is imported to handle the problem of quality evaluation in multi-source heterogeneous data. Factor analysis and multivariate analysis of variance are used specifically. Through a data aggregation case in the field of public security, it was introduced in this paper on how to comprehensively evaluate the quality of aggregated data, and how to detect the change of data quality in four time periods and when to make evaluation adjustment. In the follow-up work, we will further focus on the selection of evaluation indicators and the minimum occupancy of computing resources.

Acknowledgments
Authors wishing to acknowledge the assistance from all the colleagues.

References
[1] Huang K. T., Lee Y. W., Wang R. Y. (1998) Quality Information and Knowledge. Prentice Hall, Upper Saddle River, NJ.
[2] Wand Y., Wang R. Y. (1996) Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11):86-95.
[3] Wang R. Y., Strong D. M. (1996) Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4):5-33.
[4] Redeman T. C. (1997) Data Quality for the Information Age. Artech House, London.
[5] Wang R. Y., Henry B. K., Stuart E. M. (1993) Data quality requirements analysis and modelling. In: International Conference on Data Engineering. pp. 670-677.