# An Integrated Software System for Supporting Real-Time Near-Infrared Spectral Big Data Analysis and Management

**DOI:**
10.1109/BigDataCongress.2017.22

**Document Version**
Accepted author manuscript

Link to publication record in Manchester Research Explorer

**Citation for published version (APA):**
Zhao, L., Hu, S., Zeng, X., Wu, Y., Lin, Y., Liu, J., Fan, S., Wang, Q., Xu, Z., & Wang, Y. (2017). An Integrated Software System for Supporting Real-Time Near-Infrared Spectral Big Data Analysis and Management. In *2017 IEEE International Congress on Big Data (BigData Congress)* (pp. 97-104). IEEE. https://doi.org/10.1109/BigDataCongress.2017.22

**Published in:**
2017 IEEE International Congress on Big Data (BigData Congress)

# An Integrated Software System for Real-Time Near-Infrared Spectral Big Data Analysis and Management

Liping Zhao, Shupeng Hu, Xiaojun Zeng
School of Computer Science
University of Manchester
Manchester, United Kingdom

Yuejin Wu[1], Yanqing Lin[1], Jing Liu[1], Shuang Fan[1],
Qi Wang[1], Yu Wang[2]
[1]Hefei Institutes of Physical Science, Chinese Academy of Sciences
[2]School of Resources and Environmental Engineering, Anhui University
Hefei, China

*Abstract*—**Near-infrared spectroscopy (NIRS) is a rapid, chemical-free, easy to use, and non-destructive analytical technique that has been widely applied to a diverse range of fields. NIRS analyzes the investigated samples through their NIR spectra. However, NIR spectral data are complex and multivariate, so multivariate data analysis methods (chemometrics) are used to interpret and predict the spectra's chemical and physical information. The analysis process is also very complex, involving both data processing and modeling. This paper first introduces basic concepts of NIRS analysis with the aim to show its complexity. The paper then characterizes the NIR spectral data using the "3H" of scientific big data, with the aim to show their challenges. Finally, the paper describes our initial effort on the development of an integrated software system to support efficient real-time NIRS data analysis and management. The paper claims that this development is an important contribution to tackling the challenges of scientific big data.**

*Index Terms*— **Near-infrared spectroscopy (NIRS), near-infrared (NIR) spectral big data, software technologies for real-time NIRS big data analysis and management, scientific big data**

## I. Introduction

Near-infrared spectroscopy (NIRS) is an instrumental method for acquiring near-infrared (NIR) spectra of materials for the purpose of quantitative and qualitative analysis [1]. In contrast to most other analytical and conventional chemical methods, NIRS is rapid, chemical-free, easy to use, and non-destructive [2], [3]. More specifically, NIRS has four distinctive advantages over other analytical methods [1]: First, it requires little or no sample preparation or manipulation. Second, it is very fast as it can acquire a spectrum of a sample in as little as a tenth of a second. Third, it can perform multi-constituent analyses from a single scan, as it is not necessary to scan the sample for each chemical constituent. Finally, it is a nondestructive measurement process so the analyzed sample can be returned to the original lot with no damage. NIRS is also environmental friendly, because no sample preparation or manipulation means that there are no hazardous chemicals, solvents, or reagents involved in the analysis [4].

Due to these advantages, NIRS has experienced an ever-increasing popularity in recent years and has gained widespread acceptance in different industry sectors for new product testing, product quality control and process monitoring [5]. Its applications cover as diverse fields as agricultural [6], [7], chemical [8], fertilizer [9], food [1], [10], [11], oil  [12], environmental [13], medicines [14], and pharmaceutical industries [15], [16], [17], [18]. In addition, NIRS is also a most versatile analytical method, suitable for analyzing solid, liquid, gas, and other forms of biotechnological or pharmaceutical products [17]. The recent advancement in instrument development has resulted in portable and, more recently, miniature NIRS instruments [1], [2], [3]. Such devices make it possible to conduct NIRS spectral data analysis in the fields (in-line), on site (on-line) and at production lines (at-line) [2], [3].

This paper reports on the development of software technologies for supporting fast real-time NIRS data analysis and management. This development is part of a collaborative project undertaken by an interdisciplinary team, comprising biophysicists from Chinese Academy of Sciences and Anhui University, and computer scientists from University of Manchester. The team was brought together to develop novel software technologies capable of supporting rapid, real-time NIRS data analysis and management, to address the unprecedented challenges of scale, rate, and complexity of scientific big data processing and management.

The remaining paper is organized as follows: Section II introduces some basic concepts of NIRS data analysis. Section III then links the NIR spectral data to scientific big data through the "3H" characteristics [19]. Section IV describes our initial effort on the development of an integrated software system to support real-time NIRS analysis and management. A brief review of related work is presented in Section V before a conclusion is drawn in Section VI.

## II. Basic Concepts of NIRS Data Analysis

### *A. NIR Spectra*

The *NIR spectral region* (800 – 2500nm) is situated in between the visible light (VIS) region (400 – 800nm) and the mid infrared (MIR) region (2500 – 15000nm), as shown in Figure 1. Spectra in the NIR region result from energy absorption by organic molecules, and comprise overtones and combinations of overtones originating from fundamental bond vibrations

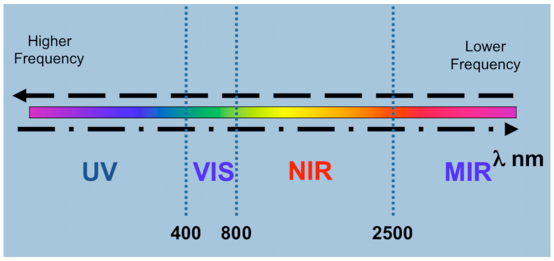


Fig. 1. The NIR spectral region (800 – 2500nm) is situated in between the visible light (VIS) region (400 – 800nm) and the mid infrared (MIR) region (2500 – 15000nm).

(stretching or bending) occurring in the MIR region of the spectrum [2].

In the application of NIRS (see Figure 2), a NIR spectrometer can obtain one single *NIR spectrum* of a tested sample by use of fiber optic probes (radiation light source). A NIR spectrum is composed of hundreds or even thousands of wavelengths within the NIR region [4]. The NIR wavelengths contain both chemical and physical information of the sample. Precisely how many wavelengths are contained with a spectrum depends on the NIRS device used. For example, our NIRS device can record a NIR spectrum with 256 wavelengths (Figure 3).

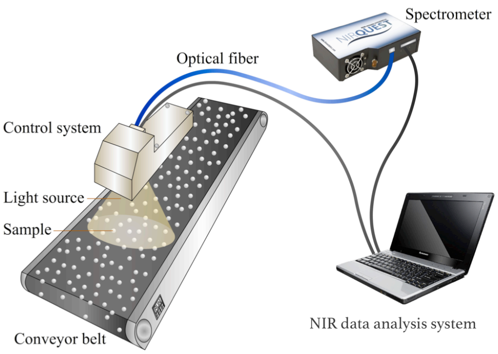


Fig. 2. NIR spectrometer scans a batch of samples one by one and records a single NIR spectrum from each sample.

*Wavelength* λ (the length of one wave) is defined as the distance between adjacent peaks (or wavelength points), and may be measured in meters, centimetres or nanometres (10$^{-9}$ meters or nm).

NIRS data are *multivariate* in nature due to a large number of data points (one at each wavelength) being collected for each sample during spectral collection [2]. They are also multidimensional, because each wavelength variable is mapped onto one point in a multivariate space (M-space), with as many axes as there are variables [20]. In other words, for a spectrum with p wavelengths (variables), the mapping will lead to a p dimensional space [5]. In our case, since each spectrum has 256 variables, the transformation of these variables will lead to a 256 M-space. For illustration purposes, Figure 4 shows a spectrum with three wavelengths and the transformation of these wavelengths to a three-dimensional space.

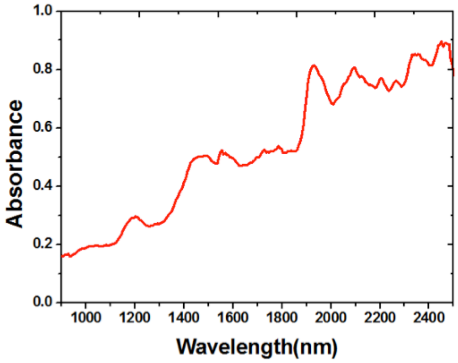
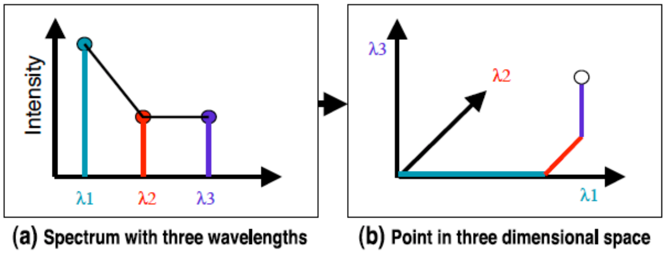


Fig. 3. A NIR spectrum of rice flour. The spectrum is composed of 256 wavelengths (variables).



(a) Spectrum with three wavelengths (b) Point in three dimensional space

Fig. 4. (a) A spectrum with three wavelengths (i.e., three variables); (b) the mapping of these wavelengths onto a three-dimensional space, with one axis for each wavelength. Both (a) and (b) are reproduced based on part of Fig. 3 of [5].

NIRS data analysis requires the use of multiple samples in order to provide an accurate analysis of the tested product. Consequently, NIRS needs to record the NIR spectra of multiple samples, not just one single sample. For example, in our case, each NIR analysis requires to use at least between 60 and 80 samples. Since each sample gives rise to one NIR spectrum, each NIR analysis needs to deal with a bundle of 60 – 80 spectra. As each spectrum contains 256 variables, each analysis needs to deal with at least between 60 × 256 = 15360 and 80 × 256 = 20480 variables or dimensions. Figure 5 shows the NIR spectra of rice flour obtained from 79 samples.

NIR spectra are very complex and normally possess broad overlapping NIR absorption bands, resulting in many overlapping peaks (referred to as "multicollinearity") [2], [4]. Consequently, it is difficult to interpret NIR spectra visually, assign specific features to specific chemical components or extract useful information contained in the spectra [2], [4], [21].

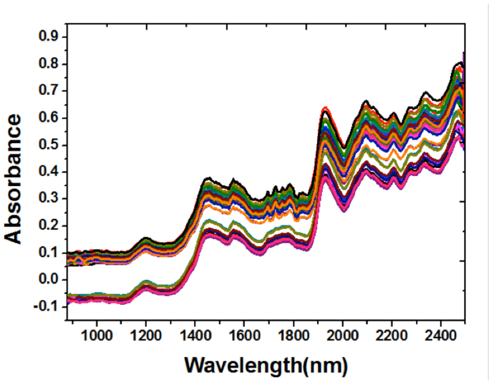


Fig. 5. NIR spectra of rice flour, obtained from 79 samples (79 equal portions). The spectra contain 79 × 256 = 20224 wavelengths or variables.

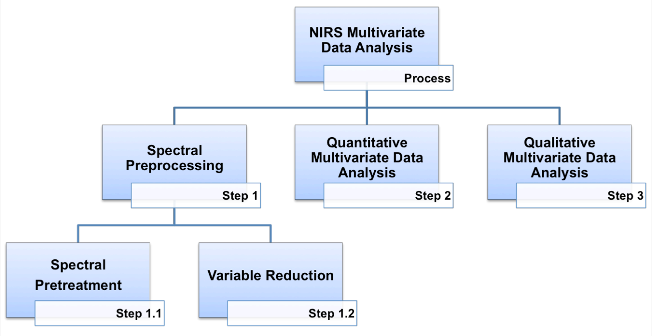

NIRS Multivariate Data Analysis
Process
Spectral Preprocessing
Step 1
Quantitative Multivariate Data Analysis
Step 2
Qualitative Multivariate Data Analysis
Step 3
Spectral Pretreatment
Step 1.1
Variable Reduction
Step 1.2

Fig. 6. The main steps in NIRS multivariate data analysis. Quantitative analysis requires the use of calibration models whereas qualitative analysis is based on classification methods.

### *B. NIRS Multivariate Data Analysis*

The purpose of multivariate data analysis is to relate the spectral variables of the investigated samples to the properties of the analyte (the reference samples whose chemical constituents are being identified and measured) [5]. This requires the use of special mathematical and statistical methods, known as chemometrics [20], [22], to extract relevant information and reduce irrelevant information in the NIR spectra of the investigated samples. NIRS data analysis is therefore fundamentally multivariate data analysis [2] or chemometric data processing [5].

The overall process of multivariate data analysis for NIRS is depicted in Figure 6, where the first step is pre-processing, which is concerned with spectral pretreatment and variable reduction. Spectral pretreatment aims to reduce noise or unwanted background information, whereas variable reduction is to reduce the number of variables to a few uncorrelated variables containing only relevant information from the samples [2], [3], [5]. The most common pretreatment methods include the moving-average method (Savitzky–Golay or SG), normalization, derivatives SG, multiplicative scatter correction (MSC) and standard normal variate (SNV) [2], [3], [5]. For variable reduction, the best known and most widely used is principal component analysis (PCA) [3], [5].

After pre-processing, the second step in multivariate data analysis is quantitative data analysis, which is then, optionally, followed by the third step of qualitative data analysis.

Quantitative multivariate data analysis uses the pre-existing knowledge about the reference samples to prognosticate the composition of the investigated samples. This knowledge is represented as a *multivariate calibration model*, which expresses a mathematical relationship between the NIR spectra of the analyzed samples and the respective reference values (i.e., chemical constituents, physical characteristics or other indirect properties) of a set of known reference samples [2]. The development of calibration models normally requires qualification by independent, reference analytical procedures [21] as well as substantial investment, although reuse of these models in future analyses can offset this development effort [3]. The most frequently used chemometric methods for calibration models are principal component regression (PCR) and partial least-squares (PLS) regression. These methods aim to construct calibration models capable of accurately predicting the chemical and physical characteristics and properties of the samples under investigation [2], [3], [4], [5]. The process of model construction contains the following basic steps:

1. Select a representative calibration sample set
2. Acquire sample spectra and determine reference values
3. Construct the calibration model to establish the spectrum–property relationship using multivariate methods
4. Validate the model.

Qualitative analysis is used to confirm the identity or the quality of the unknown samples on the basis of their physical or chemical attributes [2], [5]. The most frequently used multivariate calibration methods are principal component regression (PCR) and partial least-squares (PLS) regression [2], [5].

Qualitative analysis requires the use of a library of representative spectra to compare the spectrum of the investigated sample, to identify the similarities and differences between the sample spectrum and the spectra in the library [2], [5]. Qualitative analysis methods are based on multivariate classification methods, also known as pattern-recognition methods. These methods are divided into supervised methods, such as cluster analysis, and non-supervised methods, such as linear discriminant analysis (LDA) and PLS discriminant analysis (PLS-DA), depending on whether or not the class to which the samples belong is known [5].

This brief introduction shows that both NIR spectral data and NIRS data analysis methods are very complex.

## III. NIR Spectral Data: A Case of Scientific Big Data

Blanco and Villarroya [3] stated: “the powerful NIR instruments currently available quickly provide vast amounts of data that require speedy, efficient processing if it is to yield useful analytical information.” We argue that NIR spectral data are scientific big data. In this section, we characterize NIR spectral big data according to the “3H” characteristics of scientific big data defined by Guo et al [19], which are high dimension, high complexity, and high uncertainty.

*High dimension*. Guo et al stated [19]: “In general, the external representations of ‘scientific big data’ have high correlation and multiple data attributes. In principle, scientific big data has a high dimension.”

As stated in Section II, NIR spectral data are multivariate data, consisting of hundreds or even thousands of variables. NIR spectral data are represented as points in a multivariate space (M-space), with as many axes as there are variables. Multivariate data are known to contain a huge number of correlated variables (called multicollinearity or collinearity).

*High complexity*. Guo et al stated [19]: “Scientific big data mostly applies to complex nonlinear systems, and is accompanied by a complex data model. Therefore, the issue for scientific big data computation is not merely a matter of data processing and analysis; it is also a matter of joint modeling and computation with complex system modeling and data.”

As described in Section II, NIR spectra are very complex and difficult to interpret. NIRS data analysis requires the use of complex chemometric methods to pretreat and analyze the

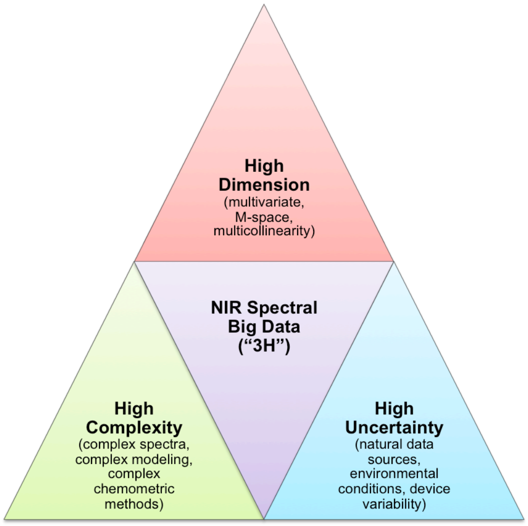

High Dimension
(multivariate, M-space, multicollinearity)
NIR Spectral Big Data ("3H")
High Complexity
(complex spectra, complex modeling, complex chemometric methods)
High Uncertainty
(natural data sources, environmental conditions, device variability)

Fig. 7. Characterizing NIR spectral big data using "3Hs" of scientific big data.

spectra, involving reduction of unwanted background information and correlated variables, and construction of calibration models to extract relevant analytical information. The whole analysis process is therefore very complex, consisting of data processing and modeling.

*High uncertainty*. Guo et al stated [19]: "Scientific big data, in general, comes from the natural process of perception and data acquisition. Because of the characteristics of these data sources, scientific big data generally has some error and incompletion, which results in data with high uncertainty. Scientific big data is often applied to the disciplines of natural systems…"

NIR spectra are obtained from natural products and often in the natural environment, rather than the controlled laboratory environment. NIR spectral data are therefore affected by sample variations and environmental conditions. In addition, NIR spectral data can also be affected by device variations (as no two devices can be exactly the same). Hence NIR spectral data are highly heterogeneous, with the high uncertainty and high variety.

Figure 7 summarizes these "3Hs" of NIR spectral big data.

NIR spectral data can also be characterized by the three defining characteristics of big data: volume, variety, and velocity [23]: *Volume*. NIR spectral data have the large sample size. *Variety*. NIR spectral data contain a large variety of samples and calibration models. *Velocity***.** NIR spectral data are generated at the high speed.

Fan et al [24] posited that the challenges of big data analysis are characterized by high dimensionality and large sample size. They explained: "(i) High dimensionality brings noise accumulation, spurious correlations, and incidental homogeneity; (ii) High dimensionality combined with large sample size creates issues such as heavy computational cost and algorithmic instability; (iii) The massive samples in Big Data are typically aggregated from multiple sources at different time points using different technologies." These challenges are clearly reflected in scientific big data in general and NIR spectral big data in particular.

## IV. Software Support for Real-Time NIRS Analysis

### *A. System Architecture and Process*

As Section II alludes, before NIRS data analysis can be performed on the investigated samples, the corresponding calibration models that provide the reference values of the analytical target property must exist. Therefore, NIRS data analysis is necessarily a two-stage process, made of the model construction (also called modeling or calibration) stage and the spectral data analysis stage. Since the modeling stage is mainly a manual process, requiring the input and actions from expert analysts, our software development effort has concentrated on the analysis stage.

Specifically, the overall aim of our system is to enable real-time NIRS data analysis by automating all the analysis process steps. Currently our system only supports quantitative multivariate data analysis, but it can be extended easily to support qualitative analysis should the needs arise. Figure 8 displays the architecture of our system as well as the process flows of the system. The main components of this system are briefly described as follows:

- *Spectral Scanner*. Used to operate the NIRS spectrometer. It is responsible for 1) configuring the spectrometer, 2) scanning the sample one at a time, 3) recording the sample spectrum, and 4) calculating the absorbance of the spectrum. This component is implemented in C programming language for efficiency.
- *Spectral Preprocessor*. Supporting spectral pretreatment and variable reduction operations. Pretreatment and variable reduction methods (e.g., SNV and SG) are implemented using both Java programming language and the MATLAB library of chemometric methods.
- *Quantitative Multivariate Analyser*. Supporting quantitative multivariate data analysis. With an imported calibration model, this component employs the regression principles of multivariate calibration methods (e.g. PCR or PLS) to predict the chemical or physical property in the sample. These methods are also implemented using Java and the MATLAB library of chemometric methods.
- *Spectral Visualizer*. Used to display the analysis results, consisting of the analyzed spectrum and the predicted chemical or physical property of the sample. This component is implemented in Java.
- *NIR Spectral Database*. This database stores and manages the output data from the aforementioned four components. Specifically, five types of data are stored in this database: 1) NIR spectrometer information; 2) sample information; 3) NIR spectra; 4) information of chemometric methods and multivariate calibration methods; and 5) calibration models. MySQL is used to implement this database system.

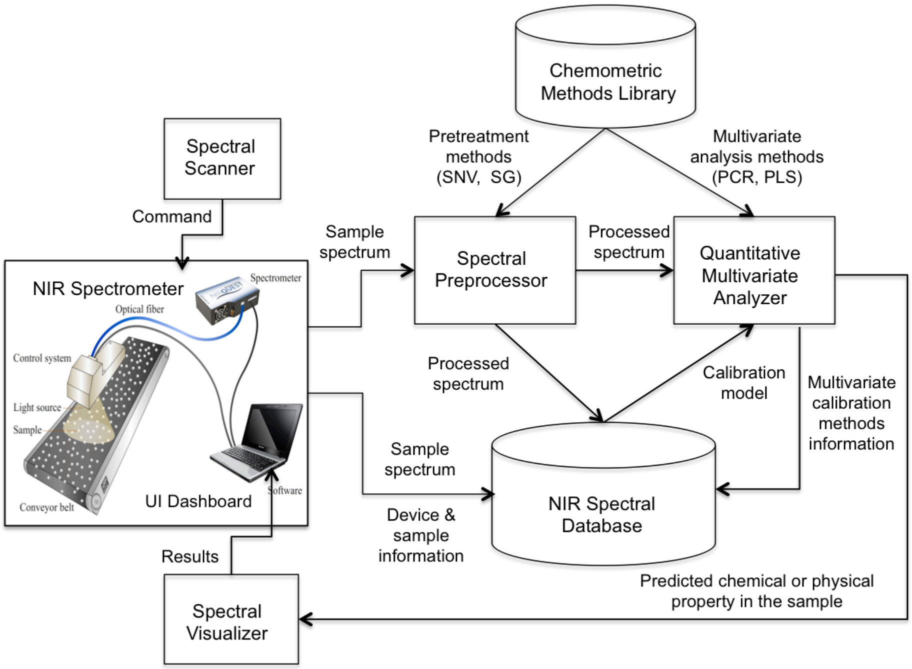


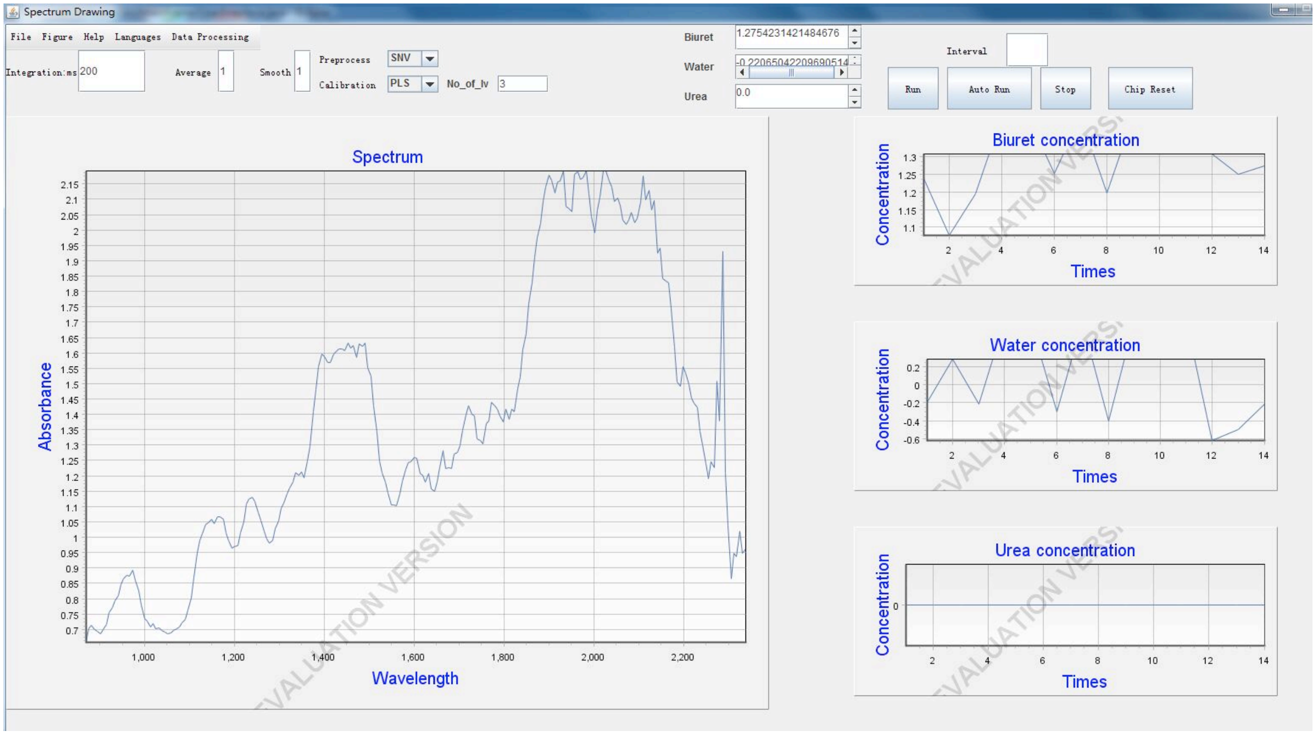
Fig. 8. Software system architecture for real-time NIRS analysis. The architecture shows the main system components and the interactions among these components. Arrows represent process data flows.

- *Chemometric Method Library*. Used to store the implementations of chemometric methods. This component is implemented using MATLAB.

Finally, the overall system control, the integration of the system components and the coordination between them are implemented using the Java programming language.

Under our system, real-time NIRS data analysis can be performed automatically in this order:

1. The conveyor belt moves the investigated sample under the light source.
2. Spectral Scanner receives the sample information and produces a spectrum for the sample.



Integration:ms 200

Average 1

Smooth 1

Preprocess SNV

Calibration PLS

No_of_lv 3

Biuret 1.2754231421484676

Water -0.22065042209690514

Urea 0.0

Interval

Fig. 9. The UI dashboard displays the spectrum of a grain sample and the analysis results (the chemical information) of the sample.

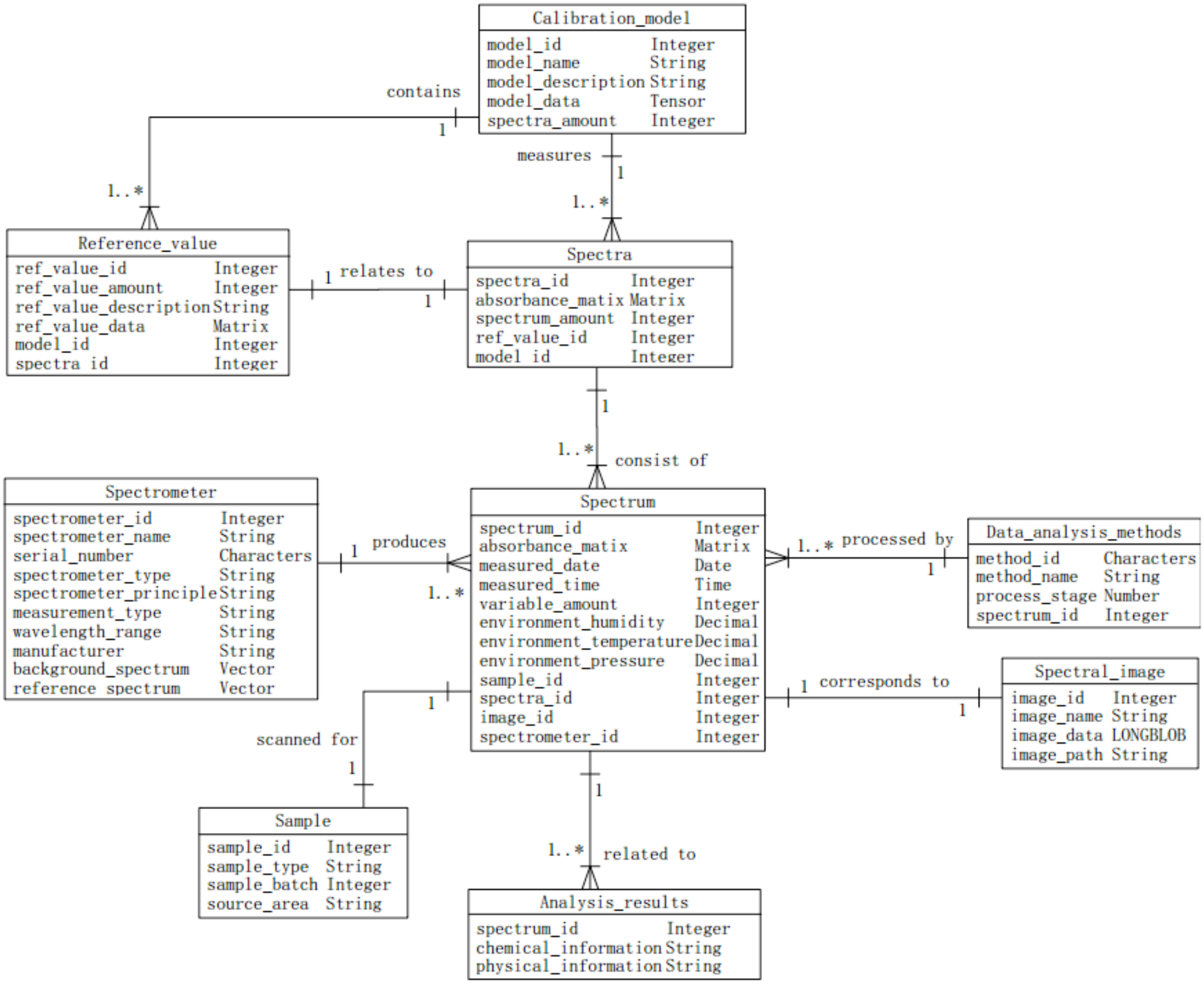


Fig. 10. The Entity-Relationship Diagram for the NIR spectral database. The diagram shows the key entities of the database.

3. Spectral Preprocessor performs pre-treatment and variable reduction on the spectrum and generates the processed spectrum.
4. Quantitative Multivariate Analyzer analyzes the processed spectrum.
5. Spectral Visualizer displays the analysis results on the user interface dashboard, comprising the predicted chemical or physical property of the sample. It also displays the sample spectrum.

Figure 9 shows the analysis results of a grain spectrum displayed by Spectral Visualizer.

### *B. A Data Model for NIR Spectral Big Data*

The NIR Spectral Database in our system intends to serve a dual purpose: 1) to support seamless and rapid real-time NIRS data analysis through efficient data input and output, and 2) to enable long-term spectral data sharing and calibration model reuse through efficient storage and management of different data types. The structure of this database is still evolving. The current version consists of nine entities and the relationships between them. Figure 10 shows the Entity-Relationship (ER) diagram of this basic data structure. The entities correspond to the concepts that we have already introduced in Section II and should therefore be self-explanatory.

Our long-term goal is to build a metadata model of NIR spectral big data. The importance of metadata for scientific big data is nicely explained by Gray et al. [25]: "Metadata is the descriptive information about data that explains the measured attributes, their names, units, precision, accuracy, data layout and ideally a great deal more. Most importantly, metadata includes the data lineage that describes how the data was measured, acquired or computed."

The use of data types Vector, Matrix and Tensor in Spectrum, Spectra and Calibration_model entities needs some explanation: As mentioned in Section II, NIRS data are multivariate in nature due to a large number of data points (one at each wavelength) being collected for each sample during spectral collection [2]. For each spectrum, the data points are related to a group of discrete values, represented as a mathematical vector. As each investigation (experiment) requires multiple samples, the spectral data of these samples thus constitute a matrix (a vector of vectors). As the mapping from the values of the investigated sample to the values of the reference sample has to be the one-to-one and onto relationship, the reference values of the spectra should also be matrix. Since a calibration model relates the relationship between the spectral values and reference values, it should be expressed as a tensor, consisting of a sequence of spectral matrices and reference values.

However, since MySQL does not support vector, matrix and tensor data types, we have worked around it by implementing vector data as a string and matrix data as a sequence of strings. We added an attribute "string_sequence" to the relevant tables to record the positions of data items in a string so that we can reconstruct a matrix or tensor from sequences during the computation. One future work will investigate alternative types of database system and build our own types to provide direct support for these NIR-specific data types.

Managing the enormous amount of scientific data being collected is regarded as the key to scientific progress [26]. Yet, as Ailamaki et al. [26] noted, although technology (e.g., NIRS) and instrumentation (e.g., NIR spectrometer) allow for the extreme collection rates of scientific data, data storage and management is still performed with stale techniques developed for small data sets (e.g., existing database management systems). Therefore, in order to exploit the value of scientific big data and support efficient processing of these data, advanced data storage and management technology are needed.

### C. Initial Validation

So far, we have used our system to analyze water concentration and protein content of rice and grains, and urine concentration in urea granules. Our predicted values for these products are very accurate, as they are very close to their reference values.

Table 1 shows the predicted results of the urine concentration in 10 urea granule samples from our system, their corresponding reference values and the standard deviation between two sets of values. The results for other types of sample are not provided here for space consideration.

TABLE 1
NIRS ANALYSIS RESULTS FOR 10 UREA GRANULE SAMPLES PRODUCED BY OUR SYSTEM

| Sample Number | Predicted Value | Standard Deviation | Reference Value |
|---|---|---|---|
| 1 | 94.806 | 0.324 | 94.200 |
| 2 | 94.446 | 0.247 | 94.500 |
| 3 | 95.908 | 0.311 | 95.700 |
| 4 | 95.881 | 0.282 | 95.900 |
| 5 | 96.304 | 0.327 | 96.200 |
| 6 | 96.669 | 0.438 | 96.800 |
| 7 | 97.617 | 0.323 | 97.300 |
| 8 | 98.566 | 0.372 | 97.900 |
| 9 | 98.884 | 0.299 | 98.200 |
| 10 | 98.429 | 0.291 | 98.400 |

## V. RELATED WORK

In spite of the wide applications of NIRS in recent years, little has been written about the development of software technologies to support NIRS data analysis or data management. Our literature review shows that there is only one persistent research project on the development of a spectral database called SPECCHIO over a period of 10 years [27], [28], [29], [30], [31]. According to Hueni et al., the researchers behind SPECCHIO [30], only three spectral database systems appeared in literature, which are SPECCHIO [27], [28], SpectraProc [29] and a free online reference spectral library [32]. However, our literature review shows that SpectraProc is in fact an incremental development of SPECCHIO, developed by the same research team. Therefore, suffice it to say that so far there are only two clear examples of spectral databases.

Yet, both SPECCHIO and the online reference spectral library store spectral signatures of images obtained from remote sensors, such as Google Earth or any Earth observation systems. They are therefore not suitable for storing and managing NIR spectral big data.

We have used two commercial NIRS data analysis software systems, *Unscrambler* and *OPUS*. But commercial confidentiality means that we can only get a glimpse of these systems from an end-user perspective, which is described (in comparison with our system) as follows.

- *Unscrambler*. This system does not support real-time NIRS analysis so it cannot scan the samples and perform the analysis instantly. In addition, it uses a special file format to store the data, which cannot be shared with other NIRS devices or systems. Unscrambler integrates the implementation of chemometric methods with the hardware device so that users cannot access the code of these methods.
- *OPUS*. This system can only be used in the laboratory environment for lab experiments. It also uses its own file format file to store data, which cannot be shared with other NIRS devices or systems. Like Unscrambler, OPUS also integrates the implementation of chemometric methods with the hardware device so that users cannot access the code of these methods.

This brief overview shows that there is an important gap in the research and development of NIRS software technologies to support efficient data storage, management and analysis. Our work represents a first step towards filling this gap.

## VI. CONCLUSION

This paper has made two important contributions to scientific big data: 1) It has characterized NIRS data as scientific big data and introduced NIRS as a new scientific big data application; 2) It has reported on the development of an integrated software system to support efficient, real-time NIRS data analysis and management.

In presenting these contributions, we sought to illustrate the complexity of the NIRS field and the importance of software technologies in addressing this complexity. The paper has also identified the inadequacy of relational database management systems and the research gap in the literature.

Due to the complexity of the topic and space constraints, this paper has necessarily placed more emphasis on the description of NIRS data analysis concepts and methods. We plan to expand on the description of software development and validation in a forthcoming journal article.

## References

[1] Y. Ozaki, W. F. McClure, and A. A. Christy, *Near-infrared spectroscopy in food science and technology*: John Wiley & Sons, 2006.

[2] M. Manley, "Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials," *Chemical Society Reviews,* vol. 43, pp. 8200-8214, 2014.

[3] M. Blanco and I. Villarroya, "NIR spectroscopy: a rapid-response analytical tool," *TrAC Trends in Analytical Chemistry,* vol. 21, pp. 240-250, 2002.

[4] M. AG, *NIR Spectroscopy: A guide to near-infrared spectroscopic analysis of industrial manufacturing processes*. CH-9101 Herisau, Switzerland, 2013.

[5] G. Reich, "Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications," *Advanced drug delivery reviews,* vol. 57, pp. 1109-1143, 2005.

[6] P. Williams and K. Norris, *Near-infrared technology in the agricultural and food industries*: American Association of Cereal Chemists, Inc., 1987.

[7] B. Khakimov, G. Gürdeniz, and S. B. Engelsen, "Trends in the application of chemometrics to foodomics studies," *Acta Alimentaria,* vol. 44, pp. 4-31, 2015.

[8] C. Fernández, M. S. Larrechi, and M. P. Callao, "An analytical overview of processes for removing organic dyes from wastewater effluents," *TrAC Trends in Analytical Chemistry,* vol. 29, pp. 1202-1211, 2010.

[9] M. L. Stone, J. B. Solie, W. R. Raun, R. W. Whitney, S. L. Taylor, and J. D. Ringer, "Use of spectral radiance for correcting in-season fertilizer nitrogen deficiencies in winter wheat," *Transactions of the ASAE,* vol. 39, pp. 1623-1631, 1996.

[10] A. L. B. Brito, L. R. Brito, F. A. Honorato, M. J. C. Pontes, and L. F. B. L. Pontes, "Classification of cereal bars using near infrared spectroscopy and linear discriminant analysis," *Food Research International,* vol. 51, pp. 924-928, 2013.

[11] C. Zhang, W. Kong, F. Liu, and Y. He, "Measurement of aspartic acid in oilseed rape leaves under herbicide stress using near infrared spectroscopy and chemometrics," *Heliyon,* vol. 2, p. e00064, Jan 2016.

[12] M. Blanco, S. Maspoch, I. Villarroya, X. Peralta, J. M. Gonzalez, and J. Torres, "Geographical origin classification of petroleum crudes from near-infrared spectra of bitumens," *Applied Spectroscopy,* vol. 55, pp. 834-839, 2001.

[13] P. Rosén, E. Dåbakk, I. Renberg, M. Nilsson, and R. Hall, "Near-infrared spectrometry (NIRS): a new tool for inferring past climatic changes from lake sediments," *The Holocene,* vol. 10, pp. 161-166, 2000.

[14] A. Bozkurt, A. Rosen, H. Rosen, and B. Onaral, "A portable near infrared spectroscopy system for bedside monitoring of newborn brain," *BioMedical Engineering OnLine,* vol. 4, pp. 29-29, 2005.

[15] J. Luypaert, D. L. Massart, and Y. Vander Heyden, "Near-infrared spectroscopy applications in pharmaceutical analysis," *Talanta,* vol. 72, pp. 865-883, 2007.

[16] M. Jamrógiewicz, "Application of the near-infrared spectroscopy in the pharmaceutical technology," *Journal of pharmaceutical and biomedical analysis,* vol. 66, pp. 1-10, 2012.

[17] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, and N. Jent, "A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies," *Journal of pharmaceutical and biomedical analysis,* vol. 44, pp. 683-700, 2007.

[18] W. Plugge and C. Van Der Vliest, "The use of near infrared spectroscopy in the quality control laboratory of the pharmaceutical industry," *Journal of pharmaceutical and biomedical analysis,* vol. 10, pp. 797-803, 1992.

[19] H. Guo, L. Wang, F. Chen, and D. Liang, "Scientific big data and digital earth," *Chinese science bulletin,* vol. 59, pp. 5066-5073, 2014.

[20] S. Wold, "Chemometrics, why, what and where to next?," *Journal of pharmaceutical and biomedical analysis,* vol. 9, pp. 589-596, 1991.

[21] E. M. Agency, "Guideline on the use of near infrared spectroscopy by the pharmaceutical industry and the data requirements for new submissions and variations," in *http://www.ema.europa.eu/ema/*, ed. London, 2014.

[22] B. K. Lavine, "Chemometrics," *Analytical Chemistry,* vol. 72, pp. 91-98, 2000.

[23] P. Zikopoulos and C. Eaton, *Understanding big data: Analytics for enterprise class hadoop and streaming data*: McGraw-Hill Osborne Media, 2011.

[24] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National science review,* vol. 1, pp. 293-314, 2014.

[25] J. Gray, D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber, "Scientific data management in the coming decade," *ACM SIGMOD Record,* vol. 34, pp. 34-41, 2005.

[26] A. Ailamaki, V. Kantere, and D. Dash, "Managing scientific data," *Communications of the ACM,* vol. 53, pp. 68-78, 2010.

[27] S. Bojinski, M. Schaepman, D. Schläpfer, and K. Itten, "SPECCHIO: a Web-accessible database for the administration and storage of heterogeneous spectral data," *ISPRS journal of photogrammetry and remote sensing,* vol. 57, pp. 204-211, 2002.

[28] S. Bojinski, M. Schaepman, D. Schläpfer, and K. Itten, "SPECCHIO: a spectrum database for remote sensing applications," *Computers & Geosciences,* vol. 29, pp. 27-38, 2003.

[29] A. Hueni and M. Tuohy, "Spectroradiometer data structuring, pre-processing and analysis: an IT based approach," *Journal of Spatial Science,* vol. 51, pp. 93-102, 2006.

[30] A. Hueni, J. Nieke, J. Schopfer, M. Kneubühler, and K. I. Itten, "The spectral database SPECCHIO for improved long-term usability and data sharing," *Computers & Geosciences,* vol. 35, pp. 557-565, 3// 2009.

[31] A. Hueni, T. Malthus, M. Kneubuehler, and M. Schaepman, "Data exchange between distributed spectral databases," *Computers & geosciences,* vol. 37, pp. 861-873, 2011.

[32] J. G. Ferwerda, S. D. Jones, and M. Reston, "A free online reference library for hyperspectral reflectance signatures," *SPIE Newsroom,* pp. 1-2, 2006.