Quality Control of Marine Big Data for Marine Physical Property Monitoring System

Chunli Chen1,*, Jie Zhou1, Xiong Luo1, Peie Zhao1, Qi Fan1, Xiaofeng Li1,2, Xinfang Li3, and Yong Chen1

1Southwest Institute of Technical Physics, 610041 Chengdu Sichuan, China
2Beijing Institute of Technology, School of Optics and Photonics, 100081 Beijing, China
3First Institute of Oceanography, Ministry of Natural Resources, 266061 Qingdao Shangdong, China

Abstract. To make good use of the ocean, we must know enough about it. Supported by the Ministry of science and technology of China, the marine physical parameters monitor is developed to obtain real-time marine physical parameters such as temperature, humidity, pressure, rainfall, wind speed, wind direction, solar radiation, water depth, salinity, conductivity and other data. The integrated intelligent service software platform for marine physical property monitoring data is used for real-time collection, storage, historical data analysis, integrated display and other work of these data. As a part of the integrated intelligent service software platform, an improved big data quality control method which comprehensively uses the meteorological industry standards test, least square polynomial fitting, multiple regression and other technologies is proposed in this paper to improve the correct and efficient use of these real-time or historical data in terms of data inspection, correction, reconstruction and quality evaluation, etc. The existing historical data prove the effectiveness of this method.

1 Introduction

There is a vast territory with both land and sea in China. In addition to the land area of 9.6 million square kilometers, according to the UN Convention on the sea, the sea area under China's jurisdiction is about 3 million square kilometers, with a coastline of 18000 kilometers. With the rapid development of society, the consumption of natural resources is increasing rapidly. People begin to focus on the ocean. The utilization of marine resources, such as tidal energy, wind energy, salinity difference energy, etc., is attracting more and more attention. At the same time, the impact of human activities on the ocean is also growing. If we want to make good use of the ocean, we must know it well. In 2018, the Ministry of science and technology of China approved the project "marine physical parameters monitor" for real-time, multiple dimensional and intensive monitoring of various meteorological and hydrological data such as temperature, humidity, pressure, rainfall, wind speed, wind direction, solar radiation, water depth, salinity, conductivity, etc. In order to meet the real-time collection, storage, historical data analysis and comprehensive display of these massive data, the integrated intelligent service software
platform for marine physical property monitoring data has been developed. The first problem the software platform faces is the quality control of these data.

National and industrial standards such as GB/T 12763.1-2007 Specification for marine survey Part 1: General, QX/T 66.22-2007 Specification for ground meteorological observation Part 22: Quality control of data provides general methods for quality control of meteorological and hydrological data, but these methods are usually only applicable to manual data processing. Moreover, Xu, Yang, etc. summarized a variety of data quality control methods [1-3], but these methods are relatively isolated, fail to give full play to the advantages of multiple data fusion applications, and are not conducive to further correction of singular data and supplement of missing data. The improved quality control method of big data proposed in this paper comprehensively uses the meteorological industry standard test, least square polynomial fitting, multiple regression and other technologies, gives the effective degree of data, improves the accuracy of data, and provides conditions for real-time processing of marine physical data and efficient application of a large number of historical data.

2 Method

2.1 Polynomial fitting

In daily statistics of meteorological and hydrological data [4], it can be found that the same set of data often has certain continuity in time or space (or other possible data of a certain type). That is, the element combination \((x_i, y_i)\) has certain correlation, when \(i = (1, 2, \ldots, m)\). We can find an approximate expression \(y = P(x)\) instead of \(y = f(x)\), where \(y\) is a kind of meteorological or hydrological data, \(y\) is time or space, or other data.

When this group of data satisfies the characteristics of polynomial curve, polynomial fitting method is used to represent this group of data:

\[
P(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n \quad (n < m)
\]

The polynomial coefficient \(a_i \quad (i = 1, 2, \cdots, n)\) solved by the least square method can be used to fit the curve.

In this paper, least square polynomial fitting can be used for the numerical analysis and quality control of time related data (such as temperature, humidity) and height related data (such as wind speed, air pressure).

2.2 Multiple regression

In the meteorological and hydrological data, there are also some special data, which have internal correlation characteristics, such as the data set of atmospheric temperature, humidity, atmospheric pressure, and the data of sea water temperature, sea water depth, salinity, conductivity. For these data, one kind of data \(y\) can be approximately represented by other kinds of data [5]:

\[
y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_p x_p + \epsilon
\]

where \(\epsilon\) is the error value. \(\epsilon\) is a random variable with a mean value of 0, which satisfies the normal distribution. The multiple regression analysis of the data set can be realized by obtaining the optimal coefficient \(\beta_i\) through mathematical solution.
In this paper, multiple regression is used for the numerical analysis and quality control of each element in the data set with internal correlation. When there are only some missing data in the data set, the effect is often due to the least square polynomial fitting method.

2.3 Quality evaluation

In order to accurately evaluate the quality of each data element, we establish a data quality measurement model for data. Each data has multiple identification bits, such as format identification bit \((b_f)\), value range identification bit \((b_d)\), and 3 association characteristic identification bit \((b_{c1} - b_{c3})\). Each identification bit corresponds to different quality evaluation scores \((s_i)\). After comprehensive quality control, each data element input by the marine physical parameter monitor will get a comprehensive quality evaluation value \((Q_c)\) with range \([0,100]\). The larger the value is, the better the quality of the input data element will be. When the comprehensive quality evaluation value is below 80, a revised value after quality control will also be given to users for reference.

The calculation method of \(Q_c\) is as follows:

\[
Q_c = 100 - b_f \times s_f - b_d \times s_d - b_{c1} \times s_{c1} - b_{c2} \times s_{c2} - b_{c3} \times s_{c3}
\]  

(3)

The quality evaluation scores are different for different kinds of data, the scores of some kinds of data as shown in the following table:

<table>
<thead>
<tr>
<th>Data</th>
<th>wind speed</th>
<th>salinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>format identification scores ((s_f))</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>value domain scores ((s_d))</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>association characteristic scores((s_{c1}))</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>association characteristic scores((s_{c2}))</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>association characteristic scores((s_{c3}))</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>

2.4 Comprehensive quality control

The realization of comprehensive quality control requires the collaborative work of data format proofreading, initial checking, polynomial curve fitting, multiple regression analysis, quality evaluation and other modules.

The specific working process of the comprehensive quality control is as follows:

**Table 1.** Scores of some data.
1. Calibrate the data format of real-time data or historical data \( D_{in} \) input by marine physical parameters monitor according to the national standards and industry standards mentioned above. Set the format identification bit \( b_f \) to 1 when the data is incorrect and correct it;

2. Formulate the valid range of input data according to standard document and historical data (the value range refer to historical data needs to be set manually in advance, and you can also only formulate the value range according to the provisions of standard document when the historical data is insufficient.). When the input data \( D_{in} \) exceeds the valid range, set the value range identification bit \( b_r \) to 1;

3. Fit each kind of data in the time dimension by least squares polynomial. When the difference between the input data \( D_{in} \) at a certain time and the fitting result data \( D_{c1} \) at the same time is significant, set the identification bit \( b_{c1} \) to 1, and temporarily save the fitting result \( D_{c1} \);

4. When the input data \( D_{in} \) conforms to the characteristic condition of polynomial curve in spatial dimension, least squares polynomial fitting is performed between the data and other data of different spatial domains at the same time. When the difference between \( D_{in} \) and the result data \( D_{c2} \) is significant, set the identification bit \( b_{c2} \) to 1, and temporarily save the fitting result \( D_{c2} \);

5. When the input data \( D_{in} \) meet the requirements of multiple regression analysis, the data and other data of the multiple regression data set are analyzed by multiple regression. When the difference between \( D_{in} \) and the analysis data \( D_{c3} \) is significant, the identification bit \( b_{c3} \) is set to 1, and the result data \( D_{c3} \) is temporarily stored;

6. Calculate the comprehensive quality evaluation value \( Q_c \) according to Formula 3. When \( Q_c \) is below 80, revise the output data \( D_{out} \) with the weighted average of valid \( D_{c1} \sim D_{c3} \) based on characteristic scores \( S_{c1} \sim S_{c3} \) (if \( S_{cx} > 0 \)), otherwise take the input data \( D_{in} \) as the data \( D_{out} \);

7. Output the comprehensive quality evaluation result of the data, which includes data \( D_{out} \), comprehensive quality evaluation value \( Q_c \), and the identification bits \( b_f, b_r, and b_{c1} \sim b_{c3} \).

### 3 Tests and results

Historical data including wind speed data measured by lidar in the East China Sea, water depth, temperature, salinity and conductivity data measured by XCTD in the East China Sea were used to verify the comprehensive quality control work in the software platform. Firstly, we selected a set of 200 wind speed data (polynomial relationship is no longer applicable when the time series is too long) which is continuous, complete and effective in a certain height from the lidar data. In this set of data, 20 elements are randomly selected, and the real value was replaced by the artificial error value (more than 50% deviation from the real value). Then, the data and other height data corresponding to the time were imported into the software platform for comprehensive quality control. In the process of integrated quality control, the wind speed data were checked by range, fitted by time-domain polynomial and fitted by space height polynomial. The results of integrated quality control were obtained. Reselected the data and carried out the test according to the above steps for 5 times in total. According to the statistics, 83 data with \( Q_c \) below 80 were obtained, among which 81 manmade data were correctly detected.

In the same way, a group of 200 salinity data from XCTD data was selected. After value set, the data and other XCTD data were imported into the software platform for comprehensive quality control. In the process of comprehensive quality control, salinity data are checked in range, fitted by time domain polynomials and multiple regression. The
similar tests repeated 5 times in total according to the above steps. According to statistics, 90 times of data with $Q_c$ below 80 were obtained, among which 87 manmade data were correctly detected.

4 Conclusions

The quality control method proposed in this paper comprehensively utilizes standard correction and mathematical model validation, and carries out comprehensive quality evaluation in the quality control process, effectively simplifies the human engineering, improves the accuracy of the data after control, and completely shows the quality of the input data to users. The validity of this method is verified by trial application of historical data on software platform, which has great promotion value and application prospect.

This research was supported by National Key Research and Development Program Key Special Projects for Development of Major Scientific Instruments and Equipment, grant number 2018YFF01014100.

References