Tillman *et al.*
*Biotechnology for Biofuels and Bioproducts* (2024) 17:112
https://doi.org/10.1186/s13068-024-02558-6

Biotechnology for Biofuels and Bioproducts

RESEARCH

Open Access

# Rapid measurement of soluble xylo-oligomers using near-infrared spectroscopy (NIRS) and multivariate statistics: calibration model development and practical approaches to model optimization

Zofia Tillman[1], Kevin Gray[1] and Edward Wolfrum[1*]

## Abstract

**Background** Rapid monitoring of biomass conversion processes using techniques such as near-infrared (NIR) spectroscopy can be substantially quicker and less labor-, resource-, and energy-intensive than conventional measurement techniques such as gas or liquid chromatography (GC or LC) due to the lack of solvents and preparation methods, as well as removing the need to transfer samples to an external lab for analytical evaluation. The purpose of this study was to determine the feasibility of rapid monitoring of a biomass conversion process using NIR spectroscopy combined with multivariate statistical modeling, and to examine the impact of (1) subsetting the samples in the original dataset by process location and (2) reducing the spectral range used in the calibration model on model performance.

**Results** We develop multivariate calibration models for the concentrations of soluble xylo-oligosaccharides (XOS), monomeric xylose, and total solids at multiple points in a biomass conversion process which produces and then purifies XOS compounds from sugar cane bagasse. A single model using samples from multiple locations in the process stream showed acceptable performance as measured by standard statistical measures. However, compared to the single model, we show that separate models built by segregating the calibration samples according to process location show improved performance. We also show that combining an understanding of the sample spectra with simple multivariate analysis tools can result in a calibration model with a substantially smaller spectral range that provides essentially equal performance to the full-range model.

**Conclusions** We demonstrate that real-time monitoring of soluble xylo-oligosaccharides (XOS), monomeric xylose, and total solids concentration at multiple points in a process stream using NIR spectroscopy coupled with multivariate statistics is feasible. Segregation of sample populations by process location improves model performance. Models using a reduced spectral range containing the most relevant spectral signatures show very similar performance to the full-range model, reinforcing the importance of performing robust exploratory data analysis before beginning multivariate modeling.

**Keywords** Near-infrared spectroscopy, Xylo-oligomers, Process monitoring, At-line monitoring, Bioenergy, Bioproducts, Multivariate statistics

*Correspondence:
Edward Wolfrum
ed.wolfrum@nrel.gov
Full list of author information is available at the end of the article

## Background

In-line or at-line monitoring of chemical processes using spectroscopic methods has been demonstrated to reliably provide real-time measurements in a wide variety of industrial processes ranging from waste management [1] to pharmaceutical manufacturing [2–6]. In this work, we use near-infrared (NIR) spectroscopy combined with multivariate statistics to demonstrate the feasibility of measuring the concentrations of xylose and soluble xylo-oligosaccharides (XOS) in an aqueous process stream. NIR has been used for multiple applications in the past, including biomass energy conversion [7], food processing [8–10], and fermentation monitoring [11, 12]. The experimental data presented in this work are derived from a process which produces and then purifies xylo-oligosaccharides (XOS) from sugar cane. There has been substantial interest in biomass-derived XOS materials because of their potential application as nutritional supplements [13–16].

A key element of rapid spectroscopic monitoring is the use of multivariate statistics applied to the spectroscopic data, a field typically referred to as chemometrics [17], to develop robust calibration models. The key steps in the development of these calibration models are spectral collection, spectral preprocessing, model development using preprocessed spectra and primary analytical chemistry of a suitable calibration population, and then testing or validation of the model [18]. There are a myriad of approaches to spectral preprocessing [19–22], including the down-selection or subsetting of the spectra used in the model [23–26]. The overall goal of variable selection techniques is to identify the key variables (for spectroscopy data the key wavelengths or spectral regions) that result in multivariate models that provide equivalent or even superior performance to a model using all available spectral variables, since multivariate models with fewer variables are computationally more efficient and may be easier for the user to interpret. In this work, we focus on a practical, data-informed approach to variable selection that combines information regarding the known chemistry of the samples with simple multivariate spectral analysis techniques to quickly identify the regions in the NIR spectra of the calibration samples that contribute the majority of the variance.

In this work, we test the feasibility of at-line measurement of xylose monomers and soluble xylo-oligosaccharides (XOS) when present simultaneously in a biomass conversion process using NIR spectroscopy and multivariate statistics. We build on previous work by demonstrating (1) the feasibility of NIR spectroscopy to measure the concentration of monomeric xylose and soluble xylo-oligosaccharides when both are present simultaneously in process streams; (2) the utility of creating subpopulations within a given population to improve the performance of multivariate calibrations; and (3) the utility of simple, data-informed approaches to identify reduced spectral ranges to effectively measure the concentrations of monomeric xylose and XOS.

## Methods

### Process description

The overall process investigated in this work was the production and purification of xylo-oligosaccharides (XOS) from a low-sugar variety of sugar cane grown in the Imperial Valley of California [16, 27]. Fresh sugar cane was harvested, shredded, and transported to a processing facility. The shredded cane was washed with moderate temperature (60–80 C) water to remove residual soil as well as extractives such as sucrose and other non-structural sugars, and then pressed to remove excess water. The cane then underwent hydrothermal treatment at temperatures of at least 160 C for approximately 2 h. This mild hydrothermal treatment resulted in the solubilization of the hemicellulose fraction of the cane producing a liquor stream containing crude xylo-oligosaccharides. Reactor conditions were chosen to favor oligomeric rather than monomer sugar formation. The initial reaction volume (directly out of the high temperature reactor) varied, but was typically several thousand gallons.

A conceptual process flow diagram of the separation process is shown in Fig. 1. Small, insoluble material ("fines") were removed from the hydrolysate using a continuous microfiltration (MF) process. The permeate from the MF unit operation was passed over a proprietary chromatography column (CHROM) to remove color bodies and other impurities. Subsequent purification steps were designed to target a specific molecular weight range of oligosaccharides. The first filtration step (F1) removed low molecular weight impurities such as sugar monomers and organic acids. The retentate fraction from F1 then passed through a second filtration step (F2) to remove high molecular weight impurities such as very long sugar oligomers and other polymers. The F2 permeate was then diafiltered and dewatered (F3) to approximately 20–25% total solids and then dried to a powder using standard industrial drying technologies. As discussed later, we found it useful to group these sampling locations into two different categories: early/waste stream samples and late stream samples (Fig. 1). The process streams prior to F1 were classified as early, and three of the four waste streams (from MF, F1, and F3) were classified as waste streams. The process streams downstream of F1 were classified as late streams, as was the F2 retentate waste stream, which contains high-molecular weight material.

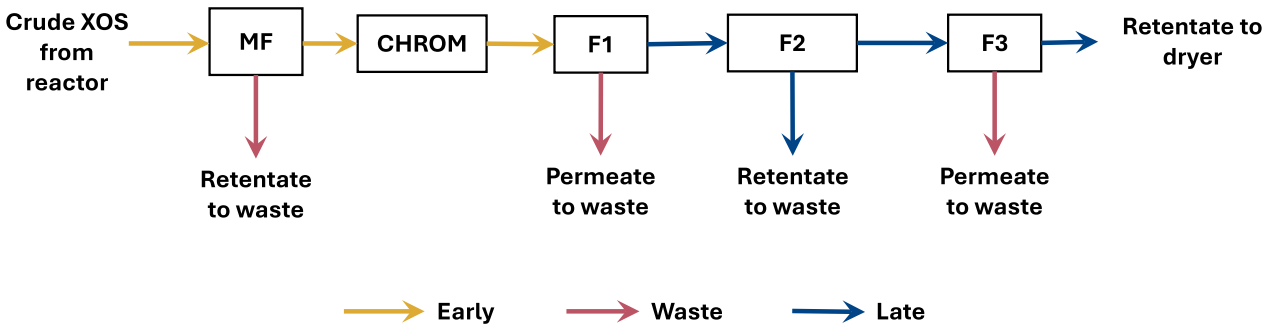


**Fig. 1** Conceptual process flow diagram of overall process showing sampling locations. Crude soluble xylo-oligosaccharides (XOS) from the reactor are first passed through a continuous microfiltration process (MF) to remove insoluble material. A series of proprietary chromatography columns are then used to remove color bodies and other impurities (CHROM). Low molecular weight impurities (organic acids and sugar monomers) are then removed via nanofiltration (F1). Next, high molecular weight impurities (very long oligomers) are removed by a second nanofiltration step (F2). Finally, diafiltration and dewatering is used to remove any final impurities and increase the solids concentration (F3)

## Sample analysis

### *Sample collection/storage method*

During normal operation of the facility, samples were taken from a total of 10 unique process locations as indicated by each of the arrows in Fig. 1. During a 4-month period of production a total of 123 samples from the 10 process locations were selected for near-infrared (NIR) spectral analysis. Approximately 20ml of each sample was frozen immediately after sampling, and groups of samples were periodically sent to NREL for NIR spectral collection. Samples were thawed at room temperature and analyzed via NIR spectroscopy in groups of approximately 20 samples.

### *Laboratory analysis*

Process samples were analyzed the next workday after collection in an on-site analytical laboratory using NREL Laboratory Analytical Procedures for total and monomeric sugar content and total solids concentration [28]. Briefly, samples were analyzed before and after analytical hydrolysis with dilute sulfuric acid using high performance liquid chromatography (HPLC, Agilent 1200 series with RID detector, an Aminex HPX-87P HPLC column, and a Bio-Rad Micro-Guard de-ashing cartridge). All HPLC calibration standards were provided by Absolute Standards Inc. Soluble xylo-oligosaccharide (XOS) concentration was defined as the difference between the soluble xylose concentration measured after analytical hydrolysis and the soluble xylose concentration prior to analytical hydrolysis. The HPLC method has a working concentration range of 0.5–36.0g/L. Process samples were diluted to ensure they were within the linear range of the HPLC method. The total solids (TS) concentration (g/L) was determined gravimetrically. The purity of XOS was defined as the XOS concentration (g/L) divided by the total solids (TS, g/L) concentration.

### *NIR spectra collection*

Near-infrared (NIR) spectra were collected using a Metrohm NIRS XDS Multivial Analyzer, controlled via Vision Software (version 4.1.1.238) [29] over a period of approximately 90 days. A workflow for exporting metadata and spectra from Vision and merging the data with laboratory data based on sample name was performed using custom R scripts. Relative humidity readings in the laboratory during the scanning period ranged from 11 to 30%. Temperature readings in the laboratory on all days of scanning ranged from 21.5 to 23.6 C.

Groups of samples were thawed and then brought to room temperature prior to analysis. Nanopure water and a designated process sample were used as controls and were brought to room temperature and scanned in parallel with a given sample set on each day of scanning. Approximately 250µL of each sample was applied to the surface of a quartz optical glass sample cup using a plastic pipette. A gold transflectance adapter with a 0.1mm lip was carefully fitted to the sample cup to ensure that no air bubbles formed between the optical glass and gold plate, thereby creating a 0.2mm pathlength for sample presentation. Sample spectra were collected in duplicate over the entire range of the instrument (400–2499.50 nm) with 0.5nm resolution. Each spectra was the average 32 unique scans, reference standardized to Metrohm Certified Reflectance Standards [30]. The sample cell and transflectance adapter were cleaned between samples by rinsing out the cell, followed by a thorough scrubbing of the surface using a cloth wetted with 70% ethanol.

## Multivariate analysis

### *Overview*

We used the open-source programming language *R* [31] for all modeling and statistical analysis. For spectral transformation and calibration population selection, we

used the *prospectr* package [32]. To build and cross-validate pls models, we used the *pls* package [33]. All data cleaning, wrangling, and visualization was done using the *tidyverse* collection of *R* packages [34].

The *R* scripts used for spectral transformation and regression modeling can be found on the NREL GitHub repository https://github.com/NREL/xos_ms_dataandcode.

#### *Spectral transformation and PLS modeling*

We used Standard Normal Variate (SNV) transformation to correct for light scattering, followed by the Savitzky–Golay filtering (SG) transformation using a 2nd order polynomial fitted 1st derivative with 7-point smoothing window for noise removal and signal amplification. After visual inspection of the two control samples (DI water and the designated process sample that were thawed and scanned with samples on each day of scanning), along with inspection of the correlation coefficients associated with an initial PLS model built on the entire data set, we truncated the spectra to remove the region below 1350nm. This region had substantial variability in the visible spectra, an artifact signal due to an instrument detector change at 1100nm, and little signal in the spectra of xylose and XOS standards. Prior to any model fitting, the transformed spectra were mean-centered.

Prior to modeling, NIR spectra and primary analytical data distributions were evaluated separately to understand relationships between constituents and look for outliers. Analytical outliers were flagged and samples reanalyzed. Principal component analysis (PCA) of the transformed NIR spectra was performed to look for spectral outliers using scores plots and looking at spectral distributions of Mahalanobis distances of the PCA scores that represented the principal components that explained the 95% of the variance. No obvious outliers were found.

PLS-2 models were fit to model XOS, monomeric xylose, and total solids concentrations as outcomes of interest. While previous literature has shown that sugars other than monomeric xylose can be measured in the NIR region [35, 36], there was not enough variability across these other sugar concentration measurements in this dataset to create sufficient spectral variability for modeling monomeric xylose. As the main outcome of interest was XOS concentration, we included monomeric xylose concentration as an outcome of interest despite its small observed range to train the model to detect the differences between the signals observed in monomeric and oligomeric xylose. Furthermore, we included total solids as an outcome of interest to explain the variability in the spectra associated with the other sugar constituents within the liquors as a summed term.

#### *Model validation*

Samples were divided into early/waste and late sample groups. Each group was split into calibration and independent validation sets using a 70–30 split provided by the Kennard–Stone algorithm from the *prospectr* package using the PCA component scores that described 95% of the variation in the transformed spectra from 1350nm-2450nm.

Leave-one-out (LOO) cross validation was used to determine the appropriate number of principal components (PCs) for the model, which typically corresponded with the lowest root mean squared error of cross validation (RMSECV). To avoid overfitting a given outcome of interest, each constituent was evaluated individually for the number of components necessary to fully explain the relationship between spectra and primary analytical measurements.

We evaluated model performance by comparing the root mean squared errors (RMSE) associated with the predictions of the calibration set (RMSEC), the cross validated models (RMSECV), and the independent validation set (RMSEP). In addition to these measurements, we also evaluated the correlation coefficient ($R.^2$) associated with each prediction set. We calculated each performance parameter using the entire data set (full model basis) and after subsetting the dataset according to the different groupings evaluated (early/waste, late). We tested for heteroscedasticity in each validation, cross validation, and independent validation set by visually evaluating the predicted vs residual plots for fanning or funneling. To test the significance of differences observed between model performance parameters, we used a Fisher z-transform followed by a Studentized t-test to compare the difference between the transformed correlation coefficients ($\alpha = 0.05$). We used an F-test to compare RMSE values ($\alpha = 0.05$). [37]

## Results and discussion

### Primary analytical data

The primary analytical data consisted of soluble xylo-oligosaccharide (XOS), monomeric xylose, and total solids (TS) concentration measurements, all with units of grams per liter (g/L), measured at 10 different locations in the process. These analytical data (organized by sample processing location) are displayed in Fig. 2 and summarized in Table 1. The concentrations of XOS and TS are substantially higher in the late streams than in either the early or waste streams, while the concentrations of monomeric xylose were quite low and similar for all sampling locations.

The mean XOS concentration in samples increased roughly 23 times from the early to the late streams. The

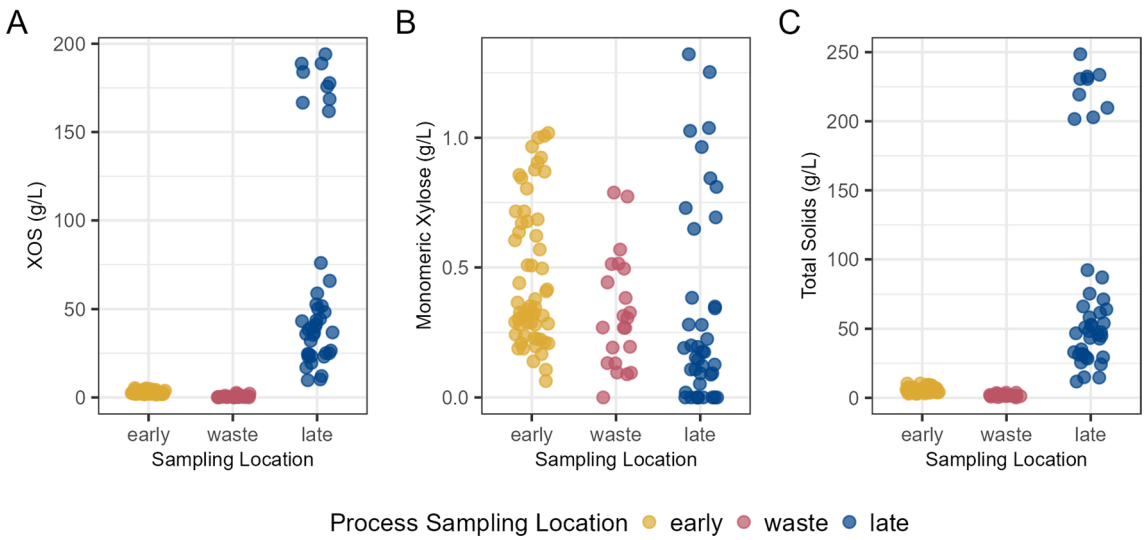


**Fig. 2** Distribution of primary analytical data separated by process stream location. **A** Soluble xylo-oligosaccharide (XOS) concentration (g/L), **B** monomeric xylose (g/L), and **C** total solids concentration (g/L) in filtered process liquor samples. Samples are taken from early process streams (early), waste streams (waste), and late process streams (late). Late stream samples show higher XOS concentrations than early and waste stream samples. Note the differences in ordinate axis range among the three plots

**Table 1** Summary of compositional analysis data for the samples used in this work

| | | | Soluble xylo-oligosaccharides (XOS) (g/L) | | | | Xylose (g/L) | | | | Total solids (g/L) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | Mean | Min | Max | SD | Mean | Min | Max | SD | Mean | Min | Max | SD |
| Total | | 123 | 23.79 | 0.00 | 194.05 | 46.74 | 0.39 | 0.00 | 1.32 | 0.31 | 31.23 | 0.26 | 248.53 | 58.02 |
| Grouped by stream location | Early | 60 | 2.87 | 1.59 | 5.35 | 0.85 | 0.46 | 0.06 | 1.02 | 0.27 | 5.78 | 2.90 | 10.35 | 1.90 |
| | Waste | 22 | 0.64 | 0.00 | 2.61 | 0.71 | 0.33 | 0.00 | 0.79 | 0.22 | 1.77 | 0.26 | 3.69 | 0.94 |
| | Late | 41 | 66.85 | 9.76 | 194.05 | 61.73 | 0.32 | 0.00 | 1.32 | 0.38 | 84.28 | 11.80 | 248.53 | 77.00 |

Summary statistics for soluble xylo-oligosaccharides (XOS), monomeric xylose, and total solids concentrations (g/L). XOS and monomeric xylose concentrations are determined via HPLC, whereas total solids concentration is determined gravimetrically. Samples are grouped by process location into early, waste, and late streams

waste streams have average XOS concentrations approximately 4 times lower than the early and approximately 100 times lower than the late streams. The monomeric xylose concentration remained low across the entire process, indicating that the production reactor conditions favored XOS production over monomeric xylose production, and that the purification process does not substantially degrade XOS. TS concentration increased from early to late stream largely (but not exclusively) due to the increase in XOS concentration. Other monomeric sugars were seen in these samples (e.g., glucose, arabinose) at much lower levels than either XOS or monomeric xylose. A summary of the concentrations of these other sugars found in the samples is provided in the supplemental material.

Figure 3 shows the correlation between XOS and TS concentrations in the sample set. A strong, positive, linear relationship exists between XOS and TS concentration in all samples, but the relationship is different between early/waste stream samples and late stream samples. The slope of the linear fit is an estimation of the XOS purity in the sample population. The early and waste streams have approximately 50% XOS purity, while the late stream samples have approximately 80% purity. XOS purity, therefore, appears to occur as a step change at the nanofiltration process step (F1 in Fig. 1).

### NIR spectroscopy

Figure 4 depicts the average NIRS spectra collected from samples taken from the three different processing locations—early, waste, and late. The y-axis of the spectra from the three locations has been shifted slightly in order to allow for better comparison of the spectral features. Figure 4A depicts the raw spectra, which are dominated by a broad water absorbance band at 1950nm. Figure 4B shows the three average spectra after mathematical

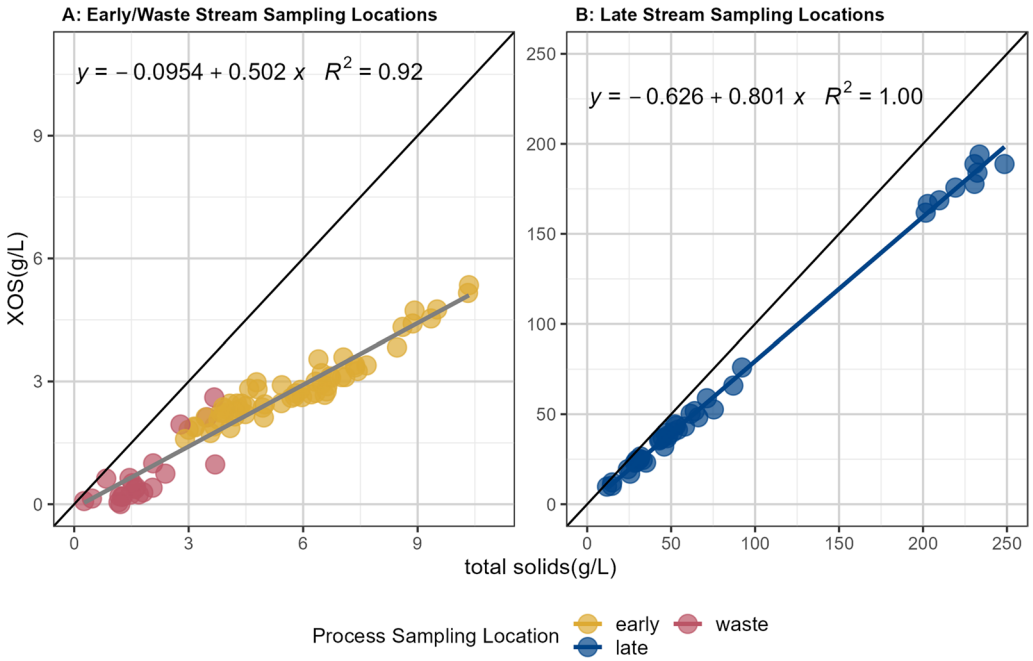


**Fig. 3** Soluble xylo-oligosaccharide (XOS) concentrations vs total solids concentration of process samples. **A** Early and waste samples, **B** late stream samples. Solid black lines in each plot represent the line of equivalence, while the solid lines through data points represent the linear regression. The slopes of the linear regression fits are a measure of the average XOS purity in these samples (50.2% for early and waste stream sampling locations and 80.1% for the late stream sampling locations)

transformation via standard normalization and Savitzky–Golay smoothing/derivatizing. The late stream spectrum (blue) has visibly unique peaks in both the first overtone (1650nm-1750nm) and the combinational overtone (2100nm-2450nm) regions. No visible differences can be observed between the early stream (yellow) and waste stream (red) spectra.

To better understand how the individual analytes contribute to the NIR spectra of the process samples, we collected spectra of solutions of deionized (DI) water containing 1-20g/L of either XOS or monomeric xylose. Figure 5A and B shows that increasing the concentration of either monomeric xylose **(A)** or XOS **(B)** increases the signal in both the first overtone (1650nm-1750nm) and the combinational overtone (2100nm-2450nm) regions. Close inspection of these plots shows that the spectral signatures of XOS and xylose in the combinational overtone region are visually distinct. In addition, the XOS spectral signature is very similar to those observed in the late stream process samples (Fig. 4), indicating that the peaks observed in these samples are the result of increased XOS content in the late stream samples. Figure 5C and D shows PCA score plots of the data in Fig. 5A and B**;** Fig. 5C shows PCA results using the full model spectral range (1350-2450nm), and Fig. 5D shows PCA results using only the combinational overtone range (2100-2450nm) of the spectra. These PCA score plots show that the majority of the variability (PC1) is driven by the concentration of either monomeric xylose or XOS while PC2 primarily differentiates monomeric xylose from XOS. Using only the combinational overtone range (Fig. 5) did not substantially reduce this discrimination ability compared to the full range (Fig. 5C). These results suggest that a model using a reduced spectral range may provide similar results compared to a model using the full spectral range.

To better investigate spectral differences between the process sampling locations, we performed PCA on the entire sample population.

Figure 6A depicts a PCA score plot showing the first two components, which together describe over 95% of the total spectral variance in the samples set, colored by process location. Late stream samples have higher PC1 values than early and waste stream samples. Interestingly, while most of the variance observed across PC1 in the late stream samples can be explained by XOS concentration (Figure 6B), ($R^2 = 0.99$), no correlation exists between PC 1 and XOS concentration in the early/waste stream samples ($R^2 = -0.01$).

The differences between the early and waste stream samples in both XOS and TS concentration (Fig. 2A and C), in the correlation between XOS and TS concentration (Fig. 3), and in the relationships between XOS concentration and PC 1 (Figure 6B) collectively indicate that the

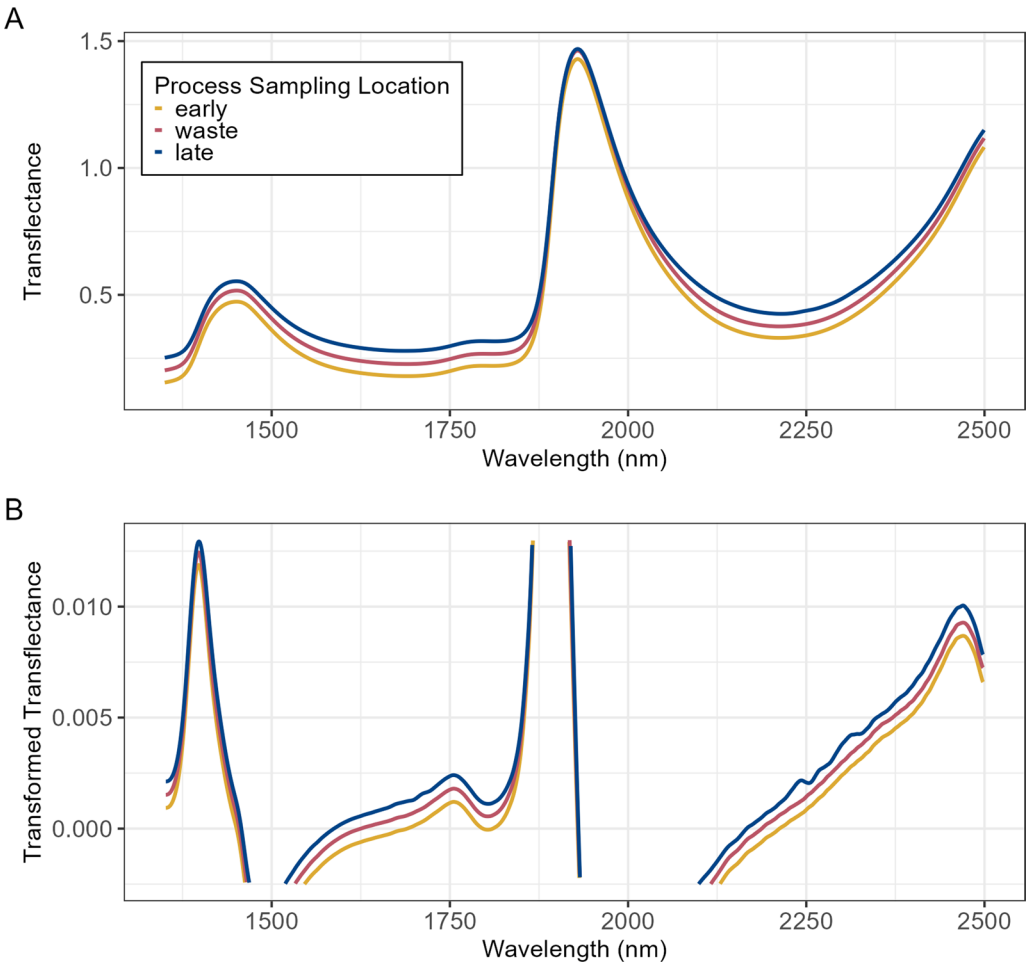


**Fig. 4** NIR spectral differences by process sample location. **A** Average near-infrared (NIR) transflectance spectra of filtered process liquor samples collected using the Metrohm NIRS XDS Multivial Analyzer, grouped by process location. The spectra are offset by process sampling location to facilitate comparison. The x-axis has been truncated to 1500-2500nm for better visibility of the spectral features of interest. **B** Average NIR transflectance spectra after transforming via standard normal variance and Savitzky–Golay smoothing. The x-axis has been truncated to 1500-2500nm for better visibility of the spectral features of interest. The y-axis has been truncated to remove the minimum and maximum peaks at 1900nm, which are associated with water and have little variance. Differences between the average late spectrum and the average early and waste spectra are clearly seen between 1650 and 1750nm and between 2100 and 2450nm

variability in the late stream samples is large and significant enough to drive most of the variability in the sample set. Thus, we anticipated that a separate PLS-2 model to predict XOS concentration for the early/waste and late stream sample groups will likely provide superior results to a single model.

## Modeling results

### *Impact of process location-splitting*

To test the effectiveness in subsetting the samples by sampling location, we compared a PLS-2 model made with the full sample set to models made after splitting the samples into two groups—an early/waste stream subset and a late stream subset. Figure 7 shows the independent validation results for both models for both XOS and TS concentrations. Summary statistical data for these models are shown in Table 2. Splitting the early/waste stream samples into a separate model statistically significantly improves the model performance for XOS as measured by cross validation and independent validation RMSE, and for total solids for RMSECV ($\alpha = 0.05$). In contrast, no statistically significant differences were found in model performance among the late process samples with subsetting ($\alpha = 0.05$). This is consistent with our hypothesis that it is the variability in the late stream samples that is driving the variability of the overall set, so sample splitting would be unlikely to affect the ability to model these samples.

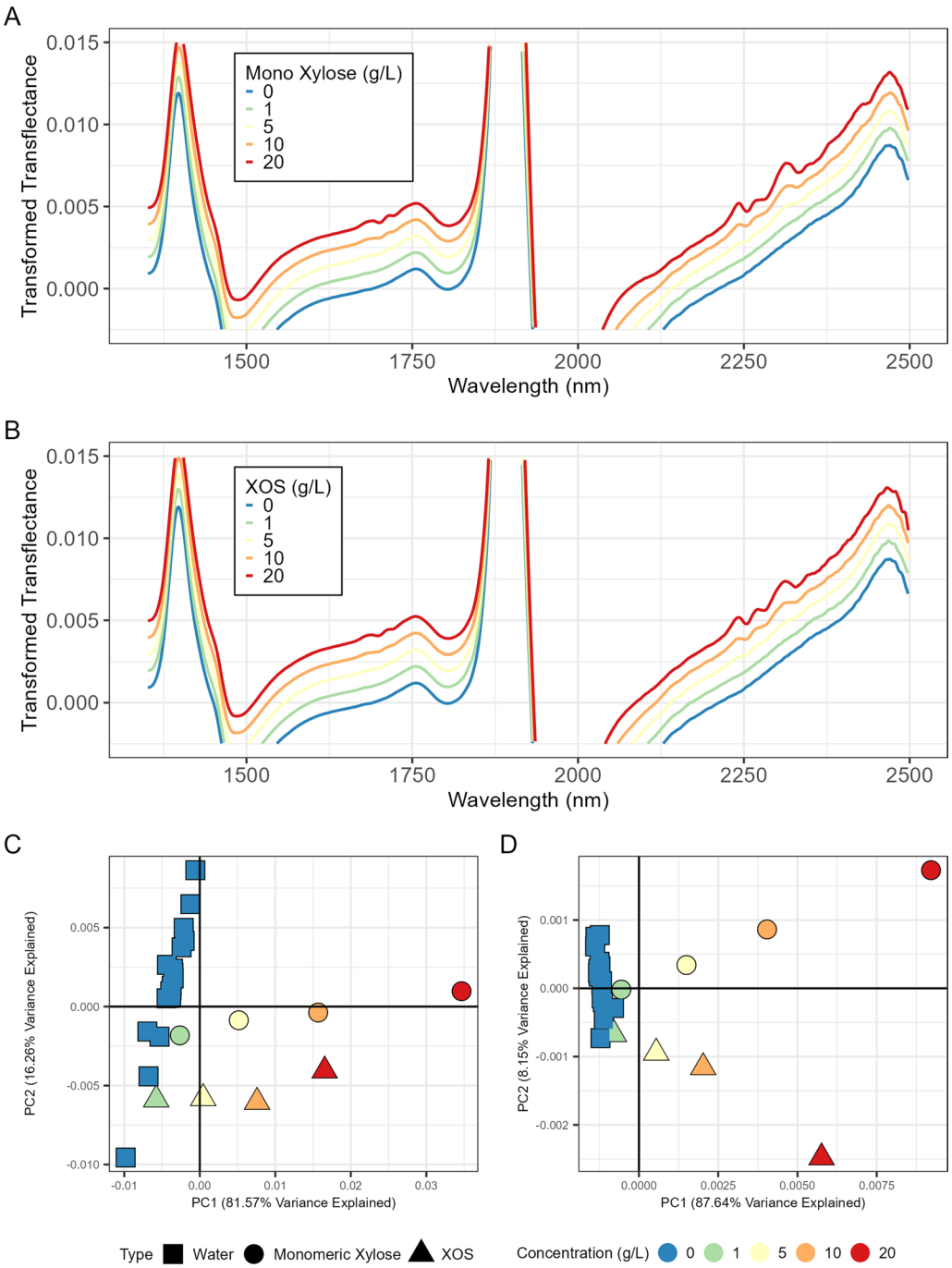


**Fig. 5** NIR spectral differences between monomeric xylose and soluble xylo-oligosaccharide (XOS). Average NIR transflectance spectra (truncated to 1350-2500nm) collected using the Metrohm NIRS XDS Multivial Analyzer, of deionized (DI) water spiked with different concentrations of **A** monomeric xylose or **B** soluble xylo-oligosaccharides **(**XOS) after transforming spectra via standard normal standard normal variance and Savitzky–Golay smoothing. Sample spectra were offset corresponding to the concentration of either monomeric xylose or XOS added. Increasing the concentration of either XOS or monomeric xylose causes a visible alteration in the combination overtone region between 2000 and 2500nm, and in the first overtone region between 1650 and 1750nm. The spectral signature of monomeric xylose and XOS are notably different. To better visualize the effect of constituent and oligomer on spectra, principal component analysis (PCA) was performed on the transformed spectra. **C** and **D** are scores plots from principal component analysis on **C** the full model spectral range (1350–2500 nm) or **D** only the combinational overtone range (2100–2450 nm)

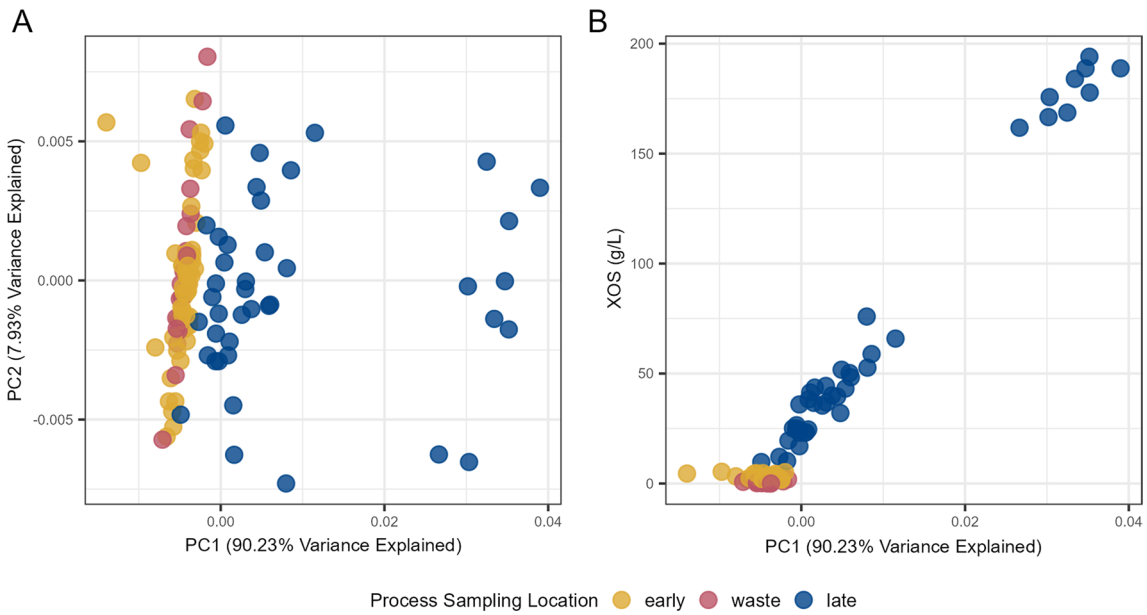


**Fig. 6** Principal Component Analysis of process samples. **A** Principal Component 2 (PC 2) vs PC 1 of transformed near-infrared (NIR) spectra (1350-2500nm) of liquor samples colored by process sampling location. The first two PCs explain over 95% of the total transformed spectral variance. The early and waste stream samples appear more similar to each other than to the late stream samples. The samples with PC1 scores greater than 0.02 are from the F3 retentate stream (Fig. 1) and contain the highest concentrations of XOS. **B** Measured soluble xylo-oligosaccharide (XOS) concentration (g/L) vs PC1 of transformed NIR spectra of liquor samples colored by process sampling location. PC1 is highly correlated with the XOS concentration in the late stream samples. No relationship exists between XOS concentration and PC1 in early and waste stream samples, which are less refined and have substantially lower XOS concentrations than the late stream samples

All model performance measures for monomeric xylose calibration were poor, as expected with such a small range in sample concentration (Fig. 2, Table 1). Predicted vs measured plots of the calibration and cross validation results for xylose concentration, along with predicted vs residual plots for both validations, are provided in the supplemental material.

As noted previously, location-splitting substantially decreased the population size of the training sets. While these modeling results act as feasibility tests for the ability of NIR to monitor total solids and XOS concentration in this process, we do not believe that either split model is robust enough for deployment at its current size. Sample sizes of at least an order of magnitude greater for each location that span the variability expected in the process should be obtained to allow for better model fitting, robust significance testing, and the implementation of a robust method for outlier determination.

### *Impact of reduced spectral range*

To further our investigation into which sections of the spectra were important for characterizing XOS and monomeric xylose, we performed a PCA on transformed spectra of water controls taken throughout the scanning campaign, along with monomeric xylose standards (1-20g/L) and XOS standards (1-20g/L) shown earlier. Figure 8A shows the score plot of the PCA from the combined dataset full spectral range PCA. PC1, which explains 81.6% of the variance in the dataset, differentiates between pure water and presence/concentration of an analyte. PC2, which explains another 16.3% of the variance, explains the inherent variability in the sampling of the water controls over time as well as differentiating between monomeric xylose and XOS. The large variability in the water controls makes the differentiation between monomeric xylose and XOS very difficult.

To identify the spectral regions contributing to the PCA results in Fig. 8A, we performed three additional PCA analyses of the water controls, the XOS standards, and monomeric xylose standards separately. Figure 8C–E clearly shows the differences in the first principal component (PC1) loadings for each subgroup. In all three subgroups, water peaks at 1450nm and 1900nm dominate the first principal component. These results suggest these regions will be of little use in modeling either monomeric xylose or XOS concentrations.

Both the monomeric xylose and XOS controls show spectral signatures in the 1st overtone region (1650nm-1750nm) and the combinational overtone region (2100nm-2450nm). In contrast, no spectral signatures of the water control spectra are evident in these regions. Furthermore, the combinational region shows a larger

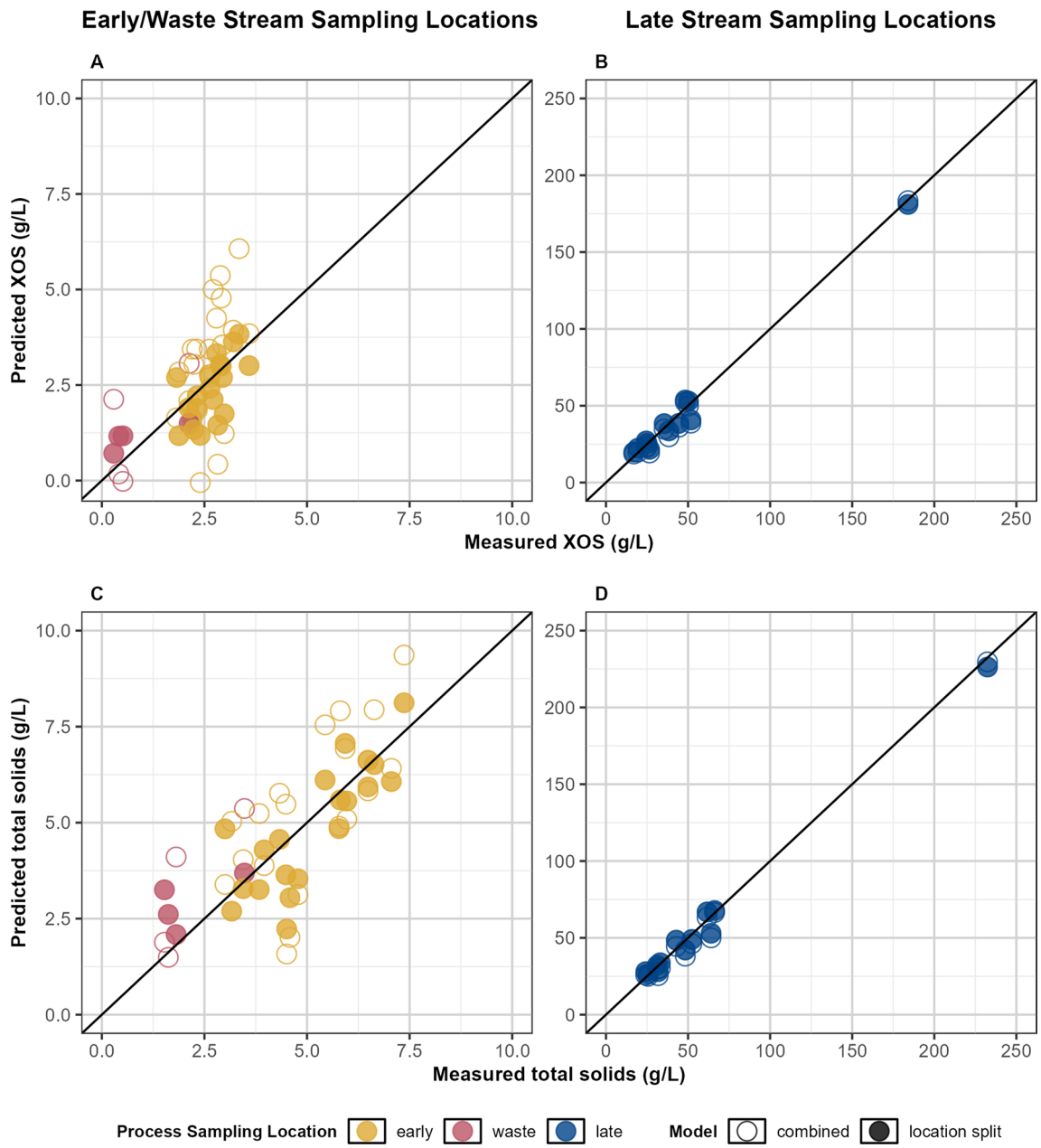


**Fig. 7** Impact of sample splitting on the performance of predictive models. Predicted vs measured soluble xylo-oligosaccharide (XOS) and total solids concentrations (g/L) for the independent validation of models with and without splitting the late stream samples and the early and waste stream samples into different calibration sets. The combined location model results are depicted with open circles, while the subset location model results are depicted with closed circles. **A** Predicted vs. measured XOS concentration in the early and waste stream samples; **B** predicted vs. measured XOS concentration in the late stream samples; **C** predicted vs. measured total solids concentration in the early/waste stream samples; and **D** predicted vs. measured total solids concentration in the late stream samples. Separating the samples into two groups by processing location significantly improves model performance in the early and waste stream samples. The impact of sample splitting on the performance of the late stream model is much less significant, suggesting the variability in these samples drives the variability of the whole population

difference in loadings signature between the XOS and monomeric xylose spectra, indicating that this region of the spectra alone may be sufficient to produce a predictive model to distinguish XOS from monomeric xylose.

To test whether using the combinational overtone spectral range (2100-2450nm) would lead to more stable spectral controls, we performed a PCA on the combined dataset (standards plus water controls) spectra using the

**Table 2** Summary of model performance results by process stream location and model sample process splitting technique

| Performance parameter | | XOS (g/L) | | | | Total solids (g/L) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Early/waste stream sampling locations | | Late stream sampling locations | | Early/waste stream sampling locations | | Late stream sampling locations | |
| | | Without location grouping | With location grouping | Without location grouping | With location grouping | Without location grouping | With location grouping | Without location grouping | With location grouping |
| Training | $R^2$ | 0.49 | 0.97* | 1.00 | 1.00 | 0.64 | 0.99* | 1.00 | 1.00 |
| | RMSEC | 1.93 | 0.24* | 4.64 | 4.17 | 2.07 | 0.30* | 3.29 | 3.72 |
| Cross Validation | $R^2$ | 0.15 | 0.46* | 0.99 | 0.99 | 0.26 | 0.46 | 1.00 | 1.00 |
| | RMSECV | 3.31 | 1.16* | 5.08 | 4.7 | 3.67 | 2.26* | 3.49 | 4.08 |
| Independent Validation | $R^2$ | 0.35 | 0.52 | 0.99 | 0.99 | 0.53 | 0.69 | 0.99 | 0.99 |
| | RMSEP | 1.44 | 0.67* | 5.72 | 4.81 | 1.50 | 0.97 | 5.81 | 5.02 |

Leave-one-out cross validation was used to tune models. Results are shown for predicted soluble xylo-oligosaccharide (XOS) concentration (g/L) and total solids concentration (g/L). Splitting the samples by location substantially improves model performance for the early/waste stream models for XOS concentration and total solids concentrations. The impact of sample splitting on the performance of the late stream model is much smaller, and not statistically significant ($\alpha = 0.05$). Asterisks (*) indicate statistically significant differences in the models using location-splitting compared to the model using all samples

reduced spectral range. Figure 8B shows the resulting score plot. When spectral range is reduced to a region showing unique spectral signatures of the components of interest, PC1 still explains the difference between water and the presence of analytes at increasing concentrations, but PC2 now differentiates between monomeric xylose and XOS, and also effectively explains the presence of XOS or monomeric xylose–increasing PC2 correlates with increasing monomeric xylose concentration, and decreasing PC2 correlates with increasing XOS concentration. Furthermore, the variability in the water control scores that masked differentiation of XOS from monomeric xylose in the full range spectra PCA (Fig. 8A), is now smaller compared to the variability of the XOS and monomeric xylose scores.

Since the spectral signature from the combinational overtone region (2100-2450nm) in the process samples showed substantial signal (Fig. 4B), could differentiate between monomeric xylose and XOS standards effectively (Fig. 5D), and also reduces the contribution of environmental variability in spectra inherent to the sampling method (Fig. 8B), we decided to test whether this reduced spectral range could be used to build useful prediction models. Figure 9 compares the independent validation results for models made with the full (1350nm-2450nm) or reduced (2100nm-2450nm) wavelength ranges for XOS and total solids concentration, and Table 3. describes the model performance results. Small but statistically significant improvements were observed in early/waste stream model performance for XOS (RMSECV) and total solids (RMSECV) when the spectral range was truncated to 2100–2450 nm. No statistically significant differences were found in independent prediction results between models (Table 3).

## Conclusions

In this work, we show the promise of multivariate calibration models for predicting the concentrations of soluble xylo-oligosaccharides (XOS) and total solids (TS) at multiple points in a biomass conversion process which produces and then purifies XOS compounds from sugar cane bagasse using near-infrared spectroscopy. A single model using samples from multiple locations in the process stream showed acceptable performance for both XOS and TS as measured by standard statistical measures. However, compared to the single model, separate models built by splitting the calibration samples according to expected XOS concentration and purity show improved performance. A calibration model with a limited spectral range that contains substantial signal for process samples provided essentially equivalent performance to the model using the full spectral range. Thus, simple, data-informed approaches can provide practically significant improvements in model performance for this dataset while maintaining use of the traditional and easily accessible partial least squares regression model. By providing this dataset and the associated modeling scripts as open source, we invite others to contribute alternative modeling solutions to the practicalities of real-time bioprocess monitoring of aqueous solutions using NIRS.

While at-line modeling is a useful, low-risk step in understanding the feasibility of NIRS for process monitoring in a given system, this work would be best followed up with the development of a an on-line NIRS process monitoring system. Fiber-optic transflectance probes could be installed in the process flow streams after each unit operation. Models predicting XOS and total solids concentrations at each location could provide operators the rapid inputs necessary to optimize each step towards

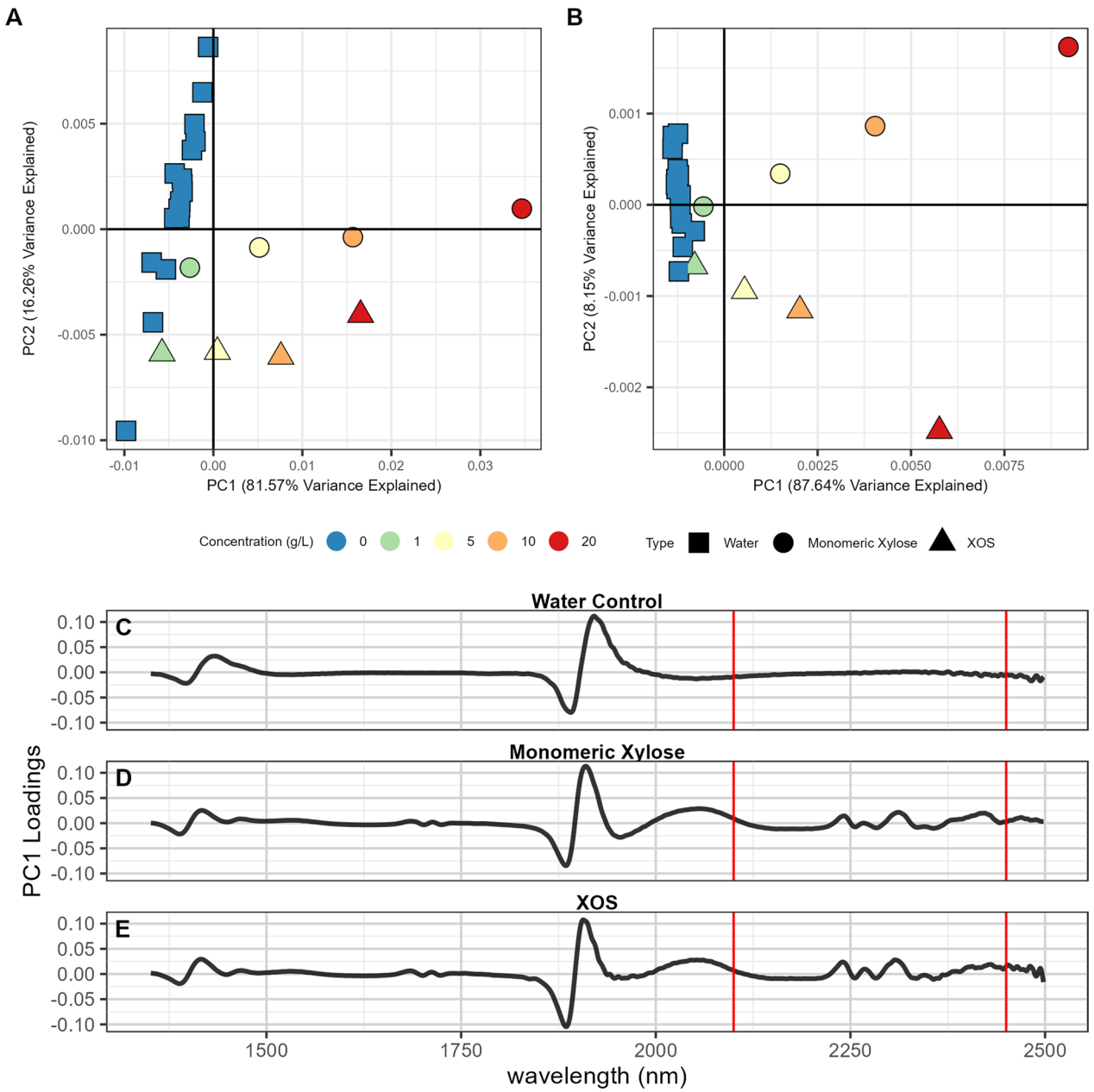


**Fig. 8** Principal component analysis of pure monomeric xylose, XOS, and water control samples. Score and loading plots from principal component analyses (PCA) of monomeric xylose and XOS (0-20g/L) standards combined with 16 water control spectra taken throughout the scanning campaign. **A** Score plot of first two principal components (PCs) of a PCA of water controls, monomeric xylose, and XOS standards using the full spectral range (1350-2500nm). PC1 differentiates between pure water and the presence and concentration of either XOS or monomeric xylose. PC2 shows the spectral variability of the water controls and differentiates between XOS from monomeric xylose. **B** Score plot of first two principal components (PCs) of a PCA of water controls, monomeric xylose, and XOS standards using only the combinational overtone spectral range (2100–2450 nm). In contrast with Fig. 5B, PC2 shows the spectral variability associated with differentiating XOS from monomeric xylose relative to the innate variability of the sampling method (e.g., the variability in the water controls). **C** Loadings plot of the 1st PC of a PCA of the 16 water control spectra. **D** Loadings plot of the 1st PC of a PCA of the monomeric xylose standards with concentrations ranging from 0 to 20 g/L. **E** Loadings plot of the first principal component in a PCA of the XOS standards with concentrations ranging from 0 to 20 g/L. Red lines depict the combinational overtone range (2100-2450nm), which show substantial signal for both monomeric xylose and XOS not associated with water absorption

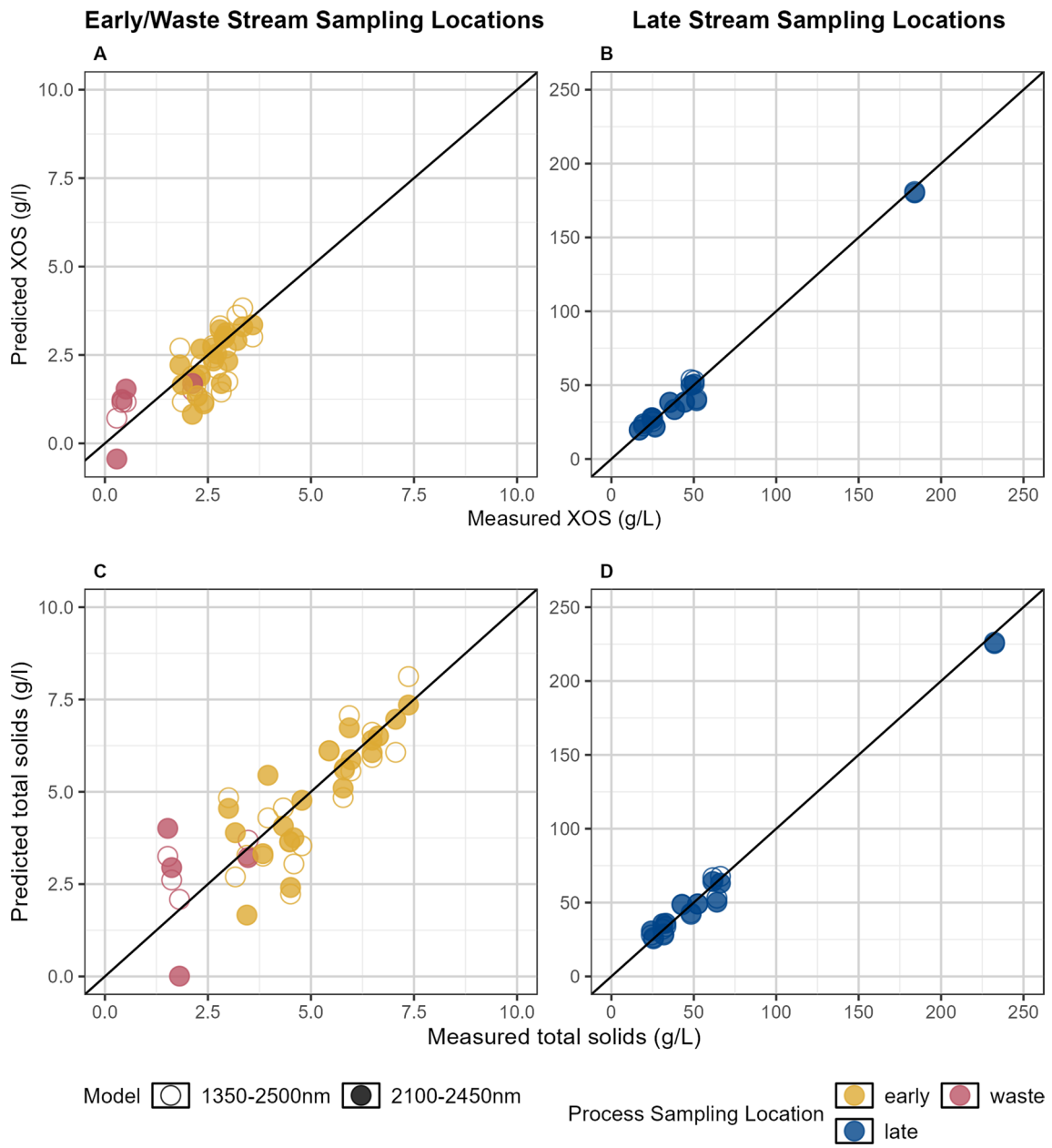


**Fig. 9** Impact of reduced spectral range on the performance of predictive models**.** Predicted vs measured soluble xylo-oligosaccharide (XOS) and total solids concentration (TS), for the independent validation of models using the full spectral range (filled circles, 1350-2500nm) or the reduced spectral range (open circles, 2100-2450nm). The color of the sample corresponds with the process stream location. **A** Predicted vs. measured XOS concentration for the early/waste stream model; **B** predicted vs. measured XOS concentration for the late stream model. **C** Predicted vs. measured total solids concentration (g/L) for the early/waste stream model; and **D** predicted vs. measured TS concentration (g/L) for the waste stream model. Reducing the spectral range used in the model to 2100-2450nm does not substantially affect the quality of either the XOS concentration or the total solids concentration models

**Table 3** Summary of model performance results by process stream location and spectral range used in modeling

| Performance parameter | | XOS (g/L) | | | | Total solids (g/L) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Early/waste stream sampling locations | | Late stream sampling locations | | Early/waste stream sampling locations | | Late stream sampling locations | |
| | | 1350–2500 nm | 2100–2450 nm | 1350–2500 nm | 2100–2450 nm | 1350–2500 nm | 2100–2450 nm | 1350–2500 nm | 2100–2450 nm |
| Training | $R^2$ | 0.97 | 0.96 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 |
| | RMSEC | 0.24 | 0.28 | 4.17 | 3.76 | 0.30 | 0.22* | 3.72 | 4.52 |
| Cross validation | $R^2$ | 0.46 | 0.71 | 0.99 | 1.00 | 0.46 | 0.76* | 1.00 | 1.00 |
| | RMSECV | 1.16 | 0.78* | 4.70 | 4.08 | 2.26 | 1.37* | 4.08 | 4.88 |
| Independent validation | $R^2$ | 0.52 | 0.60 | 0.99 | 0.99 | 0.69 | 0.66 | 0.99 | 0.99 |
| | RMSEP | 0.67 | 0.63 | 4.81 | 4.98 | 0.97 | 1.07 | 5.02 | 5.86 |

Results are shown for prediction of XOS concentration (g/L) and total solids concentration (g/L). Reducing the spectral range used in the model to 2100-2450nm does not substantially affect the quality of either the XOS concentration or the total solids concentration models. Asterisks (*) indicate statistically significant differences in truncated-range model compared to corresponding full-range model

maximum purity and concentration of XOS, which could then be automated into process control loops.

#### Abbreviations

| | |
|---|---|
| NIR | Near Infrared |
| PCA | Principal component analysis |
| PLS | Partial least squares |
| RID | Refractive index detector |
| RMSEC | Root mean square error of calibration |
| RMSECV | Root mean square error of cross validation |
| RMSEP | Root mean square error of prediction |
| XOS | Xylo-oligosaccharide |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13068-024-02558-6.

Supplementary Material 1.

#### Acknowledgements

The authors gratefully acknowledge Natalia Thompson and Evelyn Canales, who provided the primary analytical chemistry for all liquor samples, and Bi Nguyen for help with NIR scanning of the liquor samples.

#### Author contributions

EW and ZT planned the modeling and spectroscopy work. ZT performed NIR spectroscopy and multivariate modeling. KG provided insight and expertise into the metadata. EW organized the manuscript. All authors contributed to writing the manuscript and approved the final version of manuscript.

#### Funding

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Bioenergy Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

#### Availability of data and materials

All NIR spectroscopy, primary analytical data, and modeling codes are available at https://github.com/NREL/xos_ms_dataandcode. The following supplementary materials are included with this manuscript. Table S1 contains the summary statistics explaining the distribution of the measured minor sugars. Table S2 contains the summary performance statistics in measuring monomeric xylose. Figure S3 contains predicted vs measured and predicted vs residual plots for the full sample set and spectra model's measurement of monomeric xylose in the calibration, cross validation, and independent validation sets. Figure S4 contains predicted vs measured and predicted vs residual plots for the full sample set and spectra model's measurement of XOS in the calibration, cross validation, and independent validation sets. Figure S5 contains predicted vs measured and predicted vs residual plots for the full sample set and spectra model's measurement of total solids in the calibration, cross validation, and independent validation sets. Figure S6 contains predicted vs measured and predicted vs residual plots for the early sample set only full spectra model's measurement of monomeric xylose in the calibration, cross validation, and independent validation sets. Figure S7 contains predicted vs measured and predicted vs residual plots for the early sample set only full spectra model's measurement of XOS in the calibration, cross validation, and independent validation sets. Figure S8 contains predicted vs measured and predicted vs residual plots for the early sample set only full spectra model's measurement of total solids in the calibration, cross validation, and independent validation sets. Figure S9 contains predicted vs measured and predicted vs residual plots for the late sample set only full spectra model's measurement of monomeric xylose in the calibration, cross validation, and independent validation sets. Figure S10 contains predicted vs measured and predicted vs residual plots for the late sample set only full spectra model's measurement of XOS in the calibration, cross validation, and independent validation sets. Figure S11 contains predicted vs measured and predicted vs residual plots for the late sample set only full spectra model's measurement of total solids in the calibration, cross validation, and independent validation sets. Figure S12 contains predicted vs measured and predicted vs residual plots for the early sample set only reduced spectra model's measurement of monomeric xylose in the calibration, cross validation, and independent validation sets. Figure S13 contains predicted vs measured and predicted vs residual plots for the early sample set only reduced spectra model's measurement of XOS in the calibration, cross validation, and independent validation sets. Figure S14 contains predicted vs measured and predicted vs residual plots for the early sample set only reduced spectra model's measurement of total solids in the calibration, cross validation, and independent validation sets. Figure S15 contains predicted vs measured and predicted vs residual plots for the late sample set only reduced spectra model's measurement of monomeric xylose in the calibration, cross validation, and independent validation sets. Figure S16 contains predicted vs measured and predicted vs residual plots for the late sample set only reduced spectra model's measurement of XOS in the calibration, cross validation, and independent validation sets. Figure S17 contains predicted vs

measured and predicted vs residual plots for the late sample set only reduced spectra model's measurement of total solids in the calibration, cross validation, and independent validation sets. Figure S18 contains a plot of the centered and transformed control spectra.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
ZT and EW declare no competing interests. KG was formerly Chief Technology Officer (CTO) of Prenexus, which was developing a process to extract and purify xylo-oligosaccharides from sugar cane bagasse. Prenexus is no longer in existence.

**Author details**
[1]National Renewable Energy Laboratory, Golden, USA.

Received: 22 April 2024 Accepted: 26 July 2024
Published online: 14 August 2024

**References**

1. Vrancken C, Longhurst PJ, Wagland ST. Critical review of real-time methods for solid waste characterisation: Informing material recovery and fuel production. Waste Manag. 2017;61:40–57.
2. Rolinger L, Rüdt M, Hubbuch J. A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing. Anal Bioanal Chem. 2020;412:2047–64.
3. Biancolillo A, Marini F. Chemometric methods for spectroscopy-based pharmaceutical analysis. Front Chem. 2018. https://doi.org/10.3389/fchem.2018.00576.
4. Wasalathanthri DP, Rehmann MS, Song Y, Gu Y, Mi L, Shao C, et al. Technology outlook for real-time quality attribute and process parameter monitoring in biopharmaceutical development—a review. Biotechnol Bioeng. 2020;117:3182–98.
5. Skibsted E, Engelsen SB. Spectroscopy for Process Analytical Technology (PAT). In: Lindon JC, Tranter GE, Koppenaal DW, editors. Encycl Spectrosc Spectrom Third Ed [Internet]. Oxford: Academic Press; 2017, p. 188–97. https://www.sciencedirect.com/science/article/pii/B9780128032244000261. Accessed 25 Mar 2024.
6. Cogoni G, Liu YA, Husain A, Alam MA, Kamyar R. A hybrid NIR-soft sensor method for real time in-process control during continuous direct compression manufacturing operations. Int J Pharm. 2021;602: 120620.
7. Skvaril J, Kyprianidis KG, Dahlquist E, Skvaril J, Kyprianidis KG. Applications of near-infrared spectroscopy (NIRS) in biomass energy conversion processes: a review. Appl Spectrosc Rev. 2017;52:675–728.
8. Grassi S, Alamprese C. Advances in NIR spectroscopy applied to process analytical technology in food industries. Curr Opin Food Sci. 2018;22:17–21.
9. Qu J-H, Liu D, Cheng J-H, Sun D-W, Ma J, Pu H, et al. Applications of near-infrared spectroscopy in food safety evaluation and control: a review of recent research advances. Crit Rev Food Sci Nutr. 2015;55:1939–54.
10. Kumaravelu C, Gopal A. A review on the applications of Near-Infrared spectrometer and Chemometrics for the agro-food processing industries. 2015 IEEE Technol Innov ICT Agric Rural Dev TIAR [Internet]. 2015, p. 8–12. https://ieeexplore.ieee.org/document/7358523. Accessed 21 Apr 2024.
11. Nascimento RJA do, Macedo GR de, Santos ES dos, Oliveira JA de. Real time and *in situ* Near-Infrared Spectroscopy (NIRS) for Quantitative Monitoring of Biomass, Glucose, Ethanol and Glycerine concentrations in an alcoholic fermentation. Braz J Chem Eng. 2017;34:459–68.
12. Vann L, Layfield JB, Sheppard JD. The application of near-infrared spectroscopy in beer fermentation for online monitoring of critical process parameters and their integration into a novel feedforward control strategy. J Inst Brew. 2017;123:347–60.
13. Chen Y, Xie Y, Ajuwon KM, Zhong R, Li T, Chen L, et al. Xylo-oligosaccharides, preparation and application to human and animal health: a review. Front Nutr. 2021;8: 731930.
14. Gupta M, Bangotra R, Sharma S, Vaid S, Kapoor N, Dutt HC, et al. Bioprocess development for production of xylooligosaccharides prebiotics from sugarcane bagasse with high bioactivity potential. Ind Crops Prod. 2022;178: 114591.
15. Lian Z, Wang Y, Luo J, Lai C, Yong Q, Yu S. An integrated process to produce prebiotic xylooligosaccharides by autohydrolysis, nanofiltration and endo-xylanase from alkali-extracted xylan. Bioresour Technol. 2020;314: 123685.
16. Saville S, Saville BA. High fiber cane: pathway to a novel xylooligosaccharide prebiotic and human health. Agro Food Ind Hi-Tech. 2018;29:36–8.
17. Beebe KR, Pell RJ, Seasholtz MB. Chemometrics: a practical guide. Wiley-Intersci Ser Lab Autom. New York: Wiley and Sons; 1998
18. Wang H-P, Chen P, Dai J-W, Liu D, Li J-Y, Xu Y-P, et al. Recent advances of chemometric calibration methods in modern spectroscopy: algorithms, strategy, and related issues. TrAC Trends Anal Chem. 2022;153: 116648.
19. Konkol JA, Tsilomelekis G. Porchlight: an accessible and interactive aid in preprocessing of spectral data. J Chem Educ. 2023;100:1326–32.
20. Bian X. Spectral preprocessing methods. In: Chu X, Huang Y, Yun Y-H, Bian X, editors. Chemometric methods in analytical spectroscopy technology. Singapore: Springer Nature; 2022 p. 111–68. https://doi.org/10.1007/978-981-19-1625-0_4
21. Rinnan Å, van den Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal Chem. 2009;28:1201–22.
22. Beebe KR, Pell RJ, Seasholtz MB. Chapter 3: preprocessing. Chemom pract guide. New York: John Wiley & Sons; 1998.
23. Westad F, Marini F. Variable Selection and Redundancy in Multivariate Regression Models. Front Anal Sci. 2022. https://doi.org/10.3389/frans.2022.897605.
24. Mehmood T, Sæbø S, Liland KH. Comparison of variable selection methods in partial least squares regression. J Chemom. 2020;34: e3226.
25. Esquerre C, Gowen AA, Downey G, O'Donnell CP. Selection of variables based on most stable normalised partial least squares regression coefficients in an ensemble Monte Carlo procedure. J Infrared Spectrosc. 2011;19:443–50.
26. Shi L, Westerhuis JA, Rosén J, Landberg R, Brunius C. Variable selection and validation in multivariate modelling. Bioinformatics. 2019;35:972–80.
27. Saville, Bradley. Liquid co-extraction process for production of sucrose, xylo-oligosaccharides and xylose from feedstock [Internet]. Melbourne; 2015. p. 1. https://ipsearch.ipaustralia.gov.au/patents/2015252695
28. Sluiter A. Determination of sugars, byproducts, and degradation products in liquid fraction process samples: laboratory analytical procedure (LAP); Issue Date: 12/08/2006. Tech Rep. 2008
29. NIRS XDS MultiVial Analyzer [Internet]. https://www.metrohm.com/en/products/2/9211/29211210.html. Accessed 9 Jan 2023.
30. NIRS reflection standard, set of 2 [Internet]. https://www.metrohm.com/en/products/6/7450/67450000.html. Accessed 9 Jan 2023.
31. The R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; https://www.R-project.org/
32. Stevens A, Ramirez-Lopez L. An Introduction to the prospectr package [Internet]. 2022. https://cran.r-project.org/package=prospectr
33. Mevik B-H, Wehrens R, Liland KH. Introduction to the pls package [Internet]. https://cran.r-project.org/package=pls
34. Wickham H. Welcome to the TidyVerse. J Open Source Softw. 2019;4:1686.
35. Chen S-F, Danao M-GC, Singh V, Brown PJ. Determining sucrose and glucose levels in dual-purpose sorghum stalks by Fourier transform near infrared (FT-NIR) spectroscopy. J Sci Food Agric. 2014;94:2569–76.
36. Chen J, Arnold MA, Small GW. Comparison of combination and first overtone spectral regions for near-infrared calibration models for glucose and other biomolecules in aqueous solutions. Anal Chem. 2004;76:5405–13.

37. Roggo Y, Duponchel L, Ruckebusch C, Huvenne J-P. Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data. J Mol Struct. 2003;654:253–62.

## Publisher's Note



**Zofia Tillman** is a Research Scientist at NREL

**Kevin Gray** is Chief Technology officer of ZensoLabs LLC, 95 Woodrock Rd., Weymouth MA.

**Edward Wolfrum** is a Principal Researcher and Group Manager at the National Renewable Energy Laboratory (NREL).