Rapid Noninvasive Measurement of Sugar Content in Crops via NIR Spectroscopy

A Design Project Report

Presented to the School of Electrical and Computer Engineering of Cornell University
in partial fulfilment of the requirement for the Degree of Master of Engineering,
Electrical and Computer Engineering

Submitted by: Shreyas Renganathan

MEng Field Advisor: Dr. Nils Napp, Dr. El-Ghazaly

Degree date: January 2022

Abstract

Master of Engineering Program

School of Electrical & Computer Engineering

Cornell University

Design Project Report

Project Title

Rapid Noninvasive Measurement of Sugar Content in Crops via NIR Spectroscopy

Author

Shreyas Renganathan

Abstract

A key competitive advantage for sugarcane growers is being able to continuously monitor the sugar content of their crops in the field, in order to troubleshoot crop growth and optimize yield at harvest. However, this has traditionally relied on invasive sampling, which is labor-intensive and damages crops. Recently, near-infrared (NIR) spectroscopy has shown promise as a modality to monitor sugar content noninvasively and automatably. Several challenges exist: separating the spectral signature of sucrose from the other species within sugarcane, extracting a strong enough signal through the highly-scattering fibrous matter, and filtering out variation from lighting and positioning samples. With these challenges in mind, this project sought to develop a proof-of-concept model to predict sucrose content (Brix) in sugarcane stalks using NIR. Two prediction models were explored: dual-wavelength linear regression and Partial Least Squares Regression (PLS-R). Both were optimized to fit the data with R^2 >0.9 on sucrose solutions. On sugarcane stalks, a PLS-R model was able to achieve a cross-validation R^2 of 0.902. Given this promising model performance, and an optical assembly serving as a proof-of-concept, the next stage of this project would be to validate the model on a larger dataset, and translate the model into a portable embedded system, e.g. a handheld device or automated platform to perform in-situ NIR measurements of sugarcane in the field.

Executive Summary

With the goal of predicting sucrose concentration using NIR absorbance data, the proof-of-concept consisted of a predictive model and an optical assembly. Both were developed in parallel, and iteratively tested on sucrose-water solutions and then sugarcane stalk cuttings. The target specification for this project was a cross-validation R^2 of at least 0.9 and a mean absolute error within ±1° Brix.

An incremental development approach was pursued. First, the wavelengths most highly correlated with sucrose absorbance were identified in literature. Then, a publicly-available sugarcane dataset was used to explore the implementation of two predictive modeling approaches: dual-wavelength linear regression (similar to the traditional pulse oximeter algorithm), and Partial Least Squares Regression (PLS-R). Both implementations were tested on sucrose-water solutions of known concentrations, using an off-the-shelf NIR spectrophotometer. After both prediction models were able to fit the data with an R^2 greater than 0.9, they were applied to spectral data collected on sugarcane stalk cuttings. "Ground truth" measurements of sucrose concentration within sugarcane were taken using a manual juicer and a commercially-available handheld Brix refractometer. A PLS-R model developed demonstrated a cross-validation regression coefficient of 0.902 and mean absolute error of ±0.97° Brix.

A proof-of-concept optics assembly was developed and refined in concert with these experiments. This initially consisted of a sample holder and NIR spectrophotometer, housed in a light-blocking cardboard box. Eventually, the box was redesigned and rebuilt using 3D-modeling software and experimentation to accommodate other features: an external light source, lenses, internal baffles, internal reflective surfaces.

Future work on this project would include further refining the prediction performance of the PLS-R model and validating the model on a larger dataset. The eventual outcome of this project is the translation of the model and optical assembly into a custom device prototype capable of measuring sugar content in sugarcane in the field.

Table of Contents

1. Introduction

1.1. Background

Sugarcane plays a key role in the global economy, supplying around 86% of global sugar production (Voora et al 2019), as well as 40% of global biofuel production (Lam et al 2009). Sugar consumption is expected to rise by about 1.6% per year over the next few years, while production has slowed due to market volatility and increasingly-high production costs (Voora et al 2019). While sugar prices have always been volatile, often dropping below production costs, production costs have generally increased in recent years due to increasing scarcity of manual labor, land, and water (Research and Markets 2020; Hess et al 2016). Sugarcane growers are faced with the challenge of staying competitive while risking greater debt to cover production costs.

In this landscape, automation has shown great promise in maximizing productivity while minimizing production costs. For example, the sugarcane harvester has been shown to halve harvesting costs, while recovering a greater weight per volume of cane, compared to manual harvesting (Ahmed and Alam-Eldin 2015). And the recent introduction of yield mapping technology has begun to allow data-driven decisionmaking on optimal locations based on the GPS coordinates in the field at which the yield at harvest (in sugar weight per hectare) is greatest (Momin et al 2019).

However, the advent of automation in the sugarcane farming industry has been slow to tackle monitoring of crop productivity during growth. Generally, commercially-available automation tools tend to only measure crop yield at harvest time and are unable to determine sugar content while the crop is growing (Bramley 2009). As a result, by the time the yield of the crop is determined, it is too late for the farmer to take steps to take preventative actions on underperforming crops for this harvest. Additionally the results from any actions would only be visible on the next harvest--which leads to potentially wasted resources and slow troubleshooting due to the lack of real-time feedback, as well as lack of time resolution on factors influencing yield over the growth phase.

Continuous monitoring in the growth phase is a step towards rapid, efficient crop management. However, current methods for this are manual-labor-intensive, destructive, and time- consuming. Typically, whole stalks are collected from the field and defoliated (all by hand), and then sent to a laboratory, which returns the productivity measurement (sugar content) within a few days (Nawi et al 2014). Additionally, the accuracy of this measurement is confounded by the notorious tendency of sugarcane samples to deteriorate rapidly – losing about 20% of their sugar content even within 48 hours (Solomon et al 2006).

Therefore, there is a need for a non-destructive, rapid, and automated version of this process. One candidate modality is spectroscopy-- i.e. with the right optics, the sugar content can be determined by analysing optical absorbance data in sugarcane at certain wavelengths of light. Spectroscopy in the near-infrared (NIR) range in particular has emerged as a promising method to obtain data on sugar content from sugar-rich crops (Nawi et al 2014). Although spectroscopy is a common laboratory technique, implementing it in the field poses a number of challenges: variability in lighting conditions, hardware size constraints, and need for high-throughput processing despite the noisy signal expected in an outdoor environment.

Several metrics exist to quantify sucrose content in crops, including total sugar percent (TS%) and commercial cane sugar content (CCS). The primary sugar in sugarcane is sucrose, however, which makes up 94–98.5% of the sugar content (Koltuniewicz 2010). Therefore sucrose concentration is used as a standard indicator to quantify sugar content, typically in the form of Brix. Brix is a unit system represented in degrees, with an increase in one degree being equivalent to an additional gram of sucrose dissolved in 100 grams of solution. Brix does not include any corrections for impurities and fiber content (like CCS does, for example). However the advantage of Brix is that it is easily measured using a handheld Brix refractometer. In this context, the device that is the eventual end-goal of this project would serve as an upgrade to the handheld Brix refractometer. The limitation of the refractometer is that it is invasive, i.e. it requires a sample of juice extracted from within the stalk. The end-goal device offers the same functionality (i.e. Brix output) *noninvasively*. In this way, the device offers a step up from a technology already widespread in the sugarcane farming, that farmers are familiar with.

1.2. Scope

The initial scope of this project encompassed three goals:

I. Development of an algorithm to predict sucrose concentration in whole sugarcane from NIR spectroscopy data

II. Prototyping of a novel embedded system designed with custom optics and programmed with the model

III. Testing of the system on sugarcane crops in the field

However, constructing and optimizing a high-performing prediction model proved to be highly time-consuming and laborious. This was the rate-limiting step of this project, especially given that only one person was working on the project. Specific difficulties included: navigating spectrophotometer settings, developing an experimental setup with optimal optics, determining experimental procedure, and raising the predictive strength of the prediction model. Consequently, over the course of this project the scope was refined. Most of the project focused on Goal I, above, and Goal II was only explored to the extent of arriving at a proof-of-concept test setup for experimental data collection. Translating the proof-of-concept to a prototype with refined optics, along with hardware and software (i.e. the rest of Goal II), was designated future work. Testing the prototype on sugarcane in the field (Goal III) was also designated future work. With my decision to pursue a third semester at Cornell, continuing development of Goals II and III is currently planned to be completed by me as independent research in the next semester.

The eventual outcome from the work in this project is a custom handheld device capable of taking spectroscopic measurements of sugarcane in the field, and using the data to predict sucrose concentration in Brix. A key feature of this device is that measurements are taken non-invasively; that is, the device has to be able to determine sucrose concentration through whole unmodified sugarcane--i..e through the rind. As a result, experiments were designed and

conducted with the aim of creating a prediction model with its input data being from samples of whole sugarcane.

Additionally, for the prediction model to be competitive with other work in the space, the target specification for performance in this project was considered to be a regression coefficient of 0.9 or above, with a mean absolute error of ±1° Brix.

2. NIR Spectroscopy

2.1. Sucrose Spectrochemistry

The chemistry underlying spectroscopy is that bonds between atoms vibrate when they absorb electromagnetic radiation of a certain frequency. A given bond vibration is excited when the frequency of the incoming electromagnetic radiation is equal to the vibrational frequency ("fundamental frequency"), or at a frequency equal to an integer multiple thereof ("overtone"). Sucrose is composed of C–H, O–H, C–C, and C–O bonds, all of which have fundamental or overtone frequencies in the mid-infrared (2500nm - 25000nm), and near-infrared (780nm - 2500nm) ranges (Golic et al 2003).A bond between a given pair of atoms can have various vibration modes. For example, "stretching" vibrations are periodic changes in interatomic distance along the axis of a bond. And "deformation" vibrations, also known as "bending" vibrations, are periodic changes in the angle between a bond and a reference plane. When more than one vibration is excited at the same time, this is called a "combination" vibration. Figure 1 provides an overview of the vibration-exciting frequencies (presented as wavelengths) for the bonds in sucrose.

Tentative assignment		Fundamental	Vibrational frequency overtones			References
			1st	2nd	3rd	
OH stretching	nm	2860–3120	1410–1440	970	738	8, 16, 44, 45
	cm^{-1}	3200–3500	6950–7100	10 300	13 550	
OH combinations	nm	1920–2080	1100	840		14, 16
	cm^{-1}	4800–5200	9090	11 900		
CH stretching	nm	3300–3470	1600–1800	1100–1230	910	15–17
	cm^{-1}	2880–3000	5550–6250	8100–9100	11 000	
CH combinations	nm	2100–2352				16, 18
	cm^{-1}	4250–4750				
CH_2 stretching	nm	3460–3500	1720–1765	1215	930	16, 19, 20
	cm^{-1}	2880–2910	5670–5820	8230	10 750	
CH_2 combinations	nm	2310–2325				16
	cm^{-1}	4300–4330				
OH, CH, and CH_2 deformations	nm	6900–8330	2250–2320	2400–2600	1850–2120	14, 16
	cm^{-1}	11 111–25 000				
	nm	1200–1450	4310–4440	3840–4170	4720–5400	
	cm^{-1}	400–900				

Figure 1. Vibration-exciting wavelengths (fundamentals and overtones) of various O-H, C-H, and CH2 bonds, specified using wavelength. All values are from a literature review performed by Golic et al 2003. Image reproduced from Table I in Golic et al 2003.

It can be observed that several vibration-exciting wavelengths exist for each bond vibration type in sucrose. Notably, higher wavelengths are generally shown with larger ranges, because vibration-exciting wavelengths are known within narrower tolerances in the near-infrared range than in the mid-infrared range (Golic et al 2003).

It is also established in literature that at wavelengths beyond the first overtone of the O-H combination vibration (1100nm), the absorption of O–H bonds in water tends to dominate spectral measurements of aqueous substances (Golic et al 2003).

We can see the increased absorbance of water at higher wavelengths in Figure 2.

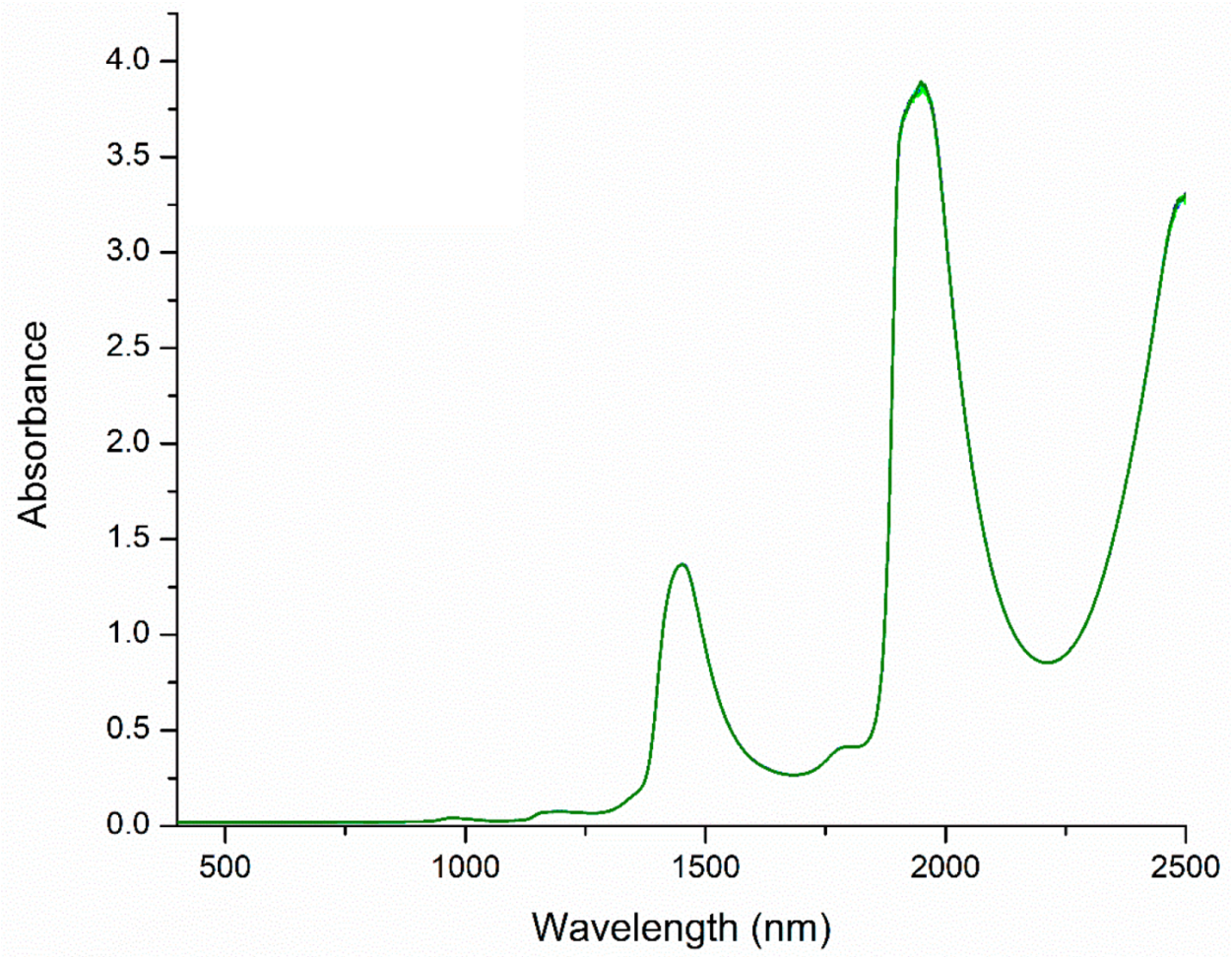


Figure 2. Absorbance of liquid water from 400nm to 2500nm. From Muncan & Tsenkova (2019)

Figure 2 confirms that 1100nm indeed appears to be a suitable threshold, marking the highest-absorbance region where absorbance seems to be relatively stable (constant) as a function of wavelength before the absorbance substantially rises starting at about 1300nm. Note a slight local maximum in absorbance visible around 990-1000nm.

Sugarcane stalks have high water content. Therefore, while sucrose also contains O-H bonds, the signal from the O-H bonds in sucrose may be attenuated or eclipsed substantially by the O-H bonds in water if taking measurements >1100nm. Specifically, choosing a wavelength with high water absorbance risks the possibility that cells in the non-sugar-containing exterior of the

stalk would absorb and attenuate the light reaching the interior sugar-containing regions. This suggests keeping the upper limit as low as possible so as to minimize signal attenuation due to water.

Since vibration-exciting wavelengths are more precisely known in the near-infrared range (780nm-2500nm) and the absorbance of O-H bonds in water tend to dominate >1100nm, a wavelength range that is optimally sensitive to sucrose might be 780nm-1100nm.

While sucrose does not have any fundamental frequencies in this subset of the near-infrared (NIR) spectrum, sucrose does have several overtone frequencies in this range. Figure 3 outlines these wavelengths, organized by vibrational mode, for sucrose solutions.

		Molecular environment		
Tentative assignment	Overtone	Sucrose in H_2O	Sucrose in D_2O	Sucrose-d_8 in D_2O
CH/CH_2 combination	. . .		1040	1040
OH stretching (more H bonded)	second	984	976	
OH stretching	second	960	954	
CH stretching	third	910	906	914
OH combination	. . .	840		
CH stretching	fourth			772
OH stretching (more H bonded)	third	770		
OD combination	. . .		762	762
CH_2 stretching	fourth		730	742
OH stretching	third	740		

Figure 3. Wavelengths associated with bond vibration within three types of sucrose solutions. H_2O refers to water, and D_2O refers to deuterium oxide. All values are from experiments performed by Golic et al (2003).

Sucrose in sugarcane occurs as an aqueous sucrose solution. Between the experimental data in Figure 3 and the literature data in Figure 1, we can see that wavelengths that excite C-H or C-H2 bond vibration in sucrose-water solutions include 910nm and 930nm. Similarly, wavelengths

that excite O-H bond vibration include 740nm, 770nm, 840nm, 960nm, 970nm, and 984nm. Golic et al (2003) recommends that spectroscopic analyses for determining sucrose concentration be based on wavelengths that O-H and C-H bond vibration, emphasizing C-H bond vibration wavelengths.

At this point, while we have established the reasoning for the upper limit of 1100nm based on water absorbance, the discussion so far has assumed a lower limit of 780nm. There is a case to be made for ignoring wavelengths much below 780nm as well, based on other species in the sugarcane stalk.

In observing sugarcane stalks, it is immediately apparent that they come in a range of colors, suggesting a variety of chromophores in the stalks to correct for. See Figure 4.

Figure 4. Reproduced from Ekpélikpézé et al (2016).

However, despite their seeming multitude, these colors are primarily a result of two pigments: chlorophyll and anthocyanin (Sandhu et al 2016). Additionally, the highest-absorbance

wavelengths for these pigments occur only within a relatively narrow range. The absorbance spectra of chlorophyll (both a and b forms) and anthocyanin are shown in Figure 5.

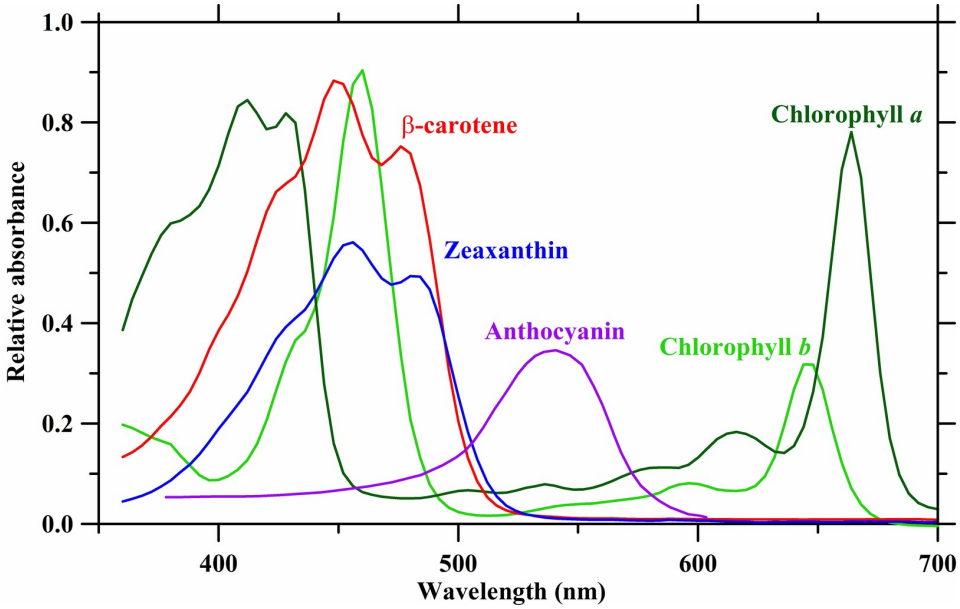


Figure 5. Absorbance Spectra of Plant Pigments (Bilodeau et al 2019)

We can see there are substantial relative absorbance peaks in anthocyanin and and in both chlorophyll forms within the 400nm-700nm range. Generally, none of these pigments are considered to have appreciable absorbance at wavelengths outside this range. One approach from this point would be to focus on wavelengths below 400nm; however, these would be in the ultraviolet (UV) range. Using an external light source in the UV spectrum would have introduced safety risks from operator exposure to UV radiation.

Therefore, the safety risks associated with >400nm wavelengths and the high absorbance of chromophores in the 400nm-700nm range support a lower wavelength limit of 700nm for the waveband of interest.

With putative limits of 700nm (lower limit) and 1100nm (upper limit) in mind, existing sugarcane spectral data were explored. Specifically, these data were observed for the visibility of clear trends in absorbance with respect to sucrose concentration (i.e. Brix value). One study

took spectroscopic measurements in the range 700nm-100nm on cross-sections of fresh sugarcane (Nawi et al 2013), as plotted in Figure 6.

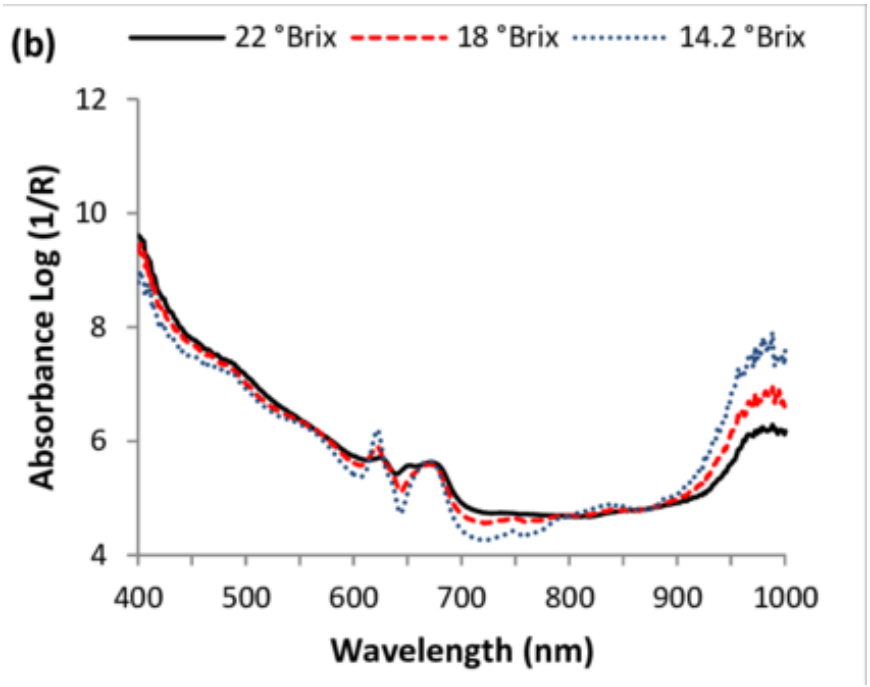


Figure 6. Transmission through sugarcane samples with three different concentrations. Reproduced from Nawi et al (2013).

We can see that for a waveband spanning from 700nm to 1000nm, there are wavelengths at which monotonic trends between absorbance and Brix seem clear.
Namely, there appears to be clear positive correlations between absorbance and Brix within 700nm-750nm. Similarly there appears to be clear negative correlations between absorbance and Brix over 900nm-1000nm.

Since areas of negative and positive correlation are visible in this range, this further supports the suitability of the spectrum of interest in this project being **700nm-1000nm**. Notably, this is also the ideal waveband recommended by Sanseechan et al (2018) for measuring NIR spectra of sugarcane for sugar content prediction.

2.2. Pretreating Spectral Data

In the context of this project, "spectral dataset" refers to a dataset containing one or more absorbance spectra, each spectrum representing a sample with a different sucrose concentration. One common challenge in analyzing NIR spectroscopy spectral datasets is the presence of scattering effects. These effects tend to differ from sample to sample-- for example, there may be slight variations in sample position or geometry relative to incident light from the spectrophotometer, leading to differences in how light scatters through the sample. This ultimately varies the angles and magnitudes of light reflecting back to the spectrophotometer.

A common technique to correct for variable scattering effects is Multiplicative Scatter Correction (MSC). MSC treats scattering effects as multiplicative and/or additive transformations from an ideal "reference spectrum" that is treated as free of scattering effects. MSC occurs in three steps:

1. Find the reference spectrum. Commonly, this is computed as the mean across the spectral dataset, i.e. the spectrum consisting of the mean absorbance at each wavelength
2. Perform a least-squares linear regression for each spectrum X_i against the reference spectrum X_r. This will generate an expression of X_i as a function of X_r, i.e. $X_i \approx a_i + b_i X_r$. Here the a_i and b_i terms represent (respectively) the additive and multiplicative effects of scattering.
3. Now that the scattering terms are known for each spectrum, calculate the corrected spectrum Xmsc = (Xi-ai)/bi. This removes the additive and multiplicative scattering effects.

MSC is applied to the dataset as a whole. This means that, once a model is generated using MSC-corrected training spectra, any new test spectra must be added to the training spectra set to re-calculate the mean spectrum ("reference spectrum") and subsequently MSC must be re-applied to the whole dataset.

Another technique to correct for scattering effects is Standard Normal Variate (SNV) correction. Unlike MSC, SNV is performed on each individual spectrum. SNV occurs in two steps:

1. Mean-center each spectrum by subtracting the mean absorbance in the spectrum from every value in the spectrum.
2. Divide each mean-centered spectrum by its own standard deviation

Since SNV is not dependent on other spectra, once a model is generated using SNV-corrected training spectra, SNV can be applied to each new test spectrum as it arrives. Using an example spectral dataset (F750 spectra in Chaix 2020), we can observe the effects of MSC and SNV. Figure 7 compares the original spectra with the MSC-processed and SNV-processed spectra.

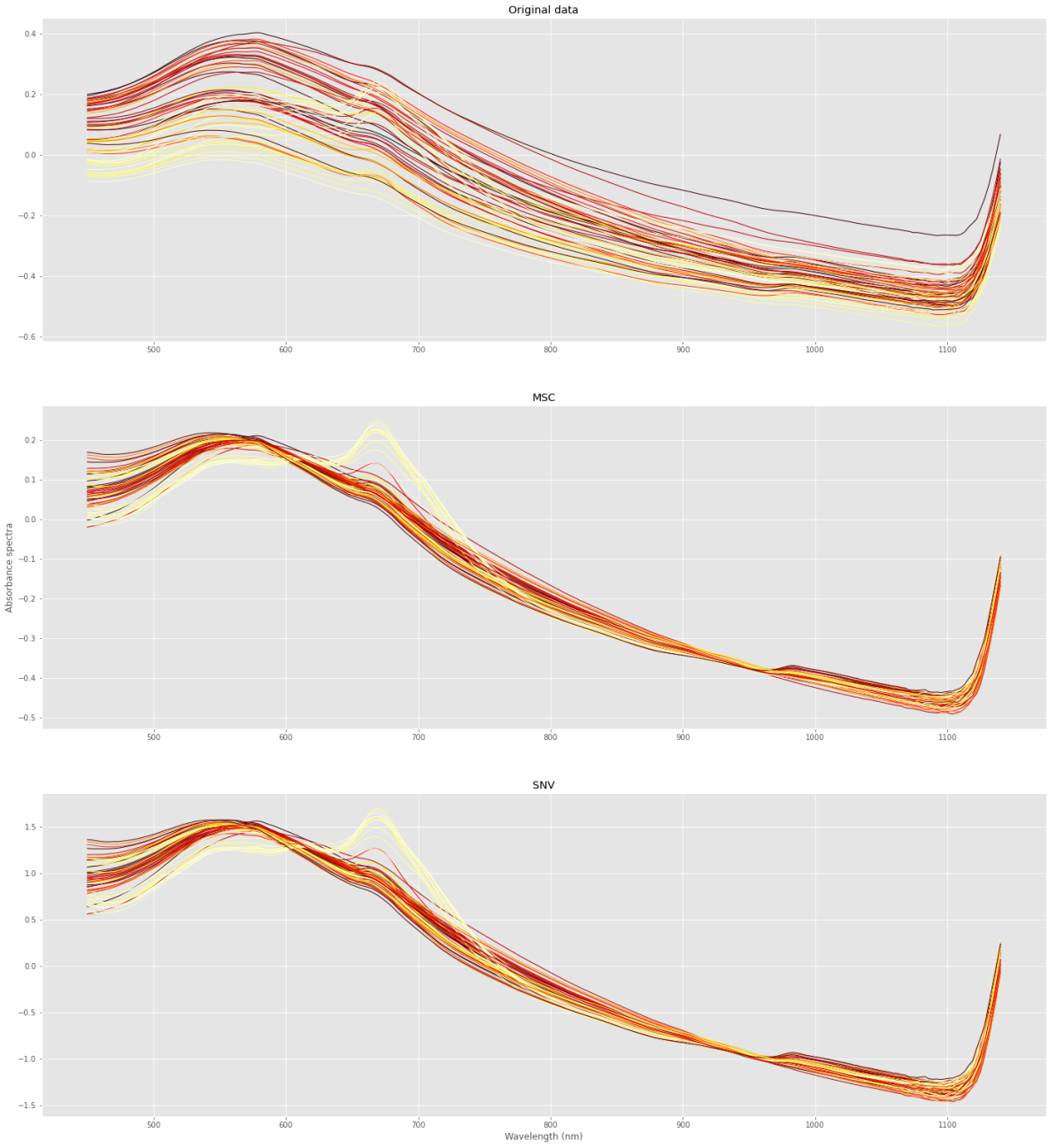


Figure 7. (from top to bottom) (a) original dataset from (F750 from Chaix 2020) (b) MSC-processed dataset (c) SNV-processed dataset

We can see that MSC and SNV yield similar results. Both techniques generally appear to reduce the variation in absorbance among the spectra at each wavelength, while preserving the overall shape of each spectrum.

One difference between the two outputs is that SNV rescales the data more substantially than does MSC. After MSC, the spectra are rescaled to a vertical axis scale close to that of the original. After SNV, however, the spectra are rescaled on the vertical axis to within approximately -2 and +2.

The output of the predictive model created using spectral data is a Brix value. Therefore the fact that SNV rescales the data to a greater extent is not a disadvantage, since the model can just be trained using those rescaled values. MSC and SNV are both commonly used in preprocessing NIR spectroscopy data. One advantage SNV offers over MSC is the lack of a need to generate a mean spectrum. This means that processing a test spectrum with SNV is not dependent on other spectra, while MSC correction is. Therefore SNV is algorithmically simpler to apply.

In terms of performance, MSC has a tendency to overemphasize outliers, because each spectrum is divided by a slope that may be small for certain spectra (Fearn et al 2009). This, and the relative simplicity of applying SNV, led to SNV being the main scattering correction used for analysis in this project.

Once scattering effects are corrected using MSC or SNV, a common next step is to apply smoothing. Smoothing attenuates variation in the data from high-frequency noise (high frequency with respect to wavelength). Savitzky-Golay ("Savgol") filtering is a widely used smoothing technique in NIR spectroscopy. This technique applies a convolution kernel to a moving window along the spectral data. The net effect is that for each point $p(\lambda)$ in a spectrum, a polynomial f of specified order is fit (using least-squares) to the data points within a specified window centered on that point. That point is then replaced with the value $f(\lambda)$. In this way, each point along a spectrum is replaced with a polynomial fit intersecting that point within a moving window. The strength of this technique is that features (peaks and troughs) consisting of a few

points are flattened to a lesser extent than other smoothing techniques, viz. moving-average filtering.

2.3. Predicting using Spectral Data

After preprocessing to remove scattering effects and high-frequency noise, the spectral dataset can then be analyzed to construct predictors. Each predictor is a mathematical function that is fit to the relationship between sucrose concentrations (independent variable) and absorbance values (dependent variable) at a particular wavelength. The relationship can be assumed to be linear, since according to the Beer-Lambert law, the absorbance of a substance (after correcting for scattering effects) is linearly proportional to the concentration of that substance within the sample. Naturally, the linear relationship will not be strong for every wavelength -- that is to say, if a predictor is created for every wavelength, not every predictor will have a high level of linear fit between absorbance and sucrose concentration. Therefore, some means to identify the best-fitting predictors is necessary, and then these predictors can be combined in some way to create a prediction model.

A naive approach to create a prediction model would be to look for a single predictor in the NIR spectral data. However, this is likely to be insufficient to capture enough signal (variation in absorbance due to sucrose concentration) against noise (variation in absorbance due to other factors). Due to the tendency of several substances to have fundamental (or overtone) frequencies in the NIR range, absorbance at a particular wavelength generally depends on more than one source of variation (NIRSystems 2002) -- i.e. not just sucrose concentration . Additionally, small sample-to-sample variations (e.g. differences in ambient light or sample dimensions) can also contribute to variations in absorption spectra.

One approach that corrects for non-sucrose-related sources of variation would be to identify two predictors, and then construct a prediction model using the ratio between them. With the selection of the right predictors, this could remove noise. Taking the ratio of the two predictors

would correct out sources of variation common to both wavelengths (i.e. noise). Additionally, the ratio also outputs a prediction model normalized to water, rather than to all matter within the sample. In other words, if one predictor shows strong positive correlation between absorbance and sucrose concentration, and the other shows strong negative correlation, then this suggests the major source of variation at the first wavelength is the concentration of sucrose in sample matter, while the major source of variation at the latter wavelength is the concentration of water in sample matter. Therefore, taking the ratio of the two predictors would yield the concentration of sucrose in water, which is the desired output of the prediction model for this project.

Notably, this is the approach underlying traditional pulse oximetry, which is a ubiquitous technology in health monitoring across inpatient, outpatient, and consumer settings. Pulse oximetry measures the oxygen saturation in blood (SpO2) noninvasively, using spectral information at (typically) two wavelengths gathered by a sensor placed on the finger. Oxygen saturation is the ratio of oxygenated hemoglobin to total hemoglobin. Pulse oximetry shares several similarities with NIR spectroscopy of sucrose in sugarcane -- it also noninvasively measures the concentration of a species (oxygenated hemoglobin) in a biological medium (tissue) in which scattering effects are present, using spectral data in the near-infrared spectrum.

Pulse oximetry relies on the principle that oxygenated hemoglobin (HbO2) and deoxygenated hemoglobin (Hb) differ in their absorbance-vs-wavelength curves. Specifically, two predictors are identified such that the absorbance of blood at a particular wavelength is positively correlated with oxygen saturation, and the absorbance of blood at another particular wavelength is negatively correlated with oxygen saturation. Figure 8 shows the absorbance spectra of Hb and HbO2.

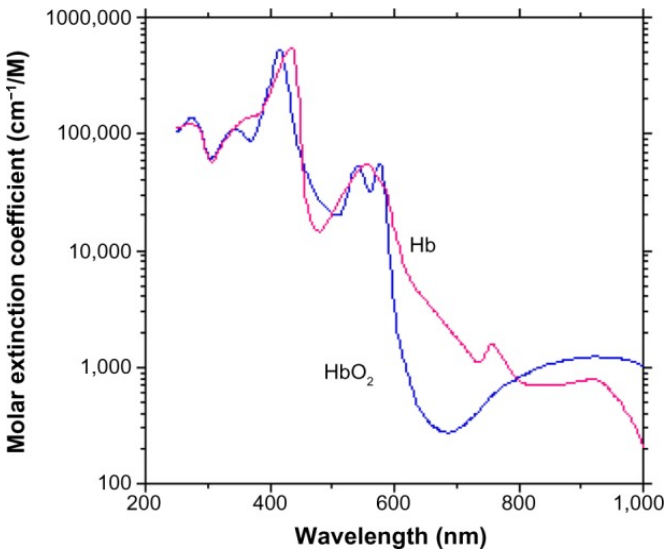


Figure 8. Absorbance spectra of oxygenated (Hb) and deoxygenated (HbO2) hemoglobin. Note that absorbance is represented by the molar extinction coefficient, which normalizes absorbance to path length and molar concentration. Reproduced from Figure 1 in Nitzan et al (2014)

We can see in Figure 8 that Hb and HbO2 have similar spectral behavior up to about 580nm. Beyond that, oxygenation of hemoglobin is correlated with a decrease in absorbance between 580nm and 780nm, and with an increase in absorbance between 780nm and 1000nm. Different manufacturers use slightly different predictors to characterize Hb and HbO2, but consistently one wavelength is chosen in the 580-780nm range and the other in the 780nm-1000nm range (Nitzan et al 2014). A common pair is 660nm and 940nm (Dobyns 2011), which in Figure 8 correspond respectively to the wavelengths of greatest negative correlation and greatest positive correlation of absorbance with oxygenation. Using these two predictors, a ratio *R* is then calculated. This ratio can be summarized as the signal at 660nm divided by the signal at 940nm (Dobyns 2011).

Next, a prediction model for SpO2 is created using the ratio *R* by passing it into a calibration equation, traditionally:

$$SpO2 = 110 - 25R.$$

This equation is derived from performing a least-squares linear regression on SpO2 vs *R* data from clinical studies.

Translating this approach to the project at hand, a spectral dataset would first be collected on sugarcane samples with various Brix values. At each wavelength, the absorbance values would be regressed (using ordinary least-squares linear regression) against the corresponding Brix values. Two wavelengths would be identified: one where absorbance is most negatively correlated with Brix with the best fit, and another where absorbance is most positively correlated with Brix with the best fit. Taking the ratio of the absorbance at the first wavelength to the absorbance at the second wavelength would produce an index representative of the ratio of sucrose to water. These indices would then be regressed (again using ordinary least-squares linear regression) against the Brix values, to yield a calibration equation that expresses the sucrose concentration (S) in terms of the index (R), a slope coefficient (m), and a vertical intercept (b):

$$S = mR + b$$

The advantage of this approach is hardware simplicity. Since there are only two predictors, spectral data is necessary only at two wavelengths. Instead of a spectrophotometer, a broad-spectrum photodetector with two LEDs (each with peak emission at one of the wavelengths) can be used to collect this data. Indeed, this is the typical hardware setup found in pulse oximeters (Dobyns 2011) (Nitzan et al 2014).

While the pulse oximetry approach was explored in this project, it may be the case that a prediction model containing just two predictors is not enough to achieve high prediction accuracy. In NIR spectroscopy, a standard method to identify multiple predictors (typically more than two) and use them to construct a prediction model is Partial Least-Squares Regression (PLS). PLS reduces the amount of data to be considered via dimensionality reduction. The complete set of predictors, one for each wavelength, is converted to a set of new variables ("latent variables") constructed using linear combinations of predictors. The variable

to be predicted (sucrose concentration in Brix) is then regressed against these latent variables in N-dimensional space, where N is the maximum number of latent variables considered. A more detailed description of the PLS method is provided in Appendix C.

A PLS regression algorithm is provided in the `scikitlearn` Python module as a callable object. This was used to create a function that would perform a basic PLS regression on a given spectral dataset `X` and a variable to be predicted `y`, with a specified number of latent variables `n_comp`. The implementation of this function was based on the tutorial provided by Pelliccia (2018), and the code is shown in Appendix A. To summarize, the function first fits a PLS regression model to the data with the given number of latent variables, then determines the PLS-predicted `y` values ("calibration"). The R^2 coefficient of this calibration model is then calculated. Cross-validation is then performed, and the cross-validation score is used to generate an R^2 coefficient for the PLS model with respect to the original ("measured") `y` values. Linear regression is performed on the predicted `y` values from cross-validation against the measured `y` values, and this line is plotted on a graph.

To reduce the quantity of data considered even more when building a prediction model, a common add-on to PLS is variable selection. Not every predictor (absorbance series at a particular wavelength) will have strong correlation with sucrose concentration, and so a variable selection process would further reduce the amount of data considered by discarding predictors that do not improve the PLS model. This is done by discarding one predictor (absorbance series at one wavelength) at a time when running PLS regression, calculating the mean squared error (MSE) in cross-validation, and keeping the predictor within the dataset only if it improves the MSE of the PLS regression model.

In this manner, the final predictive model would be a function of only latent variables constructed using the strongest predictors in the dataset. The implementation of variable selection is provided in Appendix B, and is also based on the same tutorial (Pelliccia 2018).

3. Sensor Selection

The future end-product of this project is a custom handheld device programmed with the predictive model. This device would be constructed using custom hardware components, i.e. light sources and sensor modules. As noted earlier, if the pulse oximetry algorithm ends up sufficient, hardware design would require simply a broad-spectrum NIR photodetector (e.g. NIR photodiode) and two LEDs each with peak emission at one of the predictor wavelengths. If a greater number of predictors is necessary, hardware design would include a broad-spectrum NIR light source as well as a spectrophotometer module or integrated chip, accessible via hardware interface (e.g. SPI, I^2C).

Hardware development is outside the scope of this project. However, within the scope, a spectrophotometer was needed for experimentation to develop the prediction model, i.e. to collect spectral data on sugarcane samples and explore a proof-of-concept optics setup. Consequently, a handheld spectrophotometer was selected: LinkSquare NIR (Stratio Inc). It is important to note that this is not merely a spectrophotometer module, but a whole product. This is to say, the LinkSquare NIR is a handheld spectrophotometer comprising a built-in spectral sensor, light sources, onboard processing, and WiFi connectivity. Data is transmitted to a computer application via WiFi, using the LinkSquare NIR as a local access point (local AP).
A visual of the LinkSquare system is shown in Figure 9.

Figure 9. Overview of the Stratio® LinkSquare. The handheld is shown on the right, while the laptop in the background shows the graphical interface of the LinkSquare application. Note that the LinkSquare application is also available on smartphones.

The LinkSquare NIR was chosen for its compact form factor and its spectral sensitivity range that lined up well with the 700nm-1000nm range identified in a previous section. The LinkSquare NIR has a spectral sensitivity range of 704nm-1088nm. While there are other handheld spectrophotometers with similar ranges (e.g. Consumer Physics ® SCIO), the final selection was made based on another Cornell laboratory's positive experience with using the LinkSquare NIR.

4. Previous Work

Several studies have explored methods to predict sucrose concentration using near-infrared spectroscopy, to varying degrees of accuracy.

Nawi et al (2013) showed that a model based on Partial Least Squares (PLS) regression on reflectance NIR measurements of sugarcane stalk cross-sections was able to predict Brix values with a coefficient of determination (R^2) of 0.87. The same team was able to apply a similar model

to measurements taken through the whole stalk (i.e. through the rind), and demonstrated an R^2 of 0.90 (Nawi et al 2014).

Omar et al (2012) used a model based on Multiple Linear Regression to predict sucrose concentration in sucrose-water solutions using five predictors: 730nm, 830nm, 915nm, and 960nm. This model had an R^2 of 0.985. While this indicates high prediction accuracy, the experiment was done on sucrose solutions and not whole sugarcane, and such a model will not be pursued since it does not take the optical properties of sugarcane into account. However, the wavelengths used will be noted as a reference for the model developed in this project.

Some evidence exists that a dual-wavelength approach (like the pulse oximeter algorithm described above) may produce a strong prediction model. For example, Tang et al (2016) found that on solutions of glucose in aqueous phosphate-buffered saline, a linear regression model based on voltage responses from a custom sensor at two wavelengths (1450 and 1650 nm) yielded R^2 coefficients between 0.8359 and 0.9973.

5. Optics Assembly

5.1. Overall Construction

To collect spectral data using the LinkSquare NIR, an optics assembly was created. The primary component of the assembly was a purpose-built box, designed to block ambient light from affecting the spectral measurement, and to ensure a consistent distance and angle between the spectrophotometer head and the sample. Such a box is a common feature in NIR spectroscopy studies of sugarcane (Nawi et al 2013) (Nawi et al 2014) (Lazim et al 2016). Nawi et al (2013), Nawi et al (2014), and Lazim et al (2016) all constructed a 900mm × 600mm × 450mm black box, though the material was not specified.

For initial tests with table sugar, a cardboard shoebox was used. This shoebox measured 330mm × 190mm × 100mm. A sample holder (clear plastic Tupperware container) was placed along

one of the shorter walls, and a piece of white notepad paper was placed underneath it. Figure 10 shows this setup.

Figure 10. Initial optics assembly (shoebox setup for first table sugar test). From left to right, the following components are visible: halogen bulb, white notepad paper, plastic container, and LinkSquare. Note that there is a square slot (not shown) on the right side of the box, into which the LinkSquare is pressed to take a measurement. Also not shown: shoebox lid, placed over the box before every measurement

This was the setup used to collect initial data on table-sugar solutions. Importantly, this initial box did not have a holder for the LinkSquare, and simply relied on the user to hold the device at a consistent position and angle when manually triggering measurements. This was a source of variation that the next version of the box addressed-- it was later discovered that the computer application could be used to trigger the LinkSquare, and so the next version of the box fully enclosed and fixed the device within the box.

The box was then revised with a few modifications: a) more firmly fix the positions of the light source, sample, and LinkSquare using holders for each, b) darken the box's interior surfaces so that light reflections off the box walls are minimized, c) add provisions for inserting lenses between the light source and sample, and d) increase light incident on sample by fitting light source (bulb) into a parabolic fixture.

To create this revised box, each dimension from the original box in the studies cited above was scaled down by about two-thirds, resulting in an overall size that was 3% of the original box. The final dimensions of the box were 350mm × 190mm × 140mm. This was done to render the box highly portable between my off-campus apartment and the on-campus laboratory. To darken the interior walls, the box was constructed out of six panels cut from black foam core (Elmer's Foam Boards, 11×14 Inches, Black #950024) and then five of the panels were bound together with duct tape to create an open box. The sixth panel served as a removable lid. Holders were designed in computer-aided design software (Fusion 360) for the light source, sample, and LinkSquare. They were then 3D-printed out of PLA. In addition to these holders, to allow for the addition of lenses, custom lens holders were also designed and 3D-printed out of PLA. All holders were designed to slot into a 12-inch MakerBeam (makerbeam.com, 10mm by 10mm cross-section). Figure 11 shows the virtual model of the overall arrangement of these holders.

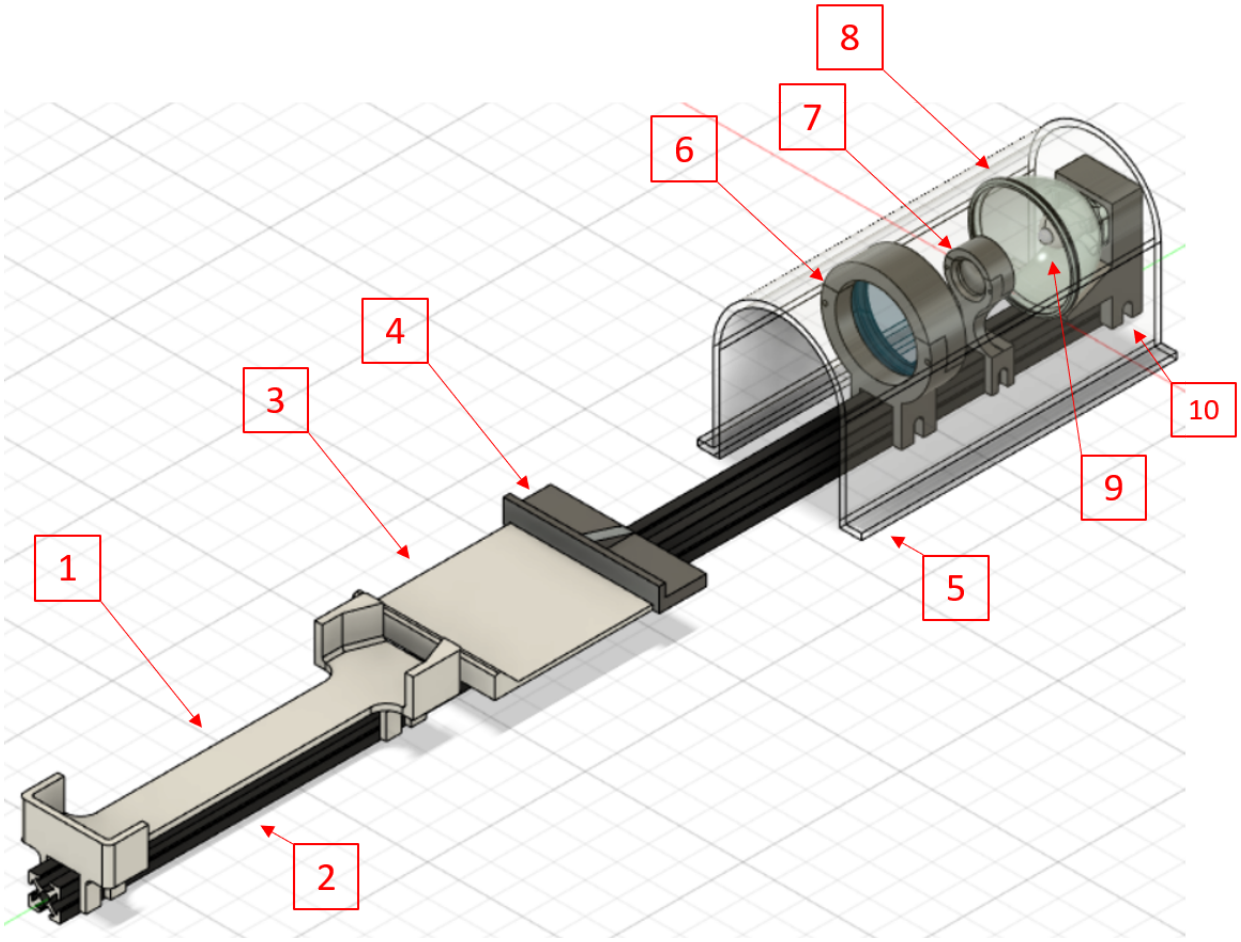


Figure 11. Internal optics assembly within box. In sequence, the components are: 1) LinkSquare holder, 2) MakerBeam, 3) sample holder 4) sliding clamp (secured to #3 by a binder clip, not shown), 5) Tunnel (interior lined with aluminum foil), 6) large planoconvex lens and holder, 7) small planoconvex lens and holder, 8) parabolic fixture for halogen bulb, 9) halogen bulb, 10) bulb fixture holder.

The internal optics assembly shown in Figure 11 was then 3D-printed and installed into the box incrementally. First, it was assembled, and then modified: one end of the Makerbeam was bolted to a black foam core panel to prevent translatory motion of the assembly at that end. Figure 12 shows the result.

Figure 12. Optics assembly bolted to black panel. Note the presence of the binder clip in the center of the image, used to secure the sliding clamp to the sample holder.

This assembly was then placed in the black foam core box. To further restrict motion of this assembly within the box, two foam inserts were created and placed at the back of the black panel shown in Figure 12. This ensured the non-bolted end of the Makerbeam was flush with the side of the box. Another two foam inserts were created and attached to the other side of the box interior with double-sided mounting tape, so as to prevent the non-bolted end of Makerbeam from sliding across the bottom surface of the box. See Figure 13.

Figure 13. Optics assembly securely fixed within box

Foam inserts were used to fix the optics assembly in this way to allow the assembly to be removed easily, while staying in position when in the box.

Two female jumper wires were attached to the terminals protruding from the back of the bulb fixture. The other ends of the wires were attached to the screw terminals of a DC barrel jack, which was secured to the exterior of the box with mounting tape. See Figure 14.

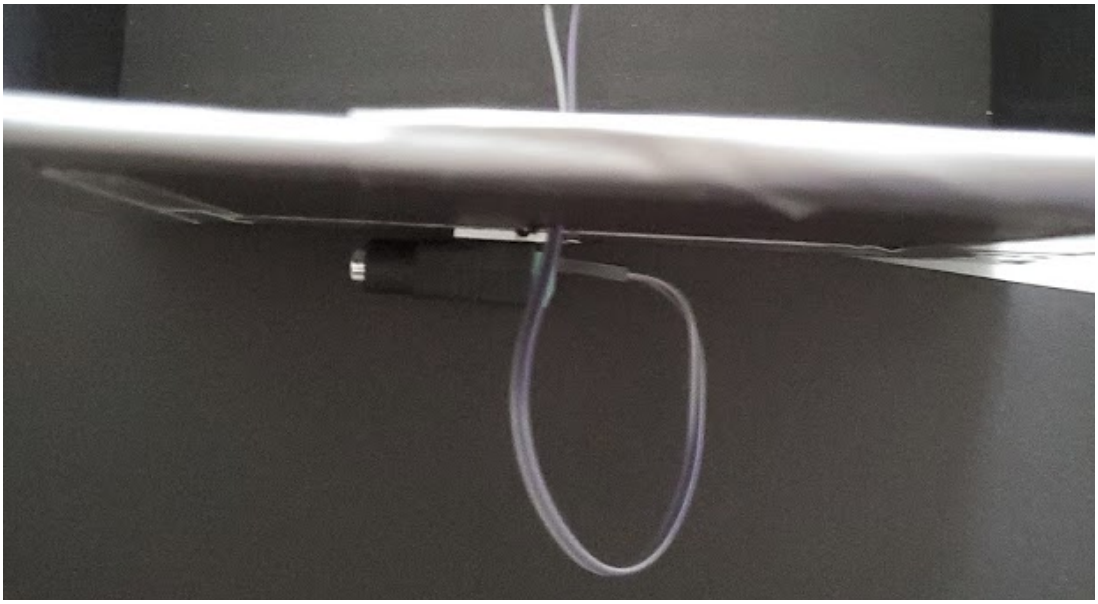


Figure 14. Detail of wire attachment to exterior DC barrel jack

This box containing the internal optics assembly was the test setup used for most of the experiments on whole sugarcane. For the final set of experiments, the lens assemblies were removed (Items 6 and 7 in Figure 11), as well as the sample holding components (Items 3 and 4 in Figure 11). The samples were simply placed directly on the Makerbeam instead.

5.2. Light Source & Light Path

The LinkSquare contains two built-in light sources: an LED and a halogen bulb. Both are designed to be used in reflectance mode, where light emitted by one (or both) of these sources reaches the sample and a portion of the light is reflected back into the LinkSquare's photodetector, located next to the light sources.

The very first attempts to obtain data from the LinkSquare were made in reflectance mode. As an initial test of whether the LinkSquare could capture sufficient data on sucrose solutions, the LinkSquare was placed with its sensor head in contact with the side of a clear plastic container filled with sucrose solutions of various concentrations. The experimental setup in Figure 10 was used, except without the external light source shown. Readings were taken with each of the

two light sources. However, the LinkSquare did not capture any observable signal with either of the built-in sources, even with exposure set to the maximum level. Figure 15 provides examples of this lack of signal with either light source.

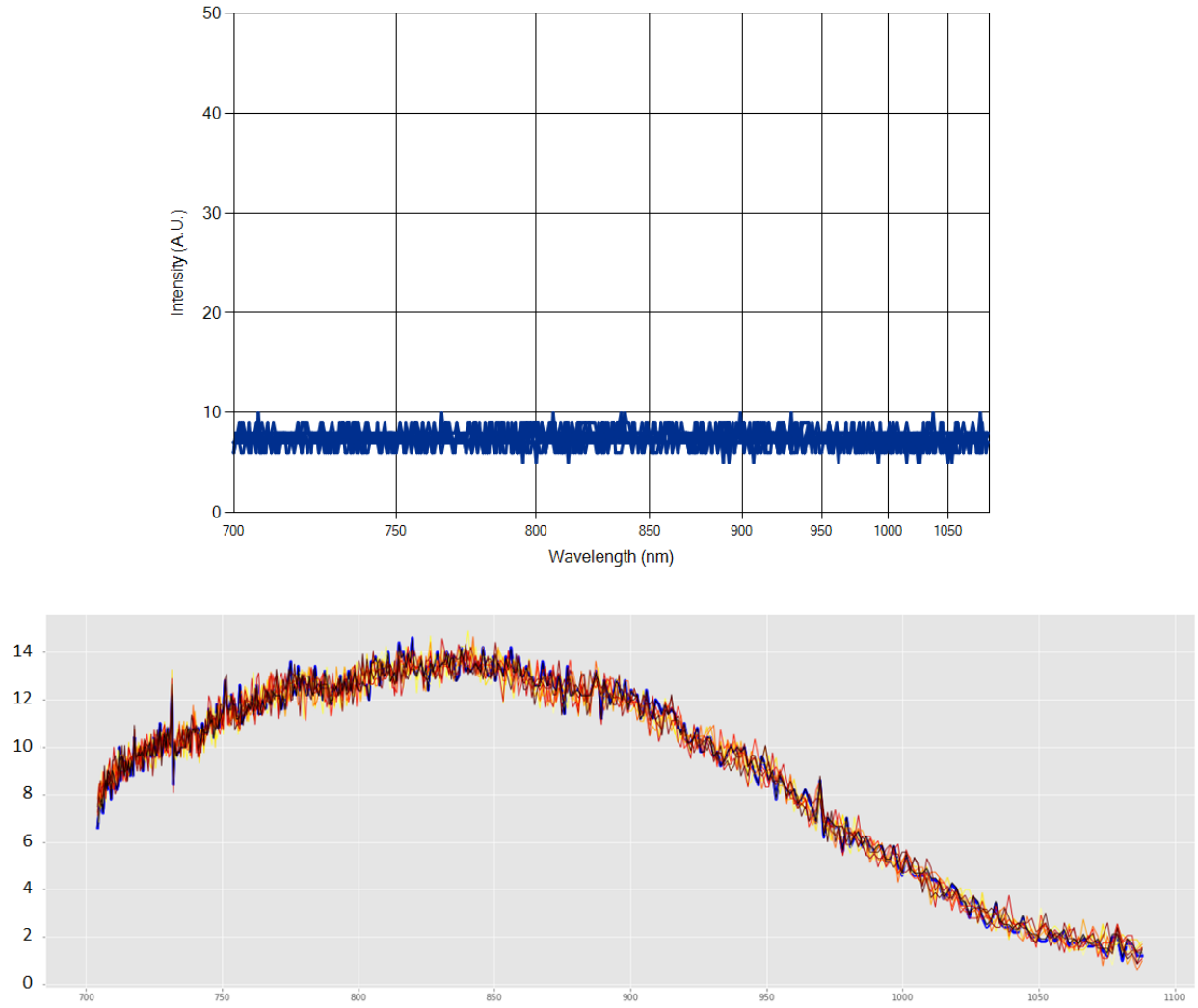


Figure 15. (Top) measurement on concentrated sugar solution (25.5° Brix) with built-in LED and LinkSquare in reflectance mode. (Bottom) measurement on sucrose solutions of various concentrations (10.3°-25.5° Brix), with the blue trace representing a solution of zero concentration.The color coding assigns lower Brix values "yellower" hues, and high Brix values "redder" hues. The blue curve represents the control solution (0° Brix). The vertical axis for both plots have the same units (intensity in the LinkSquare's arbitrary units AU).

According to the LinkSquare NIR datasheet, the minimum detectable optical power is 8 AU. Therefore with the built-in LED on, the measurement of ~8 AU across all wavelengths indicates that the reflected light is of too low power to be detected by the LinkSquare. With the built-in halogen bulb on instead, there is not much improvement. While a signal >8AU is visible, the peak absorbance observed is only around 14 AU, and the variation in absorbance for each wavelength, across all sucrose concentrations, generally appeared to span only within 2 AU. With such small variation, discerning clear correlations to build a robust prediction model seemed unlikely.

Consequently, to amplify the signal, the decision was made to use an external light source, i.e. use the LinkSquare in transmissive mode instead. In this mode, light from the external light source would pass through the sample, and the portion of transmitted light appearing at the other side of the sample would be measured.

Lab-grade transmission spectrophotometers typically use halogen bulbs, and so two halogen bulbs were purchased. Both were Xenon-tungsten bulbs rated for laboratory diagnostics applications. The first was a 10W 150-lumen bulb (OSRAM 64223 HLX). The second was a 20W 335-lumen bulb (OSRAM 64258 HLX).

Measurements were taken with both bulbs separately. It was determined that the 20W produced substantial heat, melting its holder and being painful to the touch even after being turned on for only a few seconds. At the same time, it was also determined that the 20W bulb did not substantially increase the signal, compared to the 10W bulb. Therefore the 10W bulb was used for all self-performed experiments described in the next section. Further details on bulb testing and selection are provided in Appendix D.

6. Experimental Results

6.1. Dry Mixture

Various preprocessing and prediction techniques have been described above. Before using these techniques on firsthand experimental data, these techniques were first applied to an existing dataset. This provided a means to explore and optimize the implementations, parameters, and limitations of these techniques on representative data, before getting into experiment design on real sugarcane samples. I was not very familiar with Python, especially the graphing tools in *matplotlib*, so this also served as an introductory exercise in using those tools.

A spectral dataset of sugarcane samples is provided by Chaix (2020). The dataset consisted of 480 NIR spectra representing 60 samples of dried ground-up sugarcane matter, and each sample is read by eight off-the-shelf spectrophotometers. Total sugar content (i.e. weight-percent of dry matter) in the samples ranged from 1.1 to 51.0. The properties of the samples are described in Figure 16.

Summary statistics of chemical properties of sugarcane samples used for calibration.

Chemical properties	Unit	Min	Max	Mean	SD
Total sugar content	%	1.1	51.0	23.4	17.3
Crude protein content	dry	0.9	9.6	3.1	2.1
Acid detergent fiber fraction	matter	26.0	59.3	39.2	8.7
In vitro organic matter digestibility		13.0	66.6	41.0	15.1

Figure 16. Statistics on properties of dry ground sugarcane samples. Properties include total sugar content, crude protein content, acid detergent fiber fraction, and in vitro organic matter digestibility

We can see that the mean total sugar content occurs approximately at the midpoint between maximum and minimum. The distribution of the data was plotted on a histogram, shown in Figure 17.

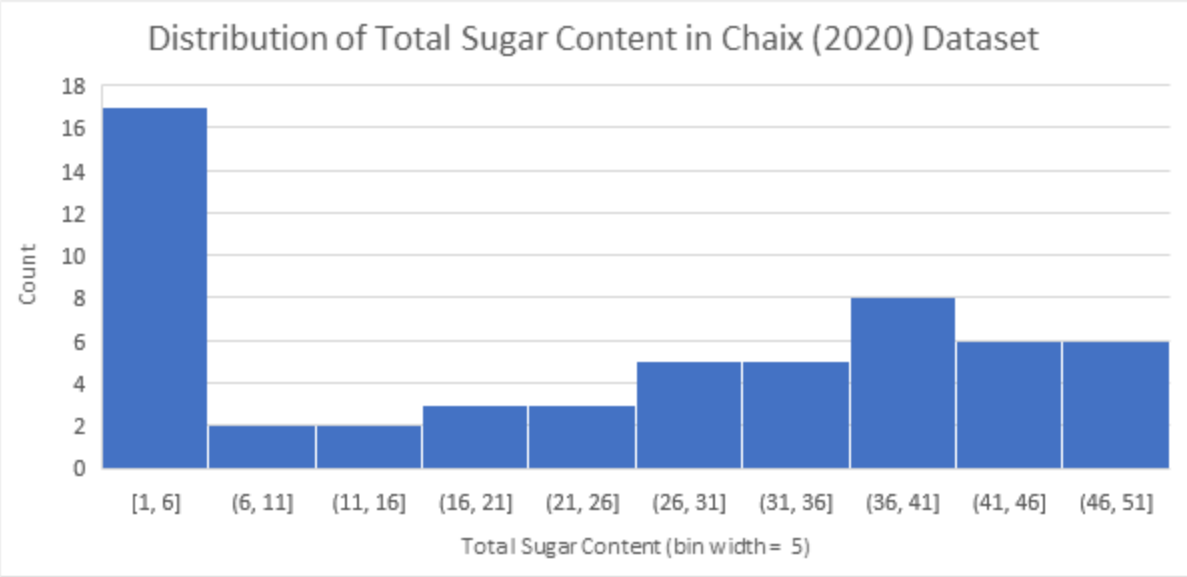


Figure 17. Histogram of total sugar content values in the Chaix (2020) dataset. Bin width is set to 5, and bin endpoints are shown to one decimal place.

We can see that the distribution of total sugar content values shows left- and right- skew -- notably values in the [1,6] bin are highly overrepresented compared to a normal distribution. That said, more than one count is present for each bin. The overrepresentation of small values with the visible right skew beyond [1,6] may be a limitation of this dataset. Any resulting prediction model may be overly reliant on low and high values, and may not be sufficiently accurate for middle values of total sugar content. However, if a linear prediction model is used, prediction accuracy at the high and low extremes would itself lead to prediction accuracy at the middle values.

Another limitation of this dataset is that each sample consisted of dried ground-up matter from a mix of cuttings of leaves and stalks. Whole sugarcane is not as optically homogeneous in structure, and does not contain parts relatively low in sugar content like leaves, roots, etc. Sugar in sugarcane typically concentrates in the stalks. Additionally, scattering/reflective properties will likely be different when light passes through the various layers (rind, flesh, etc). Another

limitation is that total sugar content (TS) is not the same metric as Brix, since TS uses total dry mass as the denominator, while Brix uses the mass of sucrose solution as the denominator.

All that said, several analyses were explored on the data. Out of the eight spectrophotometers, the SCIO (Consumer Physics) was used due to its spectral range aligning with the 700nm-1000nm waveband identified above. See Figure 18 for a comparison among the devices.

Specific characteristics for NIR spectrometers.

Device and manufacturer	Spectral range (nm)	Resolution (nm)	Technology	Lighting module	Weight
LabSpec 4 (ASD)	350–1000 1001–1800 1801–2500	3 @ 700 nm 10 @ 1400/2100 nm	Silicon array InGas photodiode array InGas photodiode array	1 halogen lamp	5.44 kg
NIRscan Nano (Texas Instrument)	901–1701	10	1 photodiode InGaAs	2 halogen lamps	85 g
F750 (Felix Instrument)	450–1140	8–13	Diode array	1 xenon tungsten lamp	1.05 kg
MicroNIR1700 (Viavi)	908–1676	6	InGas photodiode array	2 tungsten lamps	<60 g
MicroNIR2200 (Viavi)	1158–2169	8	1 photodiode InGaAs	2 tungsten lamps	<60 g
NIRONE 2.2 (Spectral Engines)	1750–2150	20–26	InGaAs photodiode array	2 tungsten lamps	15 g
SCIO (Consumer Physics)	740–1070	Not communicated	2 photodiodes	1 LED lamp	35 g
TellSpec (TellSpec)	900–1700	10	1 photodiode InGas	2 halogen lamps	136 g

Figure 18. The eight spectrophotometers in Chaix (2020) and their properties summarized.

Note that the SCIO is the only device with spectral sensitivity range (740nm-1070nm) close to the waveband of interest (700nm - 1000nm). The other salient attribute of the SCIO is that it is a handheld device--this and its spectral range make it the most similar to the LinkSquare NIR out of all the spectrophotometers shown here.

The dataset is provided in tabular text format, similar to comma-separated values format, but with semicolons as delimiters instead. As a first exploration of visualizing the data, a Python script was written, utilizing the pandas module to import the dataset as a dataframe, and then the tools within matplotlib were used to plot the data. A basic colormap was used based on a yellow-orange-red color continuum, with hues closer to yellow representing lower TS values

and hues closer to red representing higher TS values. This initial visualization served as a first foray into plotting using Python, and is shown in Figure 19.

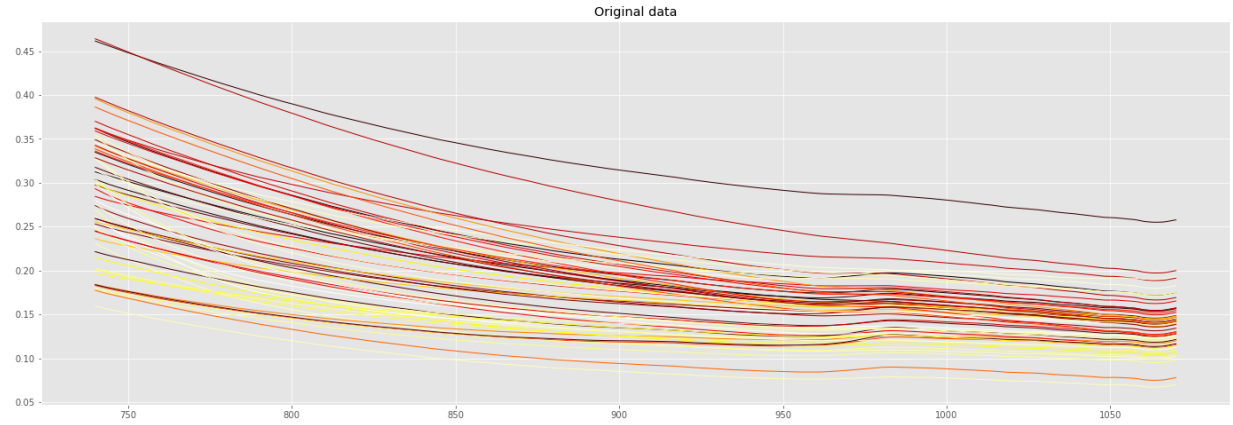


Figure 19. Raw spectra from 60 samples from SCIO spectrophotometer in Chaix (2020) dataset. Total sugar content values are represented in a warm color palette, with lower values in yellow and higher values in deep red.

It can be observed that the contour of each curve appears smooth, with no visible high-frequency elements. This suggests that the SCIO meter applies some smoothing before data output. Additionally, no trends are obvious when considering the relationship between absorbance and total sugar content.

An implementation of SNV was created, as described previously in this report. A code snippet showing the SNV function is shown below:

```
def snv(input_data):
    ''' Perform Standard Normal Variate correction'''
    output_data = np.zeros_like(input_data)
    for i in range(input_data.shape[0]):
        # Apply correction
        output_data[i,:] = (input_data[i,:] - np.mean(input_data[i,:])) /
np.std(input_data[i,:])

    return output_data
```

Here `input_data` and `output_data` are absorbance spectra (spectral datasets) in `pandas` dataframe format. This function was applied to the SCIO dataset, and then a Savgol filter (window size 15, polynomial order 3) was applied on the data. Figure 20 shows the result.

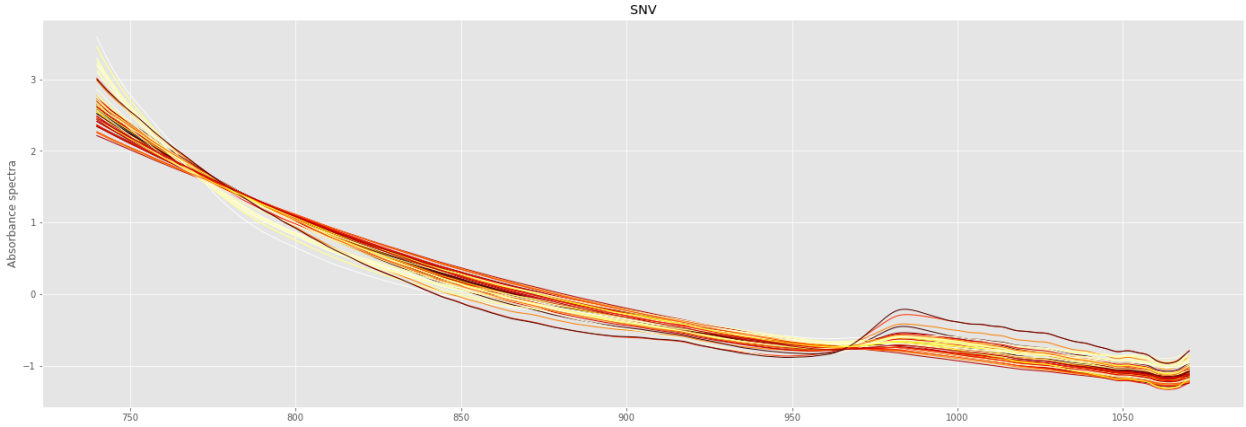


Figure 20. SCIO dataset after SNV and Savgol filtering

We can see that the variation in the spectra is generally reduced. No correlations between absorbance and total sugar content are visually clear, still. To determine correlations quantitatively, linear regression was performed on the absorbance series at each wavelength against the total sugar content values. The linear regression coefficients (R^2) were then plotted against wavelength, as shown in Figure 21.

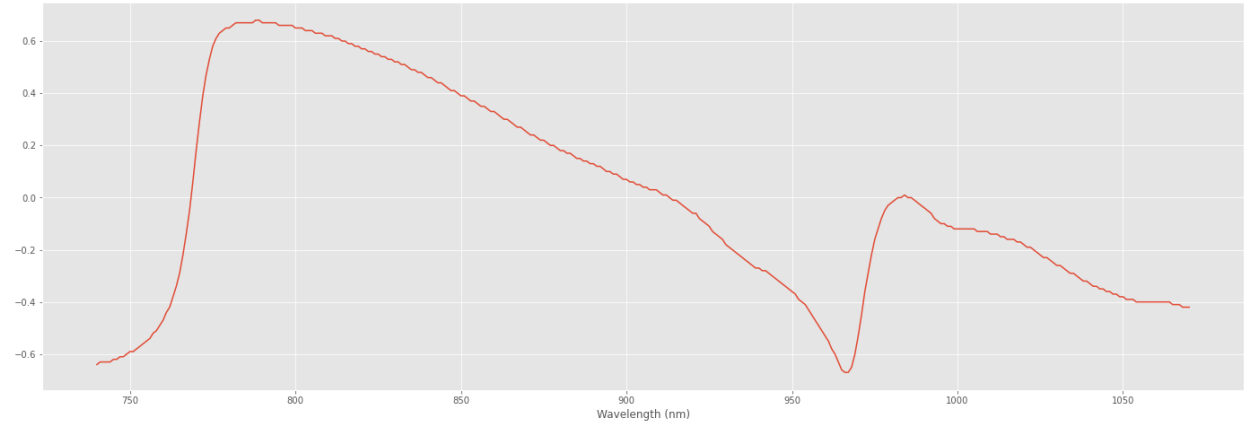


Figure 21. Regression coefficients on SCIO dataset after SNV

The most positive regression coefficient occurs at 786nm (R^2 = 0.68) and the most negative occurs at 967nm (R^2 = -0.67). Notably, 786nm is close to 770nm, the third overtone of O-H

stretching vibrations (see Figure 3). Also 967nm is close to 960nm, the second overtone of O-H stretching vibrations. While the single-predictor regression coefficients are not very high, the maximum and minimum occur close to expected high-correlation wavelengths. Oddly, even though both wavelengths correspond to O-H stretching, one wavelength shows negative correlation with total sugar content while the other shows positive correlation.

The pulse oximeter algorithm was applied on the data, with these two highest-coefficient wavelengths. First the ratio of the absorbance at 786nm against the absorbance at 967 was computed for each total sugar content (TS) value. TS values were then regressed against the ratios. The results are shown in Figure 22.

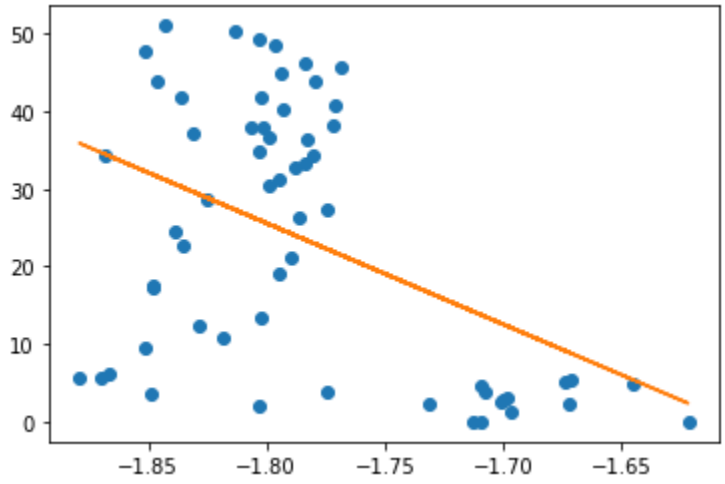


Figure 22. Least-squares regression of TS values against ratios (Abs@786nm/Abs@967nm) in SCIO dataset. The vertical axis represents TS values and the horizontal axis represents ratios. Regression coefficient (R^2) was -0.455.

We can see that the regression coefficient with this method is very low, at -0.455, and so the pulse oximetry algorithm is not suitable for analyzing this dataset.

Next, the simple PLS algorithm was applied to the dataset. The number of latent variables was incremented from 1 through 14, and the calibration and cross-validation R^2 coefficients were calculated for each number of latent variables. These calculations are plotted in Figure 23.

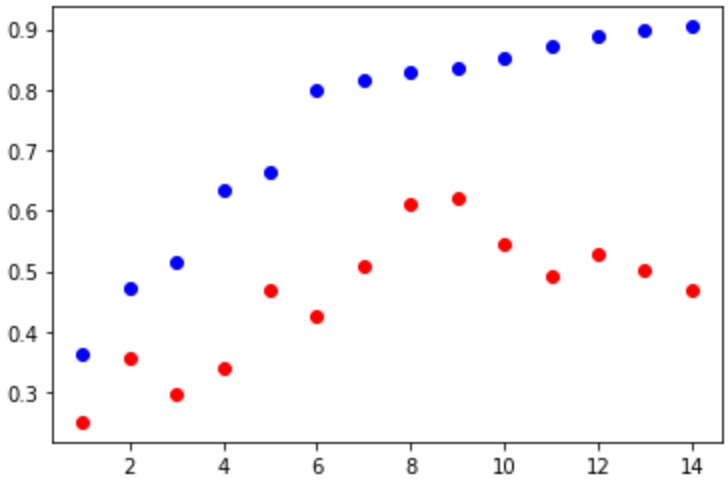


Figure 23. Regression coefficients for PLS calibration (blue) and cross-validation (red), against number of PLS latent variables

While the PLS model shows an improving fit with the dataset as the number of latent variables increases, the cross-validation performance decreases as the number of latent variables increases beyond 9. Therefore the optimal number of latent variables is 9, where the PLS model has a cross-validation score of 0.62. This is not high enough for the target specification of this project (0.9), but it is close to the cross-validation score of 0.69 in Chaix (2020)'s own PLS regression analysis of the SCIO dataset. This suggests that the implementation of PLS used in this particular project is competitive with other PLS implementations in literature.

The cross-validation score is represented as the regression coefficient of a line regressed on cross-validation TS values against the original measured TS values, as shown in Figure 23.

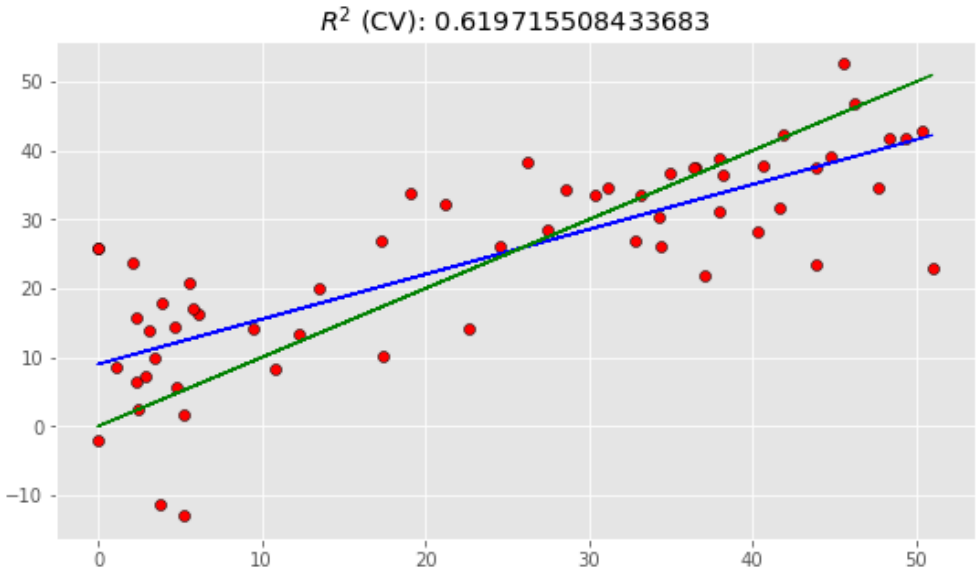


Figure 23. Linear regression of PLS-predicted TS content (vertical axis) against original measured TS content (horizontal axis). Green line represents the ideal behavior (predicted TS = original TS), and blue line represents the real behavior of predicted TS as a function of original TS.

Next, the PLS variable selection implementation in Appendix B was applied, to see if this would improve the cross-validation score. The result is shown in Figure 26.

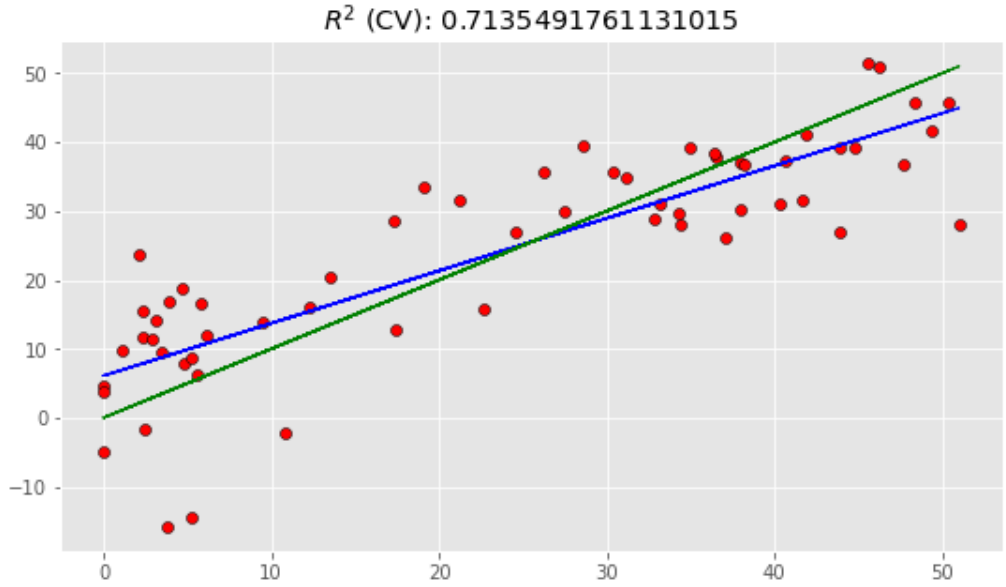


Figure 26. Linear regression of PLS-predicted TS content (vertical axis) against original measured TS content (horizontal axis). PLS was performed after variable selection.

The cross-validation score was then 0.71, which is slightly closer to Chaix (2020)'s PLS regression cross-validation score. This supports the idea that the PLS with variable selection implementation used in this particular project is comparable to PLS implementations in literature.

The variable selection algorithm discarded 281 out of 331 wavelengths, visualized in Figure 27.

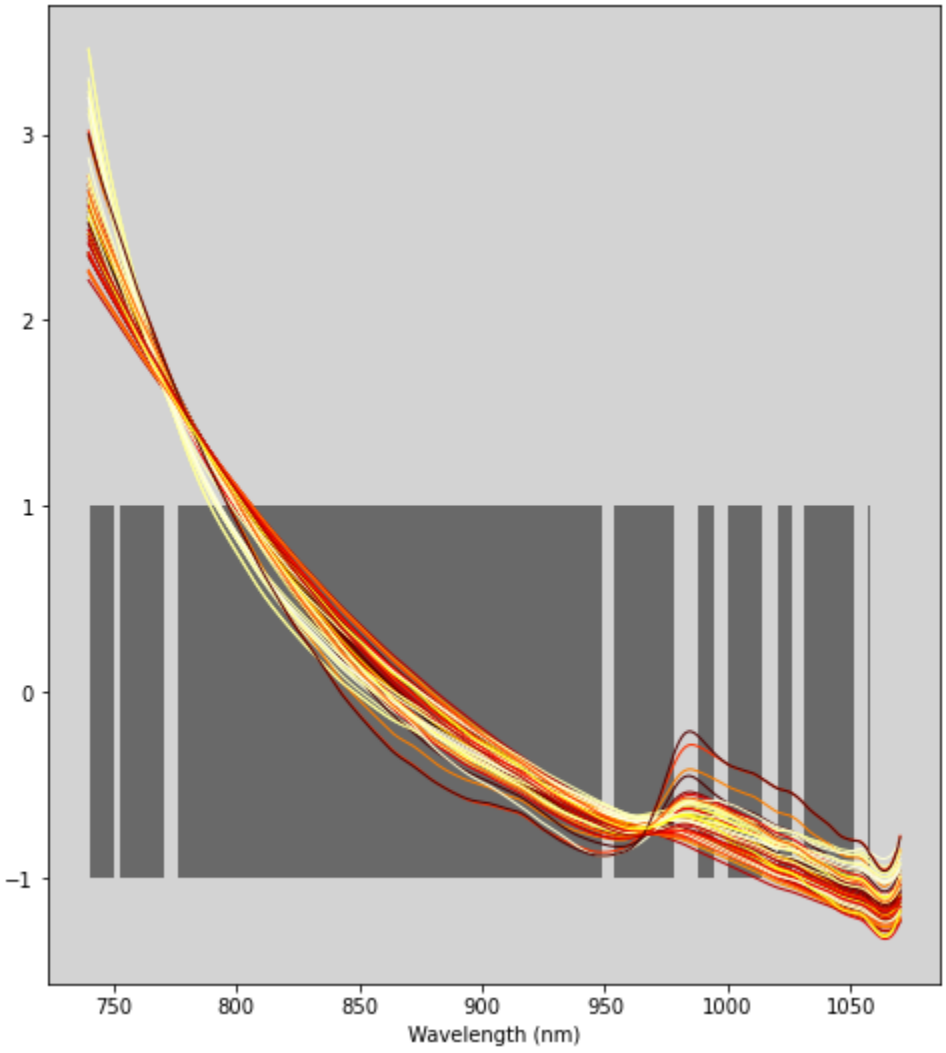


Figure 27. Wavelengths discarded by variable selection algorithm. Vertical axis shown in arbitrary units from after the data was preprocessed with SNV. Discarded wavelengths are shown by black bands.

Therefore, the final PLS regression model consisted of 50 wavelengths, while the number of PLS latent variables remained at 9.

Interestingly, Chaix (2020)'s PLS analyses generally report much higher R^2 values, up to 0.95 and 0.97 on spectrophotometers with ranges further into the IR spectrum, e.g. NIRscan Nano EVM, MicroNIR1700. This might mean that a way of improving the prediction model developed in this project would be to collect spectral data deeper in the IR spectrum. However, first, analysis on the data available in the LinkSquare's spectrum needs to be fully explored.

6.2. Sugar Solutions

With Python dataframe manipulation, plotting using matplotlib, dual-wavelength algorithm implementation, and PLS implementation explored on the Chaix (2020) dataset, experiments were then performed on table sugar (i.e. sucrose) solutions.

Twenty five sucrose solutions were prepared by dissolving known masses of sugar in known volumes of water, with the following Brix values: 10.0, 10.6, 10.6, 11.6, 12.0, 12.1, 12.5, 13.3, 13.6, 13.7, 13.8, 15.2, 15.2, 15.3, 15.9, 16.4, 16.7, 17.0, 17.6, 17.6, 19.2, 21.6, 23.5, 23.9. This range was chosen to approximate the typical range of Brix found in sugarcane (Nawi et al 2013) (Nawi et al 2014). The experimental setup from Figure 10 (initial box) was used, including the 10W 150-lumen halogen bulb. The data was preprocessed with SNV, and then a Savgol filter was applied (window size 35, polynomial order 3). Finally linear regression was performed on the absorbance series at each wavelength, against Brix values of the series at each wavelength. Linear regression coefficients were plotted against wavelength. Figure 28 illustrates the results.

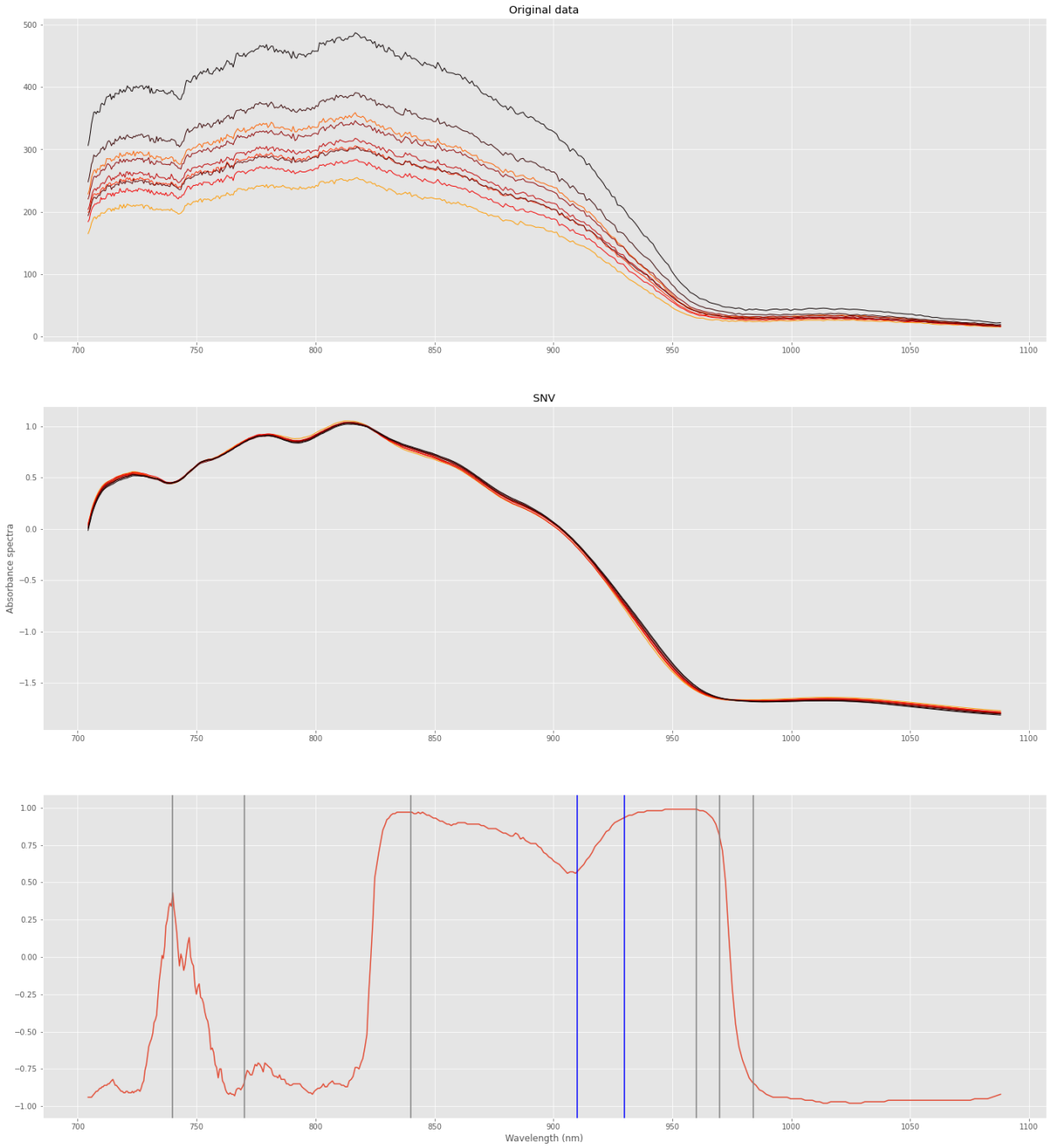


Figure 28. Results from sucrose solution tests in the initial box with 150-lumen bulb. (Top) Original spectral data, (Middle) spectral data after preprocessing with SNV (window size 15, polynomial order 3), (Bottom) Linear regression coefficients at each wavelength. Grey lines represent vibrational frequencies of O-H bonds, and blue lines represent vibrational frequencies of C-H bonds (see Figures 1 & 3).

It is immediately clear that the external light source produces much greater readings. The peak intensity of the original data is about 500AU. To put this in context, this peak intensity corresponds to about 100 microwatts at that wavelength.

The spectra appear smooth and free of visible artifacts after SNV. After the linear regression analysis, it can be seen that the sucrose signal is highly discernible at wavelengths between about 830nm and 970nm. In this region, sucrose is highly positively correlated with absorbance, evidenced by very high regression coefficients that appear stable in that region around 0.9-1.0. This high linearity region captures the C-H vibrational overtone at 930nm, as well as O-H overtones at 840nm and 970nm. Additionally, another stable high-linearity region is observable within 1000nm-1080nm, though with negative correlation. Given the O-H overtone at 984nm (see Figure 3), and the fact that there is a slight increase in water absorbance in the 990-1000nm range (see Figure 2), this latter region is likely to correspond to the proportion of water relative to the other substances in sugarcane.

Since we now have a highly-positively-correlated waveband and highly-negatively-correlated waveband, the dual-wavelength algorithm (pulse-oximetry algorithm) emerges as a promising approach. The noise floor (i.e. zero-level for the signal) is about 8AU, and the most positive correlation occurs at 953nm while the most negative correlation occurs at 1026nm. We can then formulate a ratio to predict Brix as follows:

$$\text{Brix} = \frac{Absorbance@953nm - 8}{Absorbance@1026nm - 8}$$

Regressing the Brix values against these ratios, we get the linear regression model in Figure 29.

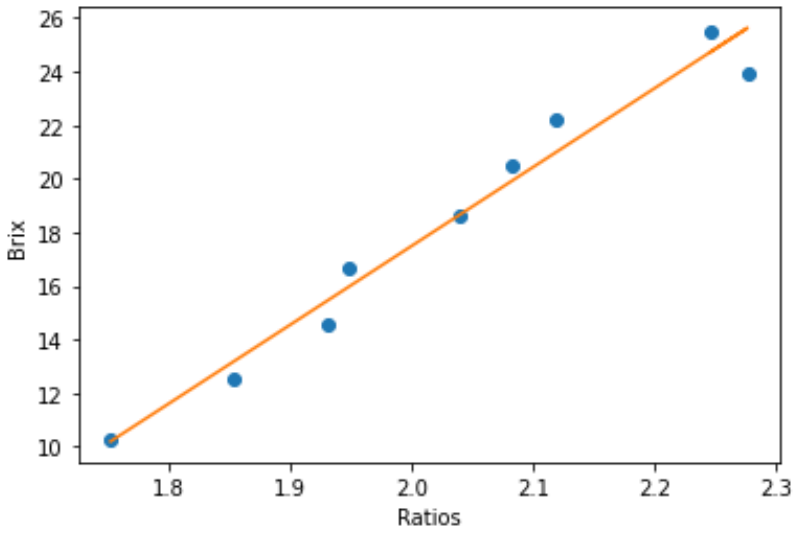


Figure 29. Dual-wavelength linear regression model for sucrose solutions in initial box

The resultant linear regression model fits the data well, with a regression coefficient of 0.98.

A PLS model (with variable selection) was also fit to the data, shown in Figure 30.

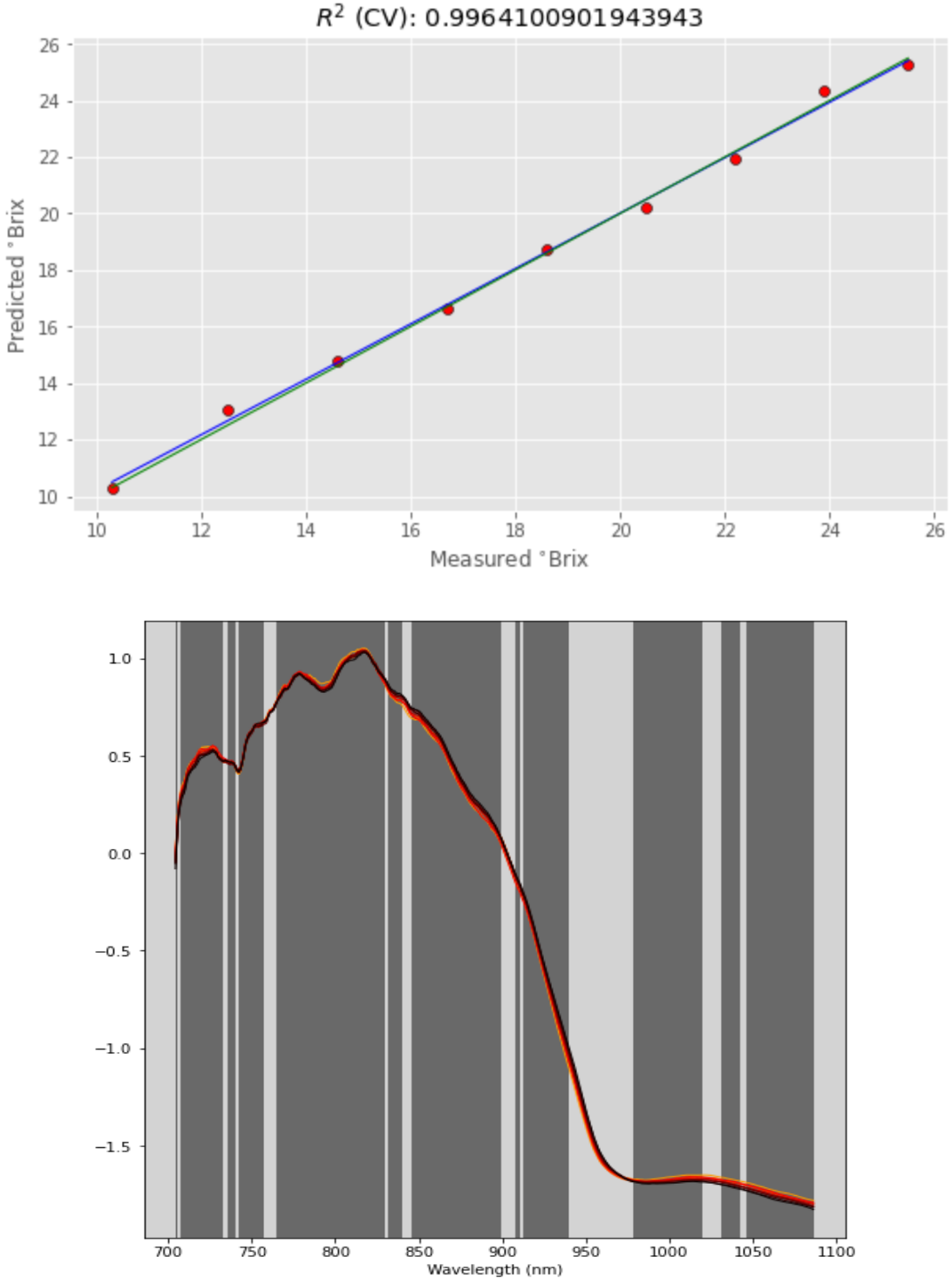


Figure 30. PLS model fit to sucrose solution dataset of initial box. (Top) linear regression of predicted versus measured Brix from cross-validation (n=9, cross-validation set of 5). (Bottom) wavelengths shown by black bands were discarded via variable selection to improve model fit.

The PLS model is able to achieve an R^2 of 0.996 in cross-validation, using just three latent variables and discarding 331 out of 400 wavelengths. We can see that the PLS model preserves wavelengths across the 930nm-960nm range, which converges with the reality that several overtone frequencies for C-H and O-H bond vibrations lie in that waveband.

Once the updated box and optics assembly were developed (Figure 13), a prediction model for sucrose solutions was developed anew. The sample holder and clamp were removed from the box to minimize the distances between light source and sample, and between sample and LinkSquare. Compared to the clear Tupperware container from the previous sucrose-solution test, in this test a small white polyethylene terephthalate container was used to hold the solutions. This container matched the visual whitish appearance of the interior surface of sugarcane rind, and also matched the approximate dimensions of the sugarcane samples obtained during this project (discussed in next section). The container is shown in Figure 31.

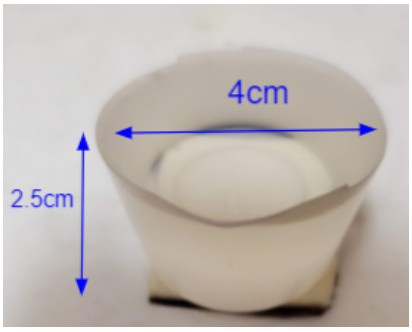


Figure 31. Polyethylene container used in second sucrose-solution experiment. Dimensions as shown are within dimensions observed in sugarcane samples.

Here, 12 sucrose solutions were prepared using table sugar and water. Brix values of these solutions were: 7.2, 10,0, 12.4, 13.1, 14.0, 15.4, 17.4, 18.6, 21.0, 21.5, 23.6, 24.1. As before, this range was chosen to match typical Brix ranges observed in sugarcane.The experimental setup from Figure 13 (final box) was used, in this case including not only the 10W 150-lumen halogen bulb as before, but also a reflective parabolic fixture used to house the bulb. As before, the data was preprocessed with SNV, and then a Savgol filter was applied (window size 35, polynomial order 3). Finally linear regression was performed on the absorbance series at each wavelength, against

Brix values of the series at each wavelength. Linear regression coefficients were plotted against wavelength. Figure 32 illustrates the results.

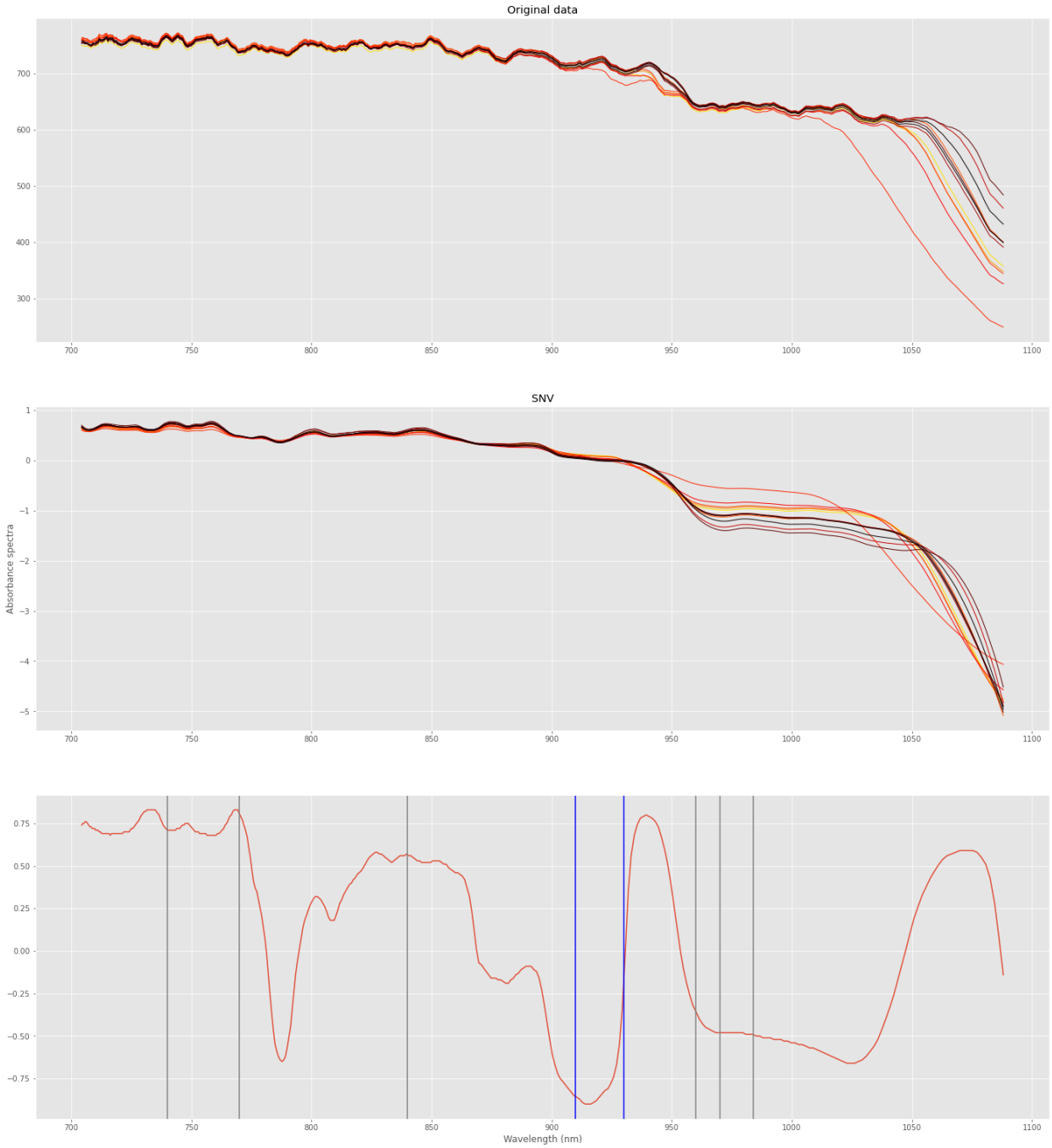


Figure 32. Results from sucrose solution tests in the final box with 150-lumen bulb in fixture. (Top) Original spectral data, (Middle) spectral data after preprocessing with SNV (window size 35, polynomial order 3), (Bottom) Linear regression coefficients at each wavelength. Grey lines represent vibrational frequencies of O-H bonds, and blue lines represent vibrational frequencies of C-H bonds (see Figures 1 & 3).

It is immediately apparent from Figure 32 that the intensities are generally much higher than in the previous sucrose solution test. This is likely due to the thinner walls of the container used in this test, and the increased concentration of light incident on the container from housing the halogen bulb in a parabolic light fixture.The peak intensity of the original data here is about 750AU, which corresponds to about 150 microwatts of power at that wavelength.

The spectra in this test, as before, appear smooth and free of visible artifacts after SNV. The variation in linear regression coefficients over wavelength is much greater here. The wide regions with stable regression coefficients are here broken into narrower regions. Additionally, the R^2 extrema are lower in magnitude here: 0.83 (maximum) and -0.9 (minimum).

The correlations appear markedly different in this data, compared to the previous sucrose solution test. The linear regression vs wavelength plots are juxtaposed for comparison in Figure 33.

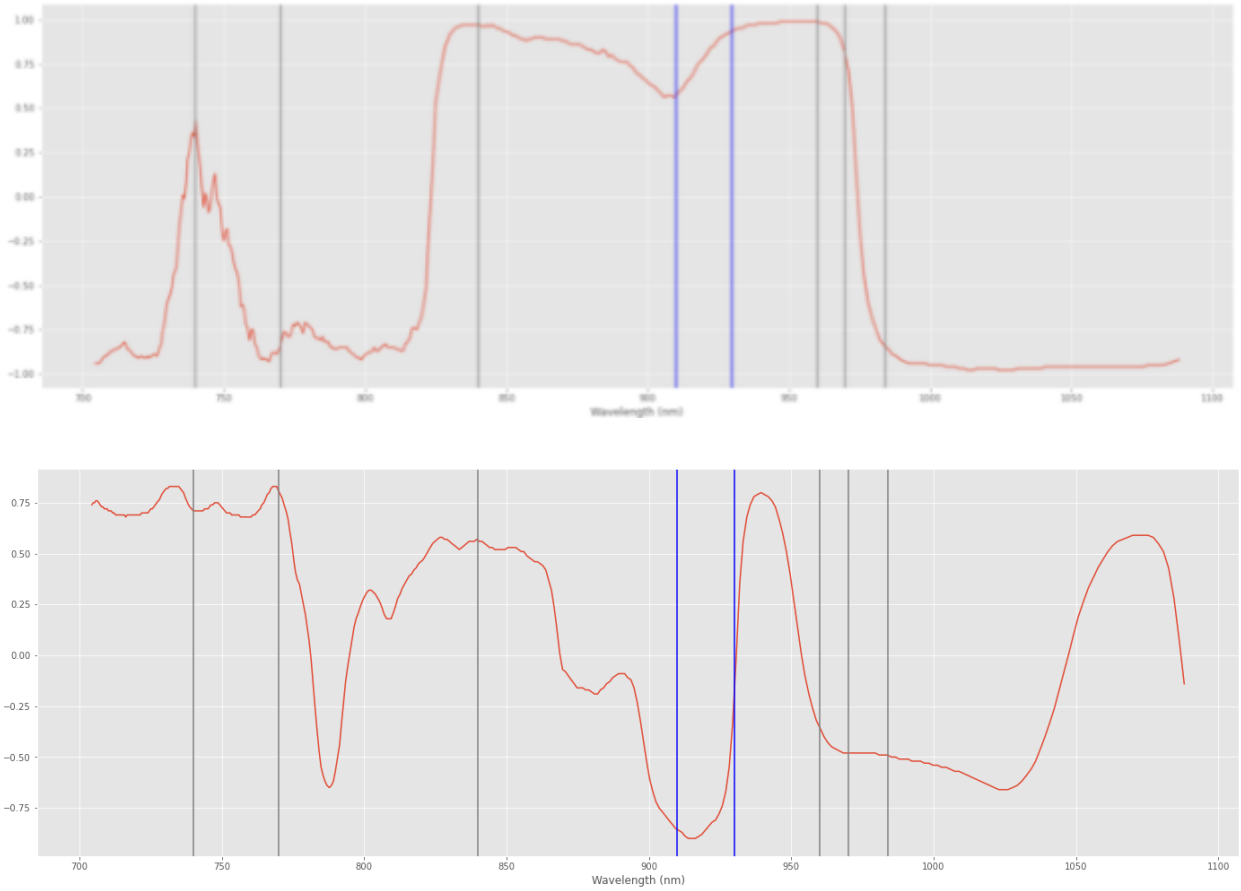


Figure 33. Linear regression coefficients plotted against wavelength for (top) sucrose solution test in initial box and (bottom) sucrose solution test in final box.

We can see from Figure 33 that the weak positive correlation at 740nm from the initial test is amplified to a stronger positive correlation in this test. However the strong positive correlation at 840nm is now weakened. 940nm still appears to be a high-correlation wavelength, and there is a steep dropoff from positive correlation to negative correlation, as seen before, from about 970nm to 1000nm. Strangely, the strongest negative correlation of absorbance with sucrose concentration appears to occur at 910nm, which is typically associated with C-H vibrations in sucrose. This, in addition to the greater fluctuation in the linear regression behavior, could be due to the LinkSquare being near saturation at such high intensity values. This may have caused distortions in the data.

In any case, a similar method as before is able to generate a strong dual-wavelength model. The most positive R^2 is observed at 770nm (a wavelength associated with a O-H stretching overtone), the most negative R^2 is observed at 915nm. Subtracting out the noise floor of 8AU and expressing Brix using the ratio yields:

$$\text{Brix} = \frac{Absorbance@915nm - 8}{Absorbance@770nm - 8}$$

Regressing the Brix values against these ratios, we get the linear regression model in Figure 34.

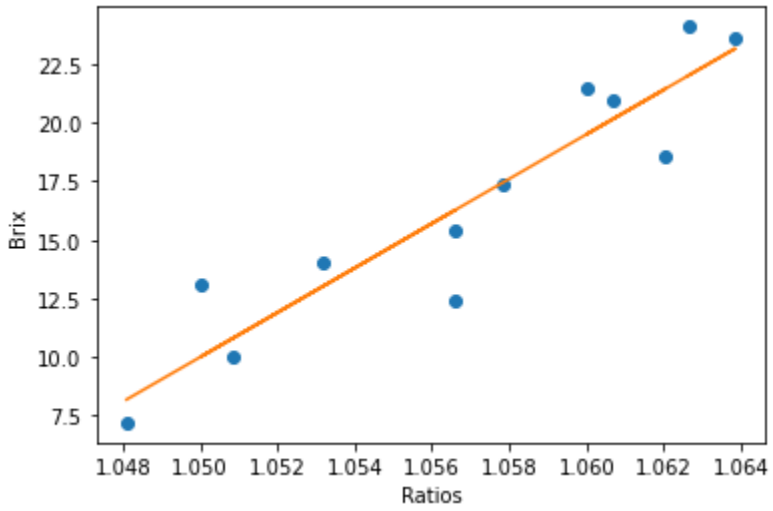


Figure 34. Dual-wavelength linear regression model for sucrose solutions in final box. Regression coefficient is 0.93.

In Figure 4, the dual-wavelength approach produces a linear regression model with an R^2 of 0.93.

A PLS model (with variable selection) was also fit to the data, shown in Figure 35.

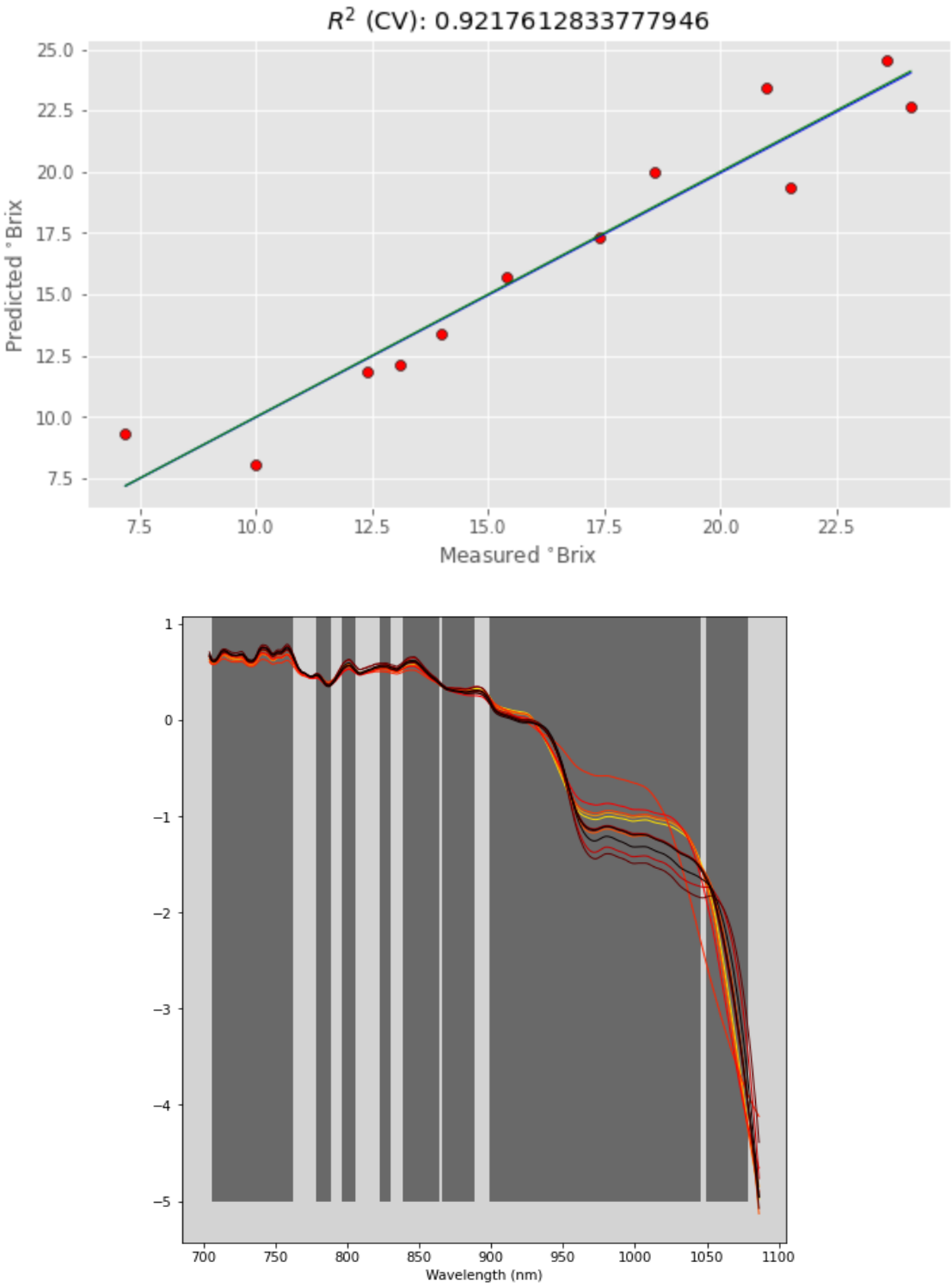


Figure 34. PLS model fit to sucrose solution dataset of finalbox. (Top) linear regression of predicted versus measured Brix from cross-validation (n=12, cross-validation set of 5). (Bottom) wavelengths shown by black bands were discarded via variable selection to improve model fit.

The PLS model is able to achieve an R^2 of 0.92 in cross-validation, using six latent variables and discarding 318 out of 400 wavelengths. We can see that the PLS model here does not preserve wavelengths across the 930nm-960nm range, contrary to the previous test. However, the model uses the 770nm and 840nm wavelengths, both of which are associated with O-H bond stretching vibrations.

6.3. Sugarcane

Now that we had prediction models with regression coefficients greater than 0.9 developed on sucrose solutions, experiments were then performed on sugarcane. A total of six pounds' worth of sugarcane stalk cuttings were obtained from two sources: Prodocshop in Inglewood CA, and Polar Bear Store in Homestead FL. Examples of stalk cuttings from a mix of both sources are shown in Figure 35.

Figure 35. Examples of sugarcane stalk cuttings used for experiments in this project

The diameters of the samples generally varied between 2.5cm and 4.5cm, and the internode lengths varied between 4cm and 12cm. Notably, substantial variation was observed in the color of the stalks, covering shades of yellow, red, green, and brown.

Along with the sugarcane samples, a sugarcane juicer (ALEEHAI Manual Fruit Juicer) was purchased to extract the juice from sugarcane after spectral measurements were taken. With the purchase of a handheld Brix refractometer (aichose SR0014-ATC), this would allow ground-truth measurements of Brix in sugarcane. This way the predictive model could be developed referenced to Brix values from the refractometer. After obtaining the refractometer, an accuracy test was performed on it, using sucrose solutions of known concentrations (see Appendix G) . The maximal error recorded was 3.2%, corresponding to a difference of 0.8° Brix.

The sugarcane stalks were stored in a freezer over the course of the experiments. Prior to each experiment, a number of stalks were taken out, thawed at room temperature for 1-2 hours, and then the condensation was wiped off with a paper towel. Samples 2.5cm in length were cut from the stalks. Each sample was placed in the experimental setup ("final box", see Figure 13), and then the LinkSquare was remotely triggered via computer application to record ten readings. The sample was then crushed with the sugarcane juicer, and 2-3 drops of the juice were applied to the Brix refractometer using a dropper. The refractometer was read, and the value was recorded. This was repeated for each sample. The refractometer head and sugarcane juicer were rinsed and dried between samples.

Before collecting the final dataset, several tests were run to optimize the experimental setup. More details on these tests are available in Appendix F. These tests led to modifications of the experimental setup: removing the lenses, removing the sample holder, lining the interior of the tunnel with aluminum foil (see Figure 11), and the addition of a baffle. All these modifications had the effect of increasing the magnitude of the signal read by the LinkSquare spectrophotometer.

The final dataset consisted of 25 whole internode sugarcane samples, each cut to a length of 2.5cm. Sample preparation was as listed above. Six of the samples were submerged in bags of

room-temperature water for 12 hours to decrease their Brix. More details on this are in Appendix H.

As before, the data was preprocessed with SNV, and then a Savgol filter was applied (window size 35, polynomial order 3). Finally linear regression was performed on the absorbance series versus Brix values at each wavelength. Linear regression coefficients were plotted against wavelength. Figure 36 illustrates the results.

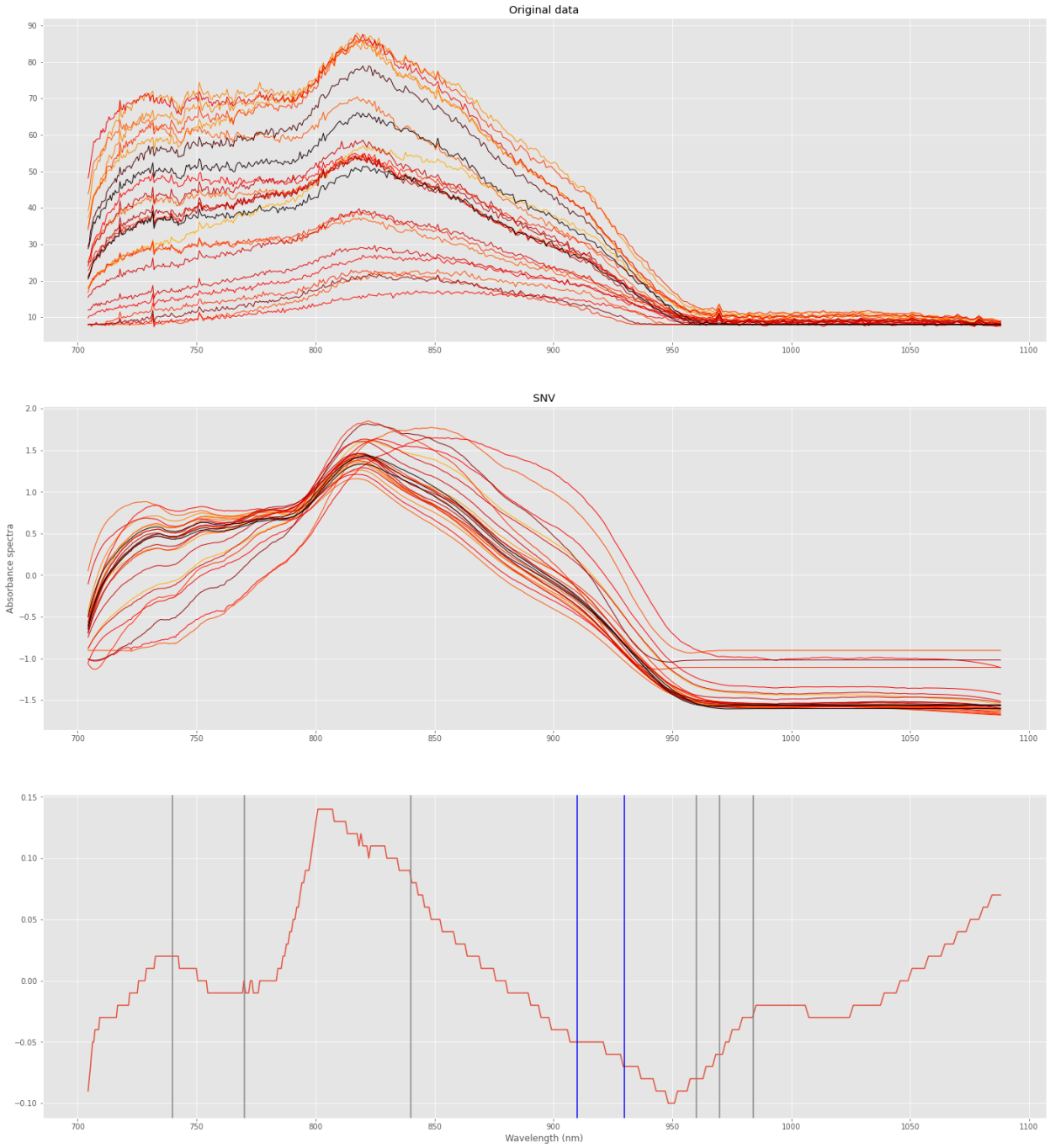
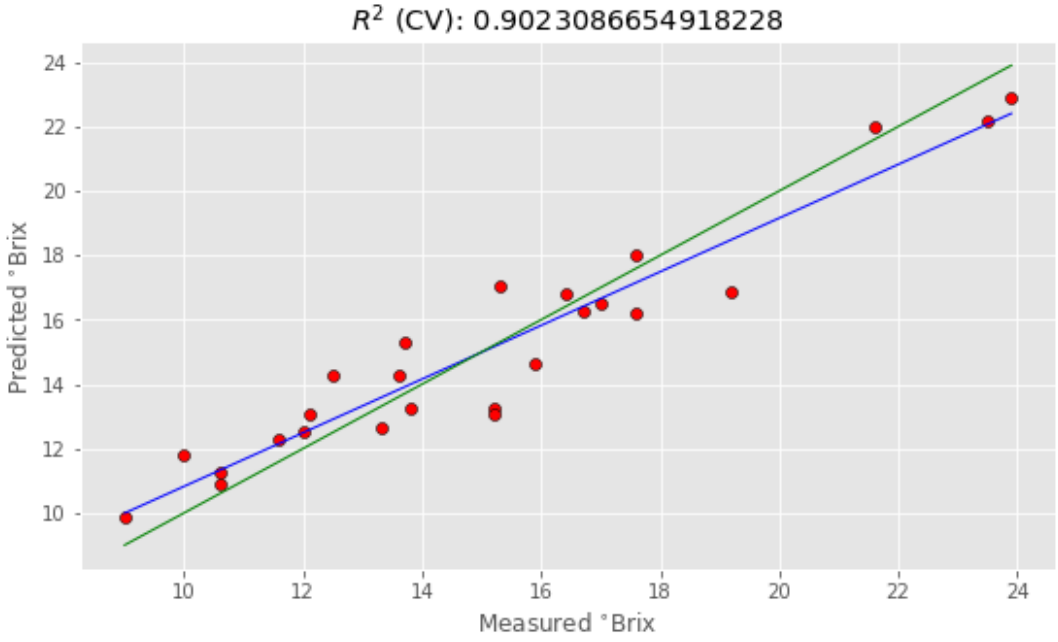


Figure 36. Results from sugarcane tests in the final box with 150-lumen bulb in fixture. (Top) Original spectral data, (Middle) spectral data after preprocessing with SNV (window size 35, polynomial order 3), (Bottom) Linear regression coefficients at each wavelength. Grey lines represent vibrational frequencies of O-H bonds, and blue lines represent vibrational frequencies of C-H bonds (see Figures 1 & 3).

The intensities are generally much lower than in the sucrose solution tests. This is likely due to the increased optical scattering and absorbing effects within the sugarcane, compared to a container of sucrose solution.The peak intensity of the original data here is about 50AU, which corresponds to about 10 microwatts of power at that wavelength.

The spectra in this test, as before, appear smooth and free of visible artifacts after SNV. However, SNV does not appear nearly as effective here at reducing variation across the absorbance spectra. Also, correlation is generally weak and does not appear to visibly concentrate at the known sucrose vibrational wavelengths (see vertical lines in Figure 36 bottom). While relative extrema appear at 810nm and 950nm, the R^2 coefficient at these wavelengths is very low: -0.14 and 0.1. Due to the lack of strong positive-correlation- and negative-correlation- wavelengths, the dual-wavelength approach (pulse oximetry algorithm) is not feasible here.

In optically complex, highly-scattering media such as sugarcane, it is common for absorbance at certain wavelengths to be dependent on each other. Therefore an approach that considers individual wavelengths as independent predictors, such as the pulse oximetry algorithm, is unlikely to be suitable for sugarcane PLS is known for its ability to account for this (Golic et al 2003). Applying the PLS with variable selection algorithm yields the plots shown in Figure 37.



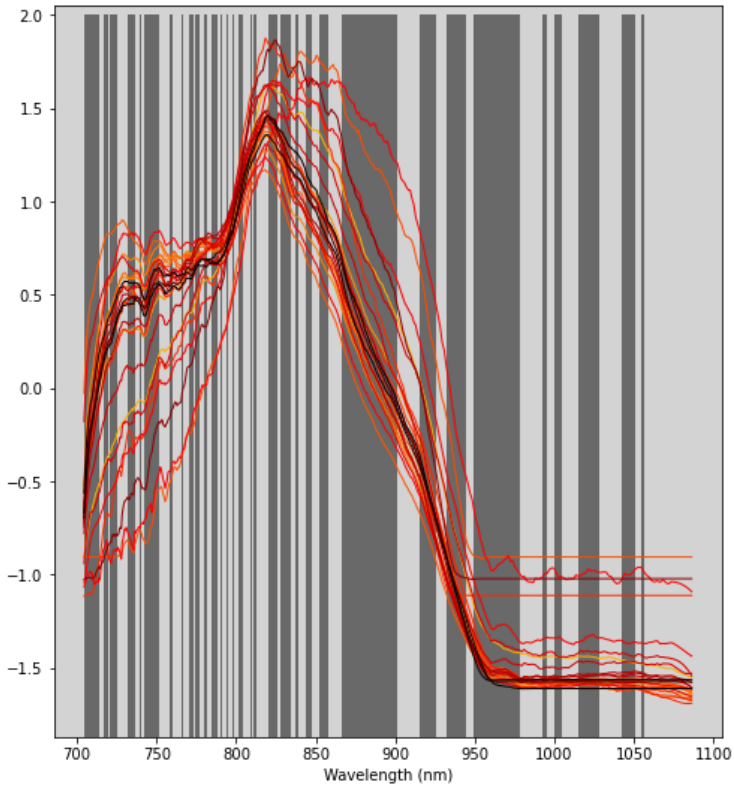


Figure 37. PLS model fit to final sugarcane dataset measured in final box. (Top) linear regression of predicted versus measured Brix from cross-validation (n=25, cross-validation set of 5). (Bottom) wavelengths shown by black bands were discarded via variable selection to improve model fit.

The PLS model is able to achieve an R^2 of 0.902 in cross-validation, using 15 latent variables and discarding 254 out of 400 wavelengths. While the number of latent variables is relatively large, we can see that nearly all vibrational overtones of sucrose are represented, including the 910nm and 930nm of the C-H bonds, as well as 770nm, 840nm, and the 960nm-970nm of O-H bonds. This stands to reason -- in the presence of a greater number of substances with overlapping absorbance bands, more wavelengths of sucrose are needed to separate sucrose absorbance from the absorbances of other substances. The error of this PLS model, quantified by mean square error (MSE) is 1.479. This means the model has a mean absolute error of:

$$\sqrt{\frac{2}{\pi}} \times \sqrt{MSE} = 0.97$$

That is, any prediction will be within ±0.97° Brix on average. This indicates strong prediction performance.

7. Discussion

Over the course of this project, two prediction models were explored: a dual-wavelength linear regression model similar to the traditional pulse oximeter algorithm, and a PLS regression model based on an existing PLS implementation in literature. While the dual-wavelength model could be optimized to fit the data well ($R^2 > 0.9$) on sucrose, there were cases when the optimal wavelengths did not match up with literature values on sucrose's high-absorbance wavelengths. PLS-R fit the data better for every experiment, with R^2 closer to 1 in cross-validation . PLS-R generally uses more than two wavelengths to construct its prediction model. Importantly, optimized PLS-R on sugarcane data used 15 latent variables constructed from 146 wavelengths, and was able to achieve cross-validation R^2 of 0.902 and a mean absolute error of 0.97 degrees Brix.

Thus, the target specification identified at the beginning of this report has been achieved. The cross-validation score is greater than or equal to 0.9, and the mean absolute error is within ±1.0 degrees Brix.

Now that we have a working PLS prediction model with fairly high prediction strength, this can easily be converted to a form that allows prediction of Brix on new spectra. In scikit-learn, `pls.coef_` is an accessible member of the PLS class. This contains all the PLS coefficients, i.e. the coefficients of the linear regression model such that Brix can in theory be approximated as Brix = Absorbance@wavelength1 + Absorbance@wavelength2 + … . However, it is not recommended to use these coefficients to predict directly. Rather, the recommended method is to save (i.e. "pickle") the `pls` object into memory, and then create a new program in which `pls.predict(X)` is called on a new spectrum X (Pelliccia 2018). Creating this prediction program falls under future work.

8. Future Work

This project has proven that a prediction model based on standard tools in NIR spectroscopy analysis (namely PLS with variable selection) can be optimized to predict Brix values noninvasively in whole sugarcane stalks with high prediction power.

From this starting point, several tasks lie ahead as future work. The first step would be to validate the PLS model on a large dataset. Once the prediction program is created, validation can be performed simply by passing in new spectra to the pls.predict(X) function, as mentioned above. Loading this program onto a laptop that is wirelessly connected to a LinkSquare NIR allows for validation to be performed in the field.

Following that, the next step would be to begin translating the algorithm and "final box" optics assembly into a custom prototype. Custom hardware would need to be selected for this

prototype, notably a spectrophotometer module. Potential candidates were shortlisted, shown in Table 1.

Table 1. Potential candidates for spectrophotometer module

Name	Manufacturer	Cost	Range Min	Range Max	NIR Resolution	Inbuilt light source?	Breakout Board?	Total Cost
C11708MA	Hamamatsu	$520.	640nm	1050nm	14nm	no	C14465 ($820)	$1,340.00
C14384MA-01	Hamamatsu	$205.	640nm	1050nm	13nm	no	C14989 ($820) + C15036 ($410)	$1,435.00
ASP-NIR-M-Reflect	Allied Scientific	$1,995.	900nm	1700nm	10nm	yes	No	$1,995.00
STS-NIR-L-10-400-SMA	Ocean Insight	$1,592.	650nm	1100nm	1nm	no	software +$243	$1,592.00

Several compact, hardware-interfaceable spectrophotometer modules are available off the shelf. The compactness of these modules enables an impressively small form factor for the device hardware; likely comparable to the LinkSquare NIR. For example the Hamamatsu C11708MA module is only 27.6 × 16.8 × 13 mm in size -- see Figure 38.

Figure 38. Hamamatsu C11708MA, with fingers for scale

However, the drawback of a spectrophotometer module is cost. Table 1 shows that even obtaining one for prototyping costs between one and two thousand dollars. At low volumes, the

modules still cost hundreds of dollars. However, the PLS model developed uses 146 wavelengths; collecting the data for this requires a spectrophotometer.

One avenue that can be explored by future work is attempting to improve the dual-wavelength algorithm. Even if more than two wavelengths end up being necessary, a multi-wavelength algorithm that relies on just a few specific wavelengths offers greater hardware simplicity and far lower cost. A spectrophotometer is no longer required, and instead, pulse-oximeter-like sensing hardware can be employed. For example, a broad-spectrum photodiode and a small number of NIR LEDs can be used. LEDs generally have narrow emission spectra, so each NIR LED would be selected so its emission peak aligns with one of the wavelengths used for prediction by the multi-wavelength algorithm. There are in fact off-the-shelf modules called "spectral sensors" that offer these components packaged together. Table 2 provides a couple of examples.

Table 2. Potential candidates for spectral sensors

Name	Manufacturer	Cost	Range Min	Range Max	NIR Resolution	Inbuilt light source?	Breakout Board?	Total Cost
AS7262	ams	$25.95	450nm	650nm	40nm	no	Sparkfun ($25)	$50.95
AS7341	ams	$24.96	350nm	1000nm	40nm	no	Sparkfun ($25)	$49.96

The AS7262 is a 6-channel spectral sensor, while the AS7341 is a 11-channel spectral sensor. The AS7341 covers the target spectral range of 700nm-1000nm for this project. Note that the prototyping costs and low-volume costs here are a small fraction of what they are in Table 1.

One way to improve the prediction strength of the multi-wavelength algorithm is to consider sensing deeper into the IR spectrum. There is some evidence to support that this would improve prediction strength. For instance, in Chaix (2020)'s work, the PLS analyses on data from spectrophotometers with ranges deeper into the IR spectrum tended to yield higher R^2 coefficients. Also, looking at Figures 1 and 3 in this report, we can see that there are several overtone frequencies, especially for C-H vibrations, at higher wavelengths in the IR spectrum.

References

Ahmed, Adam E., and Alam-Eldin, Amna O.M (2015). "An Assessment of Mechanical vs Manual Harvesting of the Sugarcane in Sudan – The Case of Sennar Sugar Factory." Journal of the Saudi Society of Agricultural Sciences, vol. 14, no. 2, 2015, pp. 160–166., doi:10.1016/j.jssas.2013.10.005.

Bilodeau, S.E., Wu, B. S., Rufyikiri, A. S., MacPherson, S., & Lefsrud, M. (2019). An update on plant photobiology and implications for cannabis production. Frontiers in plant science, 10, 296.

Chaix, Gilles (2020), "Data for: Data set of Visible-Near Infrared handled and micro- spectrometers - comparison of their accuracy for predicting some sugarcane properties.", Mendeley Data, V1, doi: 10.17632/rnvvftvmh7.1

Comprehensive Membrane Science and Engineering, Elsevier, 2010, Pages 109-164

Dobyns E. L. (2011). Chapter 39 - Assessment and Monitoring of Respiratory Function. In Fuhrman, B. P., & Zimmerman, J. J. (Eds.). Pediatric Critical Care (pp 515-519). Mosby. https://doi.org/10.1016/B978-0-323-07307-3.10039-4.

Ekpélikpézé, O. S., Agre, P., Dossou-Aminon, I., Adjatin, A., Dassou, A., & Dansi, A. (2016). International Journal of Current Research in Biosciences and Plant Biology. Int. J. Curr. Res. Biosci. Plant Biol, 3(5), 147-156.

Fearn, T., Riccioli, C., Garrido-Varo, A., & Guerrero-Ginel, J. E. (2009). On the geometry of SNV and MSC. Chemometrics and Intelligent Laboratory Systems, 96(1), 22-26.

Golic, M., Walsh, K., & Lawson, P. (2003). Short-wavelength near-infrared spectra of sucrose, glucose, and fructose with respect to sugar concentration and temperature. Applied spectroscopy, 57(2), 139-145.

Johnson, R. M., & Richard, E. P, Jr. (2005). Sugarcane yield, sugarcane quality, and soil variability in Louisiana. Agronomy Journal, 97, 760–771.

Kołtuniewicz, A. (2010). 4.05 - Integrated Membrane Operations in Various Industrial Sectors,

Lam, E., Shine, J., Silva, J. D., Lawton, M., Bonos, S., Calvino, M., . . . Ming, R. (2009). Improving sugarcane for biofuel: Engineering for an even better feedstock. GCB Bioenergy, 1(3), 251-255. doi:10.1111/j.1757-1707.2009.01016.x

Lawes, R. A., Wegener, M. K., Basford, K. E., & Lawn, R. J. (2002). Commercial cane sugar trends in the Tully sugar district. Australian Journal of Experimental Agriculture, 40, 969–973.

Lazim, S. S. R. M., Nawi, N. M., Chen, G., Jensen, T., & Rasli, A. M. M. (2016). Influence of different pre-processing methods in predicting sugarcane quality from near-infrared (NIR) spectral data. International food research journal, 23, S231.

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 118, 62-69.

Momin, M. A., Grift, T. E., Valente, D. S., & Hansen, A. C. (2019). Sugarcane yield mapping based on vehicle tracking. Precision Agriculture, 20(5), 896- 910. https://doi.org/10.1007/s11119-018-9621-2

Muncan, J., & Tsenkova, R. (2019). Aquaphotomics—From innovative knowledge to integrative platform in science and technology. Molecules, 24(15), 2742.

Nawi, N. M., Chen, G., & Jensen, T. (2013). Visible and shortwave near infrared spectroscopy for predicting sugar content of sugarcane based on a cross-sectional scanning method. Journal of Near Infrared Spectroscopy, 21(4), 289-297.

Nawi, N.M., Chen, G. & Jensen, T (2014). In-field measurement and sampling technologies for monitoring quality in the sugarcane industry: a review. Precision Agric 15, 684–703 (2014). https://doi.org/10.1007/s11119-014-9362-9

Nawi, Nazmi & Jensen, Troy & Chen, Guangnan. (2012). The Application of Spectroscopic Methods to Predict Sugarcane Quality Based on Stalk Cross-sectional Scanning. Journal of American Society of Sugar Cane Technologists. 32 (2012).

NIRSystems, M. (2002). A guide to near-infrared spectroscopic analysis of industrial manufacturing processes. United States: Silver Spring.

Nitzan, M., Romem, A., & Koppel, R. (2014). Pulse oximetry: fundamentals and technology update. Medical devices (Auckland, N.Z.), 7, 231–239. https://doi.org/10.2147/MDER.S47319

Omar, A. F., Atan, H., & MatJafri, M. Z. (2012). Peak response identification through near-infrared spectroscopy analysis on aqueous sucrose, glucose, and fructose solution. Spectroscopy Letters, 45(3), 190-201.

Pelliccia, D. (2018). Partial Least Squares Regression in Python. https://nirpyresearch.com/partial-least-squares-regression-python/

Rattey, A. R., Jackson, P. A., Hogarth, D. M., & McRae, T. A. (2009). Selection among genotypes in final stage sugarcane trials: Effects of time of year. Crop and Pasture Science, 60, 1165–1174.

Research and Markets Pvt Ltd (2020). "Sugarcane Harvesters - Global Market Outlook (2018-2027)." Research and Markets, 6 Apr. 2020, www.researchandmarkets.com/reports/5008330/sugarcane-harvesters-global-market-outlook

Sandhu, H. S., Singh, M. P., Gilbert, R. A., & Odero, D. C. (2016). Sugarcane botany: A brief view. Gainesville, FL: Institute of Food and Agricultural Sciences, Florida Cooperative Extension Service, University of Florida, Electronic Data Information Source SS-AGR-234.

Sanseechan, P., Panduangnate, L., Saengprachatanarug, K., Wongpichet, S., Taira, E., & Posom, J. (2018). A portable near infrared spectrometer as a non-destructive tool for rapid screening of solid density stalk in a sugarcane breeding program. Sensing and bio-sensing research, 20, 34-40.

Solomon, S., Banerji, R., Shrivastava, A.K. et al. (2006) Post-harvest deterioration of sugarcane and chemical methods to minimize sucrose losses. Sugar Tech 8, 74–78 (2006). https://doi.org/10.1007/BF02943746

Sukhchain, S. D., & Saini, G. S. (1997). Inter-relationships among cane yield and commercial cane sugar and their component traits in autumn plant crop of sugarcane. Euphytica, 95, 109–114.

T.M. Hess, J. Sumberg, T. Biggs, M. Georgescu, D. Haro-Monteagudo, G. Jewitt, M. Ozdogan, M. Marshall, P. Thenkabail, A. Daccache, F. Marin, J.W. Knox (2016). A sweet deal? Sugarcane, water and agricultural transformation in Sub-Saharan Africa. Global Environmental Change, 39(2016), 181-194. https://doi.org/10.1016/j.gloenvcha.2016.05.003.

Tang, J. Y., Chen, N. Y., Chen, M. K., Wang, M. H., & Jang, L. S. (2016). Dual-wavelength optical fluidic glucose sensor using time series analysis of d (+)-glucose measurement. Japanese Journal of Applied Physics, 55(10), 106601.

Voora, V., Bermudez, S., & Larrea, C. (2019). Global Market Report: Sugar (Rep.). Retrieved December 1, 2020, from International Institute for Sustainable Development website: https://www.iisd.org/system/files/publications/ssi-global-market-report-sugar.pdf

Appendices

Appendix A: Simple PLS algorithm

The implementation of a simple PLS algorithm is shown below. While the execution of PLS is performed using PLS regression tools built into the scikit-learn module, the implementation of cross-validation and model visualization was developed based on the tutorial provided by Pelliccia (2018).

```
def simple_pls_cv(X, y, n_comp):
    # Run PLS with suggested number of components (latent variables)
    pls = PLSRegression(n_components=n_comp)
    pls.fit(X, y)
    y_c = pls.predict(X)
    # Cross-validation
    y_cv = cross_val_predict(pls, X, y, cv=10)
   #^change from 5 to 10 for large dataset
    # Calculate scores for calibration and cross-validation
    score_c = r2_score(y, y_c)
    score_cv = r2_score(y, y_cv)
    # Calculate mean square error for calibration and cross
validation
    mse_c = mean_squared_error(y, y_c)
    mse_cv = mean_squared_error(y, y_cv)
    print('R2 calib: %5.3f'  % score_c)
    print('R2 CV: %5.3f'  % score_cv)
    print('MSE calib: %5.3f' % mse_c)
    print('MSE CV: %5.3f' % mse_cv)
    # Plot regression
    z = np.polyfit(y, y_cv, 1)
    with plt.style.context(('ggplot')):
        fig, ax = plt.subplots(figsize=(9, 5))
```

```
ax.scatter(y, y_cv, c='red', edgecolors='k') #reversed
ax.plot(y, z[1]+z[0]*y, c='blue', linewidth=1) #reversed
#print("Equation: " + str(z[0]) + "s + " + str(z[1]))
ax.plot(y, y, color='green', linewidth=1)
plt.title('$R^{2}$ (CV): '+str(score_cv))
plt.ylabel('Predicted $^{\circ}$Brix')
plt.xlabel('Measured $^{\circ}$Brix')
plt.show()
```

Appendix B: PLS Variable Selection algorithm

There are several approaches to variable selection for PLS described in literature. A relatively simple method that is used commonly in NIR spectroscopy is called Feature Selection by Filtering, and is outlined in Mehmood et al (2012).

The specific implementation of this variable selection method was based on Pelliccia (2018)'s tutorial and example code, which implements Feature Selection by Filtering in Python, using tools within scikit-learn.

The overall approach of this algorithm is:

1) Run PLS using 1 latent variable, then 2, then 3… up to some specified maximum number. Sort the spectra in ascending order of their PLS coefficients--i.e. ascending order of their weightage in the PLS model
    a) For each iteration, ignore the wavelength with the lowest PLS coefficient and rerun PLS.
    b) Record mean squared error (MSE) using cross-validation on the spectral data.
    c) If MSE is improved, leave that wavelength out of future iterations of PLS, and in the next iteration, try ignoring that and the wavelength with the next-highest PLS coefficient
2) Return the combination of number of latent variables and number of discarded wavelengths that minimize MSE

The code for this is shown below:

```
def pls_variable_selection(X, y, max_comp):

    # Define MSE array to be populated
    mse = np.zeros((max_comp,X.shape[1]))
    # Loop over the number of PLS components
    for i in range(max_comp):

        # Regression with specified number of components, using full spectrum
        pls1 = PLSRegression(n_components=i+1)
        pls1.fit(X, y)

        # Indices of sort spectra according to ascending absolute value of PLS
coefficients
        sorted_ind = np.argsort(np.abs(pls1.coef_[:,0]))
        # Sort spectra accordingly
        Xc = X[:,sorted_ind]
        # Discard one wavelength at a time of the sorted spectra,
        # regress, and calculate the MSE cross-validation
        for j in range(Xc.shape[1]-(i+1)):
            pls2 = PLSRegression(n_components=i+1)
            pls2.fit(Xc[:, j:], y)
```

```
        y_cv = cross_val_predict(pls2, Xc[:, j:], y, cv=5)
        mse[i,j] = mean_squared_error(y, y_cv)

    comp = 100*(i+1)/(max_comp)
    stdout.write("\r%d%% completed" % comp)
    stdout.flush()
stdout.write("\n")
# # Calculate and print the position of minimum in MSE
mseminx,mseminy = np.where(mse==np.min(mse[np.nonzero(mse)]))
print("Optimised number of PLS components: ", mseminx[0]+1)
print("Wavelengths to be discarded ",mseminy[0])
print('Optimised MSEP ', mse[mseminx,mseminy][0])
stdout.write("\n")
# plt.imshow(mse, interpolation=None)
# plt.show()
# Calculate PLS with optimal components and export values
pls = PLSRegression(n_components=mseminx[0]+1)
pls.fit(X, y)

sorted_ind = np.argsort(np.abs(pls.coef_[:,0]))
Xc = X[:,sorted_ind]
return(Xc[:,mseminy[0]:],mseminx[0]+1,mseminy[0], sorted_ind)
```

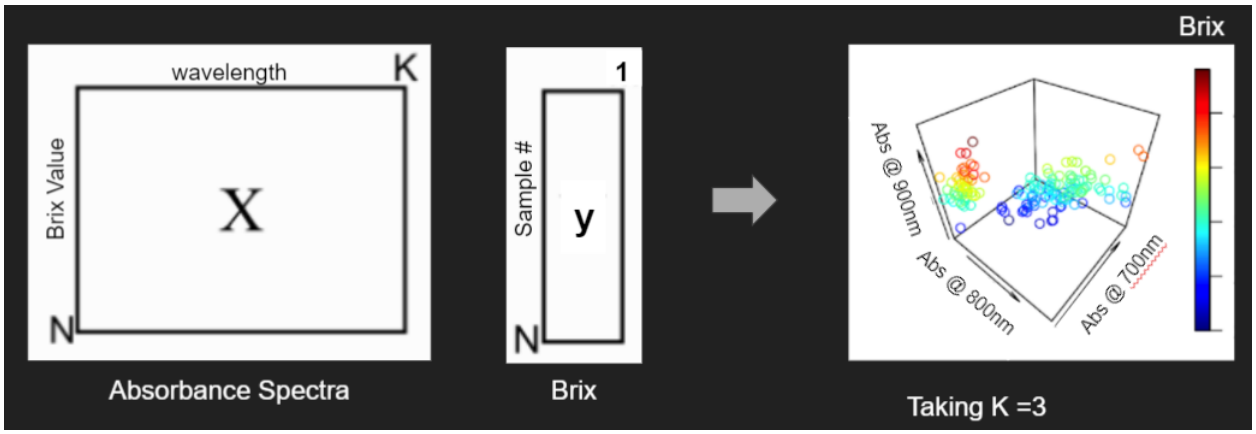
Appendix C: Partial Least Squares (PLS)

Broadly, PLS constructs a prediction model using linear combinations of predictors (i.e. absorbance series at each wavelength). The data is regressed against the linear combinations that most fully capture the covariance with respect to the variable to be predicted (i.e. Brix).

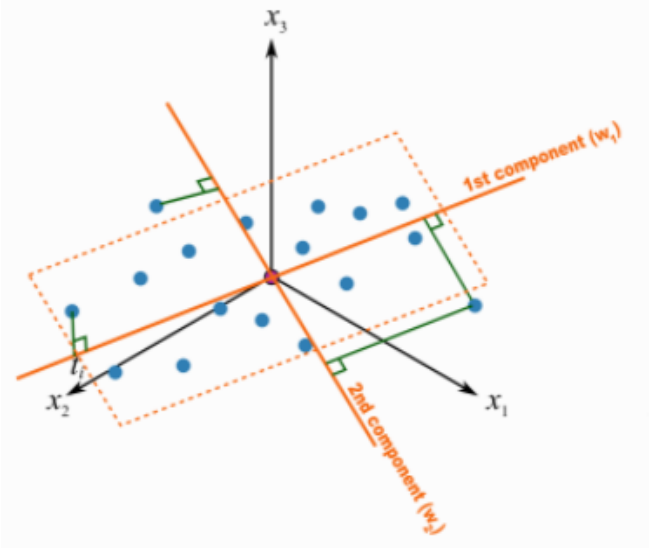
This process can be considered akin to placing axes in K-dimensional space.

Absorbance spectra X is a mean-centered N x K matrix & **Brix values y** is an N x 1 matrix.

If we take K=3 just as an example (3 PLS components aka 3 "latent variables")...



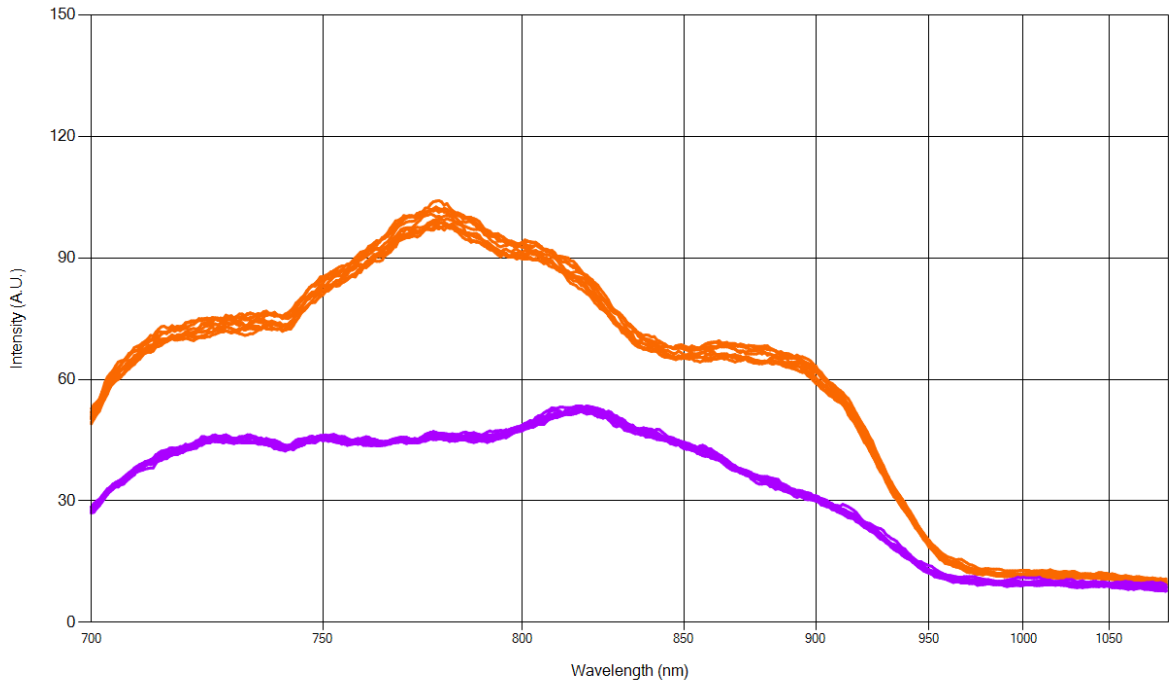
PLS constructs axes in this K-dimensional space, and then the data is regressed against these axes to form PLS regression coefficients, which then result in axes (independent variables) that maximally vary with the data.



Appendix D: External Halogen Bulb Testing

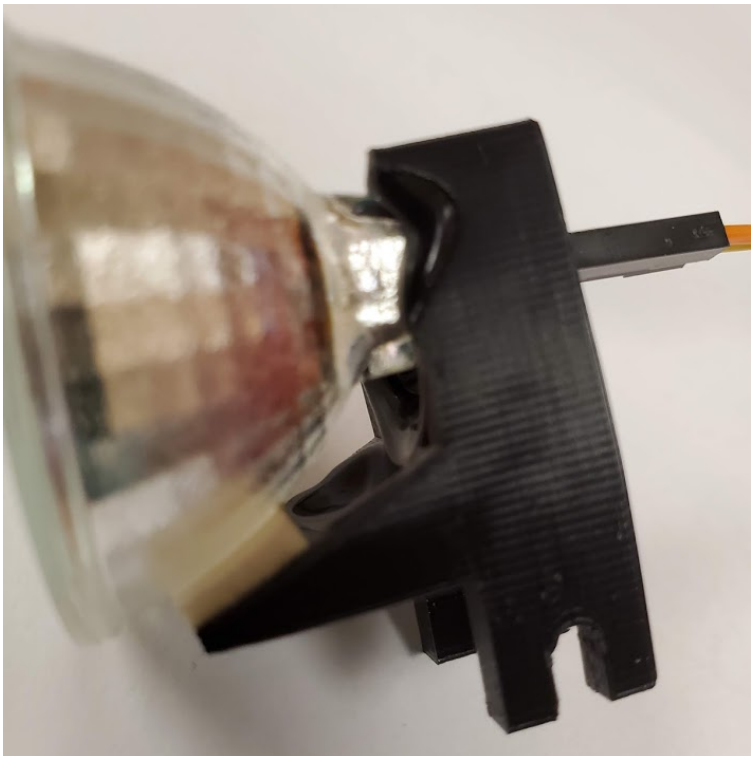
Two halogen bulbs were explored as external light sources: a 10W 150-lumen bulb, and a 20W 313-lumen bulb.

Spectra were measured on the same sugarcane sample (with the foil-lined tunnel described in Appendix F) using each of the bulbs separately. See below.



The orange curve represents the reading with the 20W bulb and the purple curve represents the reading (all else being equal) with the 10W bulb. We can see that in general, the spectra are much higher in intensity when the 20W bulb is used.

However, within seconds of turning on the 20W bulb, the 3D-printed holder had melted quite substantially, and smoke was emanating from the area of the box near the bulb. The melted bulb holder is shown below.

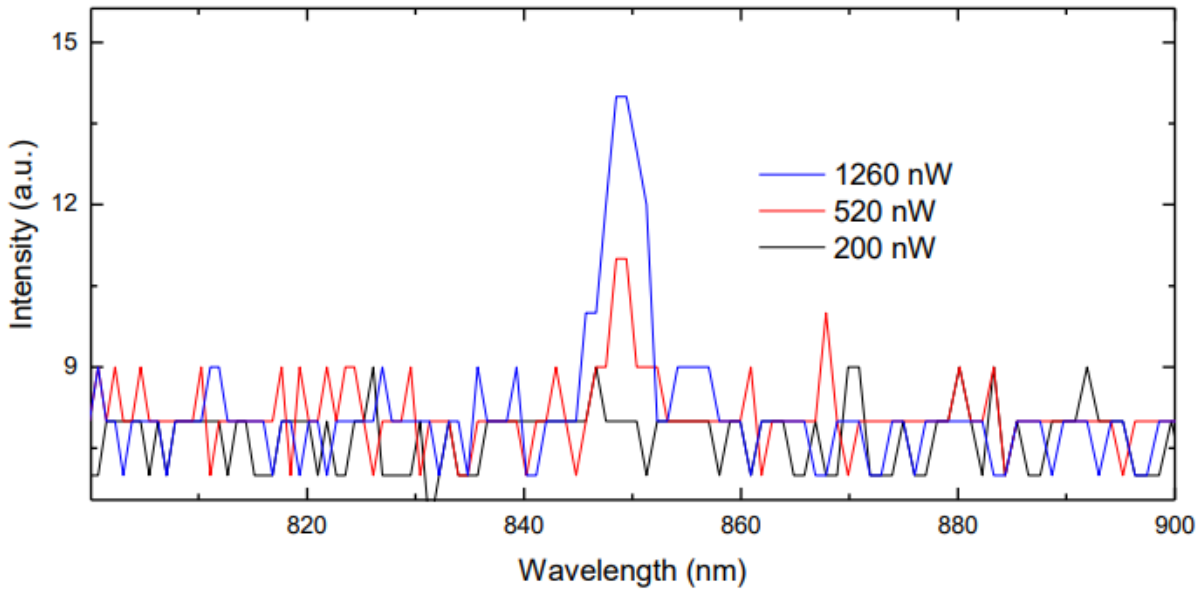


According to Golic et al (2013), the key wavelengths a sucrose-prediction model should prioritize are 910nm and 930nm, since these are where the C-H bond vibration shows maximal absorbance. We can see that the 10W bulb is able to produce a signal intensity in the 20-30AU range for those wavelengths. Additionally, the 10W bulb is able to produce even higher signal intensities at the O-H wavelengths (e.g. 740nm, 770nm, 840nm, etc).

Therefore, the 10W bulb seemed to provide a substantial enough signal while not presenting any heat-related damage. Due to this, and to minimize power consumption for a device eventually intended to become a handheld device, the 10W bulb was selected.

Appendix E: LinkSquare NIR Sensitivity

The LinkSquare NIR datasheet provides this graph:



< Minimum detectable optical power of LinkSquare NIR (850 nm LD) >

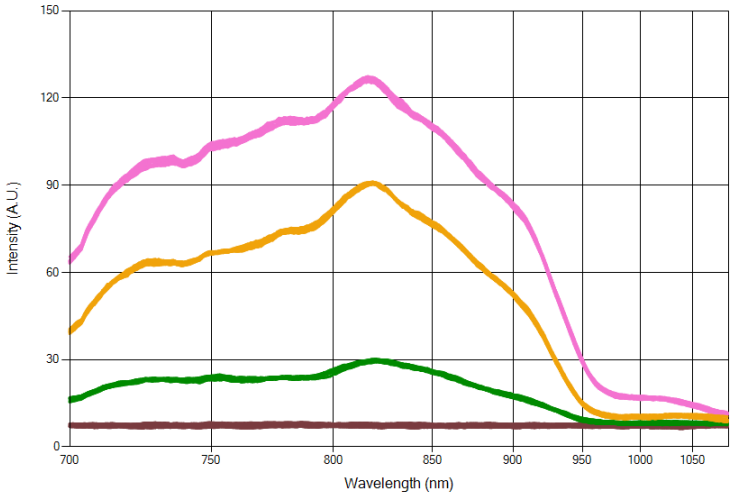
Based on this data, intensity in the LinkSquare's arbitrary units (AU) can be used to roughly approximate power (wattage). Note that the resolution of the signal appears to be around 1 AU.

We can see that between 200nW and 520nW, no signal is detected. Meanwhile, the difference between the peaks at 1260nm and 520nW curve resembles an increase from 11AU to 14AU. Due to the narrowness of the light emission (small full-width half maximum), this increase can be approximated as localized to ~850nm. Therefore, roughly speaking, an increase of +640nW (+0.64uW) corresponds to an increase of about 3AU at a given wavelength. So, each AU corresponds to a step in wattage of around 213nW (0.2uW)

Appendix F: Optimization of optics

One optimization on the optics was the addition of a tunnel lined with aluminium foil. See below.

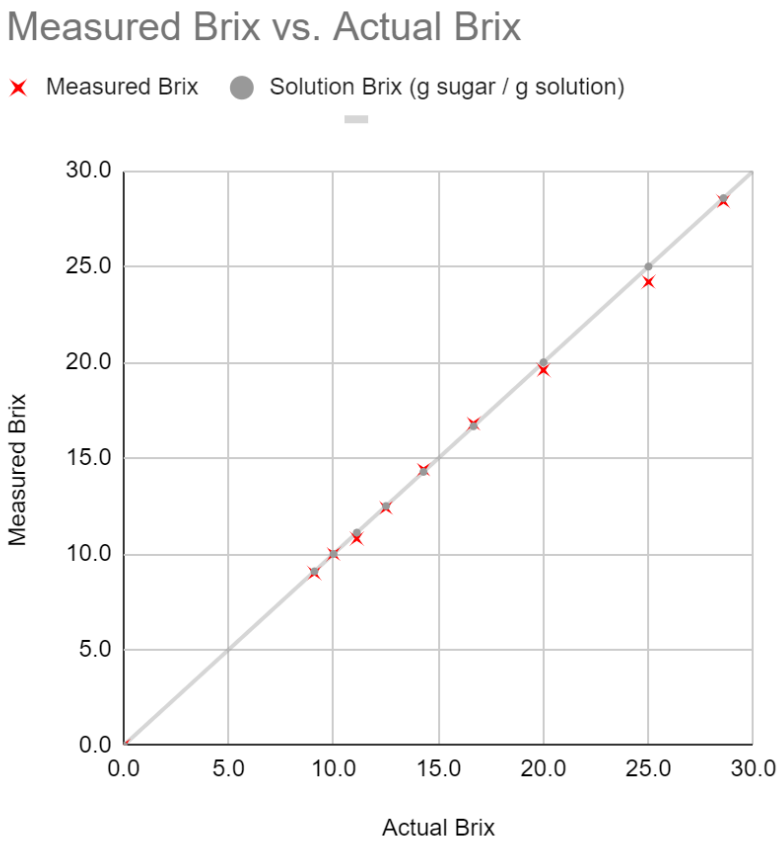
This was done after confirming that adding a tunnel increased the overall measurement intensity, and that lining the tunnel with aluminum foil substantially increased the overall measurement intensity. See below.



The brown curve represents a dark measurement (no light). The green curve represents a measurement on a sugarcane sample with a tunnel partially over the sugarcane sample (in the image above, the arch-shaped panel is placed so that it is roughly centered over the sample). The yellow curve represents the measurement with the tunnel partially over the sample and lined with foil. The pink curve represents the tunnel lined with foil placed fully over the sample (i.e. without the arch-shaped panel). The arch-shaped panel was deemed necessary to prevent light emitted by the bulb from reflecting off the sides of the box and entering the LinkSquare.

Appendix G: Brix Meter Accuracy Check

Water (g)	Sugar (g)	Solution Brix (g sugar / g solution)	Measured Brix	Difference	%Error
10.00	0.00	0.0	0.0	0.0	n/a
10.00	1.00	9.1	9.0	-0.1	-1.0%
9.00	1.00	10.0	10.0	0.0	0.0%
8.00	1.00	11.1	10.8	-0.3	-2.8%
7.00	1.00	12.5	12.4	-0.1	-0.8%
6.00	1.00	14.3	14.4	0.1	0.8%
5.00	1.00	16.7	16.8	0.1	0.8%
4.00	1.00	20.0	19.6	-0.4	-2.0%
3.00	1.00	25.0	24.2	-0.8	-3.2%
2.50	1.00	28.6	28.4	-0.2	-0.6%



Appendix H: Decreasing Brix by Soaking

The naturally-occurring range of Brix values in sugarcane is well known (Nawi et al 2013): from around 9 degrees Brix to about 25 degrees Brix. A prediction model should focus on accurate prediction within this range. However, it is not possible to guarantee that the distribution of a sample set of sugarcane will be representative of the natural distribution of sugarcane. This is a lack of foreknowledge of a sample's Brix before ordering sugarcane.

Therefore, some way to manipulate the Brix within a sugarcane stalk would provide a way to adjust a sample size distribution to be representative of the population at large. One method explored was submerging sugarcane samples (cut to a length of 2.5cm) in sucrose solutions of various concentrations.

As an initial test, six samples were cut and then submerged for 24 hours in watertight bags. The concentration of the solution, as well as the initial and final Brix within the sugarcane sample, are provided in the table below.

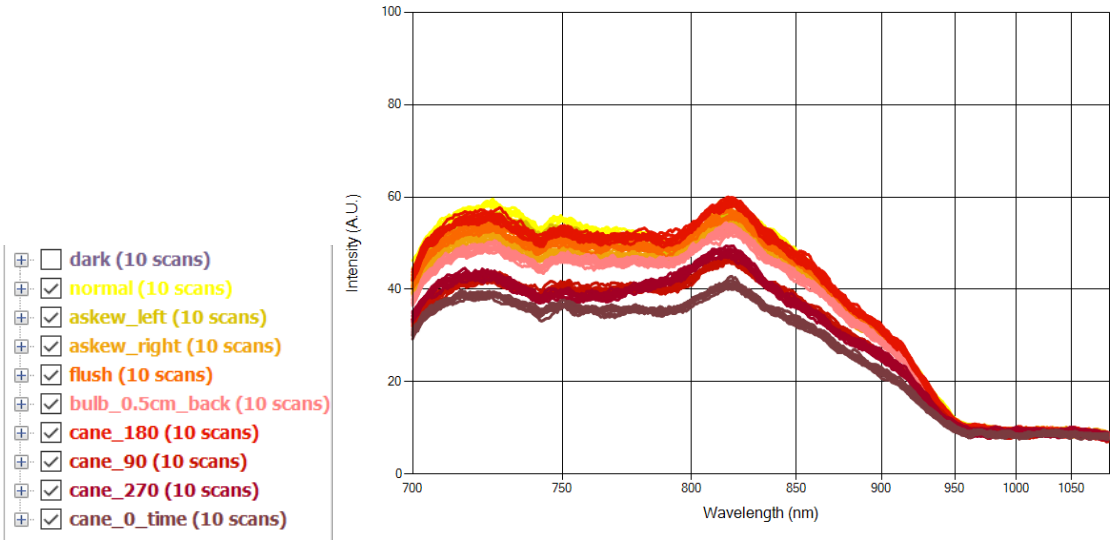
Solution	Initial Sample Brix	Final Sample Brix
0°	15.4	9.0
0°	15.4	10.0
2.6°	14.9	10.6
2.6°	12.8	10.6
5.1°	17.6	13.3
5.1°	18.6	15.9

Brix was determined using a Brix refractometer.

Appendix I: Position Variation and Heat Effects

A simple test was performed on the final box (Figure 13) to determine the effects of small variations in positioning of the various components on the measurement. Several variations were set up, including a slight askewing of the tunnel left and right -- see example below.

Other variations included rotating the sugarcane sample by increments of 90 degrees about its long axis (cane_90, cane_180, cane_270)



While some variation in signal is visible among the tests, the largest trend observed appeared to correlate with time. It can be observed that there is a general decrease in intensity as the colors shift towards darker shades. Most tellingly, the identical measurement made at the start (“normal”) and at the end (“cane_0_time”) are at either extreme of the variation. This suggests that the largest source of variation is time-based. One explanation for this is that this “droop” in intensity across the spectrum is due to temperature, and that as the sugarcane heats up with repeated “on” cycles of the light source from repeated measurements, there is a drop in overall transmittance.