Prediction of Sweetness and Nitrogen Content in Soybean Crops from High Resolution Hyperspectral Imagery

Sildomar Takahashi Monteiro*, Yohei Minekawa*,Yukio Kosugi*, Tsuneya Akazawa† and Kunio Oda‡
*Department of Mechano-Micro Engineering, Tokyo Institute of Technology
4259 Nagatsuta, Midori-ku, Yokohama, 226-8502, Japan
Email: monteiro@pms.titech.ac.jp
†University Farm, Faculty of Agriculture, Yamagata University
5-3 Takasaka, Tsuruoka-shi, Yamagata, 997-0369, Japan
‡Yamagata General Agricultural Research Center
25, Yamanomae, Fujishima, Tsuruoka-shi, Yamagata, 999-7601, Japan

***Abstract*— In this paper, we investigate a hyperspectral imagery data processing method to predict the sweetness and amino acid content of green vegetal soybean crops. Regression models based on neural networks were developed in order to calculate the level of sucrose, glucose, and nitrogen concentration, which can be related to sweetness and amino acid concentration of vegetables. We demonstrate the method using hyperspectral data of wavelengths ranging from the visible to the near infrared acquired from an experimental field of green vegetal soybeans. A performance analysis is reported comparing regression models built using datasets pre-processed using the first and second derivative analysis. The second derivative transformed dataset presented the best performance overall. Glucose could be predicted with greater accuracy.**

I. Introduction

Green vegetable soybeans (edamame, in Japanese) are a popular appetizer or side-dish cultivated throughout East Asia and Japan, where they are harvested while in the green stage, and cooked and served in the pods [1]. The taste of vegetables can be related to their sweetness and amino acid contents [2]. Sweetness varies according to, among other factors, sucrose and glucose concentration. Amino acids taste sweet or delicious (umami) to humans, and its content can be estimated by measuring nitrogen concentration. The non-destructive and immediate prediction of sweetness and amino acid content would provide useful information for farmers and producers, facilitating crop selection and the determination of the proper harvest time.

Spectroscopy of the near infrared region of the spectrum has been used for performing analytical measurements in a variety of applications, providing both qualitative and quantitative results on solids and liquids [3]. The determination of the flavor quality of vegetables has been studied using near infrared spectroscopy [4]. Conventional methods usually rely on fiber optic spectrometers and utilize few wavelengths to perform calculations of chemical contents in vegetables [5]. Calibration methods commonly utilize multivariate mathematics and result in a calibration equation, which is a linear combination of spectral data [6]. Furthermore, sugar content measurement methods that have been developed need to be performed destructively on a fruit-by-fruit basis [7]. Considering those limitations, conventional methods are not directly suitable for sweetness and nitrogen prediction in leguminous crops, such as green vegetable soybeans.

In the remote sensing field, sensors capable of acquiring hundreds of narrow contiguous bands of information for spectral signature analysis are referred to as hyperspectral sensors [8]. Recent advances in sensor and lens technology for hyperspectral imaging have improved their spatial and spectral resolution while reducing the size and cost of the equipment, adding up to the conventional spectroscopy the advantage of simultaneous and accurate data acquisition of a wide sample area. Hyperspectral sensors have been applied mainly for land cover classification [9]. Nevertheless, applications for material estimation are still challenging due to the low resolution of conventional airborne hyperspectral data acquisition devices which limit the amount of sample data [10].

In this paper, we investigate an artificial neural network (ANN) based approach to model high-resolution hyperspectral imagery data in order to non-destructively predict the sweetness and nitrogen content in soybean crops. We conducted an experimental performance analysis comparing regression models obtained using raw reflectance, first, and second derivatives of reflectance spectra.

II. Materials and Methods

The problem in which one tries to predict a dependent variable (sucrose, glucose and nitrogen) by combining a number of independent variables (reflectance spectral bands) can be defined in statistics as a regression problem [11]. In spectroscopy, this problem can also be referred to as a calibration problem. The calibration involves searching for predictive relationships between spectral data and reference values, using either laboratory standards, field standards, or modeling [12]. Regression models are main tools used in

0-7803-9510-7/06/$20.00 (c) 2006 IEEE

spectrum calibration. ANNs have been widely applied as regression models [13].

A data analysis scheme as shown in Fig. 1 is proposed. The hyperspectral imagery data is initially preprocessed and prepared to serve as input data for the regression analysis. ANN regression models were obtained using input datasets processed with different methods of hyperspectral data transformation, providing a comparative performance analysis. The algorithms assessed were the first and second derivative analysis.

Ground truth data was collected to provide the actual concentration of interesting chemical substances in labeled regions of the crop field, each one corresponding to a different variety of soybean. For each labeled region, the concentration values of sucrose, glucose, and nitrogen were measured in the laboratory using liquid chromatography of freeze-dried soybean samples. The actual contents measurements were used as target data for the supervised training of ANN regression models.

A. Hyperspectral Imagery Data

We utilized a hyperspectral sensor coupled with a CCD camera and a computer controller mounted on the tip of a crane in order to allow data acquisition at a suitable spatial resolution [14]. The crane based system presents several advantages for agricultural data analysis of localized crop fields compared to satellite or airplane based systems, e.g., higher data accuracy and spatial resolution, and reduced atmospheric effects. The hyperspectral data comprises the visible to the near-infrared range of the spectrum (400nm-1000nm) with 5nm of resolution, producing 121 bands as output [15].

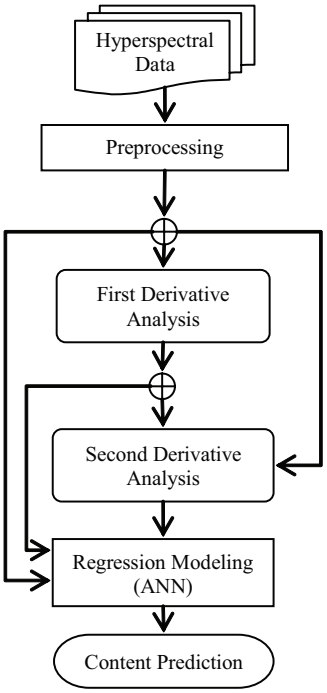


Fig. 1. Diagram for the ANN-based Hyperspectral data processing

The hyperspectral imaging device captures the radiance from the sample. The raw radiance data needs to be converted to reflectance by using a standard reference white board. Another auxiliary data is the dark current, measured by keeping the camera shutter closed at the moment of the data acquisition. The conversion observes the following equation

$$I_{ref} = \frac{I_{raw} - I_{dark}(\lambda)}{I_{white}(\lambda) - I_{dark}(\lambda)}, \quad (1)$$

where I_{ref} is the calculated reflectance value; I_{raw} is the raw data radiance value of a given pixel; and I_{dark} and I_{white} are, respectively, the dark current and the white board radiance for each spectral band λ.

In order to reduce or attenuate not only the spatial but the spectral noise of the dataset, a special averaging filter cube was designed. The filter is applied on the hyperspectral dataset as a 3D cubic window, in which each pixel is assigned the mean value of its surrounded pixels, including neighboring spectral bands. This filter can be formulated as

$$y[n_1, n_2, n_3] = \frac{1}{D^3} \sum_{i=0}^{D-1} \sum_{j=0}^{D-1} \sum_{\lambda=0}^{D-1} I[n_1 - i, n_2 - j, n_3 - \lambda], \quad (2)$$

where $[n_1, n_2, n_3]$ refer to the current window position at the hyperspectral dataset, corresponding to $[row, column, band]$; D is the window size; and I is the raw pixel value.

After the reflectance correction, the areas in the hyperspectral images of the crop field corresponding to the labeled regions of different varieties of green vegetable soybeans were then identified and separated manually.Image regions containing vegetation were then identified using the normalized difference vegetation index (NDVI), which is alleged less affected by variations on the absolute value of the raw dataset [16]. The NDVI index is calculated for each pixel of the spatial plane of the hyperspectral dataset, and the vegetation corresponds to regions that present NDVI value greater than a specified threshold. The NDVI equation is defined as

$$\text{NDVI} = \frac{NIR - R}{NIR + R}, \quad (3)$$

where NIR and R are selected bands from the near infrared $\text{NIR} = I(\lambda \cong 830nm)$ and red region $R = I(\lambda \cong 650nm)$ of the spectra, respectively.

Additionally, the vegetation region of the hyperspectral dataset was scanned for the presence of extreme or discrepant NDVI values (outliers), calculated according to Chebyshev theorem. Mathematically, outliers are regions presenting value outside the interval $[\bar{x} - 3s, \bar{x} + 3s]$ where $\bar{x}$ is the mean, and s is the standard deviation.

B. Derivative Analysis

Derivative analysis of reflectance spectra has been used in hyperspectral remote sensing and in analytical chemistry to increase the estimation accuracy of target information [17]. Derivatives are relatively less susceptible to variations

in illumination intensity in the remote sensing field [18]. Nevertheless, a secondary effect of the derivative process is that it accentuates the noise present in the data, thus the necessity of a careful preprocessing phase to reduce the noise.

The derivative of hyperspectral data can be calculated by finite approximation using suitable difference schemes. Given a finite band separation $\Delta\lambda$, the first derivative at wavelength λ_v can be estimated as

$$\left.\frac{dI}{d\lambda}\right|_v \approx \frac{I(\lambda_u) - I(\lambda_v)}{\Delta\lambda}, \tag{4}$$

where $\Delta\lambda$ is a constant gap, constrained by $\Delta\lambda = \lambda_v - \lambda_u$ and $\lambda_v > \lambda_u$.

The second derivative can be calculated from the first derivative as follows

$$\left.\frac{d^2I}{d\lambda^2}\right|_v = \frac{d}{d\lambda}\left(\frac{dI}{d\lambda}\right)\Bigg|_v \approx \frac{I(\lambda_u) - 2I(\lambda_v) + I(\lambda_w)}{(\Delta\lambda)^2}, \tag{5}$$

where $\Delta\lambda$ is the same constant gap, and the constraints for the calculation are $\Delta\lambda = \lambda_w - \lambda_v = \lambda_v - \lambda_u$ and $\lambda_w > \lambda_v > \lambda_u$. The derivation interval was determined experimentally to stand around 20nm.

C. Artificial Neural Networks

One of the most common kinds of ANN is the multilayer perceptron (MLP) network, which provides a nonlinear model that can, in principle, represent almost any function. The basic MLP network architecture is composed of the input layer, one hidden layer, and the output layer, sequentially interconnected in a feed-forward way [19]. The outputs of one layer are fed forward to the next layer through the network. The output of the MLP can be expressed mathematically as

$$y = f(x) = B\varphi(Ax + a) + b, \tag{6}$$

where x is the input vector; y is the output vector; A and a are, respectively, the weight matrix and the bias vector of the hidden layer; B and b are, respectively, the weight matrix and the bias vector of the output layer; and φ is the activation function.

A network architecture composed of 10 neurons in the hidden layer is proposed, using the hyperbolic tangent sigmoid as transfer function. The output layer is composed by only one neuron, using a linear transfer function. The number of neurons in the input layer is proportional to the number of spectral bands, thus 121.

The training method implemented was the Levenberg-Marquardt backpropagation, which experimentally has superior performance for regression problems [20]. Early stopping is used to improve generalization. The training set is split into a new training set, a validation set and a test set. The network is evaluated on the validation set periodically during training. Training is stopped when the validation error rate starts to grow. The test set is not used during training. It serves only to compare the models. For example, if the error of the test set has a configuration different than that of the validation set through a number of iterations, this may indicate a poor division of the datasets.

D. Performance Evaluation Metrics

The performance of the regression models built using ANNs was evaluated using two metrics [21].

1) Mean of Squared Errors (MSE)*:* One measure of network performance is the MSE between corresponding elements of the network response (prediction) and the target (ground truth measurements). The MSE is defined by

$$\mathrm{MSE}(y) = \frac{1}{N}\sum_{i=1}^{N}(y^p - y^m)^2, \tag{7}$$

where N is the total number of prediction comparisons, and p and m indicate, respectively, the predicted and measured values.

2) Correlation Coefficient (R-value)*:* (2) Another way of evaluating network performance is to calculate R-value between the network response and the target. As the strength of the relationship between the predicted values and measured values increases, so does the correlation coefficient value. A perfect prediction would give a coefficient of 1. The R-value is calculated by

$$\mathrm{R}(y) = \frac{C(y^p, y^m)}{\sqrt{C(y^p, y^p).C(y^m, y^m)}}, \tag{8}$$

where, C is the covariance matrix.

III. Results

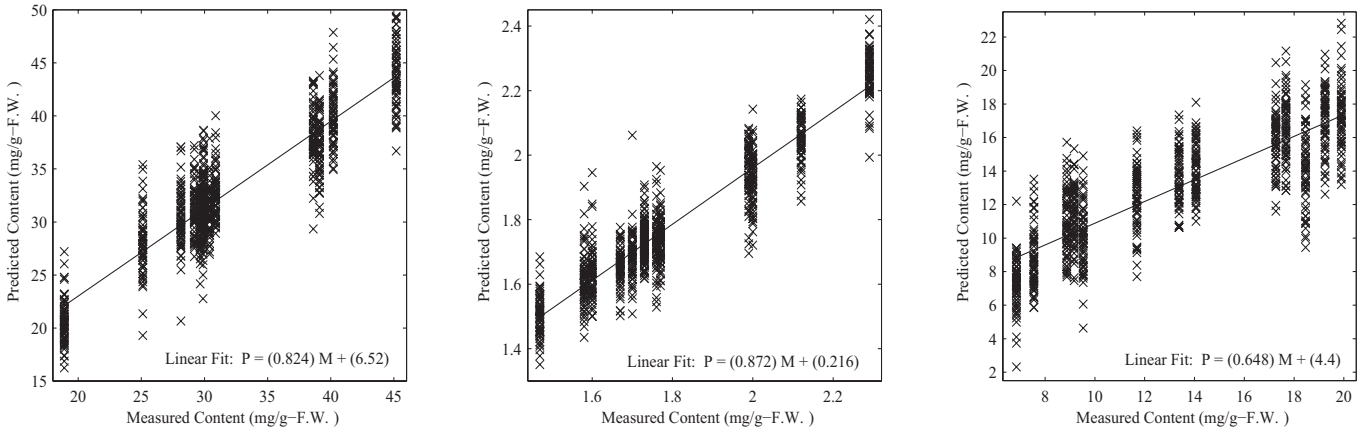
The hyperspectral imagery dataset sample consisted of diverse varieties of green vegetable soybeans of an experimental crop field at Yamagata University, Japan. The hyperspectral data was acquired in the middle of the summer on a sunny day, from noon to early afternoon. A total of 13 different varieties of green vegetable soybeans were analyzed.

To account for the stochastic nature of the MLP network training, each regression model was tested over 50 independent runs, each time starting the weights of the network with random values generated by a different random seed. Only the best networks were kept, i.e., the networks that present lower MSE value and higher R value. Table I summarizes the results of the best regression models for each chemical substance based on each different dataset: the whole of raw reflectance wavelength bands, the first derivative transformed data and the second derivative transformed data.

A linear regression analysis was performed between the network predicted output and the ground truth measurements for the best regression model of each chemical substance. The results are shown in Fig. 2 for the second derivative transformed datasets of sucrose, glucose, and nitrogen.

IV. Conclusion

The ANN-base approach provided a reasonably accurate regression model for the prediction of sucrose, glucose and nitrogen content using hyperspectral imagery data. In the range of wavelengths investigated, greater accuracy was obtained for the calculation of glucose, followed by sucrose and nitrogen. The derivatives contributed to a slight and gradual improvement on the ANN model prediction accuracy. Therefore, the



(a) Sucrose (b) Glucose (c) Nitrogen

Fig. 2. Regression analysis between the ANN prediction and the measured ground truth. The cross markers indicate the corresponding observations, and the best linear fit of the data given by the network model is shown by the solid line.

TABLE I
ANN REGRESSION MODEL RESULTS

Chemical	Dataset	MSE	R
Sucrose	Whole	0.075	0.849
	1^{st} Deriv.	0.069	0.866
	2^{nd} Deriv.	0.055	0.901
Glucose	Whole	0.062	0.893
	1^{st} Deriv.	0.045	0.924
	2^{nd} Deriv.	0.039	0.934
Nitrogen	Whole	0.224	0.736
	1^{st} Deriv.	0.179	0.799
	2^{nd} Deriv.	0.158	0.825

use of the second derivative transformed dataset is rather recommended instead of the raw dataset. The ability of the ANN-based regression model to provide an accurate prediction to other seasonal conditions needs to be verified in further experiments.

The proposed method permits to exploit the capabilities of high-resolution hyperspectral imagery data for estimating the sweetness and amino acid concentration of leguminous crops before harvesting and without interfere in the growth process.

REFERENCES

[1] K. Liu, *Soybeans: Chemistry, Technology, and Utilization*. New York: Chapman and Hall, 1997.

[2] K. Toko, "A taste sensor," *Meas. Sci. Technol*, vol. 9, pp. 1919–1936, 1998.

[3] F. W. Koehler IV, E. Lee, L. H. Kidder, and E. N. Lewis, "Near infrared spectroscopy: the practical chemical imaging solution," *Spectroscopy Europe*, vol. 14, no. 3, pp. 12–19, 2002.

[4] D. C. Slaughter, D. Barrett, and M. Boersig, "Nondestructive determination of soluble solids in tomatoes using near infrared spectroscopy," *J. Food Science*, vol. 61, no. 4, pp. 695–697, 1996.

[5] M. Zude, "Comparison of indices and multivariate models to non-destructively predict the fruit chlorophyll by means of visible spectrometry in apples," *Analytica Chimica Acta*, vol. 481, pp. 119–126, 2003.

[6] B. G. Osborne and T. Fearn, "Near infrared spectroscopy in food analysis," in *Encyclopedia of Analytical Chemistry*. New York: John Wiley & Sons, 1986, pp. 86–103.

[7] M. Tsuta, J. Sugiyama, and Y. Sagara, "Near-infrared imaging spectroscopy based on sugar absorption band for melons," *J. Agric. Food Chem.*, vol. 50, no. 1, pp. 48–52, 2002.

[8] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis, An Introduction*, 3rd ed. New York, USA: Springer-Verlag, 1999.

[9] J. Cihlar, "Land cover mapping of large areas from satellites: status and research priorities," *Intl. Journal Remote Sensing*, vol. 21, no. 6-7, pp. 1093–1114, 2000.

[10] R. A. Schowengerdt, *Remote Sensing: Models and Methods for Image Processing*. San Diego, USA: Academic, 1997.

[11] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego, USA: Academic Press, 1990.

[12] H. W. Siesler, Y. Ozaki, S. Kawata, and H. M. Heise, *Near-Infrared Spectroscopy Principles, Instruments, Applications*. Weinheim, Germany: Wiley-VCH, 2002.

[13] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, USA: Oxford University Press, 1995.

[14] Y. Minekawa, K. Uto, Y. Kosugi, and K. Oda, "Development of crane-mounted hyperspectral imagery system for stable analysis of paddy field," in *Proc. Intl. Symp. on Remote Sens.*, Jeju ,Korea, 2004.

[15] Spectral Imaging Ltd., *ImSpector Imaging Spectrograph User Manual*, 2nd ed., 2003. [Online]. Available: http://www.specim.fi/

[16] P. M. Mather, *Computer Processing of Remotely-Sensed Images, An Introduction*. Chichester: John Wiley & Sons, 2004.

[17] C. Petisco, *et al.*, "Use of near-infrared reflectance spectroscopy in predicting nitrogen, phosphorus and calcium contents in heterogeneous woody plant species," *Analytical and Bioanalytical Chemistry*, vol. 382, no. 2, pp. 458–465, 2005.

[18] F. Tsai and W. Philpot, "Derivative analysis of hyperspectral data," *Remote Sens. Environ.*, vol. 66, pp. 41–51, 1998.

[19] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs: Prentice-Hall, 1999.

[20] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE Trans. Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[21] A. H. Murphy and H. Daan, "Probability, statistics, and decision making in the atmospheric sciences," in *Forecast Evaluation*, A. Murphy and R. Katz, Eds. Boulder: Westview Press, 1985, pp. 379–437.