RESEARCH ON VISUALIZED DATA QUALITY CONTROL METHODS OF GROUND OBJECT SPECTRUM IN YANZHOU MINING AREA

Jun-fu Fan [a, *], Min Ji [a], Ting Li [a], Zhuo Li [a]

[a] Geomatics College of Shandong University of Science and Technology, 579 Qianwangang Road Economic & Technical Development Zone, Qingdao, China, 266510

KEY WORDS: Ground Object Spectrum, Data Quality Control, Cluster Analysis, Box-and-Whisker Plots, Gross Error Detection

ABSTRACT:

Errors or outliers are prone to be made on account of various accidental factors or system errors in the observation process of ground object spectrums. It is necessary to carry on some rigorous gross error detection and quality control measures on field spectroscopy data before which is conducted to further spectral analysis. To this end, in this paper, in accordance with measured data of several typical crops in Yanzhou mining area, a theory of cluster analysis for field spectroscopy data quality controlling was proposed and 4 different cluster methods included Statistical distance, Aitchison distance, Pearson's correlation coefficient and Multidimensional Vector Cosine were used in the gross error visualized detection. For the common characteristic bands of different spectrum data, the goal of visualized detection and identification of outliers was achieved by means of the statistical method of box-and-whisker plots. Outliers which were identified can be getting rid of in the use of several self-developed graphic interactive controls based on GDI+ technology. The theory proposed in this paper provided effective quality assurance for in-depth spectroscopy analysis.

1. INTRODUCTION

Study on spectral characteristics of ground objects is an important part of modern remote sensing technology. It is not only the accordance of sensor design and band selection, but also the basis for interpretation of remote sensing data analysis (Edward J. Milton, et al., 2009). The accuracy of field spectral measure results is affected by many factors, such as measure time, instrument FOV, observation geometry, solar azimuth and altitude angle, atmospheric environmental factor, etc (He Ting, et al., 2003). Therefore, the raw data of field spectral observations need to go through rigorous data quality control process to identify and get rid of records that contain errors or outliers before the whole dataset are used for in-depth spectral analysis. This paper on 2 groups of spectral data got from Yanzhou mining area as examples, Self-consistent accuracy (SCA) calculation method is used to evaluate the stability of the spectrum instrument, a theory that cluster analysis for field spectroscopy data quality controlling is proposed and 4 different cluster methods include Statistical distance, Aitchison distance, Pearson's correlation coefficient and Multidimensional Vector Cosine are used for gross error visualized detection in a same batch of spectral dataset and the pros and cons of them are discussed too. For some characteristic bands of spectral data we implemented the detection and identification of errors and outliers in a visual way by using the box-and-whisker plots. The GDI+ technology is used to draw plots automatically based on the results of 4 different cluster analysis methods and box-and-whisker plots models. The goal of visualized gross error detection and outlier identification on field spectroscopy data was achieved and thus provided effective quality assurance for in-depth spectroscopy analysis.

2. SPECTRAL CHARACTERISTICS AND QUALITY CONTROL METHODS

The spectral data of ground objects got by same spectral instruments in the same conditions ought to have same or similar characteristics, such as spectral resolution, band width and curve shape features, which can be used as the basis for the classification and anomaly detection. Only when the instruments are in a relatively stable state, do the measured data have the availability. The goal of evaluating the stability of spectral instruments can be reached by calculating self-consistent accuracy (SCA) of the data.

Clustering usually refers to grouping data or objects into a number of classes or clusters. Data records or objects in the same cluster have high similarity and different data records or objects in different clusters low (Kaufman, L., et al., 1990). The goal of cluster analysis is collecting data on the basis of similarity to classification, and can identify the one contains large differences in a group of similar dataset. Therefore, for data records which might contain gross errors and outliers in a batch of spectral datasets got in a same period of time can be found by using cluster analysis methods. The clustering results do not contain any detail information about data errors and outliers and some other measures are needed for further inspection and viewing. As a statistical method which have the characteristic of robustness of the median and quartile, box-and-whisker plots can provide detail information on changes in data range and extreme values (Wang Jian, et al., 2002). Whether data with full or specify bands scope can use box-and-whisker plots to find outliers with detail information and view the whole comparison of data records.

We evaluated the stability of spectral instruments by calculating self-consistent accuracy (SCA). Cluster analysis methods were used to detect gross errors and outliers. Box-and-whisker plots were used as further measure to compare the data records to troubleshoot out with errors and extreme values. The results of

[*] Corresponding author: Jun-fu Fan; E-mail address: yeahgis@yeah.net.

cluster analysis and box-and-whisker plots were visualized by self developed controls based on GDI+ technology.

3. ALGORITHM PRINCIPLE AND APPLICABILITY ANALYSIS

3.1 Methods on Spectral Instruments Stability Testing

To check the stability of spectral instruments, a certain number of repeated observations in the same conditions should be taken. The repeat measured data records can be tested and checked by calculating the mean square error of them to assess the accuracy and stability of the spectral instruments.

$$\varepsilon_j = \pm\sqrt{\frac{\sum_{i=1}^{n}\delta_{ij}^2}{n}}, (j = 1,2,\cdots,m) \tag{1}$$

$$\delta_{ij} = x_{ij} - x_i, (i = 1,2,\cdots,n; j = 1,2,\cdots,m) \tag{2}$$

$$x_i = \frac{\sum_{j=1}^{m} x_{ij}}{m}, (i = 1,2,\cdots,n) \tag{3}$$

$$\varepsilon = \pm\sqrt{\frac{\sum_{j=1}^{m}(\sum_{i=1}^{n}\delta_{ij}^2)}{m\times n}} \tag{4}$$

In Eq. (1-4), *i* is band number, *j* is data record (curve) number, *m* is the count of data records (curves), *n* is the count of bands. ε_j is the mean square error of curve *j*, x_{ij} is the reflectance value of curve *j* on band *i*, x_i is the average reflectance value of all curves on band *i*, δ_{ij} is the difference of x_{ij} and $x_{j,}$ and ε is the total self-consistent accuracy (T-SCA) of all curves. The results of Eq. (3) and Eq. (4) got from raw spectral data can be used as indicators of the stability of instruments. Similar curves have similar values and the smaller the better.

3.2 Cluster Analysis Methods on Gross Error Detection

There are 2 similarity measurements between the observational data records which are processed by number normalization, the distance and similarity coefficient (Yu Xiu-lin, et al., 2002). Supposed that, within the selected feature bands range there are *p* sampling points of a reflectivity curve. According to distance-based methods, curves can be regarded as points in a *p*-dimensional space, and the distance is defined in the space, points with short distance in a certain range fall into same classes and the ones with long distances fall into different classes. The methods based on similarity coefficient get the categories by calculating the similarity coefficient between data records (curves). The closer the absolute value of a similarity coefficient to 1 between curves, the more similar of them. Similarity coefficient must close to 0 if they are different with each other.

3.2.1 Classification Based on Distance Measurement:

For a given spectral instrument, if the selected characteristic bands scope contains *p* sampling points, reflectance curves can be seen as a series of points in *p*-dimensional space. This space is a simplex space of non-Euclidean space, but also can be approximated as Euclidean space. Then the degree of similarity between two observational data records (curves) can be measured by the distance between the two points in the *p*-dimensional space. Two methods as below are used to calculate the distance in this paper.

a) Statistical Distance

The Euclidean distance equation as below:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{5}$$

where *p* = sampling points count
i, j = Number of 2 data records (curves)
x = Coordinates in *p*-dimensional space
d_{ij} = Euclidean distance

Euclidean distance is a commonly used method in cluster analysis, but there are some shortcomings of its own. One is that Euclidean distance is related to the dimensions count of the statistic index, but there is no such problem for spectral reflectance curves. The other one is that the Euclidean distance does not take into account of the correlation between the various indicators. Some effective approaches must be taken to revise this problem (Yu Xiu-lin, et al. 2002). We introduced the method of weighted index variance to achieve the purpose. Eq. (6) is the improvement of Eq. (5).

$$d^s{}_{ij} = \sqrt{\sum_{k=1}^{p}\frac{(x_{ik} - x_{jk})^2}{s_{kk}}}, (k = 1,2,\cdots,p)$$
$$s_{kk} = \frac{1}{n}\left[(x_{1k} - \bar{x}_k)^2 + \cdots + (x_{nk} - \bar{x}_k)^2\right] \tag{6}$$

where S_{kk} = Variance of the No. *k* index
n = Spectral reflectance curves count
x = reflectance values
$d^s{}_{ij}$ = The statistical distance between curve *i* and *j* in *p*-dimensional space

The $d^s{}_{ij}$ equals to Euclidean distance when S_{11}=S_{22}=⋯=S_{kk}=⋯=S_{pp}. Euclidean distance can only be used in the situation of the indicators have the same deviation and equal contribution to the distance, variance weighted statistical distance is not subjected to this restriction and the results of practical application are better.

b) Aitchison Distance

Aitchison Distance is a calculation method which is used to measure the distance between objects defined in simplex space. A very good natural example is the distribution of probabilities

$P_1+\cdots+P_d=1$ for an event with d possible outcomes (Vêncio RZ, et al, 2005). The vector cluster, having p continuous sample points formed by n spectral reflectance curves which are observed in the same situations, can be seen as n points in non-normalized p-dimensional simplex space. To measure physical distance between objects in astronomical scales one should not use the regular Euclidean distance as shown in Eq. (5) but rather use proper relativistic distance measurements. This complication aroused because our world is not a Euclidean world. It is meaningless to calculate distance between 2 objects using the Euclidean distance model in a simplex and non-Euclidean space. John Aitchison proposed a meaningful calculation model to measure the distance between objects, such as u and U, in the simplex space (Aitchison, J., 1986).

$$d_{uU}=\sqrt{\sum_{j=2}^{p}\sum_{i=1}^{j-1}\left[\ln(\frac{u_i}{u_j})-\ln(\frac{U_i}{U_j})\right]^2} \tag{7}$$

where d_{uU} = The Aitchison distance between u and U
i, j = Number of sample points (bands' No.)
p = Count of sampling points
$u_{i/j}$, $U_{i/j}$= The reflectance value of curve u & U

Aitchison distance can be converted to the equivalent form of Euclidean distance by the transformation from simplex space to Euclidean space, the transformation methods may include stretching, expanding, or inflation. Aitchison distance is a complexity theory dealing with distance problems in simplex spaces. It provides a method for distance measuring and reasonable classification basis derived from series of mixed data in cluster analysis applications (Aitchison, J., 2001). The advantage of Aitchison distance are higher reliability than statistical distance because the former can show the topological space relationships between objects based on the method of algebraic topology. The disadvantage of it is the lower computing efficiency than other distance methods. We used the Aitchison distance to cluster analysis for gross error detection and obtained satisfactory results.

3.2.2 Classification Based on Similarity coefficient:

Similarity coefficient describes the similarity degree between samples. We used two different similarity coefficient methods included multidimensional vector cosine and Pearson's correlation coefficient to calculate the Similarity coefficient.

a) Multidimensional Vector Cosine

Cosine of the angle between multidimensional vectors is inspired by similar figures (Yu Xiu-lin, et al. 2002). For a spectral reflectance curves cluster contains n curves, if the length of the curves is not the major object of study, multidimensional vector cosine represents a kind of similarity coefficient which can show the corresponding similarity on the shape between the curves involved.

$$\cos\theta_{ij}=\frac{\sum_{\alpha=1}^{p}x_{i\alpha}x_{j\alpha}}{\sqrt{\sum_{\alpha=1}^{p}x_{i\alpha}^2\cdot\sum_{\alpha=1}^{p}x_{j\alpha}^2}},(0\le\cos\theta_{ij}\le 1)$$

(8)

where $\cos\theta_{i,j}$ = The cosine value between vector i and j
i, j = Number of vectors (spectral curves)
p = Count of sampling points (vector dimensions)
x = The reflectance value

In normal circumstances, spectral curves have the same sampling points in the same scope of bands and as a result, vectors based on the curves are in the space with same dimension. But if spectral instruments with different bands width or spectral resolution, spectral curves got by the instruments may have different count of sampling points and the vectors based on these curves have different dimensions, distance-based methods can not be used for the classification of such spectral curves. In this case, vectors with fewer dimensions ought to be interpolated to add dimensions as well as sampling points, or the ones with more dimensions should discard some to make all vectors having the same dimensions. The interpolation or discard process can reduce the impact given by raw data on the classification results, one serious problem is that the process on raw data may make errors or outliers in raw data amplified or neglected. The effect of this method on the gross error detection and outlier identification is not as good as that of the methods based on distance but we kept and used it as an ancillary method for cluster analysis in this paper.

b) Pearson's Correlation Coefficient

Pearson's correlation coefficient is known as the best method of measuring the correlation, because it is based on the method of covariance (Li Xi-qiang, et al., 2008). It gives information about the degree of correlation as well as the direction of the correlation. Pearson's correlation coefficient, r in Eq. (9), is a covariance-based theory about correlation measurement. It not only gives the correlation between samples, but also shows the direction relevance. Different classes can be divided by comparing a series of similar spectral curves' correlation coefficient and in turn data records or curves with gross errors or outliers can be identified.

$$r=\frac{\sum_{\alpha=1}^{p}x_{i\alpha}x_{j\alpha}-\frac{1}{p}\sum_{\alpha=1}^{p}x_{i\alpha}\cdot\sum_{\alpha=1}^{p}x_{j\alpha}}{\sqrt{\Delta}}$$

$$\Delta=\left[\sum_{\alpha=1}^{p}x_{i\alpha}^2-\frac{\left(\sum_{\alpha=1}^{p}x_{i\alpha}\right)^2}{p}\right]\left[\sum_{\alpha=1}^{p}x_{j\alpha}^2-\frac{\left(\sum_{\alpha=1}^{p}x_{j\alpha}\right)^2}{p}\right] \tag{9}$$

where i, j = Number of curves
p = Count of sampling points
x = The reflectance value

3.3 Box-and-Whisker Plots for Error and Outlier Viewing

Box-and-whisker plots, also known as schematic plots, can provide the information about variation range and extreme values of data (Wang Jian, et al., 2002). This statistic method is used for gross error detection and outlier identification because the significant character of robustness of the median and quartile. The box of a box-and-whisker plot represents the 50% values in the most middle of a data record. The upper and bottom edge of the box represent the value at 75% and 25% location of a data record which sorted from small to large, called the upper and lower quartile. The values of upper and bottom whisker are the max and min values which are smaller than max outlier limit and larger than min outlier limit in a data record. The max outlier limit is the value of upper quartile plus the interquartile range or *IQR*, similarly, the min outlier limit is the value of lower quartile minus *IQR*. The value of *IQR* is the absolute value of difference between the upper and lower quartile. All values larger than the max outlier limit or smaller than the min outlier limit are regarded as outliers (D.L. Massart, et al., 2005).

$$\begin{gathered} IQR = \left| Q_3 - Q_2 \right| \\ E_{\max} = Q_3 + 1.5 \times IQR; \quad E_{\min} = Q_2 - 1.5 \times IQR \end{gathered} \tag{10}$$

where IQR = The interquartile range
Q_2 / Q_3 = The lower/upper quartile
E_{max}/E_{min} = The max/min outlier limit

3.4 Pros and Cons

Cluster methods can find out the one which contain errors or outliers in a series of similar spectral curves. Practical applications on spectral data found that the effect of distance-based methods is superior to the methods based on the similarity coefficient. The Aitchison distance is more excellent than other methods on the detection results, but it is on the cost of calculation efficiency.

4. APPLICATION EXAMPLES AND VISUALIZATION

We select field spectral data got by a series of parallel experiments of 2 kinds of crops in Yanzhou mining area, winter wheat (*95128*, 4-April-2009, 7 curves, coded as *95128-1~7*) and single cropping rice (*Xiushui 110*, 13-September-2009, 9 curves, coded as *A002~9* and *A110*) as our experimental data. The results of classification, data plots and box-and-whisker plots are drawn on a self-developed graphic control based on GDI+ technology using C# language.

There is one data record contains measurement noises, the codes of them were *95128-1* and *A110*, in each of the 2 data groups. The spectral curve plot of each data record is shown in Figure 1.

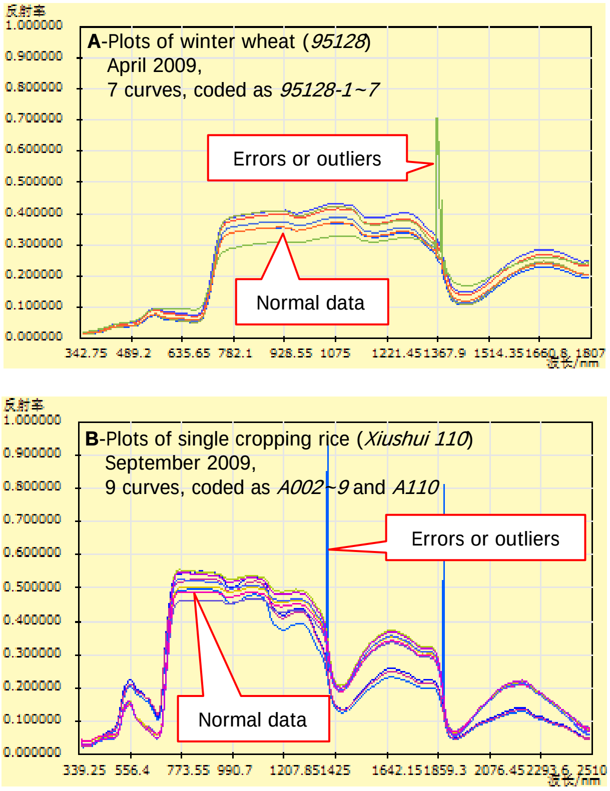

Figure 1. Plots of the experimental data

We calculated the self-consistent accuracy (SCA) of the 2 groups of spectral curves and the results are shown in table 1 and table 2.

Curve No. of winter wheat	Calculate start wavelength (nm)	Calculate end wavelength (nm)	SCA
95128-1	350	1800	±0.0446
95128-2	350	1800	±0.0194
95128-3	350	1800	±0.0144
95128-4	350	1800	±0.0104
95128-5	350	1800	±0.0297
95128-6	350	1800	±0.0212
95128-7	350	1800	±0.0176
T-SCA			±0.6344

Table 1. SCA results of winter wheat (*95128*)

Curve No. of single cropping rice	Calculate start wavelength (nm)	Calculate end wavelength (nm)	SCA
A002	1067	2189	±0.0258
A003	1067	2189	±0.0291
A004	1067	2189	±0.0178
A005	1067	2189	±0.0127
A006	1067	2189	±0.0169
A007	1067	2189	±0.0138

A008	1067	2189	±0.0342
A009	1067	2189	±0.0306
A110	1067	2189	±0.0503
T-SCA			±0.6665

Table 2. SCA results of single cropping rice (*Xiushui 110*)

In table 1 and table 2, each of the SCA value of the spectral curves is small and it represented that the spectral instruments in a stable state. Even though, we can see that the SCA values of *95128-1* and *A110* are relatively greater than others in their groups, this shows that there may be errors or outliers in the data of the curves.

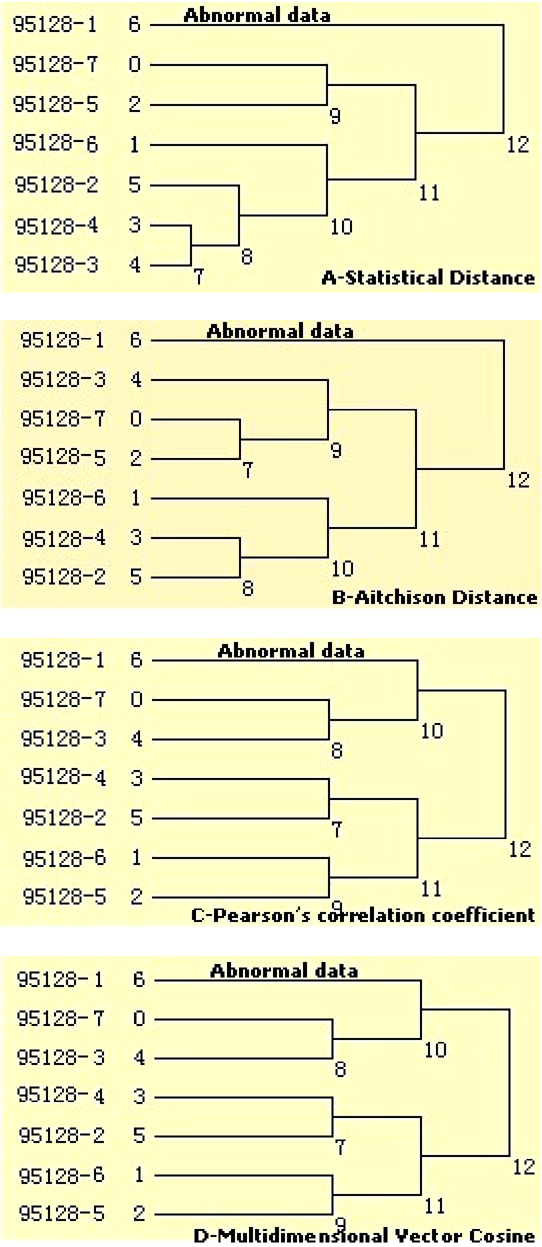

Figure 2. Cluster analysis plots of winter wheat (*95128*)

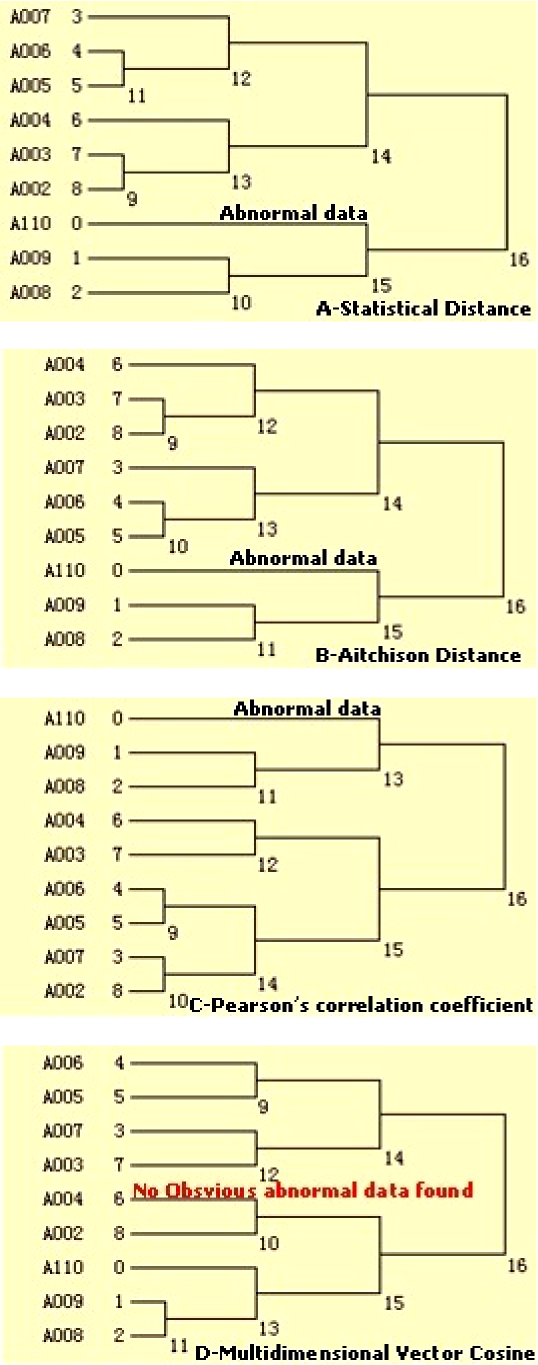

Figure 3. Cluster analysis plots of single cropping rice (*Xiushui 110*)

As shown in figure 2 and figure 3, spectral curves with errors or outliers can be identified by cluster analysis methods. Compared with spectral classification methods based on similarity coefficient, distance-based methods are more sensitive to abnormal data and give better results. If there are a finite number of sharp noise points in a spectral record (curve), the classification method based on multidimensional vector cosine may be not able to find out the abnormal curve contains errors or outliers. But this method can be used at the fuzzy classification analysis of ground object spectrum. Once the abnormal data record is identified in a series of similar data records, the detail information about the outliers or errors can be shown by box-and-whisker plots as in the figure 4.

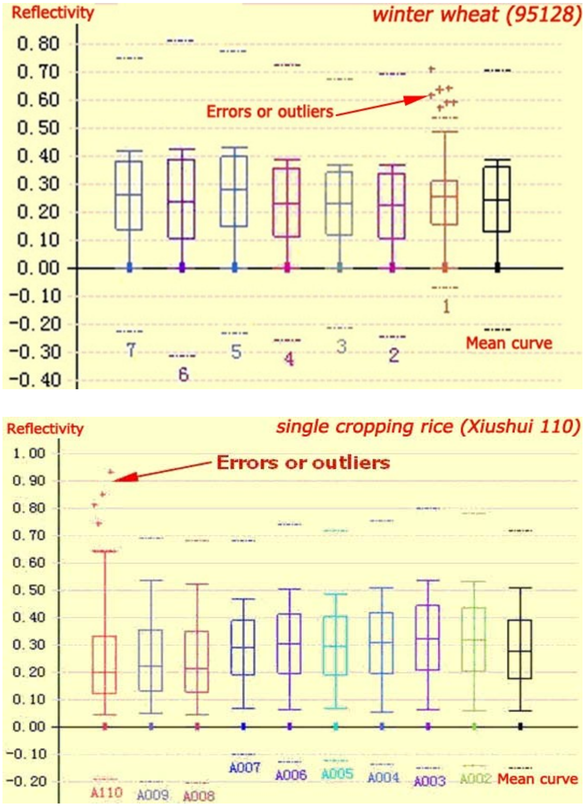

Figure 4. Box-and-whisker plots of the experimental data

Errors and outliers in a data record can be found in the box-and-whisker plots as shown in figure 4. There are several sharp noise points in the *95128-1* and *A110* spectral curve. The abnormal data records must be discarded or take some smoothing measures before being used for in-depth spectral analysis.

5. CONCLUSION

The precision of field spectroscopy data is affected by various factors and it is difficult to give out the priori statistics information of them. It is a venture that adopts some traditional gross error detection methods blindly. The process of studying on the statistical properties of research data according to the actual situation before appropriate methods are selected for data quality controlling is considered necessary and essential. This paper on 2 groups of field spectroscopy data, one was winter wheat and the other was single cropping rice. We used self-consistent accuracy (SCA) model to evaluate the stability of the spectrum instrument, proposed the theory that cluster analysis can be used for field spectroscopy data quality controlling and 4 different cluster methods were used for gross error visualized detection. For the characteristic bands of spectral data we implemented gross error detection and outlier identification in a visual way by the use of box-and-whisker plots. The GDI+ technology is used to draw plots automatically based on the calculate results of 4 different cluster analysis and box-and-whisker plots models. The practical application results show that abnormal data can be identified and removed before in-depth analysis is taken on. The goal of visualized gross error detection and outlier identification for field spectroscopy data got in Yanzhou mining area is achieved and thus provide quality assurance on spectrum data for in-depth spectroscopy analysis.

REFERENCES

Aitchison, J., 1986. *The Statistical Analysis of Compositional Data*. Wiley, NY.

Aitchison, J., 2001. *Algebraic Methods in Statistics and Probability: Contemporary Mathematics Series.* American Mathematical Society, Providence, Rhode Island, pp. 1-22.

D.L. Massart, J. Smeyers-Verbeke, X. Capron, et al., 2005. Visual Presentation of Data by Means of Box Plots. *LCGC Europe*, 18(4), pp. 215-218.

Edward J. Milton, Michael E. Schaepman, Karen Anderson, et al., 2009. Progress in field spectroscopy. *Remote Sensing of Environment*, 113(S1), pp. S92-S109.

He Ting, Liu Rong, Wang Jing., 2003. The Influences Factors on Field Spectrometry. *Geography and Geo-Information Science*, 19(5), pp. 6-10.

Kaufman, L. & Rousseeuw, P.J., 1990. *Finding Groups in Data*. Wiley, NY.

Li Xi-qiang, Wang Di, Lu She-ming, et al., 2008. Study of Fingerprint Spectra of Tobacco Flavor with Pearsonion Correlation Coefficient and the UPLC. *FINE CHEMICALS*, 25(5), pp. 475-478.

Vêncio RZ, Varuzza L, de B Pereira CA, et al., 2007. Appendix - Simcluster: clustering enumeration gene expression data on the simplex space, London, United Kingdom. http://www.biomedcentral.com/content/supplementary/1471-2105-8-246-s1.pdf (accessed 18 Aug. 2009)

Wang Jian & Jin Feng-xiang., 2002. Box-and-Whisker Plots and Correlation Model Method for Data Quality Control. *Journal of Shandong University of Science and Technology (Natural Science)*, 21(2), pp. 55-58.

Yu Xiu-lin & Ren Xue-song., 2002. *Multivariate Statistical Analysis.* China Statistic Press, Beijing, pp. 61-69.

APPENDIX

This article was sponsored by Chinese National Programs for High Technology Research and Development (2009AA12Z147); the Key Laboratory of Surveying and Mapping Technology on Island and Reed, State Bureau of Surveying and Mapping (2009B14); the Key Laboratory of Advanced Engineering Surveying of SBSM (TJES0805).