Analytica Chimica Acta 962 (2017) 15–23

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Predicting the ethanol potential of wheat straw using near-infrared spectroscopy and chemometrics: The challenge of inherently intercorrelated response functions

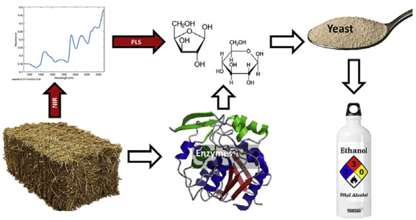
Åsmund Rinnan [a], [*], Sander Bruun [b], Jane Lindedam [b], Stephen R. Decker [c], Geoffrey B. Turner [c], [1], Claus Felby [d], Søren Balling Engelsen [a]

[a] Chemometrics and Analytical Technology, Department of Food Science, Faculty of Science, University of Copenhagen, Denmark
[b] Plant and Soil Science Laboratory, Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Denmark
[c] National Renewable Energy Laboratory, Biosciences Center, 15013 Denver West Parkway, Golden, CO 80401, USA
[d] Biomass Science and Technology, Department of Geosciences and Natural Resource Management, Faculty of Science, University of Copenhagen, Denmark

HIGHLIGHTS

- Intercorrelated response variables influence the regression models.
- Making a calibration model from NIR on straw to enzymatic released sugars is not straight forward.
- Reducing the inherent correlation makes it possible to understand the underlying dependencies in the regression models.
- Models have been validated, both through chemometric tools, as well as in detail understanding of the system.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 23 November 2016
Received in revised form
31 January 2017
Accepted 2 February 2017
Available online 7 February 2017

Keywords:
NIR
Calibration
Correlated response variables
Straw
Enzymatic sugar release

ABSTRACT

The combination of NIR spectroscopy and chemometrics is a powerful correlation method for predicting the chemical constituents in biological matrices, such as the glucose and xylose content of straw. However, difficulties arise when it comes to predicting enzymatic glucose and xylose release potential, which is matrix dependent. Further complications are caused by xylose and glucose release potential being highly intercorrelated.

This study emphasizes the importance of understanding the causal relationship between the model and the constituent of interest. It investigates the possibility of using near-infrared spectroscopy to evaluate the ethanol potential of wheat straw by analyzing more than 1000 samples from different wheat varieties and growth conditions. During the calibration model development, the prime emphasis was to investigate the correlation structure between the two major quality traits for saccharification of wheat straw: glucose and xylose release. The large sample set enabled a versatile and robust calibration model to be developed, showing that the prediction model for xylose release is based on a causal relationship with the NIR spectral data. In contrast, the prediction of glucose release was found to be

---

Abbreviations: BESC, BioEnergy Science Center; CAL, Calibration dataset; conv, Converged; CV, Cross-validated; def, Deformation; HTPH, High-throughput pretreatment and enzymatic hydrolysis system; NIR, Near Infrared; NREL, National Renewable Energy Laboratory; opt, Optimal; OT, Overtone; P, Predicted; PCA, Principal Component Analysis; PLS, Partial Least Squares; r, Correlation coefficient; RFR, Reduced Full Range; RGR, Reduced Glucose release Range; RMSE, Root-Mean-Squared-Error; RXR, Reduced Xylose release Range; SG, Savitzky-Golay; str, Stretch; VAL, Validation dataset.

[*] Corresponding author.
E-mail address: aar@food.ku.dk (Å. Rinnan).

[1] Current address: Waters Corporation, 34 Maple Street, Milford, MA 01757, USA.

http://dx.doi.org/10.1016/j.aca.2017.02.001

highly dependent on the intercorrelation with xylose release. If this correlation is broken, the model performance breaks down. A simple method was devised for avoiding this breakdown and can be applied to any large dataset for investigating the causality or lack of causality of a prediction model.

1. Introduction

To increase biofuel production without expanding the agricultural area, the efficient use of lignocellulosic biomass is imperative. In recent years, major efforts have been directed at process optimization of lignocellulosic biomass conversion to ethanol in order to develop second-generation bioethanol to a commercial level [1,2].

Wheat is a major food crop with a worldwide harvest of about 713 Tg in 2013 [3], and wheat straw is one of the largest potential sources of lignocellulosic biomass. A large amount of wheat straw is used for different purposes, including animal feed and bedding for domestic animals and fuel for heating, but the major part is either burned in the field or incorporated into the soil. Thus, unused wheat straw represents a major potential for biofuel production.

The degradability of the feedstock logically affects the potential ethanol yield of the process [4]. This variation in yield may be caused by the specific crop variety and environmental changes reflected in the growing site and year [5]. The variation caused by the variety can be used to breed new varieties with higher ethanol potential. To provide data for resolving such issues, high-throughput quality assessment systems including pretreatment and hydrolysis, such as those compared by Lindedam et al. [6], have been established.

Rapid high-throughput analytical methods increase the number of samples that can be analyzed. In addition, such methods can be applied for more practical purposes, such as to determine the price of the feedstock or in process control for the plant. Diffusive reflectance near-infrared (NIR) spectroscopy offers a number of advantages here. The method is non-destructive, highly reproducible, rapid and requires minimal sample preparation. Several studies have correlated NIR spectra with the chemical composition of feedstock used for conversion to ethanol [7–9].

A few feasibility studies have shown that NIR spectroscopy can be used to assess the potential for bioethanol production of switch grass [8] and for sugar release upon enzyme treatment of wheat straw [4,10]. The calibrations in these studies were based on sample sets in the range of 85–212 samples. However, none included a predicted independent test set (*i.e.* only cross-validated figures of merit). Calibrations based on a larger number of samples and with different geographic and variety origins are therefore in great demand to validate the findings more thoroughly. Another common phenomenon which needs to be addressed, is that the amounts of released sugars are highly correlated, *i.e.* the varieties that release more xylose upon enzymatic hydrolysis are also the ones that release more glucose, and *vice versa* [11]. This implies that prediction of glucose release could be based on features in the spectra that are associated with the signals from xylan. Such deviations from causality in the model can have detrimental effects on the prediction of future samples. If the prediction model is based on indirect correlations, the ability to predict new samples correctly will depend on the correlation structure in the quality parameters (response variables) for new samples. Thus, the problem of intercorrelated quality parameters may make calibration models unreliable (or even incorrect) when encountering new sample sets with a different correlation structure between xylose and glucose.

The purpose of the current study was to develop a robust calibration model capable of predicting the enzymatic release of glucose and xylose, with a special emphasis on the development of calibrations that were independent of correlations with the release of the other sugar. Previously, similar investigations have been done through the use of orthogonal pre-processing [12,13]. However, using orthogonal pre-processing will also remove the interesting part of the signal that is correlated to the unwanted parameter. Thus, these approaches do not answer our question. There are a multitude of sample selection methods, but they all focus on careful selection of samples based on how they span the **X**-space, and not on the correlation between references. Instead, our approach was to subdivide the dataset in order to break the correlation structure and by thorough validation, using traditional chemometric methods and by inspecting the wavelengths that are important for the prediction together with their associated molecular vibrations.

2. Materials and methods

2.1. Collection and preparation of wheat straw samples

A total of 1124 wheat straw samples were collected from nine different locations in Denmark and one location in the United Kingdom (Table 1). The samples were collected from ongoing experiments with different wheat varieties, fertilizer treatments and harvesting times. The experiments include a total of 209 different wheat varieties. An overview of the different experiments, sites and treatments are given in Table 2.

From all but one experiment in Denmark, fully senescenced air-dried straw (approx. 7% moisture) was sampled from the experimental plots after the grain had been harvested by a combine harvester, leaving the straw in the field. Approximately 80 g of straw was collected representatively from each experimental plot or Heston bales, as described by Lindedam et al. [4,14], and stored at ambient temperature. Material from the experiment with different harvest times was collected by hand three weeks before maturity, at maturity and three weeks afterwards. The plants were cut 5–7 cm from the soil and the grain was removed from the samples, before being stored at ambient temperature. Material from the UK was collected as described by Murozuka et al. [15]. All straw samples were subsequently ground in a cyclone mill (President, Holbæk, Denmark) mounted with a 1 mm screen.

Table 1
Locations from where wheat straw samples were collected for the experiment.

Name	Country	Coordinates
Abed	Denmark	54° 49′ 40″ N, 11°19′ 31″ E
Sejet	Denmark	55° 49′ 12″ N, 9° 55′ 22″ E
Holstebro	Denmark	56° 24′ 5″ N, 8° 38′ 22″ E
Tystofte	Denmark	55° 15′ 9″ N, 11° 20′ 14 E
Taastrup	Denmark	55° 40′ 36″ N, 12° 18′ 10″ E
Rothamsted	United Kingdom	51° 48′ 24″ N, 0° 21′ 49″ W
Søtoften	Denmark	56° 14′ 49″ N, 10° 6′ 1″ E
Fyn	Denmark	55° 18′ 29″ N, 10° 22′ 36″ E
Hornsherred	Denmark	55° 47′ 06″ N, 11° 57′ 56″ E