Hungary-Romania
Cross-Border Co-operation
Programme 2007-2013

European Union
European Regional Development Fund

HURO/1001/121/2.2.2/01 BIOETHANOL
A new method and system for real time fermentation process monitoring

DATA PROCESSING AND CURVE FITTING FOR OPTICAL DENSITY - ETHANOL CONCENTRATION CORRELATION

VESELLENYI TIBERIU

ŢARCĂ RADU CĂTĂLIN

ŢARCĂ IOAN CONSTANTIN

UNIVERSITY OF ORADEA
2013

DATA PROCESSING AND CURVE FITTING FOR OPTICAL DENSITY - ETHANOL CONCENTRATION CORRELATION

Vesellenyi Tiberiu, Ţarcă Radu Cătălin, Ţarcă Ioan Constantin

Abstract: The paper describes data processing techniques and methods used to find the correlation between measured optical density and ethanol concentration. These methods fall in the category of statistical data processing and curve fitting. The results of the analysis are presented and conclusions are drawn for further investigations.

Keywords: data processing, normal distribution, curve fitting, optical density, ethanol concentration.

1. Introduction

Experimental data obtained during measurement of optical density using the Jaz platform from Ocean Optics has to be analyzed in order to obtain the correlation of the measured data with ethanol concentration. The optical density data has been acquired for known concentrations of ethanol in order to calibrate the procedure. The measurement for the same concentration had been repeated for a large number of times in order to have statistical significant information. After acquiring the data the values were stored in separate files for each ethanol concentration. In order to analyze the data the MATLAB programming environment had been used. MATLAB software has toolboxes with a large number of options for statistical data processing and analysis and also for curve fitting. The statistical data analysis functions were used to acquire knowledge on measured values distribution in order to have an image on accuracy and precision of measurements. After that, curve fitting algorithms were employed to find a minimal error function which complies with the measured data and which can be a good candidate for

calibration function. We also present the results of statistical analysis and curve fitting obtained and conclusions that can be drawn from these results.

2. Statistic data processing

In probability theory, the normal (or Gaussian) distribution is a continuous probability distribution, defined by the formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \qquad (1)$$

The parameter μ in this formula is the *mean* of the distribution. The parameter σ is its standard deviation; its variance is σ^2. A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate.

If $\mu = 0$ and $\sigma = 1$, the distribution is called the standard normal distribution or the unit normal distribution, and a random variable with that distribution is a standard normal deviate.

Normal distributions are important in statistics, and are often used in data processing for real-valued variables whose distributions are not known. One reason for their usage is the central limit theorem in which the mean of a large number of random variables independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution. Thus, physical quantities that are expected to be the sum of many independent processes often have a distribution very close to normal. Another reason is that a large number of results and methods (such as least squares parameter fitting) can be derived analytically, in explicit form, when the relevant variables are normally distributed.

The normal distribution is the continuous distribution with the maximum entropy for a given mean and variance. The normal distribution is symmetric about its mean, and is non-zero over the entire real line. In some cases variables may be better described by other distributions, such as the log-normal distribution or the Pareto distribution.

If the analyzed values have a significant fraction of outliers, values that lie many standard deviations away from the mean the normal distribution may be not suitable for those measurements. In this case least-squares and other statistical inference methods that are optimal for normally distributed variables often become highly unreliable. In these cases an appropriate robust statistical inference methods had to be assumed. The normal distributions

are a subclass of the elliptical distributions. There are many other distributions that resembles normal distribution, such as Cauchy's, Student's, and logistic.

3. Curve fitting

Generally, curve fit algorithms determine the best-fit parameters by minimizing a chosen merit function. In order to optimize the merit function, it is necessary to select a set of initial parameter estimates and then iteratively refine the merit parameters until the merit function does not change significantly between iterations. One of the most used algorithm for nonlinear least squares calculations is the The Levenberg-Marquardt which is also implemented in Matlab.

In order to understand curve fitting algorithms there are two important topics that has to be defined: parametric fitting and the least square method.

Parametric fitting

Parametric fitting involves finding parameters for one or more models that can be fitted to data. The data is assumed to be statistical in nature and is divided into two components: *deterministic and random component.*

The deterministic component is given by a parametric model and the random component is often described as *error* associated with the data. The model is a function of the independent variable and one or more coefficients. The error represents random variations in the data that follow a specific probability distribution, usually a normal or Gaussian distribution. In case of optical density measurements the variations can come from many different sources, but we assume that they are present in the measured data.

Systematic variations can also exist, but they can be eliminated with a proper calibration of the instrument and careful handling of the instrument.

Least square method

Curve Fitting Toolbox software (included in the Matlab environment) uses the method of least squares when fitting data. Fitting requires a parametric model that relates the response

data to the predictor data with one or more coefficients. The result of the fitting process is an estimate of the model coefficients.

To obtain the coefficient estimates, the least-squares method minimizes the summed square of residuals. The residual for the *i*-th data point r_i is defined as the difference between the observed response value y_i and the fitted response value $\hat{y}_i$, and is identified as the error associated with the data:

$$r_i = y_i - \hat{y}_i \quad (2)$$

The summed square of residuals is given by:

$$S = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

where *n* is the number of data points included in the fit and *S* is the sum of squares error estimate. There area large number of types of least-squares fitting which includes: linear least squares, weighted linear least squares, robust least squares, nonlinear least squares.

When fitting data that contains random variations, there are two important assumptions that are usually made about the error:

- the error exists only in the response data, and not in the predictor data;
- the errors are random and follow a normal (Gaussian) distribution with zero mean and constant variance, σ^2.

As it has been shown in the previous chapter, the errors are assumed to be normally distributed because the normal distribution often provides an adequate approximation to the distribution of many measured quantities. Although the least-squares fitting method does not assume normally distributed errors when calculating parameter estimates, the method works best for data that does not contain a large number of random errors with extreme values.

The normal distribution is one of the probability distributions in which extreme random errors are uncommon.

There are several different models available for curve fitting. Some of the most used are:

- *Straight Line*- by choosing the line that minimizes the least square sum of the vertical distance *d*, of all predictor data;
- *Logarithmic* -calculates the least squares fit through points by using the following equation: $y = a + b \cdot lnx$, where *a* and *b* are constants and *ln* is the natural logarithm function;
- *Exponential* -by using the following equation: $y = a \cdot e^{bx}$, where *a* and *b* are constants, and e is the base of the natural logarithm;
- *Power* -by using the following equation: $y = a \cdot x^b$, where *a* and *b* are constants;
- *Polynomial* - by using the following equation: $y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \ldots$, where a_0, a_1, a_2, etc., are constants.

4. Experimental data processing

The experimental measurement data had been first evaluated using histogram and normal probability plot of each set of data for the same concentration of ethanol. There had been taken measurements of sets of 84 to 282 values for 10 output targets each. These are measured for ethanol concentrations that are ranging from 2% to 20% with an increment of 2%.In figure 1, the raw data is shown.

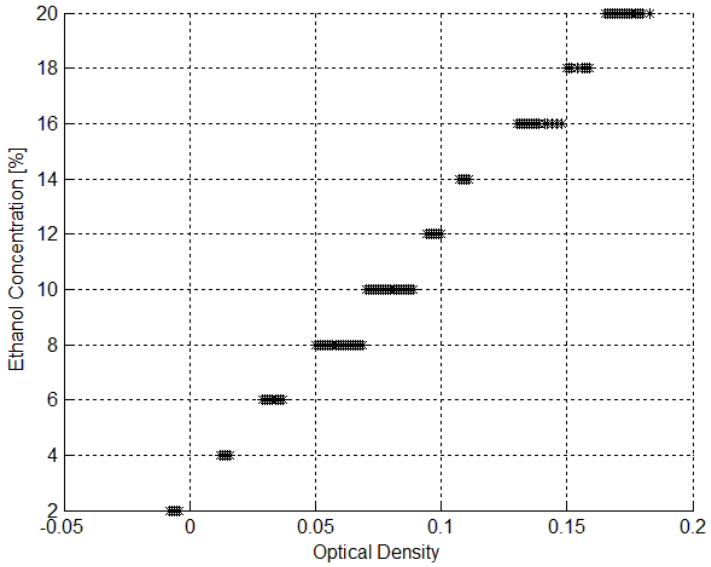
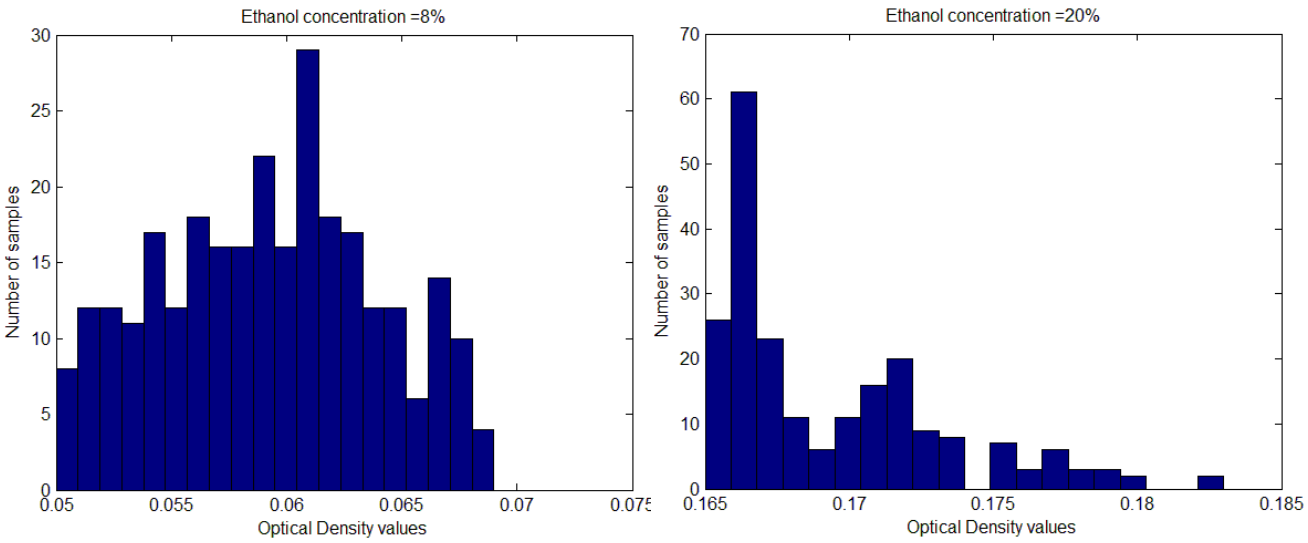


Fig. 1. Ethanol concentration versus optical density.

In figure 2 we can observe the histograms associated with optical density data for different concentrations of ethanol. In some cases the histograms are not bell-shaped showing a large number of marginal values. This tendency gives us a hint that the data are not normally distributed. In order to enhance the analysis we also computed the normal probability of the data. Some specific diagrams of normal probability are shown in figure 3.

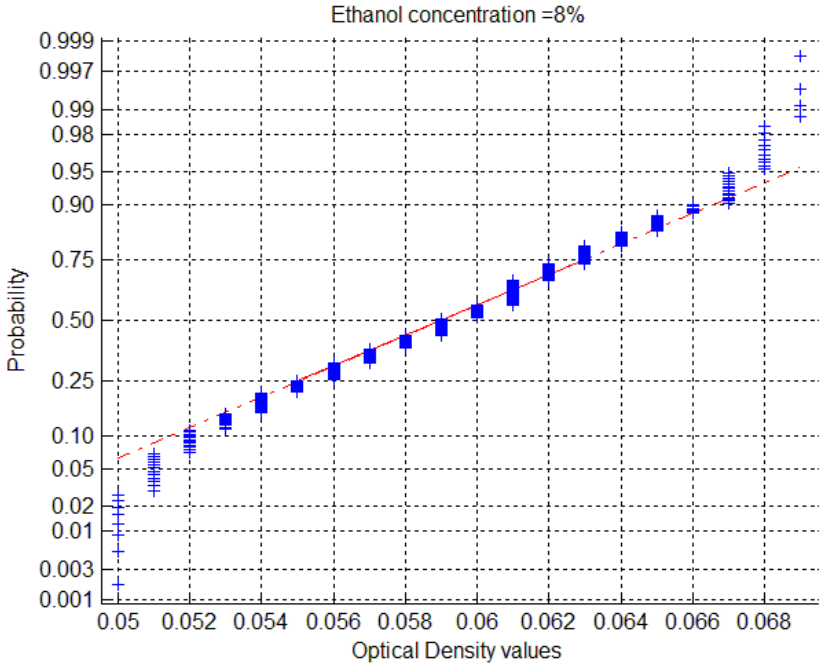
The curve fitting algorithm had been tried with several different model functions in order to investigate an optimal solution for the calibration function. For these tests the mean of each set of values corresponding to each ethanol concentration had been used. The Curve fitting toolbox from Matlab environment had been employed to test the fitting of each model. Basics of the algorithms were presented in the previous chapter.

In the followings the results for 2 model functions are presented: exponential and polynomial of 7th degree. The diagram in figure 4 shows the exponential fit and in figure 5 the polynomial fit together with their residual plots.

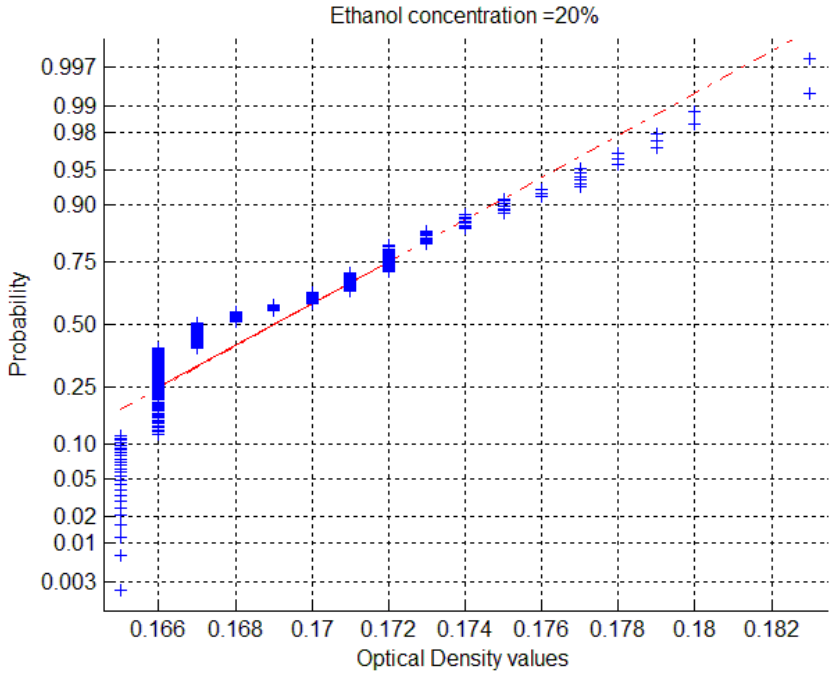


a. **b.**

Fig.2. Optical density histograms for: a. 8% ethanol; b. 20% ethanol.



a.



b.

Fig.3. Normal probability plots for: a. 8% ethanol; b. 20% ethanol.

Exponential

General model Exp2:

$f(x) = a \cdot e^{(b \cdot x)} + c \cdot e^{(d \cdot x)}$

Coefficients :

a = 9.525

b = 4.918

c = -6.817

d = -8.259

Goodness of fit:

SSE: 0.4137

R-square: 0.9987

Adjusted R-square: 0.9981

RMSE: 0.2626

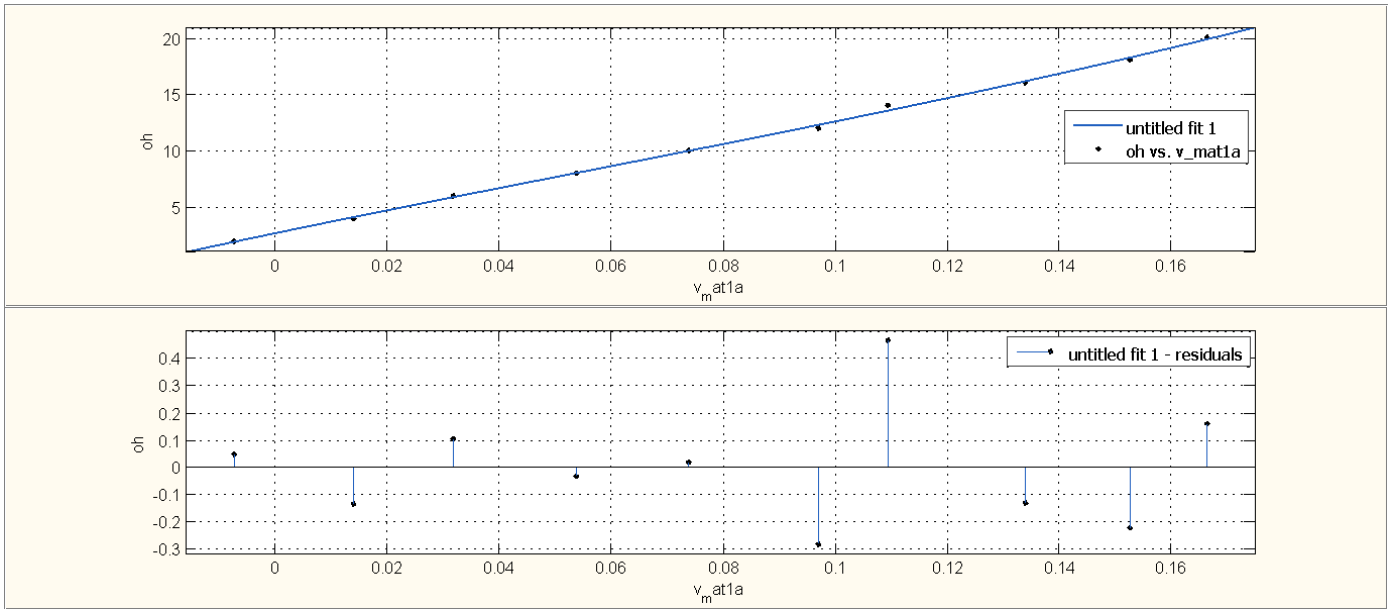


Fig.3. Exponential fit and residuals.

Polynomial

Linear model Poly7:

$f(x) = p_1 \cdot x^7 + p_2 \cdot x^6 + p_3 \cdot x^5 + p_4 \cdot x^4 + p_5 \cdot x^3 + p_6 \cdot x^2 + p_7 \cdot x + p_8$

Coefficients:

p1 = -4.658e+07

p2 = 3.853e+07

p3 = -1.144e+07

p4 = 1.578e+06

p5 = -1.039e+05

p6 = 2809

p7 = 87.55

p8 = 2.445

Goodness of fit:

SSE: 0.2395

R-square: 0.9993

Adjusted R-square: 0.9967

RMSE: 0.346

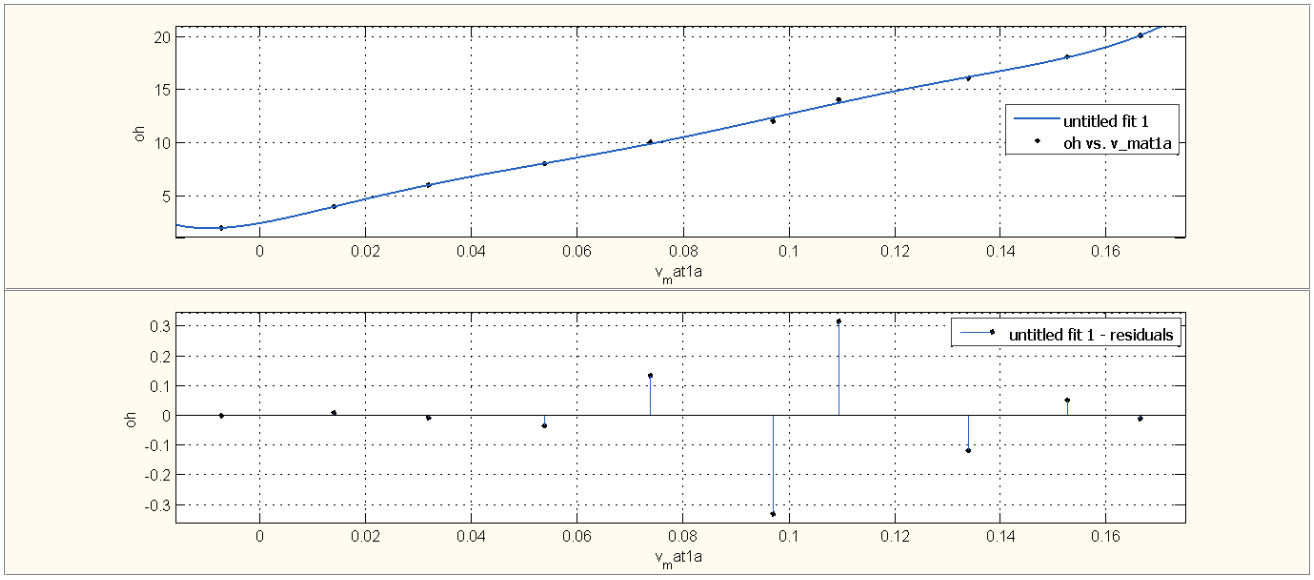


Fig.5. Polynomial fit and residuals.

5. Conclusions

From statistical analysis we could learn that more experiments have to be made in order to better analyze the process. From the results obtained until now we can say that data do not fit normal distribution which points us to the conclusion that there must be other factors in the process which influences the results. A good candidate for such a factor would be the temperature of the process which has to be measured together with the optical density.

In what regards the curve fitting tests it has been concluded that polynomial function of degree smaller than 7 have larger errors and larger than 7 degree polynomials have bad conditioned equations. We also tried to use power and rational functions as models but the algorithm do not converge in this cases.

Summarizing the results we can say that the temperature of the process has to be measured together with the optical density. In this case other algorithms must be tested for finding an optimal calibration function. Good candidate for this purpose are Artificial Neural Networks (ANN) which have multiple inputs. There is also of interest to study the relationship between other possible parameters of the process and the measured optical density. For the training of ANN larger amount of measurements have to be made.

6. Acknowledgements

This work has been funded as part of the project: “A new method and system for real time fermentation process monitoring”, HURO/1001/121/2.2.2, Hungarian-Romanian Cross-border Co-operation Programme 2007-2013.

7. References

1. ***Heath, M.T.,*** (2002), Scientific Computing: An Introductory Survey, 2nd ed., McGraw-Hill, New York.
2. ***Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D.,***

(1999), LAPACK Users' Guide, 3rd ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, ISBN = 0-89871-447-8

3. ***Frydenvang, J., Husted, S., Frydenvang, J., Kinch, K.M., Madsen, M.B.,*** (2013), An optimized calibration procedure for determining elemental ratios using laser-induced breakdown spectroscopy, Analytical Chemistry, Vol. 85, Iss. 3, pp. 1492-1500.
4. ***Prasad, D.K., Leung, Maylor K.H., Quek, C.,*** (2013) ElliFit: An unconstrained, non-iterative, least squares based geometric Ellipse Fitting method, Pattern Recognition, Vol. 46, Iss. 5, pp. 1449-1465.
5. **Su, Z.H., Zhang, Q.H., Bao, J.M., Gan, J., Yu, Q.K., Liu, Z.H.,** (2013) High-resolution fiber optic temperature sensors using nonlinear spectral curve fitting technique, Review of Scientific Instruments, Vol. 84, Iss. 4, pp. 045002.
6. ***Li, D., Zhu, R., Wu, P., Wu, J., Xu, K.,*** (2013), Measurement of glucose concentration by fiber-optic surface plasmon resonance sensor, Progress in Biomedical Optics and Imaging - Proceedings of SPIE, Volume 8576, pp. 85760X
7. ***Pien, H.P., Karl, W.C., Puff, D., Li, P., Cunningham, B.,*** (2012), Method and apparatus for biosensor spectral shift detection, Official Gazette of the United States Patent and Trademark Office Patents.
8. ***Nguy-R.A.L., Nguy-R.A.L., Tedesco, L., Li, L., Tedesco, L., Wilson, J., Soyeux, E.,*** (2012) Comparing the performance of empirical, semi-empirical, and curve fitting models in predicting cyanobacterial pigments, Photogrammetric Engineering and Remote Sensing, Vol. 78, Iss. 5, pp. 485-494.
9. ***Herman, P., Lee, J.C.,*** (2012), The advantage of global fitting of data involving complex linked reactions. Methods in molecular biology, Volume 796, pp. 399-421.