Microwave Sensing for Cranberry Crop Yield Estimation using Machine Learning

By

Alex F. Haufler

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Electrical and Computer Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON

2020

Date of final oral examination: 07/24/2020

The dissertation is approved by the following members of the Final Oral Committee:

Susan C. Hagness, Professor, Electrical and Computer Engineering
John H. Booske, Professor, Electrical and Computer Engineering
Amy E. Wendt, Professor, Electrical and Computer Engineering
Juan E. Zalapa, Associate Professor, Horticulture

Contents

1 Introduction 1

2 Background 3

3 The Microwave Sensing Characteristics of Cranberry Crops (Aim 1) 6

3.1 Introduction . . . 6

3.2 Microwave Sensing Foundation . . . 6

3.2.1 Introduction . . . 6

3.2.2 Methods and Materials . . . 7

3.2.3 Results . . . 10

3.2.4 Summary and Conclusion . . . 13

4 Prototype Microwave Sensor Development and Experimental Data Collection (Aim 2) 14

4.1 Introduction . . . 14

4.2 Simulated Prototype Sensor Characteristics . . . 14

4.2.1 Introduction . . . 14

4.2.2 Methods . . . 15

4.2.3 Simulated Data . . . 18

4.2.4 Summary . . . 23

4.3 Experimental Data Collection with Cranberry Sods . . . 25

4.3.1 Data Collection with Prototype WR975 Sensor . . . 25

4.3.2 Experimental Data . . . 25

4.3.3 Summary . . . 27

4.4 Field Data Collection with Prototype Sensor . . . 30

4.4.1 Methods and Materials . . . 30

4.4.2 Results . . . 33

4.4.3 Summary . . . 41

5 The Predictive Accuracy of a Machine Learning Approach to Cranberry Yield Estimation (Aim 3) 44

5.1 Introduction . . . 44

5.2 Statistical Yield Prediction from Microwave Data - First Attempt 44
5.2.1 Introduction 44
5.2.2 Statistical Learning Concepts 45
5.2.3 Ground Truth Data and Yield Predictions 47
5.2.4 Summary and Conclusion 49
5.3 Statistical Learning Algorithm Comparison 50
5.3.1 Introduction 50
5.3.2 Methods for Training and Testing Six Statistical Learning Algorithms 50
5.3.3 Yield Prediction and Generalization Error Results 54
5.3.4 Summary 57
5.4 The Impact of Soil Moisture and Yield Diversity on Predictive Accuracy . . 57
5.4.1 Introduction 57
5.4.2 Materials and Methods 57
5.4.3 Prediction Results 60
5.4.4 Summary 62
5.5 Averaging for Enhanced Yield Prediction 62
5.5.1 Introduction 62
5.5.2 Methods for Averaging 62
5.5.3 Averaging Results 64
5.5.4 Summary 65
5.6 Yield Prediction vs. Size of the Training Set 67
5.6.1 Introduction 67
5.6.2 Methods 67
5.6.3 Results 67
5.6.4 Summary 70
5.7 Microwave Features for Improved Yield Prediction 71
5.7.1 Introduction 71
5.7.2 Methods 71
5.7.3 Results 73
5.7.4 Summary 76

6 Data Acquisition and Data Fusion Strategies for Enhancing Yield Prediction (Aim 4) 77
6.1 Introduction 77
6.2 Simulated Data Acquisition for Improved Yield Prediction 77
6.2.1 Introduction 77
6.2.2 Methods 78
6.2.3 Results 81
6.2.4 Summary 87
6.3 Experimental Data Fusion 89
6.3.1 Introduction 89
6.3.2 Methods 89

6.3.3 Results 95
6.3.4 Summary 96

7 Summary, Conclusions, and Future Work 99

List of Figures

3.1 The setup used to measure the dielectric properties of cranberry tissue. The dielectric probe was placed in contact with the caps of each berry at four locations on the surface of each cap. 7

3.2 Debye models fit to the mean measured dielectric properties of cranberry-plant fruit and leaves. (a) Relative permittivity. (b) Effective conductivity. Variability bars span the range of measured values. A Debye model for water at room temperature [44] is shown for comparison. 8

3.3 Realizations of randomly distributed cranberry-spheres within the 12.7-cm thick canopy layer of the simulated canopy structure. (a) Spheres constitute 5% of the canopy by volume. (b) Spheres constitute 15% of the canopy by volume. Each full-canopy visualization (12.7 cm × 40 cm × 40 cm) is accompanied by a view of a 2.5-cm-thick slice from within the canopy layer to illustrate the in-plane spacing between berries that is difficult to discern in full-canopy views. 9

3.4 Normalized simulated reflected power for a plane wave incident upon our computational representation of the cranberry canopy. Pulse parameters included 3775 MHz center frequency and 2900 MHz FWHM. The thickness of each colored curve displays the variation in backscattered power due to randomness in cranberry spatial distribution across 50 simulations at a fixed berry volume fraction (BVF) of 0%, 5%, 10%, or 15% within each graph and a fixed soil moisture content (SMC) of (a) 10%, (b) 15%, (c) and 20%. The canopy thickness is four inches (10.2 cm) in all simulations. 11

3.5 The monostatic Radar Cross Section for a 1.3 cm diameter sphere with the dielectric properties of cranberry tissue. 12

4.1 Computational model of prototype WR975 sensor in Computer Simulation Technologies (CST) Microwave Studio. (a) Exterior view. (b) Interior-exposed view with coaxial-waveguide transition. The interior of the waveguide is air-filled. The transition from 50 Ohm coaxial cable to a TE_{10} mode current probe is visible in (b). 16

4.2 Illustration of two cranberry extents in the sensing volume study using the computational sensor model. (left) Eight inch cranberry extent. (right) 24 inch cranberry extent. Cranberry extent is variable and represents the side length of a square surface area of canopy in which berries are randomly distributed through the full depth of the canopy. The area in which the berries are distributed is centered along the central axis of the computational model. 17

4.3 Reflection coefficients of WR975 waveguide in laboratory vs. simulated reflection coefficients of our computational model in Computer Simulation Technologies (CST). The laboratory and simulated reflection coefficients were referenced to the 50 Ohm coaxial input to the coaxial-waveguide transition. . . 19

4.4 Simulated electric field magnitude of TE_{10} mode at 1000 MHz. (a) X-Z plane view. (b) Y-Z plane view. The magnitudes of the electric field intensities are referenced to the maximum electric field intensity at the coaxial input. The cranberries are distributed within a three inch canopy, two inches separated from the aperture. The yield shown is 315 grams of cranberry per 1-ft^2 surface area of canopy, and the soil moisture content is 20% by volume. 20

4.5 The dissimilarity metric i.e. the normalized difference between test signals and a fixed reference signal as a function of berry extent. Test signals are the reflection coefficients as a function of frequency from 606 MHz to 1300 MHz for 12 random distributions of cranberries within the area bounded by each berry extent side-length. The fixed reference signal is the mean of the 12 reflection coefficient signals at each frequency for the 24 inch berry extent case. Black dots represent the dissimilarity metric for each test signal. The red line denotes the median of the 12 dissimilarity metric calculations for each berry extent. 21

4.6 Sensitivity metric vs. (a) Yield, (b) Soil moisture content (SMC), (c) Canopy depth (Depth), and (d) Aperture separation (Sep.). The sensitivity metric (Eq. 4.2) is the norm of the forward difference between reflection coefficient signals resulting from a single increment in one of the canopy-environment variables. Statistics on the sensitivity metric are calculated by varying the other three canopy-environment variables for two fixed sequential test variables e.g. variable smc, depth, and sep. for a yield of 100 g and a yield of 120 g. 24

4.7 Experimental setup for microwave data collection using a cranberry sod-canopy and our WR975 sensor in a University of Wisconsin-Madison greenhouse. The sod is set on a square wooden board with a two foot side length. The canopy depth ranges from 3-6" above a 4-5" deep soil base. The sensor is mounted on a PVC support, and its aperture is roughly 2" above the top of the canopy. A 1-ft^2 PVC marker bounds the site in which cranberries are randomly distributed. The sensor aperture is suspended directly above the site bounded by the marker. 26

4.8 Photographs of the same 1-ft^2 site within a greenhouse sod-canopy depicting two different yields. (a) 220 g. (b) 280 g. Recently harvested cranberries are spaced along fishing line and placed within the canopy of the site. The total yield is varied by placing additional cranberry-lines with known cranberry mass in the canopy. 27

4.9 Ground truth data collected with a single sod-canopy in the greenhouse. The yield was varied from 100 g to 500 g in 20 g increments. Soil moisture was varied by adjusting the level of watering and proximity in time to data collection. 28

4.10 Magnitude of the microwave reflection coefficient as a function of frequency from 606 MHz to 1300 MHz for 21 unique yields between 100 g and 500 g for a fixed soil moisture of 11.3%. 29

4.11 A rectangular open-ended waveguide (WR975) mounted above a cranberry canopy in central Wisconsin (top, right and left). A 1-ft^2 marker delineating a site for sampling the cranberry bed yield (bottom). 30

4.12 Measurements obtained during data collection in central Wisconsin during Fall 2018 and Fall 2019. (a) Soil moisture obtained with the Field Scout TDR 350 with 4.5 inch rods inserted into the soil at three locations within each 1-ft^2 site. (b) Canopy height measured from the soil surface to the top cranberry in the cranberry canopy with a ruler. (c) A photograph of the cranberry canopy taken at each 1-ft^2 site from a height of roughly two feet. 32

4.13 Setup for measuring the harvested cranberry yield with the MARS MS 200 precision scale. 33

4.14 Distribution of yields for each of the nine distinct days of data collection from Fall 2017 to Fall 2019. (a) Testbed 17-1. (b) Testbed 17-2. (c) Testbed 18-1. (d) Testbed 18-2. (e) Testbed 18-3. (f) Testbed 18-4. (g) Testbed 19-1. (h) Testbed 19-2. (i) Testbed 19-3. 36

4.15 Distribution of soil moisture contents for each of the seven distinct days of data collection from Fall 2018 to Fall 2019. (a) Testbed 18-1. (b) Testbed 18-2. (c) Testbed 18-3. (d) Testbed 18-4. (e) Testbed 19-1. (f) Testbed 19-2. (g) Testbed 19-3. Soil moisture was not measured during Fall 2017. 37

4.16 Distribution of soil moisture contents and ground truth yields for each of the seven distinct days of data collection from Fall 2018 to Fall 2019. (a) Testbed 18-1. (b) Testbed 18-2. (c) Testbed 18-3. (d) Testbed 18-4. (e) Testbed 19-1. (f) Testbed 19-2. (g) Testbed 19-3. Each dot represents a single unique 1-ft^2 site. 38

4.17 Distribution of soil moisture contents and yields across the seven distinct days of data collection from Fall 2018 to Fall 2019. (a) Ground truth yield distribution. (b) Soil moisture content distribution. (c) Yield vs. soil moisture content. 39

4.18 Illustration of the impact of soil moisture and yield on the microwave reflection coefficients measured with our prototype sensor during Fall 2018 and Fall 2019 field data collection. There are 12 curves plotted in each color representing 12 spatially different views of the same 1-ft^2 site i.e. 12 unique positions of the sensor aperture over the site. (a) Microwave data collected on the same day and from the same bed with homogeneous soil moisture. An example of feasible microwave based yield prediction. (b) Microwave data collected on different days and from different beds with heterogeneous soil moistures. An example of complicated microwave based yield prediction. 42

4.19 Top-view canopy photographs collected with a smart-phone camera and their corresponding yields for three unique 1-ft^2 sites. (a) Canopy photograph where some cranberries are visible. (b) Canopy photograph where many cranberries are visible. (c) Canopy photograph where many cranberries are obscured. 43

5.1 Data processing used throughout this research. N is the number of measured reflection coefficient signals used in training. F is the number of discrete frequencies where the magnitude of the reflection coefficient is sampled. L is the number of distinct yield values per 1-ft^2 site attached to the training signals. $\mathbf{R}$ is a matrix with the N concatenated reflection coefficient signals. $\mathbf{P}$ is the matrix of 10 principal components. $\mathbf{M}$ is the vector of yield value labels for the N reflection coefficient signals. $\mathbf{W}$ is the matrix of linear discriminants covering all pairwise classifications for the L distinct yield values. $\mathbf{S}$ is an arbitrary reflection coefficient signal, and $\hat{m}$ is the scalar yield value estimate taken from the set of L possible yield values. 46

5.2 (a) Distribution of total-canopy cranberry mass per 1-ft^2 site encompassing 360 field measurements from 30 sites in Testbed 1. (b) Average distribution of estimation error across 40 permutations of training and testing data (180 reflection coefficient signals in each). Error bars represent the maximum and minimum bin-heights across the 40 permutations. 47

5.3 (a) Distribution of total-canopy cranberry mass per 1-ft^2 site encompassing 240 field measurements from 20 sites in Testbed 2. (b) Average distribution of estimation error across 40 permutations of training and testing data (120 reflection coefficient signals in each). Error bars represent the maximum and minimum bin-heights across the 40 permutations. 48

5.4 (a) Distribution of total-canopy cranberry mass per 1-ft^2 site encompassing 600 field measurements from all 50 sites from Testbed 1 and Testbed 2. (b) Average distribution of estimation error across 40 permutations of training and testing data (300 reflection coefficient signals in each). Error bars represent the maximum and minimum bin-heights across the 40 permutations. . 49

5.5 Training stage in evaluating generalization error using microwave data for a single testing case in leave-one-out cross validation (LOO-CV). The squiggly lines represent 12 reflection coefficient signals collected within each 1-ft^2 site. Each reflection coefficient signal is the magnitude of the reflection coefficient at frequencies from 606 MHz to 1300 MHz spaced by 1 MHz intervals. There are 161 1-ft^2 sites in total from the Fall 2018 and Fall 2019 field data collections used to evaluate the yield predictive accuracy of each statistical learning algorithm. 51

5.6 Testing stage in evaluating generalization error using microwave data for a single testing case in leave-one-out cross validation (LOO-CV). The prediction error across the 161 test cases is representative of the prediction error for an algorithm trained on 12 microwave signals from 160 1-ft^2 sites and tested using 12 microwave signals from a single new 1-ft^2 site. 52

5.7 Illustration of 32-fold cross validation used during training to optimize the parameters of each statistical learning algorithm. In 32-fold cross validation, there are 32 sets of five unique 1-ft^2 sites used to evaluate the prediction error of the statistical learning algorithm. In each of the 32 evaluations, 155 1-ft^2 sites are used to train the algorithm, and the prediction error is evaluated using the remaining five 1-ft^2 sites. The prediction error across the 32 test evaluations is the loss (function) that is minimized with respect to the algorithm parameters during training. 53

5.8 Prediction error from the Linear Discriminant Analysis algorithm described in section 5.2. (left) Mean of the absolute prediction errors (MAE) from the 12 signals at each 1-ft^2 test site. (right) Statistics on the MAE across the 161 testing sites in LOO-CV. The five statistics depicted represent the 0th, 25th, 50th, 75th, and 100th percentiles on the MAE across the 161 1-ft^2 testing sites. 55

5.9 Statistics on the generalization error using LOO-CV for six statistical learning algorithms applied to field data from Fall 2018 and Fall 2019. The 25th and 75th percentiles are shown in black, and the 50th percentile is shown in red. 'LDA' represents the Linear Discriminant Analysis algorithm from section 5.2. 'Ridge' and 'LASSO' denote linear regression with l^2 norm and l^1 norm regularization on the regression coefficients respectively. 'RF' represents random forest regression, 'SVM' is support vector machine regression, and 'GP' is Gaussian process regression. 56

5.10 Ground truth yield and soil moisture for each testbed from Fall 2018 and Fall 2019. Each dot represents a unique 1-ft^2 site with 12 microwave measurements. Each testbed corresponds to a single day of data collection within a single 2-acre cranberry bed. The number of 1-ft^2 sites per testbed ranges from 10 to 38 across the seven testbeds. 58

5.11 Training and testing scenarios for evaluating the impact of heterogeneity of soil moisture content on statistical learning based yield prediction. (a) Ground truth yield and soil moisture for Testbed 1 (T1). (b) Ground truth yield and soil moisture across all testbeds. Prediction error is evaluated using LOO-CV. A single testing evaluation is depicted in which red dots represent the 1-ft^2 sites used in training and the yellow dot represents the single 1-ft^2 site left out of training and used for prediction error evaluation. 59

5.12 Statistics on the mean absolute prediction error generated by LOO-CV. From the 12 microwave signals at each test site there are 12 yield predictions. The mean absolute error across the 12 predictions represents the site-specific prediction error. The statistics displayed are calculated across site-specific testing errors. Error bars represent the 25th and 75th percentiles. The height of each bar represents the mean of the site-specific testing errors. 61

5.13 Illustration of input averaging with six microwave reflection coefficient signals during testing. Squiggly lines represent the microwave reflection coefficient signals, which consist of reflection coefficient magnitudes at frequencies spaced by 1 MHz intervals from 606 MHz to 1300 MHz (695 discrete frequencies). Input averaging is done in an ensemble fashion for the 695 dimensional input vectors. The trained statistical learning algorithm maps these 695-dimensional input vectors to real-valued yield outputs Y. 63

5.14 Illustration of output averaging with six microwave reflection coefficient signals during testing. The squiggly lines represent microwave input signals, which consist of reflection coefficient magnitudes at frequencies spaced by 1 MHz intervals from 606 MHz to 1300 MHz. The trained statistical learning algorithm maps these 695-dimensional input vectors to yields Y. Output averaging is applied to the independent output predictions from some fixed number of input microwave signals, in this case six. 64

5.15 Absolute error in yield prediction versus the number of (a) input signals averaged or (b) output predictions averaged. Prediction error is calculated with LOO-CV and SVM regression applied to 161 1-ft^2 sites from Fall 2018 and Fall 2019. Black horizontal lines represent the 25th and 75th percentiles, and the red horizontal line represents the median absolute prediction error across 16,100 testing evaluations for a fixed number-average. 66

5.16 Illustration of training set size analysis for a training set size of three 1-ft^2 sites selected randomly from the full set of field data collected in Fall 2017, Fall 2018, and Fall 2019 encompassing 213 1-ft^2 sites. Squiggly lines represent reflection coefficient magnitudes between 606 MHz and 1300 MHz measured with the prototype sensor. There are 12 microwave signals collected within each 1-ft^2 site used to train a Support Vector Machine with a Gaussian kernel. Random selection of training and testing sites is completed 200 times for prediction error evaluation. 68

5.17 RMS prediction error vs. the number of 1-ft^2 sites used in training i.e. the training set size. The boxes correspond to the 0th, 25th, 50th, 75th, and 100th percentiles with respect to the RMS prediction error across 200 random draws of training and testing sites for each fixed training set size. 69

5.18 Mean of the RMS prediction error across 200 random draws of training and testing sites for each fixed training set size and the exponential fit of Eq. 5.1. 70

5.19 Supervised Principal Component Analysis (SPCA) applied to the magnitude of the reflection coefficient concatenated with the slope of the magnitude of the reflection coefficient. The univariate regression coefficient is calculated for each feature i.e. each random variable e.g. $|S_{11}(f_i)|$. Features with regression coefficients larger than a threshold are retained and input to Principal Component Analysis. A subset of principal components with the largest variance are retained and used as inputs to the Support Vector Machine that maps them to yield. 72

5.20 (a) Standardized or univariate regression coefficients for the magnitude of the reflection coefficient and slope between 606 MHz and 1300 MHz, calculated with Eq. 5.2 from 161 1-ft^2 sites between Fall 2018 and Fall 2019. The regression coefficient illustrates a normalized correlation between each microwave feature and yield. (b) The magnitude of the reflection coefficient at 650 MHz vs. yield. (c) The slope of the magnitude of the reflection coefficient at 800 MHz vs. yield. 74

5.21 Statistics on the prediction error across 161 1-ft^2 sites evaluated with LOO-CV. Black horizontal lines represent the 25th and 75th percentiles, and a horizontal red line represents the median. 75

6.1 Computational sensor models submerged 12.84 cm in a dielectric half-space. (a) Original WR975 prototype computational sensor model in dielectric half-space. (b) Original WR975 prototype computational sensor model shown embedded within the dielectric material. (c) Modified computational sensor model in dielectric half-space. (d) Modified computational sensor model shown embedded within the dielectric material. The modified model is a flanged section of WR975 waveguide and is stimulated with a waveguide port at the top opening. Both models are simulated at frequencies between 606 MHz and 1300 MHz. 80

6.2 Frequency dependent S-parameters in computational sensor model network described by Eq. 6.1. S_{11} denotes the reflection coefficient at the coaxial input, S_{12} denotes the transmission coefficient from the coaxial input to the aperture plane of the waveguide, S_{21} denotes the transmission coefficient from the aperture plane of the waveguide to the coaxial input, and S_{22} is the reflection coefficient at the aperture of the waveguide. 82

6.3 Simulated reflection coefficients for the computational sensor and plane wave models in a canopy environment with 20% soil moisture content and 4" canopy depth. (a) Simulated reflection coefficients measured at the coaxial input Γ_{in} with a 2" aperture separation. (b) Simulated reflection coefficients measured at the coaxial input with a 2" aperture separation and calibrated using Eq. 6.1 for Γ_{ap}. (c) Simulated reflection coefficients captured with the computational plane wave model. 84

6.4 Multi-Dimensional Scaling (MDS) applied to simulated reflection coefficients for the computational sensor and plane wave models in a canopy environment with 20% soil moisture content and 4" canopy depth. (a) MDS applied to simulated reflection coefficients measured at the coaxial input Γ_{in} with a 2" aperture separation. (b) MDS applied to simulated reflection coefficients measured at the coaxial input with a 2" aperture separation and calibrated using Eq. 6.1 for Γ_{ap}. (c) MDS applied to simulated reflection coefficients captured with the computational plane wave model. Coordinate 1 and Coordinate 2 correspond to the two coordinates on which the 695-dimensional data were projected. 85

6.5 Yield predictions from the simulated reflection coefficients measured by the (a) uncalibrated sensor model, the (b) calibrated sensor model, and the (c) plane wave model. Ground truth yield is plotted on the x-axis. There are 26 dots displayed in each plot corresponding to test predictions from each reflection coefficient signal in LOO-CV. 86

6.6 Prediction error statistics across the full set of simulation environments. Statistics are calculated from the 26 test prediction errors resulting from LOO-CV for a fixed soil moisture content, aperture separation, and canopy depth. (a) Simulation environment with 3" canopy depth and 10% soil moisture content. (b) Simulation environment with 3" canopy depth and 20% soil moisture content. (c) Simulation environment with 4" canopy depth and 20% soil moisture content. Box plots display the 0th, 25th, 50th, 75th, and 100th percentiles on absolute prediction error. 88

6.7 Mathematical representation of microwave data and the statistical mapping to yield. The magnitude of the reflection coefficient is recorded at 695 discrete frequencies from 606 MHz to 1300 MHz spaced by 1 MHz intervals. The derivative of the magnitude of the reflection coefficient with respect to frequency at each of the 695 frequencies is illustrated. The magnitude and slope constitute 695 length vectors. After concatenating the magnitude and slope vectors the resulting vector has a length of 1390. The statistical learning algorithm (T) maps the real-valued 1390 length vector to a real valued yield. 90

6.8 Conversion of optical images to optical feature vectors and the statistical mapping to yield. Raw canopy images are represented by Red, Green, and Blue (RGB) intensities at each pixel. At each pixel, 25 color indices are calculated from the RGB values. Optical features are calculated from the mean and variance of each color index across the image. Two additional features are included in the input vectors based on a leaf count and leaf segmentation fraction calculated with a Region-Based Convolutional Neural Network trained on synthetic leaf data [53]. The statistical learning algorithm (T) maps the real-valued 52 length vector to a real valued yield. 91

6.9 Optical features calculated from Red (R), Green (G), and Blue (B) intensities at each pixel in canopy image. The RGB intensities fall in the range [0, 1]. . 92

6.10 Yield prediction using a linear combination of a microwave based yield prediction and an optical based yield prediction. The parameters of each statistical learning algorithm and the final weights a, b are determined during the training step using 32 fold cross validation. 93

6.11 Combination of microwave and optical inputs through concatenation. Microwave input vectors consist of 1390 elements and optical input vectors consist of 52 elements. Concatenating the inputs results in a 1442 length input vector which the statistical learning algorithm maps to a real valued yield prediction. 94

6.12 Yield predictions and statistics for microwave and optical based prediction. (a) Microwave based yield predictions. (b) Optical based yield predictions. (c) Linear combination of microwave and optical based yield predictions. (d) Weights for the linear combination of microwave and optical based yield prediction. (e) Concatenated microwave and optical input based predictions. (f) Concatenated microwave, optical, and soil moisture input based predictions. Predictions are the result of LOO-CV applied to the 161 1-ft^2 sites from field data collection in Fall 2018 and Fall 2019. 'MAE' is the mean absolute error in prediction and 'Corr.' is the correlation coefficient between the predicted and ground truth yields. 97

6.13 Prediction error statistics across the 149 1-ft^2 sites comprising the field data from Fall 2018 and Fall 2019 with ground truth yields between 200 g and 600 g. LOO-CV was used to evaluate the prediction error using Gaussian kernel SVM regression and the hybrid approach that averages microwave and optical predictions. (a) Mean prediction error across 12 microwave refection coefficient signals at each site. (b) Absolute value of the mean prediction error as a function of the number of sites sampled for testing. (c) Absolute value of the mean prediction error as a function of the number of sites sampled for testing across a wider range of the number of sites sampled for testing. The black lines in (a) and (b) represent the 25th and 75th percentiles of prediction error, and the red line represents the mean prediction error. 98

Chapter 1

Introduction

Agricultural microwave sensing proceeds as follows. Electromagnetic energy is transmitted through a device that directs the energy towards the crop environment. The transmitting device then collects the energy scattered from the crop structure in a monostatic configuration, or a separate electromagnetic sensor collects the scattered energy in a bistatic configuration. Finally, a relationship between the measured microwave signals and the crop-environment variables such as soil moisture and above-ground biomass is established. In our work, cranberry yield is the desired crop variable. Therefore, the goal of our research within the framework of agricultural sensing is to understand the capability of both microwave sensing and signal processing algorithms to accurately estimate cranberry crop yield.

Microwave sensing is promising for rapid cranberry yield estimation. Air-borne or space-borne microwave sensors can illuminate hundreds of square meters of a crop's surface in small fractions of a second, allowing large agricultural areas to be scanned quickly. Ground-based microwave systems illuminate smaller surface areas than air- or space-based systems, but they can be mounted on vehicles and swept across crop canopies swiftly. Cranberry growers use large truck booms throughout the growing season to apply fertilizer, and each boom extends to half of the width of each cranberry bed. We envision a ground-based system where the microwave sensors are attached to the truck booms and swept across the beds. In a later implementation, it is possible to augment the fine spatial resolution of such a ground based system with microwave scattering data obtained by satellite-based sensors. There is a significant quantity of such data freely available online. For example, large swaths of the Earth are regularly covered by satellites controlled by the European Union's Copernicus Programme.

Microwave sensing is a promising approach for accurate cranberry yield estimation, and it has been successful in estimating crop parameters for a wide range of crop types over the past 40 years. Crop structure and water content in the canopy and soil are the primary variables that affect microwave backscatter in agricultural settings. Cranberries are relatively large compared to the leaves and vines that surround them in the canopy, and the water content of cranberry tissue is significant. Therefore, the size and water content of the cranberries distinguish them from the surrounding air, vines, and leaves in a microwave sensing environment. Understanding how cranberries as a whole within their crop-structure affect microwave scattering across a range of microwave frequencies is core to our work.

A major thrust in the proposed dissertation involves predicting cranberry yield from microwave backscattered power using machine learning methods, also known as statistical learning methods. Machine learning is promising in cranberry yield estimation for a couple of reasons. First, a machine learning approach is statistical in nature and can directly relate input microwave data to output cranberry yield. All of the complexity in the input-output mapping is handled by the learning algorithm. As opposed to the statistical learning approach, electromagnetic inverse scattering approaches can easily become cumbersome in trying to describe this input-output mapping. Second, a statistical learning approach does not require assumptions on the canopy element distributions, whereas some electromagnetic approaches do. For example, some electromagnetic representations impose assumed distributions on the geometries and orientations of the elements. If the imposed distributions are not exact, statistical learning algorithms can potentially outperform their electromagnetic counterparts.

This research program broadly focused on using statistical learning to estimate cranberry yield from microwave scattering signatures obtained with active sensors in commercial cranberry beds. The research encompassed computational studies of microwave scattering, experimental validation and prototype testing, and refined sensor and algorithm design. Specifically, the following fundamental aims were proposed:

1. Investigate the microwave sensing characteristics of cranberry crops through the dielectric characterization of plant components and the computed backscattered radiation from simulated crop environments.

2. Develop a prototype microwave sensor and collect computational and experimental scattering data, ground-truth cranberry yield, and measured cranberry bed parameters such as soil moisture in commercial cranberry beds.

3. Investigate the predictive accuracy of a machine learning based approach to estimating cranberry yield from crop backscatter, evaluating the impact of known cranberry bed parameters and statistics of the training data.

4. Investigate experimental data acquisition and data fusion strategies to enhance machine learning based cranberry yield estimation, specifically with regard to microwave sensor design and multi-modality sensing.

Chapter 2

Background

Accurate prediction of cranberry yield is desirable to farmers, agricultural researchers, and the industry as a whole in order to maximize supply chain efficiency and future crop yields. The prevalent method for predicting cranberry yield involves hand-picking the fruit at a few 1-ft^2 sites within each bed and estimating the total bed yield from the sampled fruit count. This method is invasive since it requires the removal of cranberries from each sampled site. The process of hand-picking and counting the fruit at a few sites within each bed is time consuming, as there can be more than 500 berries within a single site. Furthermore, sampling at a few 1-ft^2 sites within beds of roughly 100,000 ft^2 constitutes a small sample size, which is problematic given the significant variation in yield that can exist across the bed [1]. The small sample size and significant variation in sampled yields within each bed contribute to significant uncertainty and inaccuracy in the inferred total yield per bed. It is common for total bed yield estimation errors to exceed $\sim$15% of the true bed yields [2]. Although the yield estimates could be improved by sampling more 1-ft^2 sites, this would add to the invasiveness, time consumption, and labor costs of the current method. Alternative methods for yield estimation based on satellite images (optical) and standard indices such as the normalized difference vegetation index (NDVI) have poor correlation with cranberry yield, since the indices measure vegetative growth which is not a consistent predictor of yield [3–5].

Alternative yield estimation procedures that offer improved predictive accuracy and can sweep large areas would provide significant benefits to the cranberry industry. We are investigating low-power microwave sensing coupled with machine learning as an approach to enable localized, rapid, and non-invasive cranberry yield estimation. Machine learning is used as a data processing method to transform electromagnetic backscattered signals within the UHF- and L-bands to yield on a 1-ft^2 canopy-surface area basis. The technology is mobile and can be extended to nondestructive total crop yield assessment by capturing measurements within and across crop beds.

We have previously explored the physical basis for sensing cranberry crop yields using microwave backscattered radiation. Specifically, we showed in [6] that a nearly 3:1 contrast in relative permittivity exists between the berries and foliage across a wide range of microwave frequencies. The relatively large dielectric constant of the cranberry flesh is due to its high water content. In commercial cranberry beds, cranberry volume fractions within the canopy can be as high as 20% [7]. The relatively large dielectric contrast between the cranberries and leaves combined with significant fractional berry volumes motivates examining low-power microwave sensing of canopy cranberry content.

Steele-Dunne *et al.* [8] provided a review of active microwave technologies for sensing agricultural crop features using ground-based, air-borne, and space-borne radar instruments. A ground-based system is preferred over an air- or space-based system for our application because of the local control and finer spatial resolution that it can provide within agricultural fields.

The principal objective of microwave remote sensing is to relate the measurements obtained by the sensor to crop-soil variables such as fruit yield. The three basic model types for achieving this mapping are empirical, physical, and statistical. In addition, some models have been developed which are combinations of the three basic types. For example, some physical models that have been parameterized are called semi-empirical.

Empirical models are parameterized expressions that include the relevant crop-soil variables. The parameters are obtained by fitting the model to measured microwave data. Empirical models have been developed with active and passive sensors to estimate crop yield and vegetation biomass [9], soil moisture and surface roughness [10–12], and leaf area index [13]. Empirical models are generally less complicated than physical models, which makes them more accessible to a broad user base. However, empirical models do not fully describe the complex electromagnetic interactions within sensing environments, which can result in poor accuracy when applied to scattering environments that weren't represented during model development [14].

Substantial effort has been devoted to physical "forward" models that provide a fundamental description of microwave backscatter for a specific agricultural crop setting. These forward models are derived from radiative transfer [15–19] or wave equations [20, 21]. Some radiative transfer or wave approaches have been parameterized into semi-empirical models for improved agreement with measured or simulated data [22, 23]. The relevant crop-soil variables are estimated from observed microwave backscatter by forward model inversion or lookup tables. Radiative transfer [24–26], wave-based [27–29], and semi-empirical [30–32] models have been inverted. However, modeling error results in crop-soil variable estimation error during inversion. These model errors arise from incorrect or imprecise assumptions about the scattering environments. Consequently, it is difficult to generate high-fidelity physical models of complex sensing environments.

Machine learning or statistical learning provides an alternative description of the canopy scattering response and associated crop-soil parameter estimation. In a machine learning approach, microwave backscatter signals (inputs) corresponding to different crop variables (outputs) are linked by statistics of collected data. Usually, a complete machine learning algorithm involves some combination of unsupervised and supervised learning. In unsupervised learning, inputs are transformed by the machine as a result of inherent statistical structure. In supervised learning, the machine learns a mapping from inputs (electromagnetic signals) to outputs (values for crop variables such as biomass or soil moisture). Generally, the model (mapping) is found by minimizing a cost function involving losses across the training data and model complexity such that the expected loss induced by the model is small.

The machine learning approach is flexible in that no *a priori* electromagnetic assumptions are required for relating measured data to crop-soil variables. This avoids forward modeling errors. However, predictive accuracy and ability to generalize to new scattering environments are dependent on the quality and quantity of training data, and on the chosen statistical learning algorithm. There are many agricultural applications of machine learning with electromagnetic signals. A substantial body of work discusses learning crop-soil variables and classifying crops with hyperspectral imaging [33–35] and radar [36–39].

We chose to relate backscattered microwave signals directly to cranberry yield through a machine learning framework. This approach enabled us to generate a yield prediction system without making physical assumptions about the cranberry canopy or statistical assumptions about canopy element geometries or orientations. Such assumptions would be difficult to make due to the unique structure of the canopy and the particular way in which cranberries and leaves are randomly distributed and clustered within the canopy.

Chapter 3

The Microwave Sensing Characteristics of Cranberry Crops (Aim 1)

3.1 Introduction

We aim to investigate the microwave sensing characteristics of cranberry crops through the dielectric characterization of plant components and the computed backscattered radiation from simulated crop environments. Dielectric properties govern the interaction between electromagnetic waves and materials. We measure or formulate the dielectric properties of cranberry canopy components and incorporate them into simulations of microwave scattering from cranberry canopy models.

3.2 Microwave Sensing Foundation

3.2.1 Introduction

Long-range (e.g. [49–51]) and short-range (e.g. [42,52]) radar techniques have been previously explored for several crop biomass mapping applications. Here we investigate the feasibility of short-range pulsed (or swept-frequency) radar for cranberry yield estimation. We characterized the dielectric properties of cranberries and leaves to establish the physical basis of our proposed technique. We also conducted full-wave simulations of microwave scattering from random distributions of cranberries to examine the relationship between fractional berry volume and backscatter signal characteristics. We focused on the co-pol because the backscattered cross-pol is highly dependent on the randomness in cranberry distribution within the canopy. Finally, we computed the monostatic radar cross section of a 1.3 cm sphere with the dielectric properties of cranberry material. We related the monostatic radar cross section to the simulated cranberry canopy backscattered power.

3.2.2 Methods and Materials

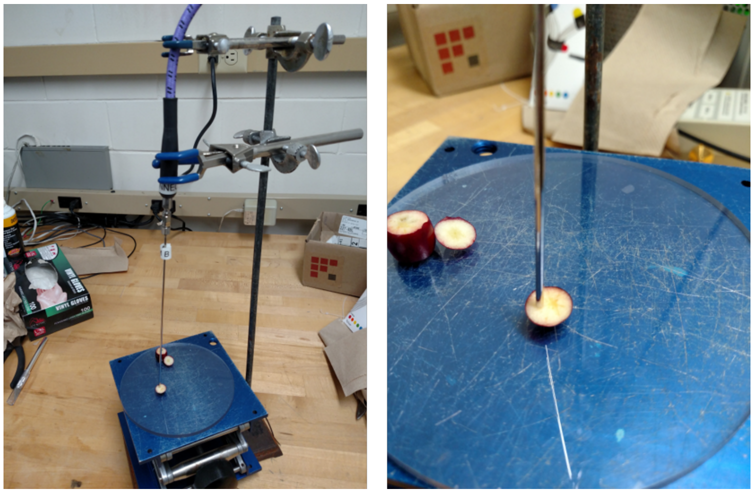
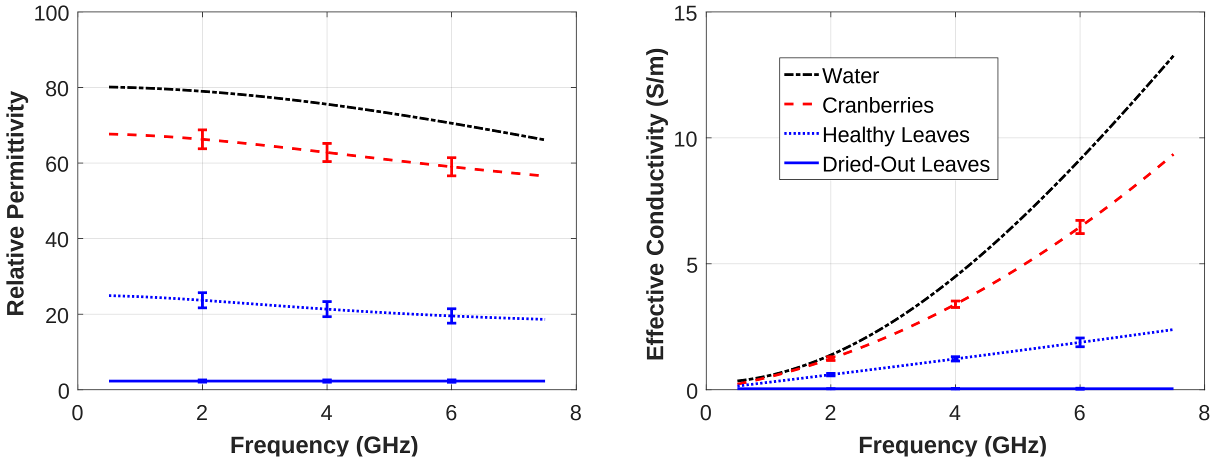


Figure 3.1: The setup used to measure the dielectric properties of cranberry tissue. The dielectric probe was placed in contact with the caps of each berry at four locations on the surface of each cap.

The dielectric properties of fresh cranberries and healthy and dried leaves from cranberry plants were measured over the 0.5-7.5 GHz range using a commercial narrow-diameter open ended coaxial probe and vector network analyzer. Figure 3.1 illustrates the probe placement normal to the surface of the sliced end caps (avoiding the small air pockets in the central part of the berries). We performed measurements at four probe locations on each of 80 berry end caps. Thus, a total of 320 cranberry dielectric data sets were collected. In addition, we removed healthy leaves from a cranberry plant, stacked them into 30 sets, and immediately conducted measurements on each stack with the same probe system setup. For comparison, we characterized an additional 30 sets of harvested leaves after drying out for three weeks. Each leaf stack was at least 2.8 mm thick to ensure a sufficient volume of material beneath the probe. Debye models were fit to the mean dielectric property measurements of the cranberry, healthy leaf, and dried leaf materials.

We created the idealized computational canopy structures in Computer Simulation Technology (CST) Microwave Studio. The computational canopy structures consisted of stacked rectangular prisms, each approximating a different layer in the cranberry crop-soil environments. Each stacked prism has a cross sectional area of 40 cm by 40 cm (roughly 30 cranberry diameters by 30 cranberry diameters), which allowed for significant randomness in spatial berry distribution within the canopy layer.

The top layer of our computational structure represented air. It consisted of a 20 cm thick homogeneous lossless dielectric with relative permittivity $\epsilon_r = 1$. The next layer represented the above-ground cranberry canopy and consisted of uniform random distributions of cranberry-dielectric spheres embedded in a homogeneous dielectric-mixture background



(a)

(b)

Figure 3.2: Debye models fit to the mean measured dielectric properties of cranberry-plant fruit and leaves. (a) Relative permittivity. (b) Effective conductivity. Variability bars span the range of measured values. A Debye model for water at room temperature [44] is shown for comparison.

of air and leaf material. The cranberry spheres were 1.3 cm in diameter, and the canopy layer ranged from 7.6 cm to 12.7 cm in thickness. These values reflect representative berry sizes and canopy thicknesses based on measured berries and canopies of commercial beds. We incorporated the Debye fits derived from our dielectric measurements of cranberries and foliage [6] into the material properties of the cranberry-dielectric spheres and the air-leaf mixture. The air-leaf mixture was defined by a two-component linear mixing formula with 90% air and 10% healthy-leaf. The Debye fits for the cranberry and healthy-leaf materials are shown in Fig. 3.2. The 10% healthy-leaf volume fraction was an average estimate derived from photographs of 10 canopies in central WI. We calculated an average individual leaf volume from weight measurements of leaves, counted the leaves in a known canopy volume within the photographs, and took the ratio of total leaf volume to total canopy volume. Illustrative distributions of the cranberry-spheres at 5% and 15% berry volume fractions (BVF) are shown in Fig. 3.3. Fig. 3.3 illustrates that cranberry distributions ranged from relatively sparse to relatively dense.

The bottom layer in the simulation structure is a dielectric half-space of moist soil. We defined the dielectric properties of the soil using the semi-empirical formula of [43]. This formula provides expressions for the real and imaginary parts of the relative permittivity of a soil medium in terms of the variables: sand, silt, and clay mass fractions, bulk density, volumetric moisture content, and the permittivity of water at a specified microwave frequency and physical temperature. We defined our soil-material with the soil parameters in [45], where the bulk density is 1.54 g/cm^3 and the mass fractions of sand, silt, and clay are 0.95, 0.025, 0.025 respectively. This soil composition describes regularly sanded cranberry beds [45] like those found in Wisconsin. The permittivity of the water is also described by the Debye-type dispersion equation of [43] at room temperature (20°C). We simulated soil moisture contents (SMC) of 10%, 15%, and 20% water by volume. These soil moisture contents cover the range of typical equilibrium values obtained with overhead irrigation and controlled water tables [45].

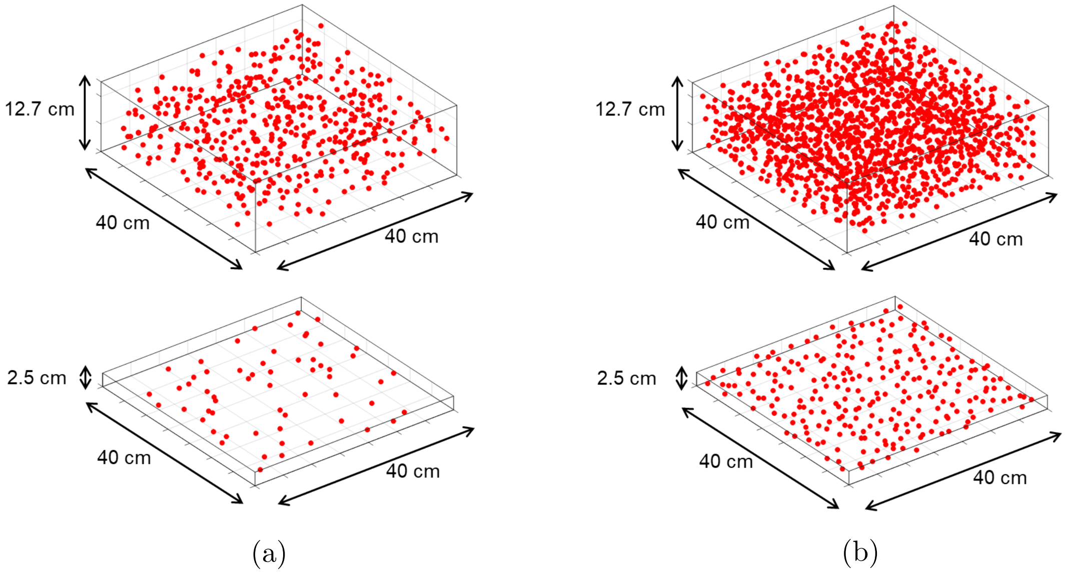


Figure 3.3: Realizations of randomly distributed cranberry-spheres within the 12.7-cm thick canopy layer of the simulated canopy structure. (a) Spheres constitute 5% of the canopy by volume. (b) Spheres constitute 15% of the canopy by volume. Each full-canopy visualization (12.7 cm × 40 cm × 40 cm) is accompanied by a view of a 2.5-cm-thick slice from within the canopy layer to illustrate the in-plane spacing between berries that is difficult to discern in full-canopy views.

The excitation in our simulations was a plane wave with horizontal linear polarization that was normally incident upon the canopy layer from the air layer. We chose plane wave excitation to evaluate canopy scattering free from antenna specific effects. An electric field probe was located within the air layer 20 cm away from the top of the canopy layer. We placed the probe directly in the cross-sectional center of the simulation domain, and recorded the time-domain incident and reflected electric field signals. We simulated excitations from 300 to 7500 MHz, but only report the results up to 2400 MHz, because, as discussed shortly, results above 2400 MHz were determined to be unsuitable for cranberry yield estimation. We simulated the canopy backscatter with 50 random spatial distributions of cranberry spheres at three berry volume fractions of 5%, 10%, 15% at each of the three soil moisture contents 10%, 15%, and 20%. For these simulations the canopy thickness was four inches (10.2 cm). In total, there were 450 simulated signals across the three soil moistures and berry volume fractions.

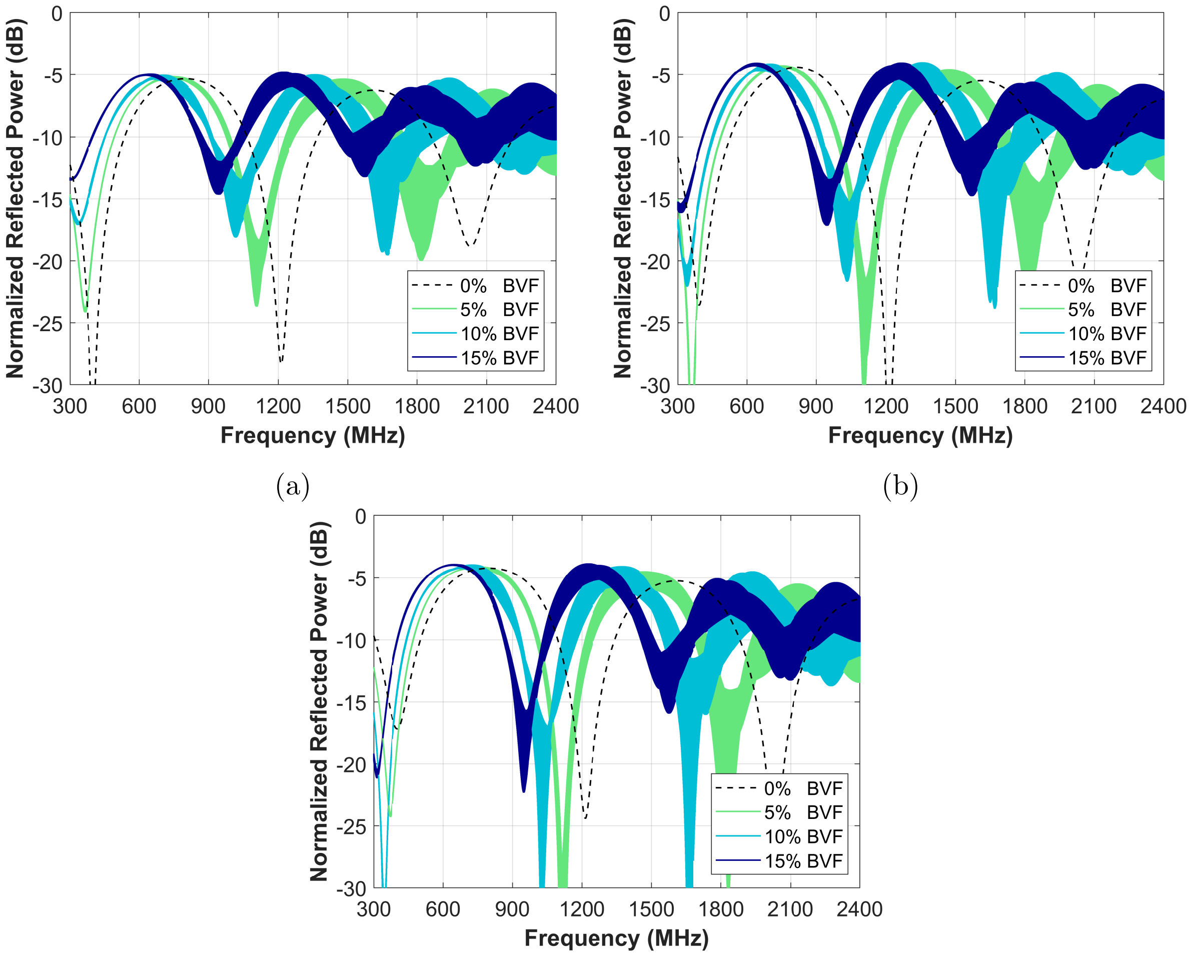
We also calculated the monostatic radar cross section for a 1.3 cm sphere with the cranberry dielectric properties described by the Debye model fit to the mean dielectric measurements of cranberry shown in Fig. 3.2. Monostatic radar cross section describes the power scattered by an object in the backwards direction for an incident plane wave. This calculation is summarized as follows. There is a linearly polarized plane wave incident upon the homogeneous dielectric 1.3 cm diameter sphere. The internal and scattered fields are described in terms of spherical wave functions that satisfy Maxwell's equations. We calculate the electric field scattered in the direction opposite to the direction of propagation for the incident plane wave. Then, we normalize the backscattered power (square of the magnitude of the electric field intensity) by the incident power (squared magnitude of the plane wave electric field intensity). Finally, we divide the normalized backscattered power by the geometrical cross sectional area of the cranberry sphere i.e. πR^2 where $R = 1.3/2$ cm. The resulting value for monostatic radar cross section represents the effective scattering size of the cranberry. A value of 1 denotes an object that is a 'strong' scatterer, while values close to zero indicate 'weak' scattering.

3.2.3 Results

Fig. 3.4 displays the normalized reflected power of the co-polarized component for cranberry volume fractions (BVF) 0%, 5%, 10%, and 15%, and volumetric soil moisture contents (SMC) of 10%, 15%, and 20%. The thicknesses of the colored lines in Fig. 3.4 represent the variations in backscattered power across 50 random spatial distributions of cranberry-spheres. A small line thickness equates to little change in backscattered power due to randomness in the spatial arrangement of cranberries. At 0% BVF there are zero berries within the canopy layer. Therefore, the 0% BVF curve represents one simulation without berries.

There are three apparent scattering regimes from 300 MHz to 2400 MHz. The first scattering regime occurs for frequencies below ∼900 MHz, where the spatial distribution of cranberries has little effect on the backscattered field. In this lower-frequency regime, the canopy layer containing the berries appears relatively homogeneous electromagnetically. As frequency increases from 900 MHz to approximately 1800 MHz the spatial arrangement of cranberries has an increasing significance on the backscattered power as seen by the thicker curves of each color. Mie scattering is most resonant at approximately 1400 MHz for a single typical cranberry. This is evident in Fig. 3.5 for the Radar Cross Section (RCS) of the cranberry sphere where the RCS sharply increases from a value of roughly -60 dB at 1200 MHz to -10 dB at 1400 MHz. Therefore, the berries begin to act as discrete scatterers above 1400 MHz and the canopy layer no longer appears homogeneous. Finally, above 1800 MHz (verified up to 7500 MHz) the spatial distribution of the cranberries has a strong influence on the backscattered power. This higher frequency regime is unsuitable for cranberry yield estimation because of the relatively large variance in backscattered power with berry spatial distribution. The large variance in backscattered power makes it prohibitively difficult to extract aggregate parameter values such as cranberry volume fraction.

As seen in Fig. 3.4, there is a significant change in simulated backscattered power between



(a)

(b)

(c)

Figure 3.4: Normalized simulated reflected power for a plane wave incident upon our computational representation of the cranberry canopy. Pulse parameters included 3775 MHz center frequency and 2900 MHz FWHM. The thickness of each colored curve displays the variation in backscattered power due to randomness in cranberry spatial distribution across 50 simulations at a fixed berry volume fraction (BVF) of 0%, 5%, 10%, or 15% within each graph and a fixed soil moisture content (SMC) of (a) 10%, (b) 15%, (c) and 20%. The canopy thickness is four inches (10.2 cm) in all simulations.

cranberry volume fractions for frequencies below roughly 2000 MHz. This is apparent when comparing the 0% BVF simulation with the 5% BVF simulations, where the structure of the backscattered power as a function of frequency is remarkably different for the two. Also, Fig. 3.4a shows a change in simulated backscattered power of roughly 15 dB between the 5% and 15% BVF at ∼1000 MHz. Therefore, in our simulations with fixed soil moisture content all practical cranberry volume fractions are discernible from the backscattered power. Importantly, almost all of the backscattered power for these configurations is at a level above -25 dB, and a large portion of the backscattered power is at a level above -15 dB. Therefore, a significant percentage of the incident power is scattered back from the canopy. The combination of a significant variation in backscattered power with BVF and a detectable value of backscattered power identifies the feasibility of low power microwave sensing for yield estimation. Finally, soil moisture content (SMC) affects the structure of the backscattered power spectral density as seen by comparing the heights of the peaks and depths of the nulls in Figs. 3.4a, 3.4b, and 3.4c, which were simulated with SMC's of 10%, 15%, and 20%, respectively. For example, at roughly 300 MHz the backscattered power at 15% BVF in Fig. 3.4a is approximately -13 db, while it is -16 dB in Fig. 3.4b and -21 dB in Fig. 3.4c. This indicates that soil moisture content has a significant influence on the backscattered power for a given cranberry volume fraction. This suggests that knowledge of soil moisture content combined with a measured backscattered power spectral density can be used to uniquely specify cranberry volume fraction at microwave frequencies below ∼2000 MHz.

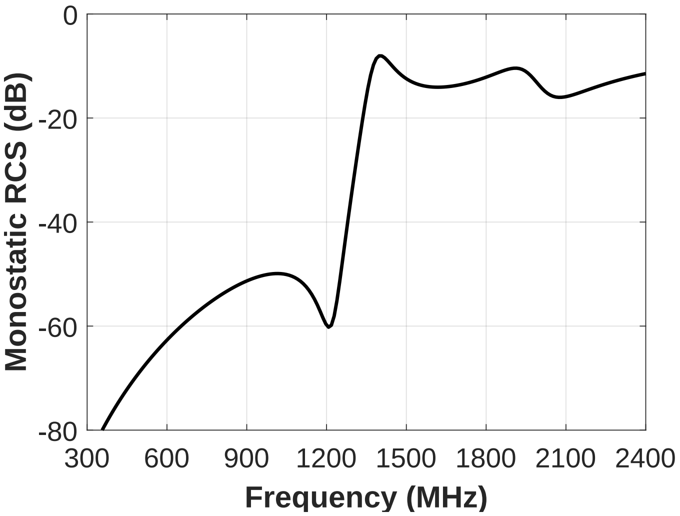


Figure 3.5: The monostatic Radar Cross Section for a 1.3 cm diameter sphere with the dielectric properties of cranberry tissue.

3.2.4 Summary and Conclusion

We presented dielectric measurements of freshly harvested cranberries and leaves measured over the 0.5-7.5 GHz frequency range. We observed a 3:1 dielectric contrast between berries and foliage. The relatively large dielectric constant of the cranberry flesh is due to its high water content. Similarly, the dielectric properties of the leaves are also dependent on water content, as seen by the difference in dielectric constant between healthy and dried-out leaves. We computed the microwave backscatter of simulated canopy structures and found that relevant cranberry volume fractions can be discriminated well at frequencies below 2 GHz. The first Mie resonance at roughly 1400 MHz explains the significant variation in backscattered power due to randomness in canopy cranberry distribution above that frequency. The relatively large dielectric contrast between the cranberries and leaves, combined with the ability to distinguish between cranberry volume fractions enables low-power microwave sensing of canopy cranberry content.

Chapter 4

Prototype Microwave Sensor Development and Experimental Data Collection (Aim 2)

4.1 Introduction

In this aim, we develop a prototype microwave sensor and collect computational and experimental scattering data, ground-truth cranberry yield, and measured cranberry bed parameters such as soil moisture in commercial cranberry beds. First, we investigate the radiation characteristics of our prototype sensor in simulation. We characterize the sensing volume of the sensor and its sensitivity to soil moisture, canopy depth, and yield. Next, we collect experimental data with our prototype microwave sensor using sods from a commercial cranberry bed in a greenhouse at UW-Madison. Finally, we collected field data in central Wisconsin with the prototype sensor in Fall 2017, Fall 2018, and Fall 2019 with the corresponding ground truth yield and soil moisture.

4.2 Simulated Prototype Sensor Characteristics

4.2.1 Introduction

Our aim is to use statistical learning to predict yield from the microwave signals collected with our prototype sensor. In order for any statistical learning algorithm to produce an accurate yield prediction, the yield needs to have an impact on the collected microwave signal. Thus, we simulate the prototype sensor in canopy environments with variable yield, soil moisture, canopy depth, and waveguide aperture separation to characterize the impact of each variable on microwave signals collected with our prototype sensor.

It is important to evaluate the sensing volume of our microwave sensor in order to have suitable data for training a statistical learning algorithm to accurately predict yield from the collected microwave signals. Yield is defined as the total mass of cranberries per unit surface area of canopy. Therefore, in training an algorithm to predict yield we have to designate the unit area in which we harvest the cranberries for the ground truth yield. Physically, microwave power is transmitted through a finite canopy volume i.e. sensing volume, and only the cranberries within this sensing volume have an effect on the collected signal.

There are two cases to consider in terms of overestimating and underestimating the sensing volume with respect to the ground truth yield. First, we are adding 'noise' to the target yield (output) if we overestimate the sensing volume and harvest cranberries outside of the true sensing volume. This is because the berries harvested outside of the true sensing volume have no effect on the microwave signal, but we try to map the microwave signal (input) to a yield (output) that is affected by those berries anyway. Second, if we underestimate the sensing volume then microwave power is transmitted beyond the berries harvested as the ground truth yield (output), which results in 'noise' added to the (input) microwave signal. In this case, the harvested berries had a definite impact on the collected microwave signal, but so did the surrounding unharvested cranberries. Thus, overestimating the sensing volume results in an unrepresentative ground truth yield, and underestimating the sensing volume results in an unrepresentative microwave signal for the corresponding ground truth yield. Consequently, having an accurate estimate of the prototype's sensing volume will enable enhanced statistical yield prediction through accurate ground truthing.

4.2.2 Methods

Computational Sensor Model

As illustrated in Fig. 4.1., we created an idealized computational model of our prototype microwave sensor in Computer Simulation Technologies (CST) Microwave Studio. The computational model has the same exterior dimensions as the actual prototype and is also constructed from lossy Aluminum material. The primary difference between the prototype and the computational model is in the coaxial-waveguide transition highlighted in Fig. 4.1b. We created a custom coaxial-waveguide transition for the model, since the prototype's transition from a 50 Ohm coaxial cable to TE_{10} mode current probe is proprietary. We optimized the custom coaxial-waveguide transition in order enhance the electromagnetic match between the prototype sensor and the computational model. In the optimization we minimized the discrepancy between the microwave signals collected with the prototype sensor opening into air in the laboratory and the computational model surrounded by an infinite air background. Specifically, we minimized the sum of squared differences between the magnitudes of the reflection coefficients measured by each sensor at the coaxial input to the coaxial-waveguide transition across the operating bandwidth. The operating bandwidth consists of the following discrete frequencies 606, 607, 608, ..., 1299, 1300 MHz corresponding to a propagating TE_{10} mode.

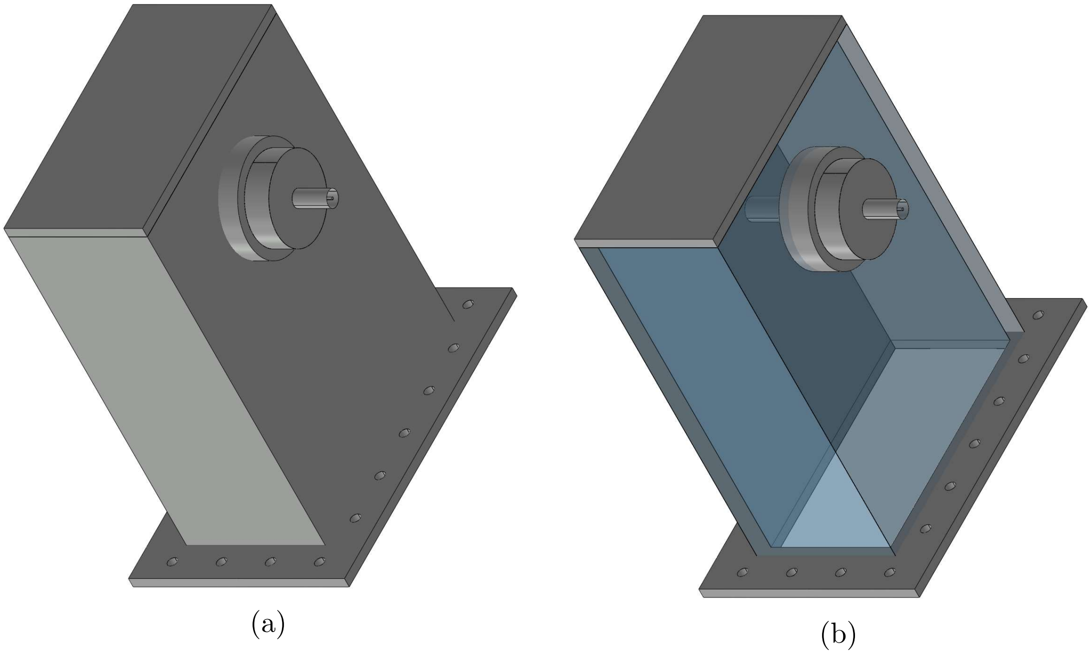


Figure 4.1: Computational model of prototype WR975 sensor in Computer Simulation Technologies (CST) Microwave Studio. (a) Exterior view. (b) Interior-exposed view with coaxial-waveguide transition. The interior of the waveguide is air-filled. The transition from 50 Ohm coaxial cable to a TE_{10} mode current probe is visible in (b).

Simulated Sensing Volume Evaluation

We evaluate the sensing volume of the computational sensor model after obtaining a good reflection coefficient match with the experimental prototype. There is consistent penetration of microwave power through the full depth of the canopy. Therefore, we are primarily interested in the lateral extent of microwave power transmission by the sensor. An illustration of the simulation setup is displayed in Fig. 4.2, in which we gauge the impact of lateral berry extent on the reflection coefficient measured at the coaxial input to the coaxial-waveguide transition. We define berry extent as the side length of a square surface area of canopy in which cranberries are randomly distributed through the full canopy depth. The canopy is a rectangular prism three inches deep by an area of 24” by 24”. There is a 10” deep soil slab directly beneath the canopy, and a 2” thick air layer between the sensor aperture and the top of the canopy. Berry extent is varied from 8” to 24” by 1” increments for a total of 17 unique berry extents. We simulated 12 random realizations of berry positions within the area bounded by the berry extent side length for each fixed berry extent. Therefore, across the 17 unique berry extents we completed $17 * 12 = 204$ simulations in total. We fixed yield at 315 g (per 1-ft^2 surface area of canopy) and soil moisture at 20% water by volume. The canopy slab consisted of 10% leaf material and 90% air material by volume.

We evaluate the impact of berry extent on the simulated reflection coefficients using a

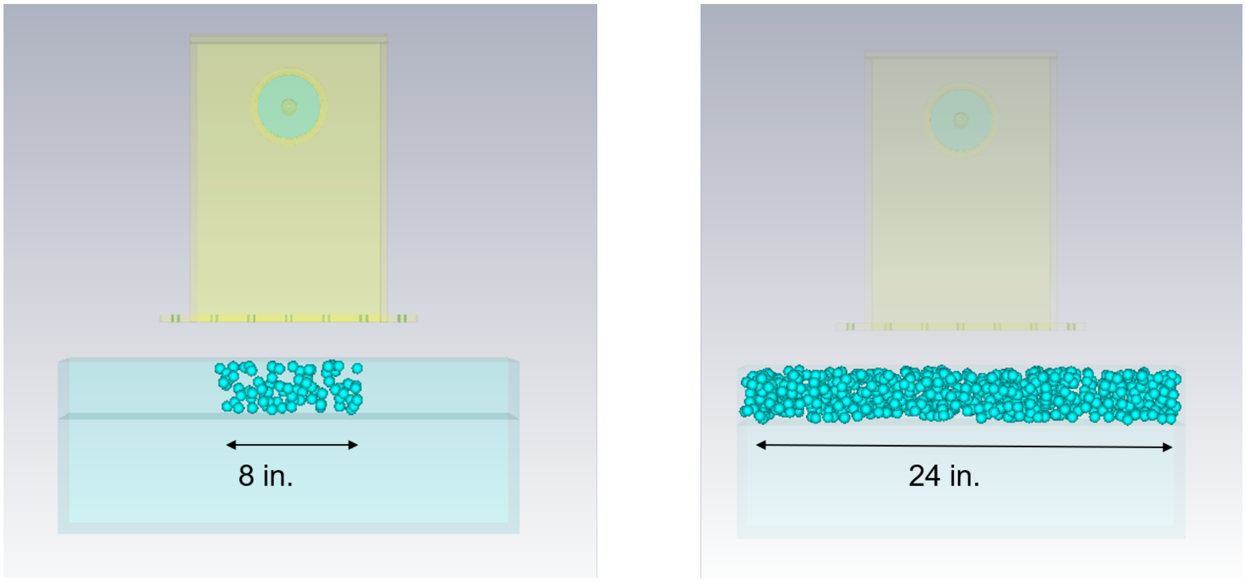


Figure 4.2: Illustration of two cranberry extents in the sensing volume study using the computational sensor model. (left) Eight inch cranberry extent. (right) 24 inch cranberry extent. Cranberry extent is variable and represents the side length of a square surface area of canopy in which berries are randomly distributed through the full depth of the canopy. The area in which the berries are distributed is centered along the central axis of the computational model.

dissimilarity metric D as defined in equation 4.1. $|S_{11}^{ref}(f_i)|$ is the magnitude of the reflection coefficient at frequency (f_i) measured at the coaxial input to the waveguide transition. There are 695 discrete frequencies from 606 MHz to 1300 MHz spaced by 1 MHz intervals. ref refers to the reference reflection coefficient and is the mean of the 12 reflection coefficient signals at the 24” berry extent. $test$ refers to the test reflection coefficient and is any of the 204 reflection coefficient signals from all of the simulations. Thus, the dissimilarity metric measures the deviation or normalized difference of any single reflection coefficient signal from the mean reflection coefficient signal at a 24” lateral berry extent. A small normalized difference corresponds to a reflection coefficient signal that is similar to the reflection coefficients at a 24” lateral berry extent.

$$D = \frac{\left[\sum_{i=1}^{695}\left(|S_{11}^{ref}(f_i)| - |S_{11}^{test}(f_i)|\right)^2\right]^{1/2}}{\left[\sum_{i=1}^{695}\left(|S_{11}^{ref}(f_i)|\right)^2\right]^{1/2}} \tag{4.1}$$

Simulated Sensitivity to Canopy Environment Variables

We evaluated the sensitivity of the reflection coefficients measured with the computational model to canopy environment variables. We used a simulation environment that was similar to the sensing volume study. Geometrically, the computational sensor model was located directly above a slab model for the canopy. There was variable separation between the sensor aperture and the top of the canopy. We simulated separations of 1.0”, 1.5”, and 2” with a slab of air material between the aperture and the top of the canopy. We simulated variable canopy depths, where the thickness of the canopy slab was 3”, 4”, and 5”. The canopy slab again consisted of a 10% leaf material and 90% air material by volume. Cranberry spheres were generated inside the canopy slab at random spatial positions. We varied yield by changing the number of cranberry spheres inside of the canopy slab. Again, yield is defined as the mass of cranberry material per 1-ft^2 surface area of canopy. We varied yield from 100 g to 600 g in 20 g increments. Finally, the canopy slab was backed by a soil half-space. We varied the soil moisture content of the soil half-space from 10% to 40% in 5% increments.

We evaluated the relative sensitivity of the sensor to each canopy environment variable using the sensitivity metric shown in equation 4.2. The relative sensitivity measured by the metric is in terms of the simulated reflection coefficient signals $\overrightarrow{S_{11}}$. Specifically, the sensitivity metric measures the impact of an incremental change in a canopy environment variable on the reflection coefficients measured at the coaxial input to the coaxial-waveguide transition. Looking at equation 4.2, $\overrightarrow{S_{11}}$ denotes the reflection coefficient signal as a 695 length vector of reflection coefficient magnitudes at frequencies between 606 MHz and 1300 MHz (695 discrete frequencies spaced by 1 MHz intervals). The x in $\overrightarrow{S_{11}(x)}$ represents the canopy variable value, and $x + 1$ represents the incremented canopy variable value. For example, $x = 100$ g and $x + 1 = 120$ g when discussing the sensitivity to yield at 100 g for a single increment of 20 g i.e. $\overrightarrow{S_{11}(100)}$. The $\|\|$ denotes the l^2 norm. Therefore, the sensitivity metric is the square root of the sum of squared differences between reflection coefficient magnitudes at each frequency.

$$S(x) = \left\| \overrightarrow{S_{11}(x)} - \overrightarrow{S_{11}(x+1)} \right\| \tag{4.2}$$

4.2.3 Simulated Data

Computational Sensor Model Results

The measured reflection coefficient signals for the experimental prototype sensor and the optimized computational sensor model are shown in Fig. 4.3. It is apparent that the optimized structure of our custom coaxial-waveguide transition did not perfectly match the experimental prototype. This physical discrepancy can be seen in the mismatch between the reflection coefficient curves in Fig. 4.3. There is a maximum difference in the magnitude of the reflection coefficient of 0.06 at roughly 940 MHz. The next largest difference in the magnitude of the reflection coefficient occurs at roughly 1170 MHz and is 0.03. Apart from narrow bands around these two frequencies, there is a 0.01 or less difference in the magnitude of the reflection coefficients across the operating band of the sensor. Therefore, the computational model is an accurate simulated representation of the experimental prototype.

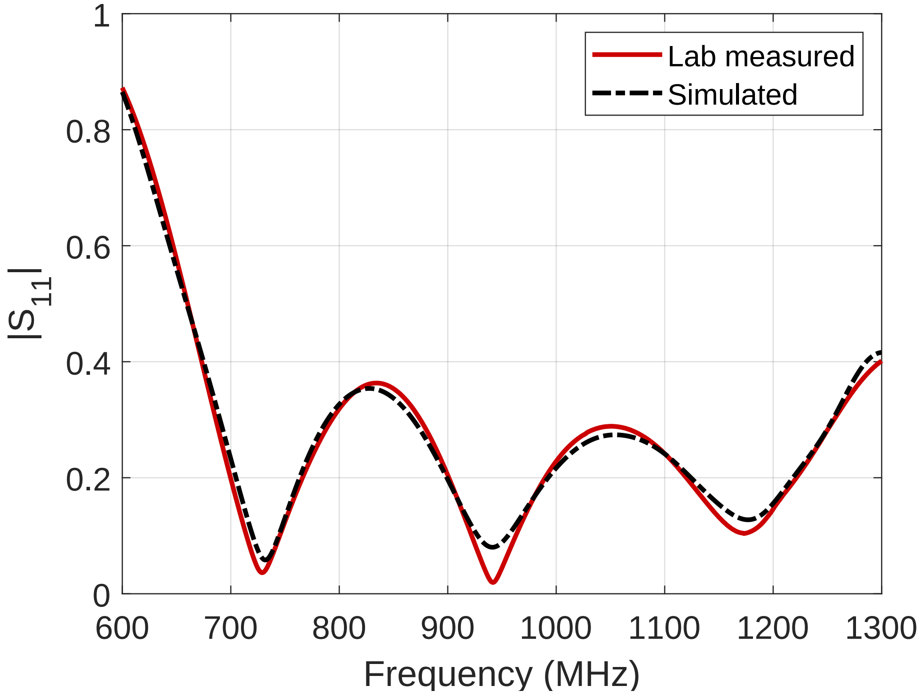
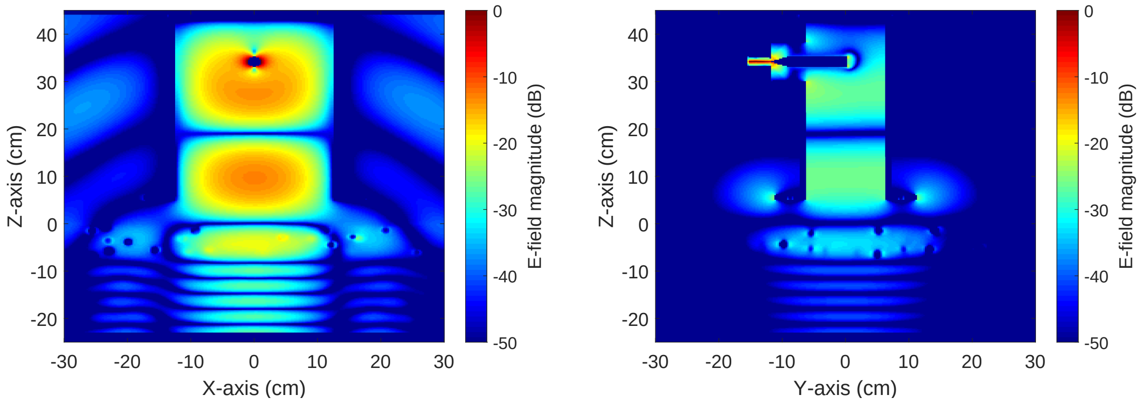


Figure 4.3: Reflection coefficients of WR975 waveguide in laboratory vs. simulated reflection coefficients of our computational model in Computer Simulation Technologies (CST). The laboratory and simulated reflection coefficients were referenced to the 50 Ohm coaxial input to the coaxial-waveguide transition.

We display the radiation characteristics of the computational sensor model at 1000 MHz in Fig. 4.4. As seen in Fig. 4.4, the aperture of the computational model is separated from the top of the canopy slab by 2” (z ∈ [0, 5] cm), the canopy slab is 3” deep (z ∈ [-7.6, 0] cm), and it is backed by a soil half-space (z ∈ [-25, -7.6] cm). The maximum magnitude of the electric field intensity occurs at the coaxial input. Importantly, there is significant microwave power transmitted through the entire depth of the canopy and across the full extent of the sensor aperture. This shows that full canopy depth sensing is achieved at 1000 MHz. We verified full canopy depth sensing across the operating bandwidth of the sensor (606-1300 MHz).

Another interesting feature of the radiation is that significant power is transmitted down into the soil. From Fig. 4.4a there are electric field intesities with magnitudes greater than -30 dB at -25 cm on the z-axis, or roughly 7” below the top of the soil. This is a negative aspect of the sensor at 1000 MHz. Scattering objects like rocks or pipes near the surface of the soil would cause a spurious reflection in the collected microwave signal which would



(a)

(b)

Figure 4.4: Simulated electric field magnitude of TE_{10} mode at 1000 MHz. (a) X-Z plane view. (b) Y-Z plane view. The magnitudes of the electric field intensities are referenced to the maximum electric field intensity at the coaxial input. The cranberries are distributed within a three inch canopy, two inches separated from the aperture. The yield shown is 315 grams of cranberry per 1-ft^2 surface area of canopy, and the soil moisture content is 20% by volume.

complicate yield prediction.

The final noteworthy feature of this plot is the scattering from individual cranberries. As shown in Fig. 4.4, there are dark blue circles apparent between -8 cm and 0 cm on the z-axis. The dark blue corresponds to an electric field intensity magnitude smaller than -40 dB, and each circle corresponds to an individual cranberry. Thus, the magnitude of the electric field intensity inside each of the cranberries in the cross sectionional view is less than -40 dB. Physically, this corresponds to scattering of electric fields incident on individual cranberries at 1000 MHz.

Scattering from individual berries has advantages and disadvantages. A disadvantage is that a high level of scattering from individual berries increases the variance of the backscattered power for fixed yields. This is because each berry scatters power significantly, and so the position of each berry with respect to the sensor aperture impacts the overall power backscattered toward the aperture. Ideally, the spatial positions of the cranberries within the canopy would be immaterial since we are interested in the aggregate mass of cranberry material in the canopy i.e. the yield. The advantage of scattering from individual berries is that the berries have a strong impact on the received signal. Consequently, as scattering from individual berries intensifies, the aggregate impact of the berries (yield) on the received signal increases. However, as scattering from individual berries intensifies so does the variance in the received signal for fixed yields due to the increasing dependence on spatial positions of the berries.

Simulated Sensing Volume Results

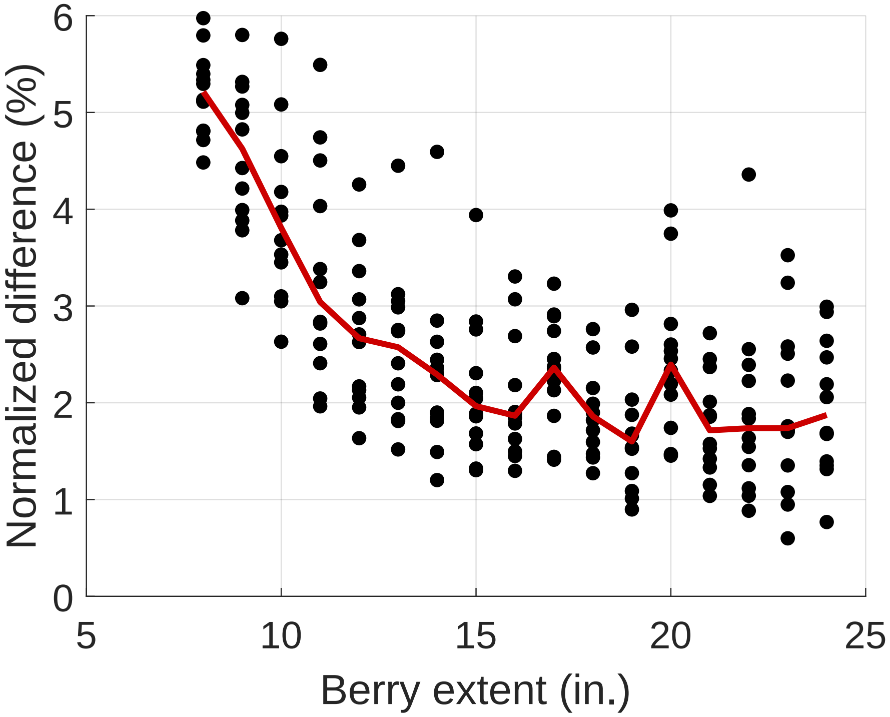


Figure 4.5: The dissimilarity metric i.e. the normalized difference between test signals and a fixed reference signal as a function of berry extent. Test signals are the reflection coefficients as a function of frequency from 606 MHz to 1300 MHz for 12 random distributions of cranberries within the area bounded by each berry extent side-length. The fixed reference signal is the mean of the 12 reflection coefficient signals at each frequency for the 24 inch berry extent case. Black dots represent the dissimilarity metric for each test signal. The red line denotes the median of the 12 dissimilarity metric calculations for each berry extent.

In Fig. 4.5, we plot the dissimilarity metric of equation 4.1 for the sensing volume simulations involving variable lateral berry extent. The black dots represent the value of the normalized difference between each reflection coefficient signature and the mean of the 12 reflection coefficient signatures at a 24” berry extent. The median of the 12 normalized differences at each berry extent is shown in red. There is a downward trend in the median normalized difference as berry extent increases from 8” to 24”. It appears that the median normalized difference converges to approximately 2% for berry extents between 15” and 24”. Convergence of the median normalized difference implies that the set of measured microwave signals for any given lateral berry extent are similar to the set of measured microwave signals for a 24” berry extent. As an example, the median normalized difference at a 15” berry extent is equal to the median normalized difference at a 24” berry extent. Thus, the berries located between 15” and 24” in terms of lateral extent had a negligible impact on the collected microwave signals on average.

The normalized difference at a 12” berry extent is roughly 2.5%. There seems to be a 0.5% variation around the 2% value as seen from the normalized differences at 17”, 19”, and 20”. This implies that the 12” berry extent could be classified as part of the convergent group of berry extents. At the very least, we are confident that a 12” berry extent is close to the true lateral sensing extent of the sensor. Based on the results depicted in Fig. 4.5, the sensing volume is reasonably defined as the full depth of the canopy by a surface area of canopy with dimensions 12-15” by 12-15”.

Simulated Sensitivity to Canopy Environment Variables

In Fig. 4.6 we display the sensitivity, as defined in equation 4.2, of the simulated reflection coefficient signal to the canopy environment variables of yield, soil moisture, aperture separation, and canopy depth. For our purposes, the sensitivity metric value is only useful for comparing the relative effects of incremental changes in canopy environment variables and whose absolute value is irrelevant. As seen in Fig. 4.6a, the sensitivity metric is plotted as a distribution (boxplot) for each value of yield. The distribution of sensitivity metric values arises from calculating the sensitivity metric across all simulated canopy environments. For example, there are simulations for a fixed yield at all seven soil moisture contents, all three aperture separations, and all three canopy depths for a total of 63 canopy environments. Therefore, the distribution of sensitivity metric values displayed for each yield in Fig. 4.6a is derived from the 63 simulations where yield was held constant and the 63 simulations where yield was incremented by 20 g. Similarly, for each fixed soil moisture content there are 26 yields, three aperture separations, and three canopy depth for a total of 234 calculated sensitivity metric values. Finally, for each fixed aperture separation there are 26 yields, seven soil moisture contents, and three canopy depths for a total of 546 calculated sensitivity metric values. Similarly for each fixed canopy depth there are 26 yields, seven soil moisture contents, and three aperture separations for a total of 546 calculated sensitivity metric values.

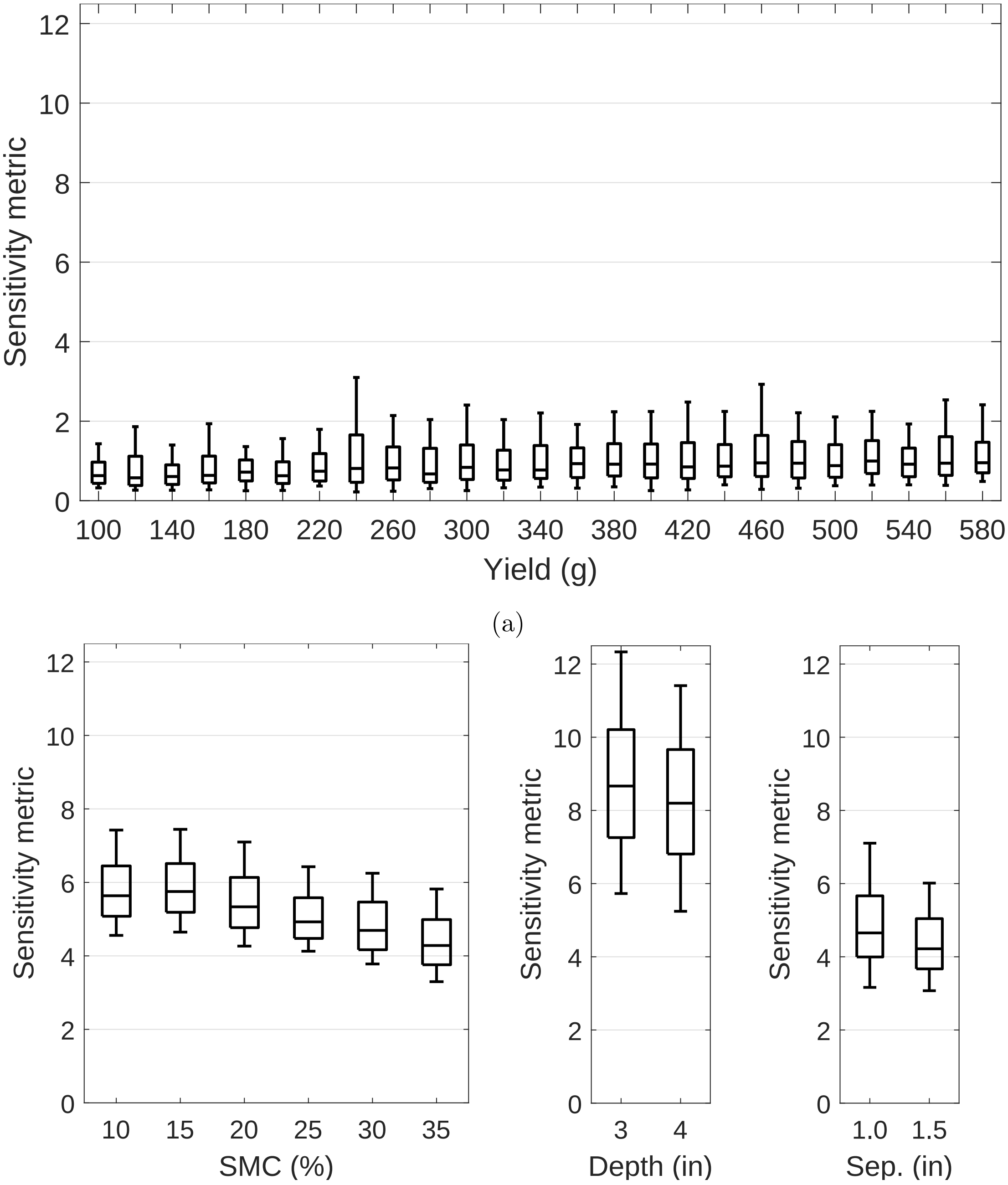
As seen in Fig. 4.6a the sensitivity metric is less than 2 for a majority of yields. Also, the median sensitivity metric value is roughly constant at one for all yield values. Therefore, the impact of yield on the simulated reflection coefficient is roughly constant across all expected yields. This means that a change in yield from 100 g to 120 g has the same effect on the reflection coefficient signal as a change in yield from 580 g to 600 g. This is advantageous because it implies that yield prediction will be equally difficult across all breeds of cranberry plant (high yielding or low yielding) and time periods in the growing season for comparable canopy structures.

Figure 4.6b illustrates the impact of a 5% increment in soil moisture on the simulated reflection coefficient signal. The sensitivity metric values range between 4 and 8, and the median has a downward trend as soil moisture increases. This is an interesting feature of the simulations. The downward trend in the calculated sensitivity metric indicates that changes in soil moisture at higher base levels of soil moisture have a diminishing impact on the simulated reflection coefficient signal. As an example, a change in soil moisture from 10-15% has a larger impact on the reflection coefficient than a change in soil moisture from 35-40%. This trend indicates that higher soil moisture levels are advantageous to yield prediction since fluctuations in soil moisture around higher base levels have a diminished impact on the measured signal compared to fluctuations around lower base levels.

Since we only simulated three aperture separations and canopy depths it is impractical to look for a trend. However, as seen in Fig. 4.6d, an increment of 0.5” in aperture separation produced an effect on the simulated reflection coefficient comparable to a 5% change in soil moisture, and two to three times larger than an increment of 20 g in yield. This is an adverse result, since in the field fluctuations of 0.5” in aperture separation are commonplace. However, it should be noted that aperture separation and canopy depth results are less reliable than soil moisture and yield since they tie in closely with the slab geometry of the simulated canopies. The real-world cranberry canopies do not have the perfect facets found in our simulated canopy slabs, which could be partly responsible for the large sensitivities of aperture separation and canopy depth.

4.2.4 Summary

We developed a computational model of our prototype WR975 waveguide sensor. The computational sensor model was optimized for a good match to the prototype in terms of measured reflection coefficients. We used the computational sensor model to estimate the sensing volume of the prototype sensor. The sensing volume extended through the full depth of the canopy and extended laterally to cover a canopy surface area of 12-15” by 12-15”. We found that a 20 g increment in yield produced the smallest effect on simulated reflection coefficient relative to soil moisture and aperture separation. However, the effect of incremental changes in yield on the reflection coefficient is consistent from 100 g to 600 g.



(a)

(b)

(c)

(d)

Figure 4.6: Sensitivity metric vs. (a) Yield, (b) Soil moisture content (SMC), (c) Canopy depth (Depth), and (d) Aperture separation (Sep.). The sensitivity metric (Eq. 4.2) is the norm of the forward difference between reflection coefficient signals resulting from a single increment in one of the canopy-environment variables. Statistics on the sensitivity metric are calculated by varying the other three canopy-environment variables for two fixed sequential test variables e.g. variable smc, depth, and sep. for a yield of 100 g and a yield of 120 g.

4.3 Experimental Data Collection with Cranberry Sods

We cut sods from Valley Corporation (a commercial cranberry farm in central Wisconsin) to be used in experimental data collection. We collected four sods in total and stored them in a greenhouse at UW-Madison. The cranberry vines growing in the sods were kept alive to ensure a realistic cranberry crop canopy during data collection. We completed an initial round of data collection using one of the sod-canopies. We collected microwave reflection coefficient data with our prototype sensor for controlled yields and soil moistures.

4.3.1 Data Collection with Prototype WR975 Sensor

We collected data with our prototype microwave sensor mounted above one of the sod-canopies in the greenhouse. Our experimental setup is shown in Fig. 4.7. As illustrated in Fig. 4.7, a custom PVC mount positioned the aperture of the sensor 2” directly above the canopy of the sod. We placed a 12” by 12” marker on the sod’s canopy to delineate a 1-ft^2 site. The sod measured 2’ by 2’ in cross section, and the soil was 3-4” deep. The canopy of the sod grew 3-6” above the top of the soil. We measured the microwave reflection coefficients at the coaxial input to the coaxial-waveguide transition using a portable vector network analyzer at 695 discrete frequencies (606, 607, ..., 1299, 1300 MHz).

We controlled the soil moisture in the sod’s soil through watering. To achieve a smaller soil moisture content we let the soil dry out for variable lengths of time after watering. To achieve a higher soil moisture content we collected data in closer proximity to watering. We measured soil moisture with a calibrated Vernier soil moisture sensor with 3.5” probe tips inserted into the center of the sod. Yield was artificially controlled by placing a known mass of freshly harvested cranberries into the sod’s canopy. We attached cranberries to nylon fishing line and spaced them throughout the canopy as depicted in Fig. 4.8. We varied yield from roughly 100 g to 500 g in 20 g increments. We measured the mass of the cranberries on each fishing line before placing them on top of and around existing berries in the canopy. As illustrated in Fig. 4.8, we filled in the space around existing cranberry-lines with new increments until we had to place the new lines on top. After placing a new cranberry-line in the canopy, we collected 12 microwave reflection coefficient measurements with the prototype sensor. We obtained 12 measurements by translating and rotating the sensor aperture above the 1-ft^2 site.

4.3.2 Experimental Data

We plot the ground truth soil moisture content and yield from the five days of greenhouse data collection in Fig. 4.9. There are 21 unique yields at each of the five soil moisture contents in the range of 10% to 20% for a total of 105 unique pairs of soil moisture and yield. We collected a total of 1260 microwave measurements across the 105 soil moisture and yield combinations.

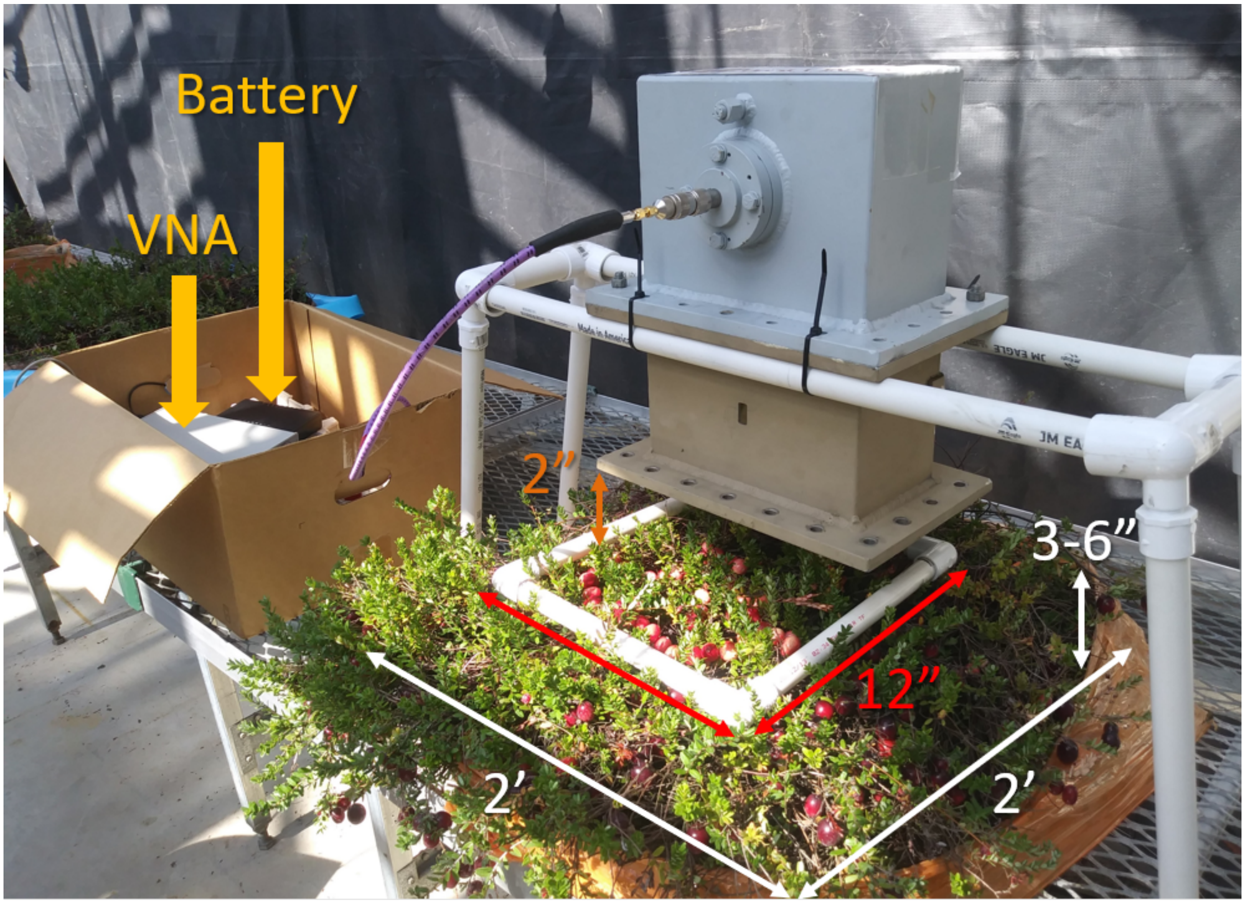
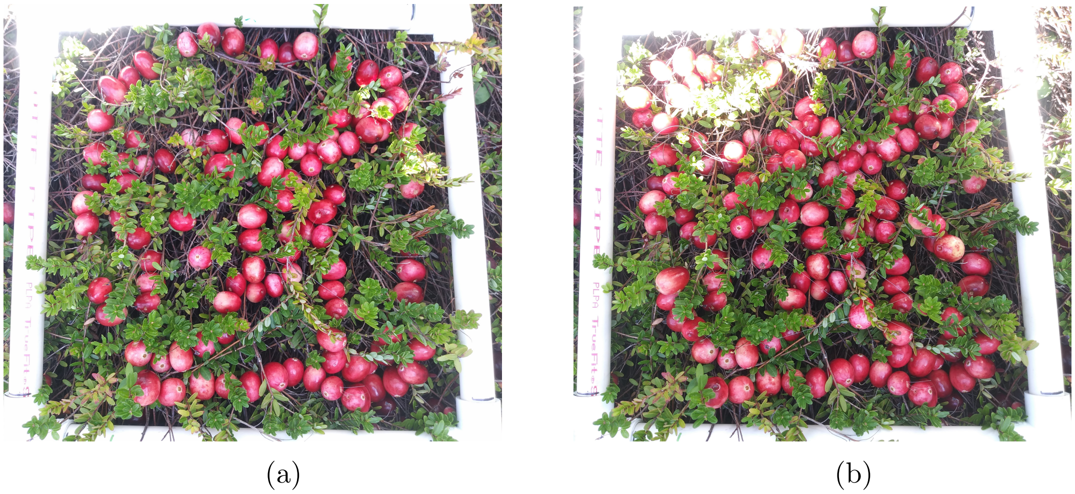


Figure 4.7: Experimental setup for microwave data collection using a cranberry sod-canopy and our WR975 sensor in a University of Wisconsin-Madison greenhouse. The sod is set on a square wooden board with a two foot side length. The canopy depth ranges from 3-6” above a 4-5” deep soil base. The sensor is mounted on a PVC support, and its aperture is roughly 2” above the top of the canopy. A 1-ft^2 PVC marker bounds the site in which cranberries are randomly distributed. The sensor aperture is suspended directly above the site bounded by the marker.

In Fig. 4.10, we show the mean of the 12 reflection coefficient measurements at each of the 21 yields between 100 g and 500 g for the data collection with 11.3% soil moisture content. From the figure, it is apparent that yield has a distinguishing effect on the measured reflection coefficients. It’s also apparent that certain frequency bands are more suitable than others for distinguishing yield. For example, the reflection coefficients between 1100 MHz and 1200 MHz are separated almost uniformly as a function of yield, whereas the reflection coefficients at 800 MHz are indistinguishable. Between 1100 MHz and 1200 MHz there is a significant change in the magnitude of the reflection coefficient from 0.1 at 100 g to 0.3 at 500 g. This plot proves experimentally that yield can be predicted to within 20 g of the



(a) (b)

Figure 4.8: Photographs of the same 1-ft^2 site within a greenhouse sod-canopy depicting two different yields. (a) 220 g. (b) 280 g. Recently harvested cranberries are spaced along fishing line and placed within the canopy of the site. The total yield is varied by placing additional cranberry-lines with known cranberry mass in the canopy.

ground truth when soil moisture content and canopy structure are held constant. This plot also illustrates the suitability of certain sub-bands for yield prediction over others.

4.3.3 Summary

We developed a prototype microwave sensor constructed from a WR975 waveguide with a PVC mounting. We used the microwave sensor to collect experimental scattering data with a sod-canopy cut from a commercial cranberry bed. We controlled yield and soil moisture within the sod and collected microwave data for 105 unique soil moisture and yield pairs. We showed that yield prediction from the measured microwave reflection coefficients is favorable for fixed soil moisture contents, and accuracy to within 20 g of the ground truth is possible.

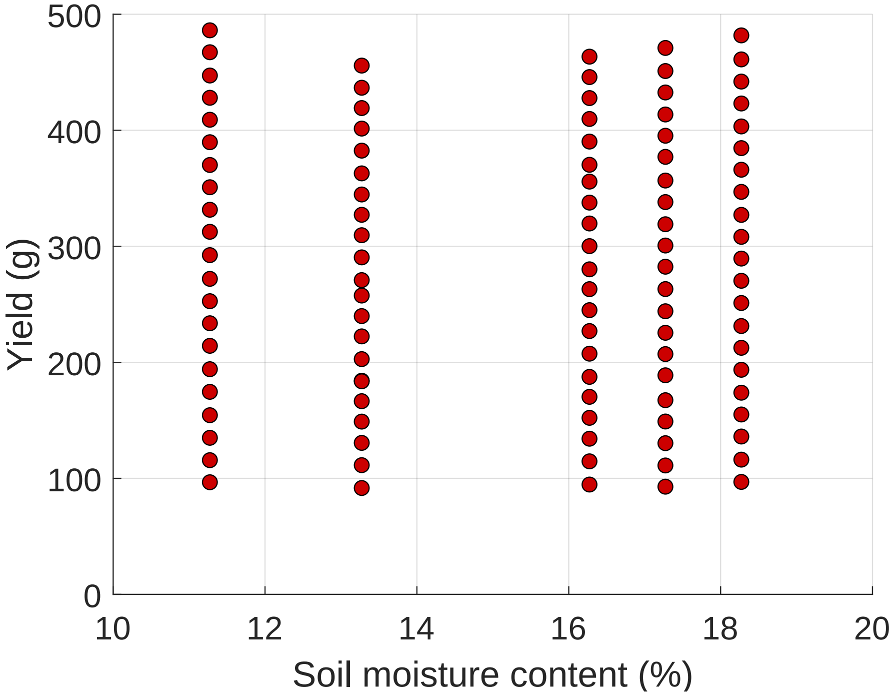


Figure 4.9: Ground truth data collected with a single sod-canopy in the greenhouse. The yield was varied from 100 g to 500 g in 20 g increments. Soil moisture was varied by adjusting the level of watering and proximity in time to data collection.

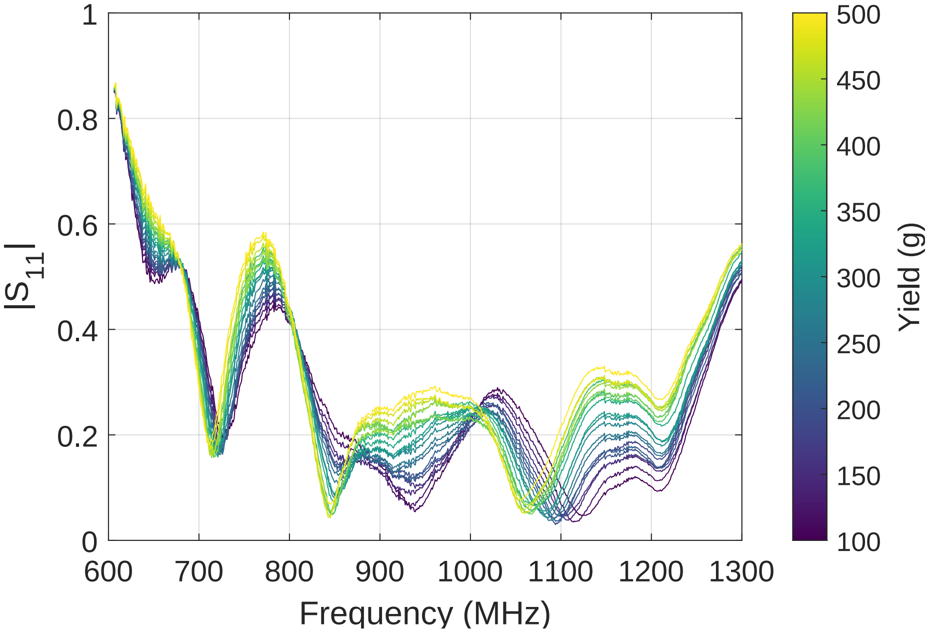


Figure 4.10: Magnitude of the microwave reflection coefficient as a function of frequency from 606 MHz to 1300 MHz for 21 unique yields between 100 g and 500 g for a fixed soil moisture of 11.3%.

4.4 Field Data Collection with Prototype Sensor

4.4.1 Methods and Materials

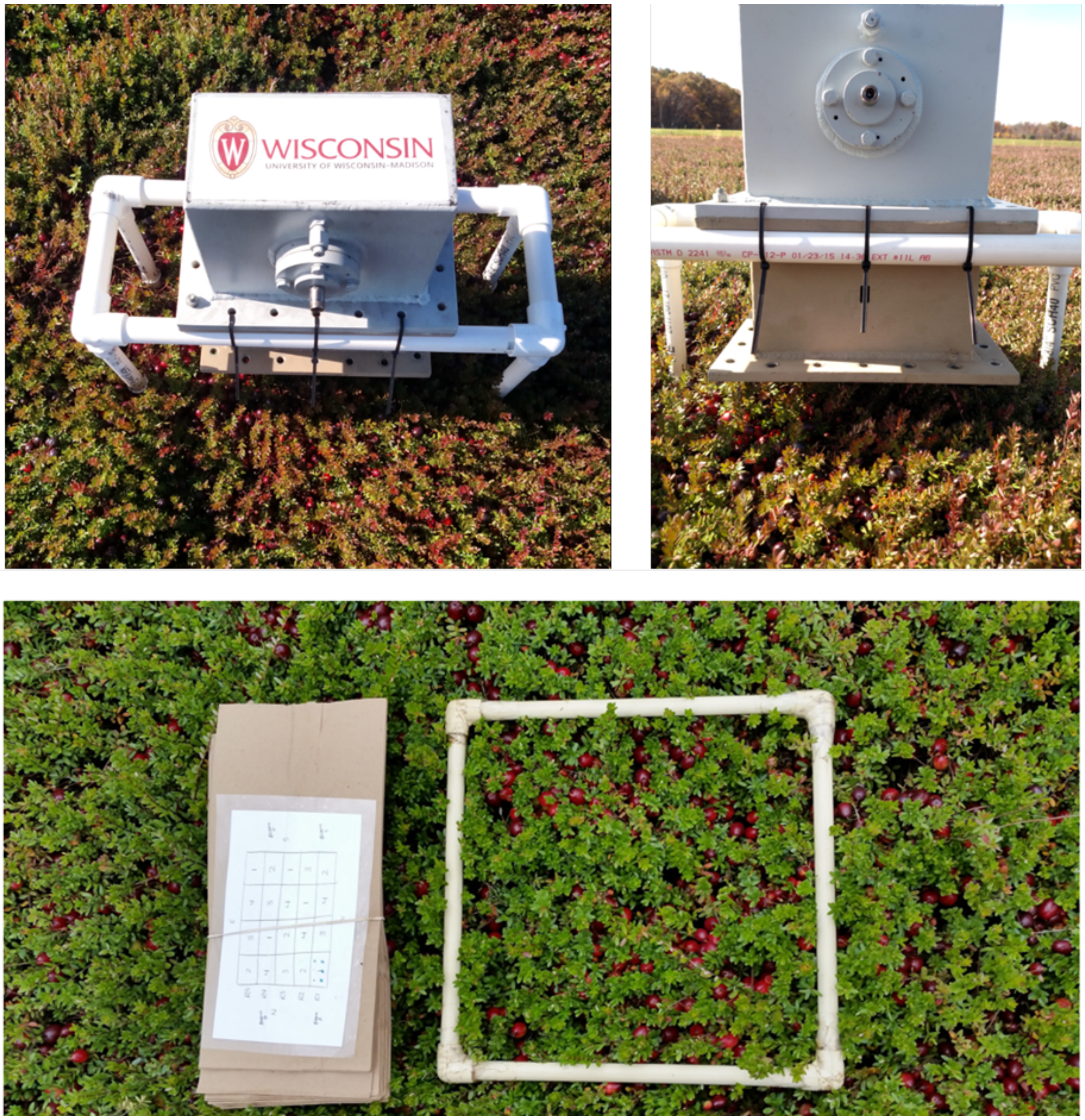


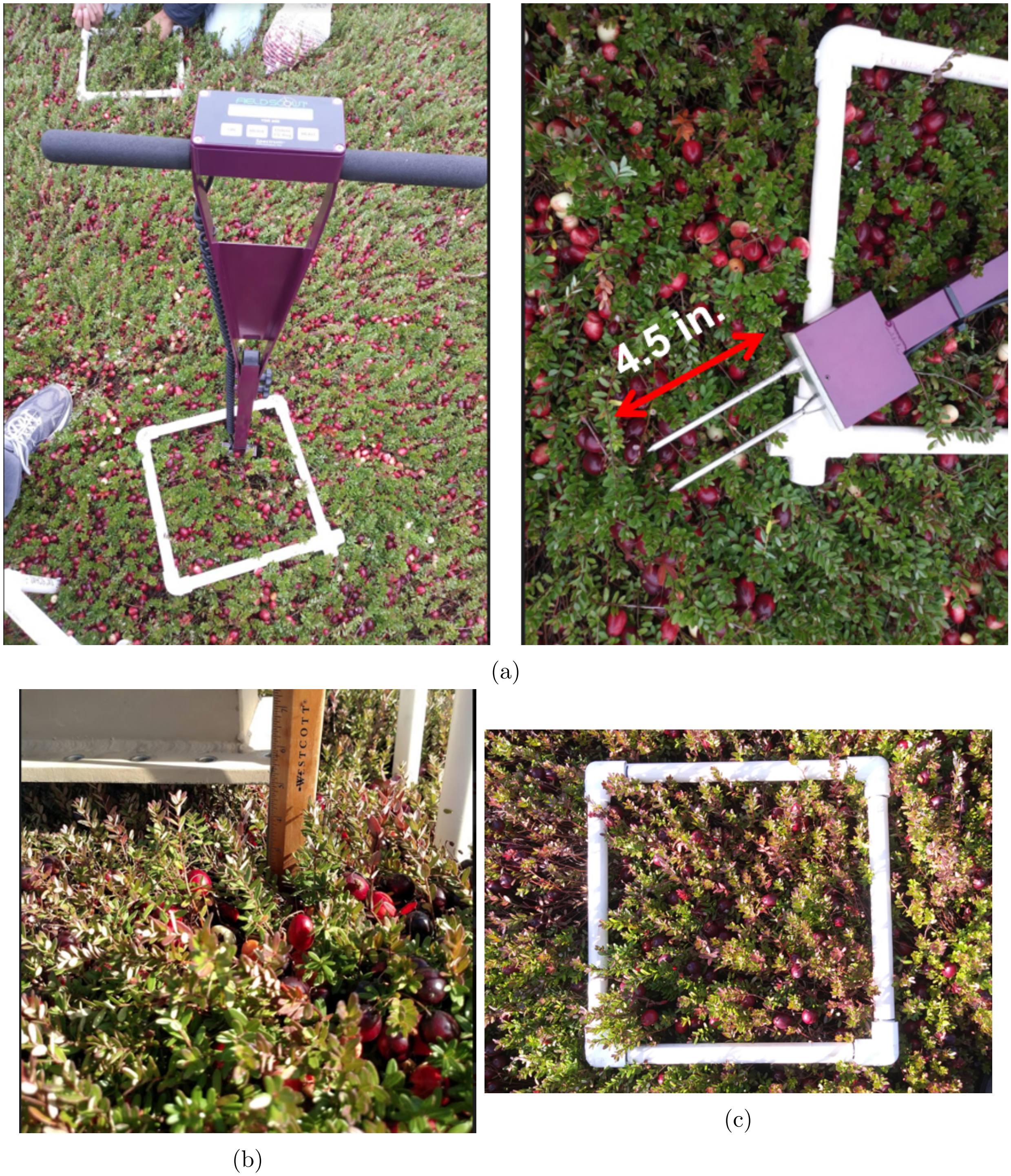
Figure 4.11: A rectangular open-ended waveguide (WR975) mounted above a cranberry canopy in central Wisconsin (top, right and left). A 1-ft^2 marker delineating a site for sampling the cranberry bed yield (bottom).

As illustrated in the previous chapter, the idealized plane wave simulations indicated that microwave backscatter for frequencies below 2000 MHz is well suited for yield discrimination. With this information, we chose an open-ended WR975 rectangular waveguide as our experimental microwave transceiver, shown in the top two photographs of Fig. 4.1. We mounted a 16 inch (40.6 cm) long section of WR975 on a polyvinyl chloride (PVC) structure for suspension above the cranberry canopy, as shown in the top two photographs of Fig. 4.11. The waveguide aperture has cross sectional dimensions of 4.875 inches by 9.75 inches (12.4 cm by 24.8 cm). The height of the waveguide aperture above the ground or soil surface was fixed at 8 inches (20.3 cm) by the PVC mounting structure, which corresponds to the higher canopies we encountered in central WI. We connected the waveguide system with a phase-stable cable to a commercially available compact vector network analyzer (VNA), powered by a portable 12 V DC battery.

Our measurements consisted of reflection coefficient magnitudes at 695 discrete frequencies spaced linearly from 606 MHz (TE_10 cutoff frequency) to 1300 MHz. Phase was not used in this study, although it could be included as an input to statistical learning algorithms as well. The output power at the VNA port was 0.5 mW for each frequency. Our reflection coefficient measurements were referenced to the termination of the phase-stable cable at the input of the type-N coaxial to WR975 transition. Our measurements were referenced to this plane using a three-measurement calibration procedure with different cable terminations (open, short, matched 50 Ohm load). Since the measurements were not referenced to the aperture of the waveguide, the effects of the type-N to WR975 transition, length of the waveguide section, and conductor loss are present in all of the reflection coefficient measurements.

The established method for estimating yield within cranberry beds was to harvest all of the cranberries within 1-ft^2 sites delineated by one square foot markers and extrapolate the berry yield to account for the total area of each bed. We also used 1-ft^2 sites to benchmark our yield estimation procedure, where each 1-ft^2 site was delineated with a one square foot PVC marker. As determined in the sensing volume study, a 1-ft^2 surface area approximates the sensing extent of the sensor. Our microwave sensor was placed directly above the surface of the cranberry canopy. As seen in the bottom illustration of Fig. 4.11, the marker lied on top of the cranberry canopy surface. We then obtained 12 independent reflection coefficient measurements (S_{11} sweeps across frequency) at each 1-ft^2 by randomly rotating and/or translating the sensor aperture above each site. After the 12 measurements at each 1-ft^2 site were logged, we harvested all of the cranberries in the canopy volume bounded by the marker.

Figure 4.12a displays the sensor we used to measure soil moisture within each 1-ft^2 site. The sensor shown in Fig. 4.12a is a Field Scout TDR 350. We used the 4.5 inch metal rods to transmit an electromagnetic wave into the soil. The Field Scout TDR 350 converts the amplitude of the reflected electromagnetic wave and the time delay of propagation in the soil to a value of volumetric water content, or the percentage of water by volume in the soil. We inserted the metal rods into the soil at three different locations within each 1-ft^2 site. We recorded the mean of the three soil moisture measurements within each 1-ft^2 site. Before data collection, we calibrated the Field Scout with a two measurement procedure. One calibration measurement was taken with the metal rods suspended in air and the other



(a)

(b)

(c)

Figure 4.12: Measurements obtained during data collection in central Wisconsin during Fall 2018 and Fall 2019. (a) Soil moisture obtained with the Field Scout TDR 350 with 4.5 inch rods inserted into the soil at three locations within each 1-ft^2 site. (b) Canopy height measured from the soil surface to the top cranberry in the cranberry canopy with a ruler. (c) A photograph of the cranberry canopy taken at each 1-ft^2 site from a height of roughly two feet.

with the rods submerged in a large beaker of distilled water.

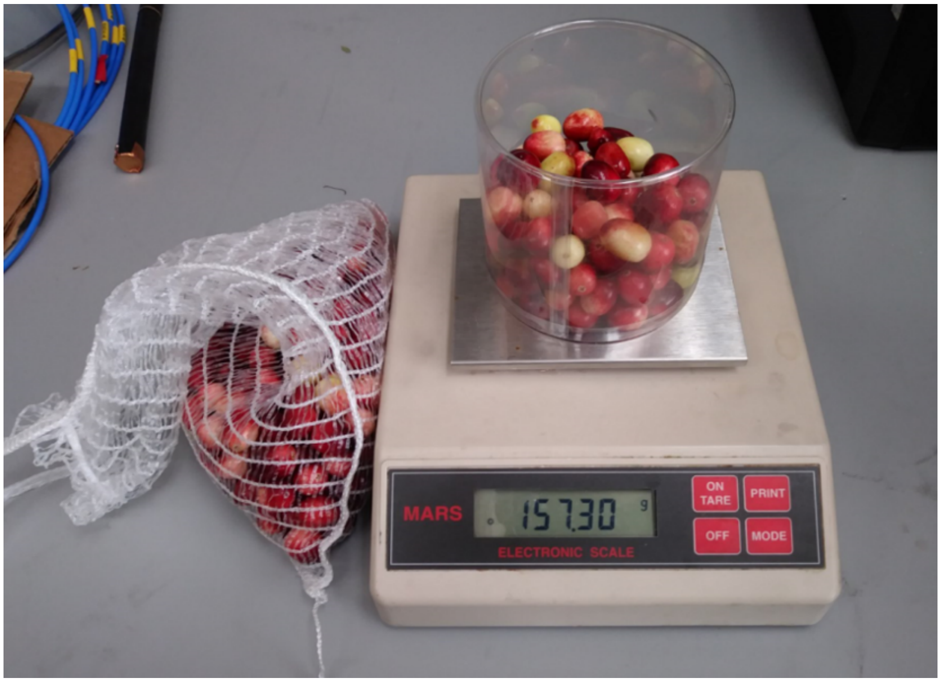


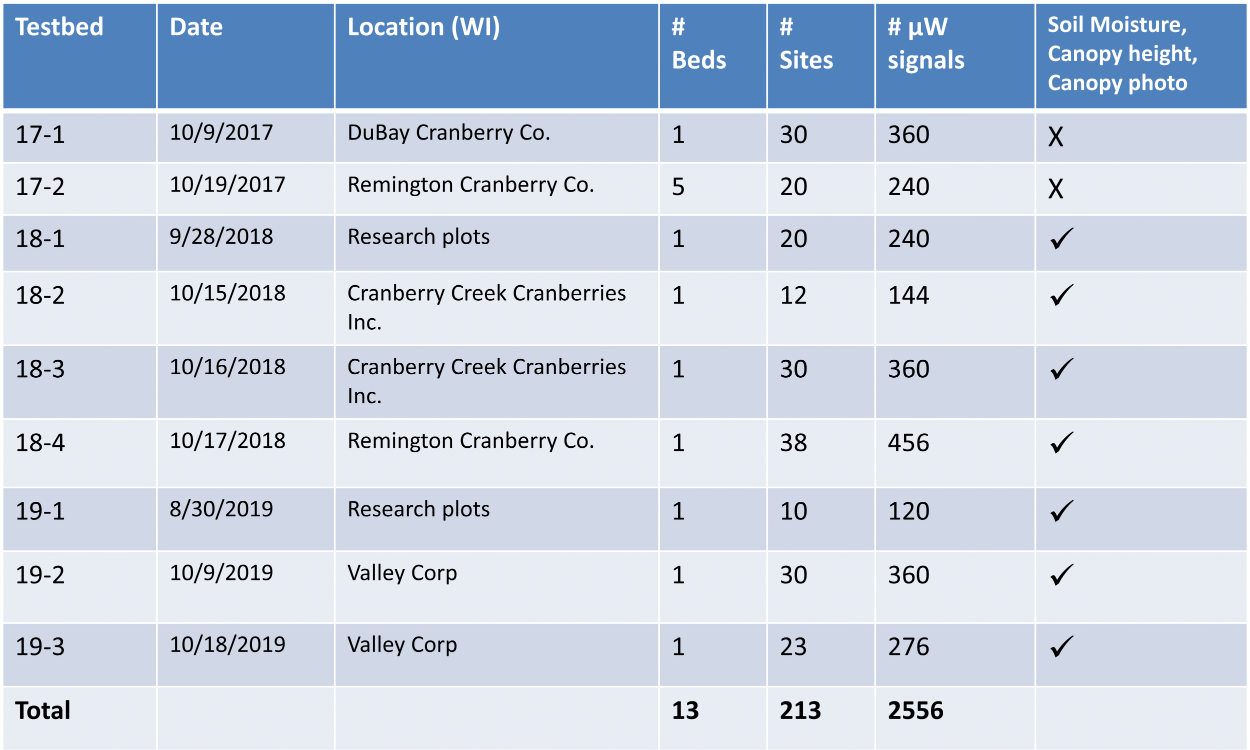
Figure 4.13: Setup for measuring the harvested cranberry yield with the MARS MS 200 precision scale.

We also measured the height of the canopy with a ruler. The height of the canopy was defined by the distance from the surface of the soil to the highest cranberry within each 1-ft^2 site. The canopy height measurement is shown in Fig. 4.12b. The canopy height measurement didn't reflect the depth of the canopy as in the simulations from section 4a, and we only measured the canopy height for 1-ft^2 sites in Fall 2017. We did not measure canopy height in 2018 and 2019 because the canopy height measurement is invasive and our ultimate goal is a non-invasive technique that can be swept rapidly across the cranberry beds.

We took a photograph of the canopy surface from about two feet above the soil surface with a cell phone as seen in Fig. 4.12c. The cell phone camera had a resolution of 4160x3120 pixels, and the model was a VS501 from LG Electronics. After collecting the microwave measurements, the soil moisture measurements, and the canopy photograph, we harvested all of the cranberries within each 1-ft^2 site. In our laboratory, we measured the mass of the harvested cranberries with a MARS precision scale as illustrated in Fig. 4.13.

4.4.2 Results

In Table 4.1 we give an overview of the data that we collected across Fall 2017, Fall 2018, and Fall 2019. In total, we collected data on nine separate days. We name the set of data



Testbed	Date	Location (WI)	# Beds	# Sites	# µW signals	Soil Moisture, Canopy height, Canopy photo
17-1	10/9/2017	DuBay Cranberry Co.	1	30	360	X
17-2	10/19/2017	Remington Cranberry Co.	5	20	240	X
18-1	9/28/2018	Research plots	1	20	240	✓
18-2	10/15/2018	Cranberry Creek Cranberries Inc.	1	12	144	✓
18-3	10/16/2018	Cranberry Creek Cranberries Inc.	1	30	360	✓
18-4	10/17/2018	Remington Cranberry Co.	1	38	456	✓
19-1	8/30/2019	Research plots	1	10	120	✓
19-2	10/9/2019	Valley Corp	1	30	360	✓
19-3	10/18/2019	Valley Corp	1	23	276	✓
Total			**13**	**213**	**2556**	

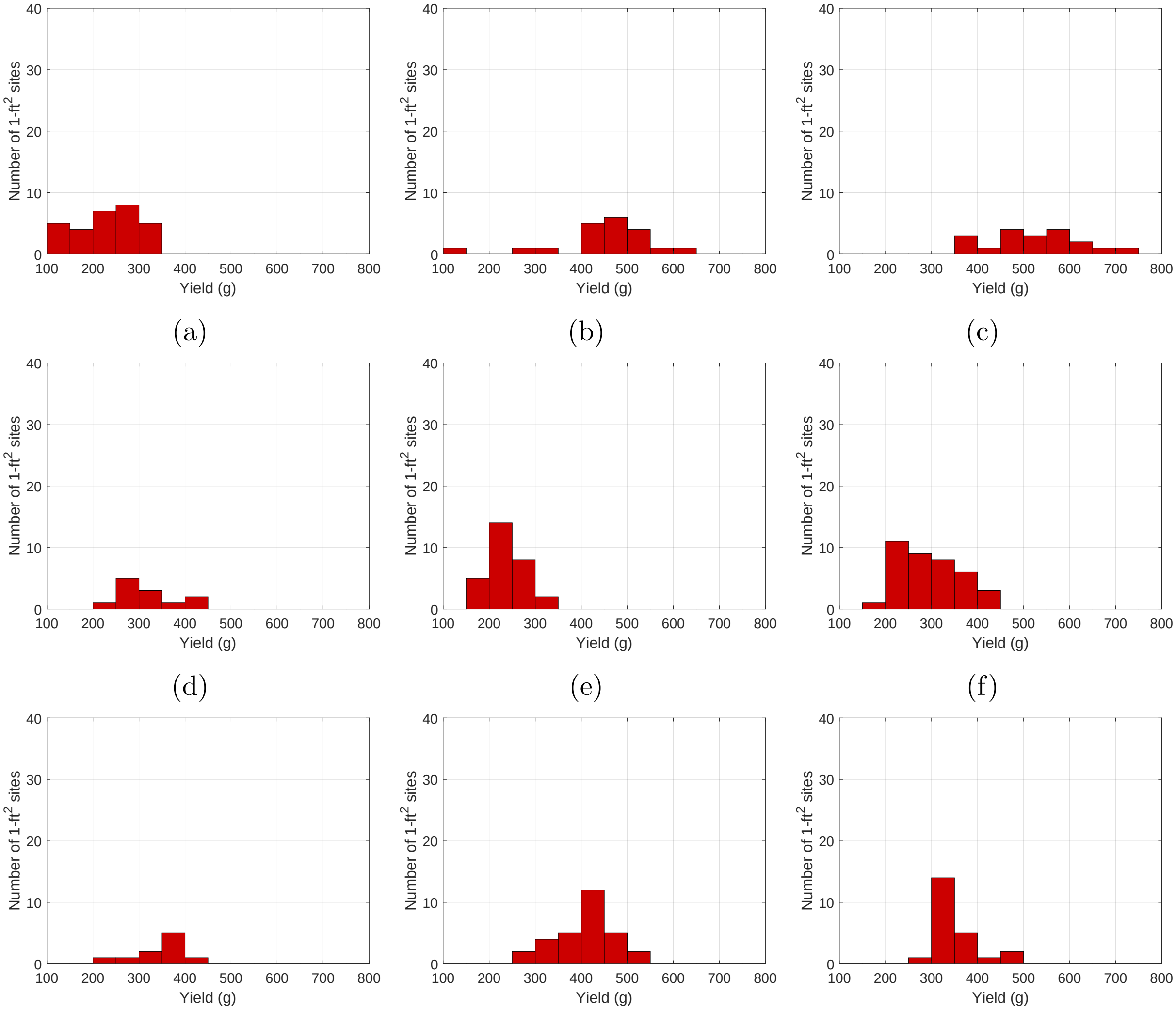
Table 4.1: Overview of data collection in Fall 2017, Fall 2018, and Fall 2019 with WR975 sensor.

collected on any given day by the year, followed by a hyphen and the day index within that year. For example, the first day that we collected data in Fall 2018 is called Testbed 18-1. Seven out of the nine days of data collection occurred within commercial beds while two out of the nine were in research test plots. All data were collected within central Wisconsin. In Fall 2018 and Fall 2019, we only collected data from within a single bed for each day. In total, we collected data from 213 1-ft^2 sites which amounted to 2556 microwave signals. We have microwave, soil moisture, and canopy photo data from 161 1-ft^2 sites between Fall 2018 and Fall 2019. Data in Testbed 18-1 were collected from Professor Juan Zalapa's research plots in Tomah, WI. Data in Testbed 19-1 were collected from research plots in Necedah, WI at Cranberry Creek Cranberries Inc.

Figure 4.14 shows the distribution of yields from the field data collection in Fall 2017, Fall 2018, and Fall 2019. Visually, the distribution of yields looks unique for each Testbed. Each Testbed in 2018 and 2019 represents data from within a single bed, and it's apparent from Fig 4.14c-i that yield can have a wide range of distributions across a single bed. For example, the standard deviation across the 20 1-ft^2 sites in Testbed 18-1 is 111 g, whereas the standard deviation across the 30 1-ft^2 sites in Testbed 18-3 is 48 g. Also, the smallest mean yield of 234 g occurred for Testbed 18-3, while the largest mean yield of 542 g occurred for Testbed 18-1. This wide range of ground truth yield distributions within each bed emphasizes the importance of bed-specific yield prediction.

Figure 4.15 shows the distribution of soil moisture contents from the field data collection in Fall 2018 and Fall 2019. The distribution of soil moistures is tightly centered around the mean for each Testbed. The standard deviation in soil moisture content within each Testbed ranges from 0.83% in Testbed 19-3 to 4.1% in Testbed 18-3. The smallest standard deviation occurs in Testbed 19-3 due to the soil being saturated with water. We collected data for Testbed 19-3 late in the season where surrounding cranberry beds were already flooded. However, even the largest standard deviation from Testbed 18-3 is relatively small at 4.1% during normal growing conditions. This suggests that soil moisture which is an invasive measurement could be collected once for each bed and incorporated as an input to the statistical learning algorithm. The mean soil moisture content ranged from 5.3% to 50.9%. During the harvest period when nearby beds are flooded, the soil moisture content is above 25% as seen in Fig 4.15b, c, d, and g. During the normal growing season it appears that soil moisture ranges from 0% to 35%.

Figure 4.16 shows the distribution of yields and soil moisture contents from each Testbed in Fall 2018 and Fall 2019. For any given Testbed, yield varies significantly while soil moisture is distributed more tightly around its mean. In Fig. 4.17, we plot the ground truth yield and soil moisture distributions across the Fall 2018 and Fall 2019 data collections. Figure 4.17a reveals a somewhat asymmetrical distribution in terms of yield. However, there is a relatively dense sampling of yields between 200 g and 600 g, which we will use to evaluate the yield-predicative accuracy of a statistical learning approach. Figure 4.17b shows randomness in ground truth soil moisture distribution. Also, Fig. 4.17c illustrates a gap in soil moisture data between 10-20%. Nonetheless, there is a relatively fine sampling of soil moistures and yields in the data we collected. Consequently, there are enough data for a preliminary



(a)

(b)

(c)

(d)

(e)

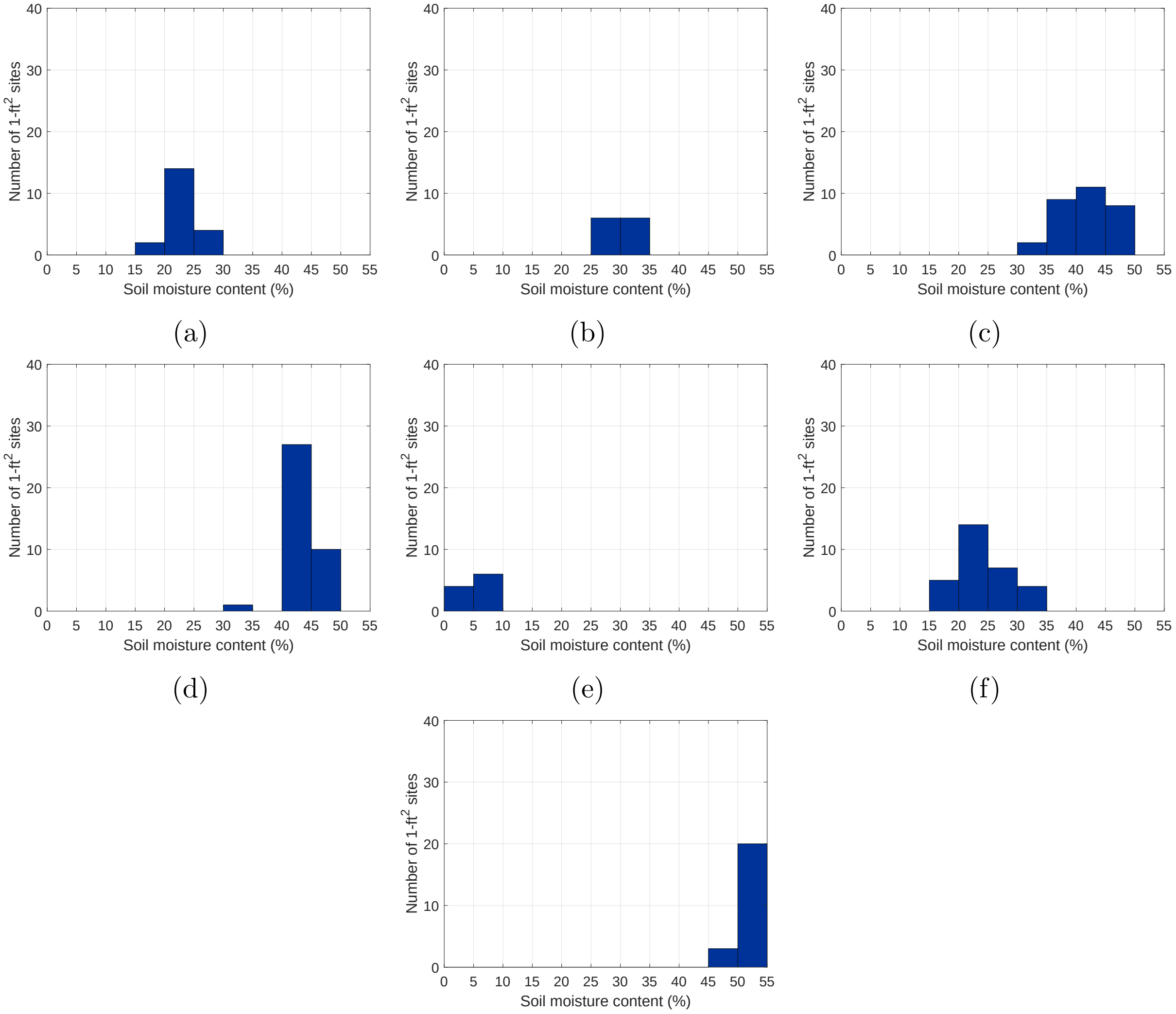
(f)

(g)

(h)

(i)

Figure 4.14: Distribution of yields for each of the nine distinct days of data collection from Fall 2017 to Fall 2019. (a) Testbed 17-1. (b) Testbed 17-2. (c) Testbed 18-1. (d) Testbed 18-2. (e) Testbed 18-3. (f) Testbed 18-4. (g) Testbed 19-1. (h) Testbed 19-2. (i) Testbed 19-3.

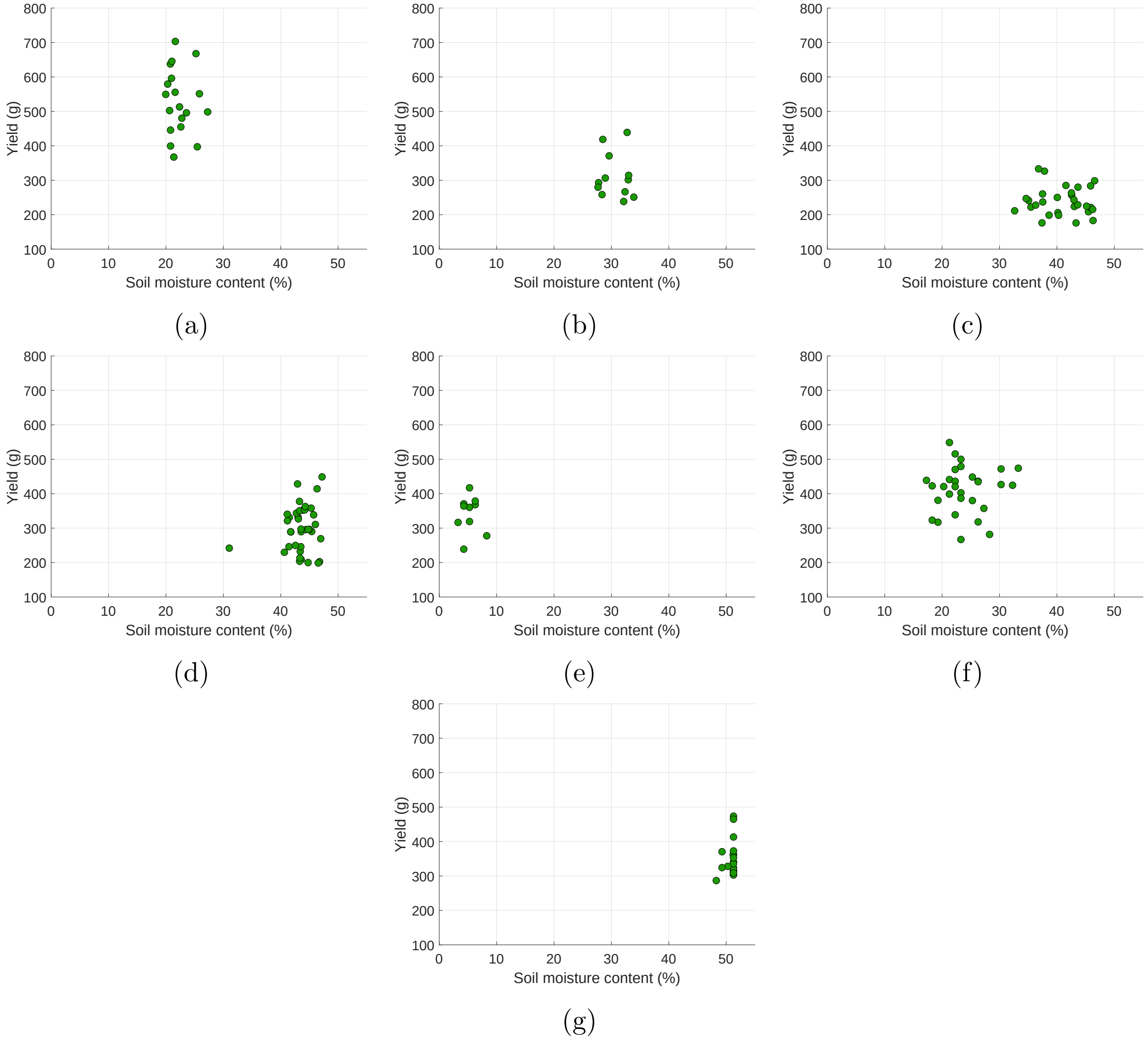


(a) (b) (c)

(d) (e) (f)

(g)

Figure 4.15: Distribution of soil moisture contents for each of the seven distinct days of data collection from Fall 2018 to Fall 2019. (a) Testbed 18-1. (b) Testbed 18-2. (c) Testbed 18-3. (d) Testbed 18-4. (e) Testbed 19-1. (f) Testbed 19-2. (g) Testbed 19-3. Soil moisture was not measured during Fall 2017.



(a)

(b)

(c)

(d)

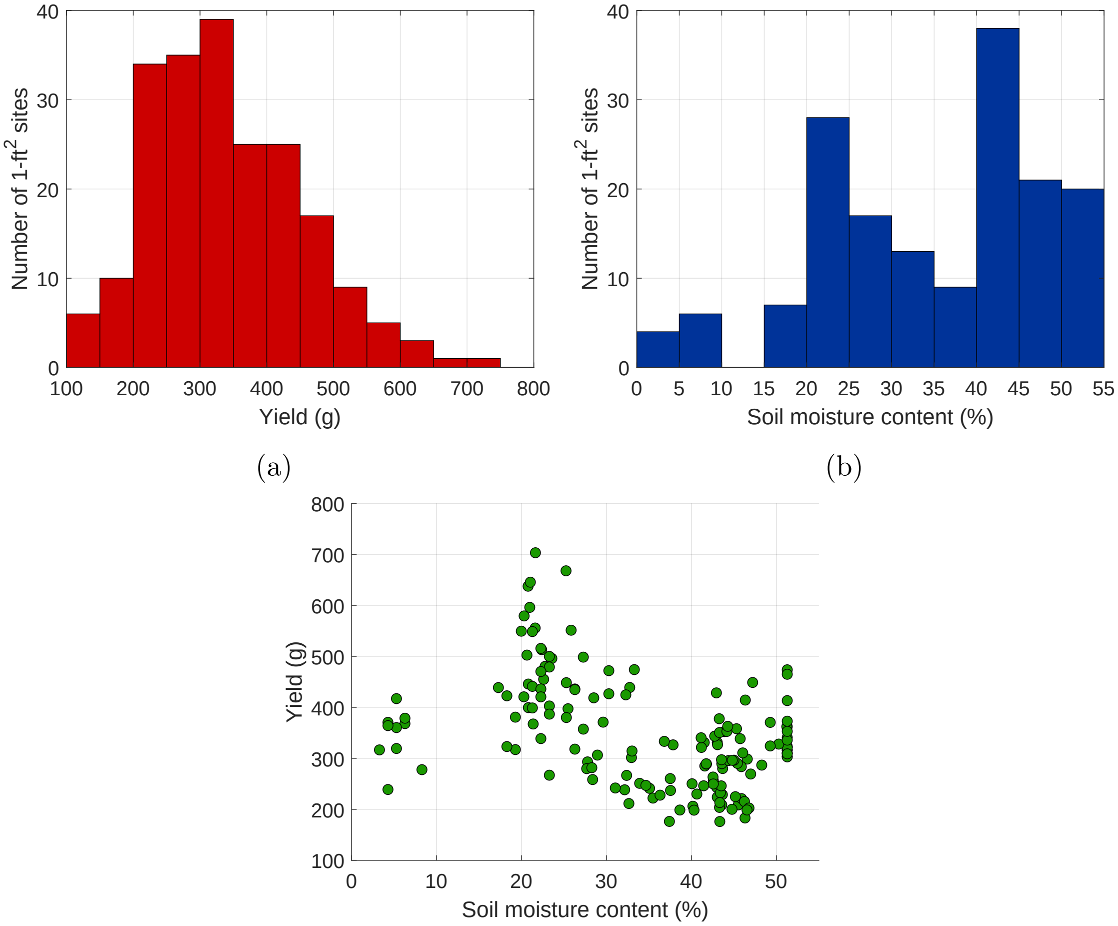
(e)

(f)

(g)

Figure 4.16: Distribution of soil moisture contents and ground truth yields for each of the seven distinct days of data collection from Fall 2018 to Fall 2019. (a) Testbed 18-1. (b) Testbed 18-2. (c) Testbed 18-3. (d) Testbed 18-4. (e) Testbed 19-1. (f) Testbed 19-2. (g) Testbed 19-3. Each dot represents a single unique 1-ft^2 site.

assessment on yield prediction using a statistical learning algorithm in diverse soil moisture and yield environments.



(a)

(b)

(c)

Figure 4.17: Distribution of soil moisture contents and yields across the seven distinct days of data collection from Fall 2018 to Fall 2019. (a) Ground truth yield distribution. (b) Soil moisture content distribution. (c) Yield vs. soil moisture content.

Fig. 4.18a displays the 12 microwave measurements collected with our experimental waveguide sensor for three separate 1-ft^2 sites in central Wisconsin chosen randomly from within the same cranberry bed on the same day. The three 1-ft^2 sites had comparable soil moisture contents, leaf densities, and canopy depths. The soil moisture content for the three sites was between 43% and 43.6%. The leaf densities were estimated from photographs of the canopies. A minimum threshold for green intensity was applied to the canopy images, and the fraction of pixels exceeding this threshold was used to estimate the leaf density. This green area fraction ranged from 15% to 22% across the three sites. These three sites had canopy depths of roughly four inches. The yields for the sites were 208 g, 246 g, and

326 g, which corresponded to average cranberry volume fractions (BVF) of 4.2%, 5.0%, and 6.6% respectively given the four inch canopy depth and 1-ft^2 canopy surface area.

From Fig. 4.18a it is apparent that yield had a significant and distinguishing impact on the microwave reflection coefficients. For example, near 700 MHz there is roughly a 5 dB difference between the measurements from the site with a 208 g yield and the site with a 246 g yield. Also, near 750 MHz there is roughly a 10 dB difference between the measurements from the site with a 208 g yield and the site with a 326 g yield. Another example occurs around 850 MHz where there is roughly a 5 dB difference between the measurements from the site with a 208 g yield and the site with a 326 g yield. These measurements indicate that yield has a strong effect on the backscatter, even for relatively small differences in yield e.g. 208 g versus 246 g. Similar to our idealized simulations, there is greater variability in the backscatter at higher frequencies in this band. Despite the larger variance in the upper end of the frequency range, the structure of the backscatter across the whole bandwidth appears useful in distinguishing between different yields.

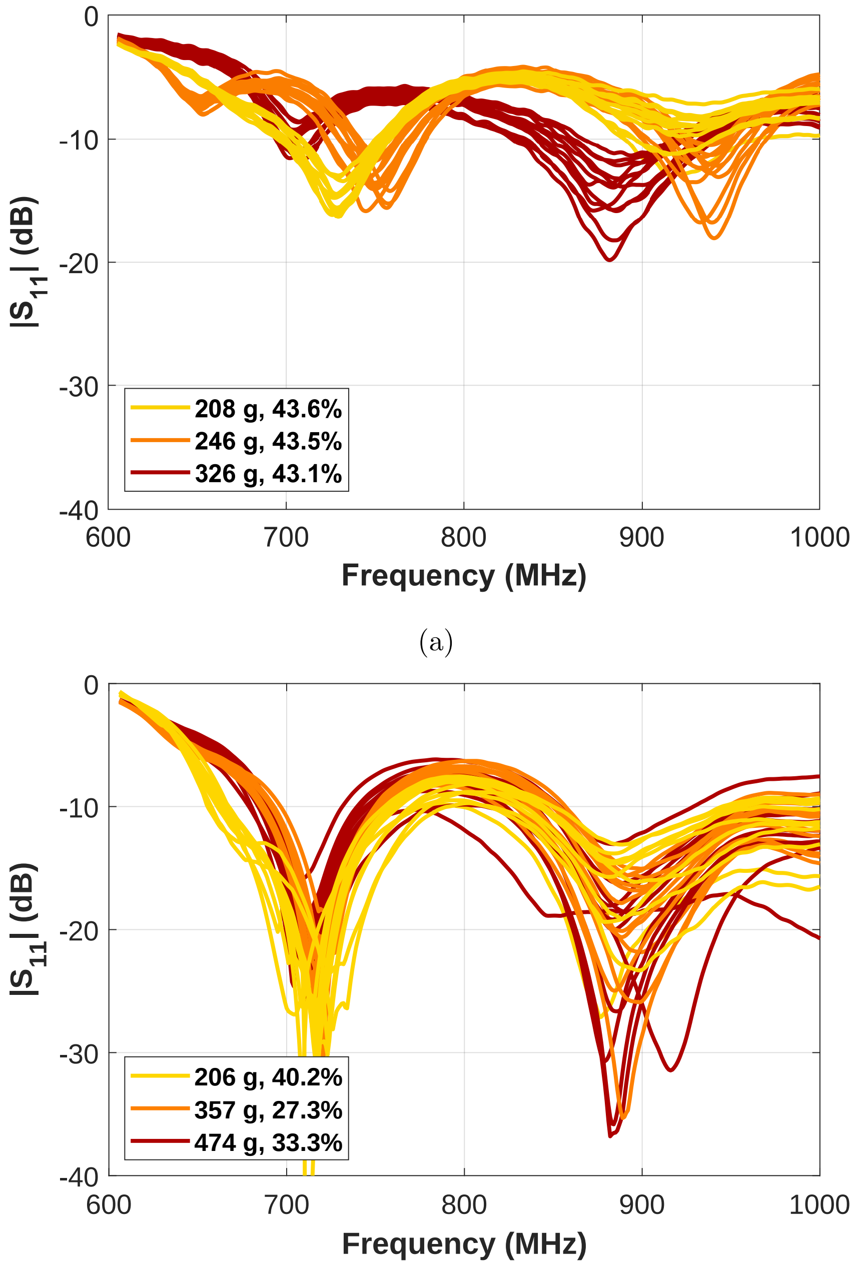
Figure 4.18b displays the 12 microwave measurements collected with our experimental waveguide sensor for three separate 1-ft^2 sites in central Wisconsin selected from different cranberry beds on different days. In this case, the three yields were significantly different, but there is significant overlap in the resulting measured reflection coefficients. The overlap between the signatures makes it difficult for a statistical learning algorithm to produce dissimilar yield predictions even though the ground truth yield spans 268 g from 206 g to 474 g. A major contributing factor in causing the reflection coefficients to overlap is soil moisture. As shown in Fig. 4.18b, the soil moisture varies significantly from 27.3-40.2% across the three sites. Fig. 4.18b highlights the importance of accounting for soil moisture in order to achieve accurate yield prediction. Unfortunately, soil moisture is an invasive measurement, which makes additional canopy-environment information necessary for a completely non-invasive approach. However, the soil moisture within a bed is distributed tightly around the mean as discussed previously. So, it is possible that a single soil moisture measurement per bed could be used to improve the predictive accuracy of statistical yield prediction from microwave signals measured with our prototype sensor.

Soil moisture variation, as we have just seen, has a negative effect on the measured microwave reflection coefficients in terms of distinguishing between different yields. However, optical photographs of cranberry canopies are not influenced by soil moisture at all. It is possible optical photographs can aid in statistical learning based yield prediction. In Fig. 4.19 we display canopy photographs of three separate 1-ft^2 sites in central Wisconsin and their corresponding ground truth yields. There are a small number of cranberries visible in the canopy photograph of 4.19a, representing a 480 g yield. There are many cranberries visible in Fig. 4.19b, and the yield is 502 g.

We can visibly detect more berries in the canopy of Fig. 4.19b than in the canopy of Fig. 4.19a. Thus, we can predict yield, or at least an increase in yield, just from the canopy photographs in this comparison. However, when comparing Fig. 4.19b with Fig. 4.19c it is clear that yield prediction is not feasible just looking at the cranberries in the photograph. There are more berries visible in Fig. 4.19b than in Fig. 4.19c, but the yield is 200 g smaller in Fig. 4.19b than in Fig. 4.19c. In Fig. 4.19c, the canopy has grown a little higher and there are more leaves obscuring the cranberries in the canopy. This comparison indicates that direct yield prediction from visible cranberries in photographs is only feasible for sparse canopies. It is possible that leaf structure is related to cranberry yield in some way, in which case yield prediction is feasible from any canopy photograph. However, there isn't strong reason to believe that this is the case.

4.4.3 Summary

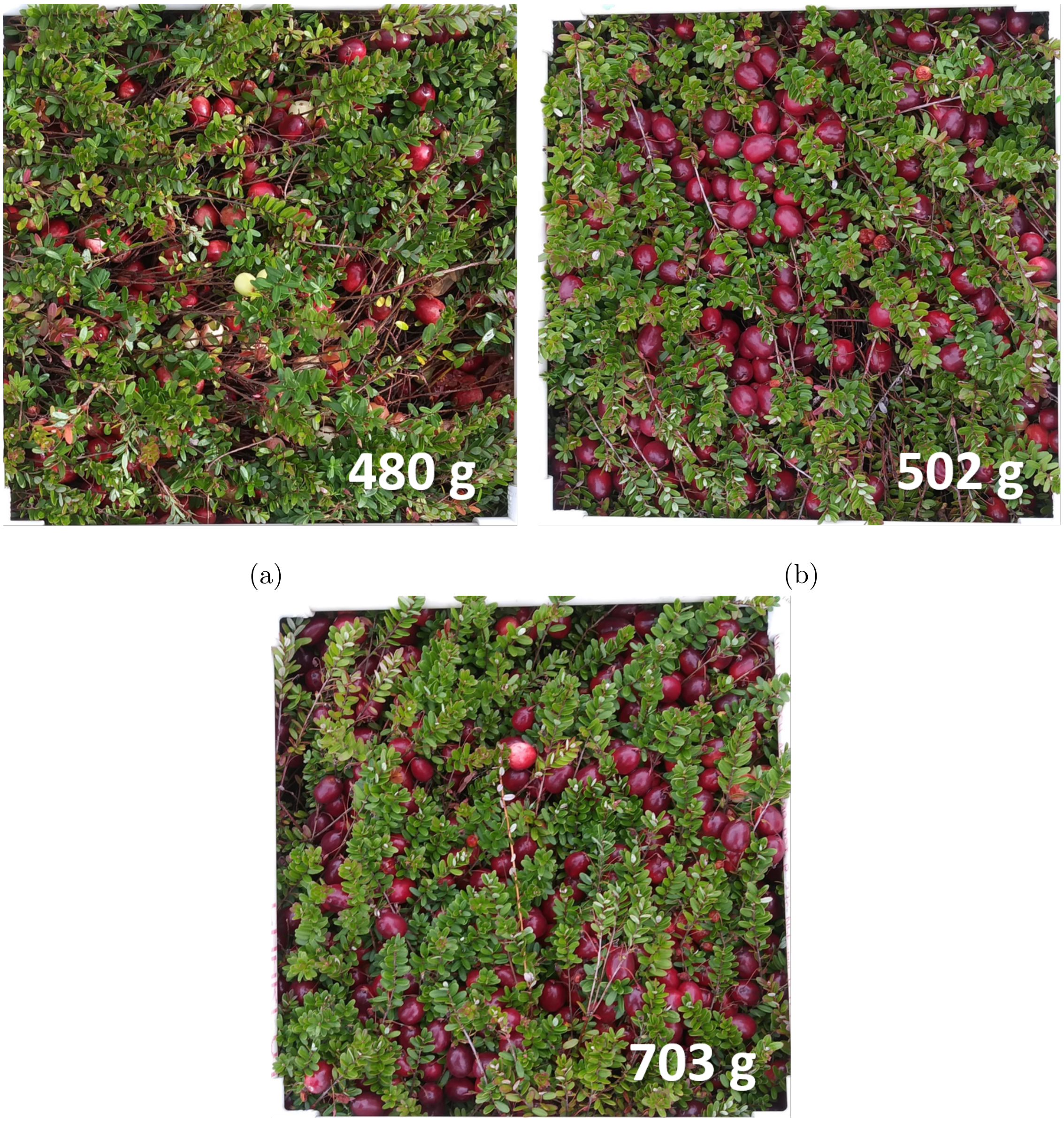
We collected data with our prototype microwave sensor in commercial cranberry beds in central WI during Fall 2017, Fall 2018, and Fall 2019. We also measured soil moisture and captured a canopy photograph in Fall 2018 and Fall 2019. We collected microwave data and ground truth yield for a total of 213 1-ft^2 sites across the three seasons. We collected microwave data, ground truth yield, soil moisture content, and canopy photographs for a total of 161 1-ft^2 sites between Fall 2018 and Fall 2019. We found that yield prediction is feasible from measured reflection coefficient signals, but variation in soil moisture content complicates yield prediction. We also found that yield prediction from canopy photographs is feasible, but dense and leafy canopies complicate yield prediction.



(a)

(b)

Figure 4.18: Illustration of the impact of soil moisture and yield on the microwave reflection coefficients measured with our prototype sensor during Fall 2018 and Fall 2019 field data collection. There are 12 curves plotted in each color representing 12 spatially different views of the same 1-ft^2 site i.e. 12 unique positions of the sensor aperture over the site. (a) Microwave data collected on the same day and from the same bed with homogeneous soil moisture. An example of feasible microwave based yield prediction. (b) Microwave data collected on different days and from different beds with heterogeneous soil moistures. An example of complicated microwave based yield prediction.



(a)

(b)

(c)

Figure 4.19: Top-view canopy photographs collected with a smart-phone camera and their corresponding yields for three unique 1-ft^2 sites. (a) Canopy photograph where some cranberries are visible. (b) Canopy photograph where many cranberries are visible. (c) Canopy photograph where many cranberries are obscured.

Chapter 5

The Predictive Accuracy of a Machine Learning Approach to Cranberry Yield Estimation (Aim 3)

5.1 Introduction

Our aim is to investigate the predictive accuracy of a machine learning based approach to estimating cranberry yield from crop backscatter, evaluating the impact of known cranberry bed parameters and statistics of the training data. This aim was completed in six parts. First, we attempt yield prediction using a preliminary statistical learning algorithm applied to a subset of the full set of field measurements obtained with the prototype microwave sensor. Then, we evaluate the yield predictive accuracy of a set of other well-developed statistical learning algorithms. Finally, we investigate methods for improving microwave based yield prediction using the most accurate predictive algorithm we evaluated.

5.2 Statistical Yield Prediction from Microwave Data - First Attempt

5.2.1 Introduction

We conducted a baseline investigation into the cranberry yield predictive accuracy of a statistical learning algorithm applied to the microwave scattering signatures collected with our prototype sensor in Fall 2017. The statistical learning algorithm was based on Principal Component Analysis and Linear Discriminant Analysis. In the baseline investigation, the microwave signals were assigned a label of cranberry yield that represented the total mass of cranberries per 1-f^2 surface area of canopy.

5.2.2 Statistical Learning Concepts

The statistical techniques we employed for predicting yield from microwave reflection coefficient signals are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in conjunction with error correcting output codes. PCA, LDA, and error-correcting output codes are briefly introduced in this section. Details on the theory of PCA and LDA can be found in many resources such as [46] and likewise for error-correcting output codes [47]. We initially chose to approach yield prediction with a classification algorithm as opposed to a regression algorithm. In classification, the algorithm outputs a yield value from the set of ground truth yield values present in the training data. For example, if we train the algorithm with only two unique yield values, then the algorithm will only be able to output one of the two unique yield values for an arbitrary input. Classification seemed appropriate in our situation because the ground truth yield values we encountered were relatively close to one another.

The first step in our data processing pipeline was PCA, used to reduce the dimension of the input data while retaining important features from the raw data. Dimensionality reduction via PCA occurs by retaining a subset of the linearly transformed random variables that exhibit the largest variances. Dimensionality reduction is an important step in avoiding the so called curse of dimensionality [46], which can negatively impact the performance of many learning algorithms. Next, LDA was chosen as the base binary classifier in the algorithm since it often produces the best results due to its simplicity and low variance [46]. We incorporated error-correcting output codes into the final algorithm to enable classification with an arbitrary number of unique yield outputs. Error correcting codes also have the potential for greater predictive accuracy than multi-class LDA alone [48]. For each unique pair of yields in the training set the algorithm performs LDA i.e. learns a linear discriminant. After training, the algorithm chooses the optimal yield from all pairwise linear discriminant comparisons for an arbitrary input via Hamming decoding. Overall, PCA and LDA together with error correcting codes comprises a statistical algorithm with a specific structure for mapping the input microwave reflection coefficient signals to yield outputs. The mapping is learned by minimizing a loss function relating to yield prediction errors across the training set.

We now discuss the details of PCA, LDA, and error-correcting output codes applied to the microwave reflection coefficient data collected with our prototype sensor in Fall 2017. Each input to our data processing pipeline was a vector of reflection coefficient magnitudes at 2401 discrete frequencies from 600 MHz to 1300 MHz. We illustrate the data processing procedure in Fig. 5.1. The number of discrete frequencies was $F = 2401$. We trained the learning algorithm with N reflection coefficient signals concatenated into the matrix $\mathbf{R}$. From the N input vectors, the principal components were derived, and ten were retained in the matrix $\mathbf{P}$. Ten principal components corresponded to at least 99% of the total variance for any given set of inputs across the study. We projected the N inputs onto the ten principal components and assigned their yield targets with the vector $\mathbf{M}$. We then performed linear discriminant analysis on the projected inputs with their targets. In the vector $\mathbf{M}$ there were

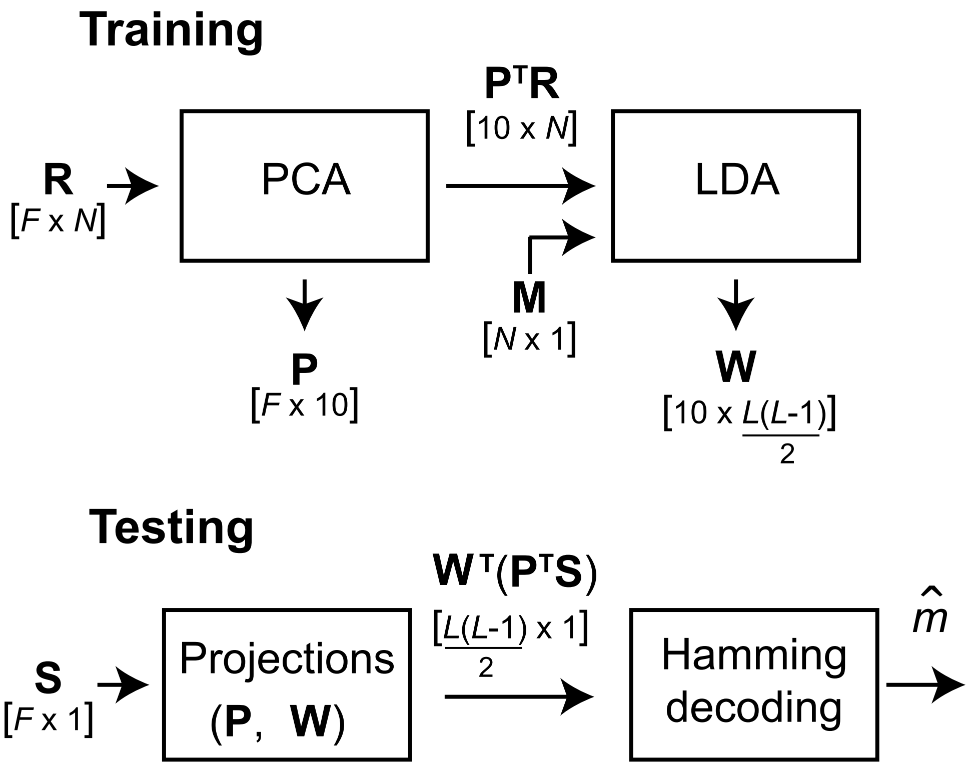


Figure 5.1: Data processing used throughout this research. N is the number of measured reflection coefficient signals used in training. F is the number of discrete frequencies where the magnitude of the reflection coefficient is sampled. L is the number of distinct yield values per 1-ft^2 site attached to the training signals. $\mathbf{R}$ is a matrix with the N concatenated reflection coefficient signals. $\mathbf{P}$ is the matrix of 10 principal components. $\mathbf{M}$ is the vector of yield value labels for the N reflection coefficient signals. $\mathbf{W}$ is the matrix of linear discriminants covering all pairwise classifications for the L distinct yield values. $\mathbf{S}$ is an arbitrary reflection coefficient signal, and $\hat{m}$ is the scalar yield value estimate taken from the set of L possible yield values.

a total of L unique yield values. Therefore, there were a total of $L(L-1)/2$ unique pairs of yields and thus $L(L-1)/2$ linear discriminants (for all unique pairwise yield comparisons). In Fig. 5.1, the linear discriminants are shown concatenated into the matrix $\mathbf{W}$. Training was complete at this point, and $\mathbf{W}$, $\mathbf{P}$ were retained for testing the trained algorithm.

We tested the algorithm by first projecting an arbitrary reflection coefficient signal $\mathbf{S}$ onto the ten principal components derived in training. Then, we projected the input signal of reduced dimension onto the $L(L-1)/2$ linear discriminants. Next, we took the sign of the resulting vector's components to convert it into a codeword. Finally, we applied Hamming decoding to the resulting codeword, which then output a yield value prediction $\hat{m}$ from the L unique yield values in the training set.

We evaluated the predictive accuracy of this data processing procedure with our experimental field data by splitting the total data set for a given testbed into two sets of equal size, one of which was used for training and the other used for testing. We generated a training set by randomly selecting six of the 12 signals per 1-ft^2 site, for each 1-ft^2 site across the testbed. We tested the algorithm on the remaining six signals per 1-ft^2 site for each 1-ft^2 site across the testbed. Finally, we generated the prediction error distribution by completing 40 random permutations on the training and testing data sets for a given testbed.

5.2.3 Ground Truth Data and Yield Predictions

The ground truth field data and machine-learning based yield prediction results for Testbed 1, Testbed 2, and combined Testbeds 1 and 2 are displayed in Fig 5.2, Fig. 5.3, and Fig. 5.4 respectively. A test case refers to a single reflection coefficient signal spanning the full 600 MHz to 1300 MHz operating band of the prototype sensor. Each case corresponds to a unique position of the waveguide aperture above an arbitrary 1-ft^2 site. For Testbed 1, there are 12 cases for each of 30 1-ft^2 sites, totaling 360 cases. For Testbed 2, there are 12 cases for each of 20 1-ft^2 sites, totaling 240 cases. There are a total of 600 cases when aggregating data from Testbed 1 and Testbed 2.

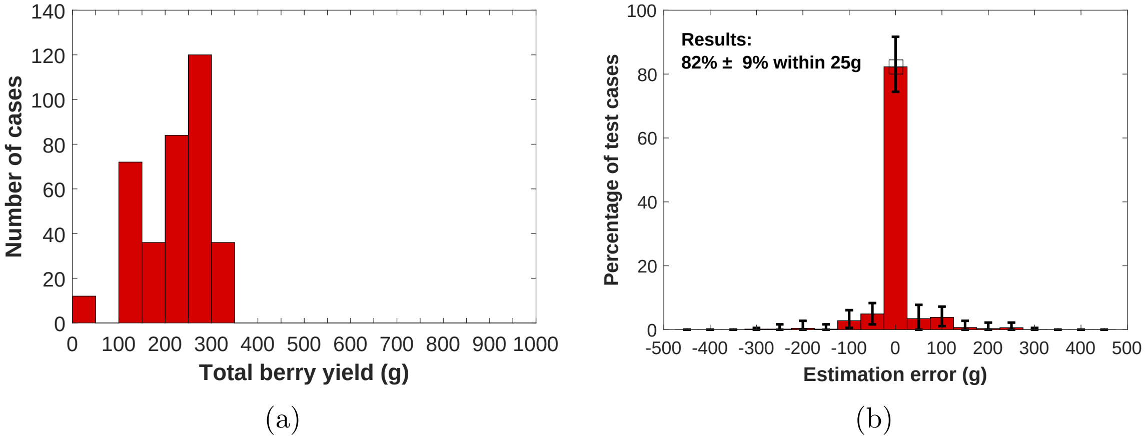


Figure 5.2: (a) Distribution of total-canopy cranberry mass per 1-ft^2 site encompassing 360 field measurements from 30 sites in Testbed 1. (b) Average distribution of estimation error across 40 permutations of training and testing data (180 reflection coefficient signals in each). Error bars represent the maximum and minimum bin-heights across the 40 permutations.

As displayed in Fig. 5.2a, the mean ground truth yield for Testbed 1 was 227 g, and the standard deviation was 71 g. As shown in Fig. 5.2b, the trained algorithm was able to predict yield with less than 25 g of error in 82% ± 9% of the test cases from Testbed 1. As shown in Fig. 5.3a, the mean ground truth yield for Testbed 2 was 452 g, and the standard deviation was 111 g. The prediction error for Testbed 2 is displayed in Fig. 5.3b, where the percentage of test cases with less than 25 g of error was 64% ± 10%. The predictive accuracy is worse for Testbed 2 than for Testbed 1 by almost 20 percentage points on average. We collected data from within five different beds for Testbed 2, whereas all measurements were taken from within the same bed for Testbed 1. Therefore, it is possible that Testbed 2 contained data from 1-ft^2 sites with variable canopy environments e.g. variable soil moisture. This could explain the degradation in performance. Also, the ground truth yield distribution for Testbed 2 differed from the ground truth yield distribution for Testbed 1. For example, the standard deviation of ground truth yield for Testbed 2 was 111 g versus 71 g for Testbed 1. The larger spread in yields for Testbed 2 might contribute to the flattening of the prediction error distribution visible in comparing Fig. 5.2b with Fig. 5.3b. Also, the mean ground truth yield for Testbed 2 was relatively high at 452 g, versus 227 g for Testbed 1. The full impact of ground truth yield distribution on predictive accuracy has not yet been determined.

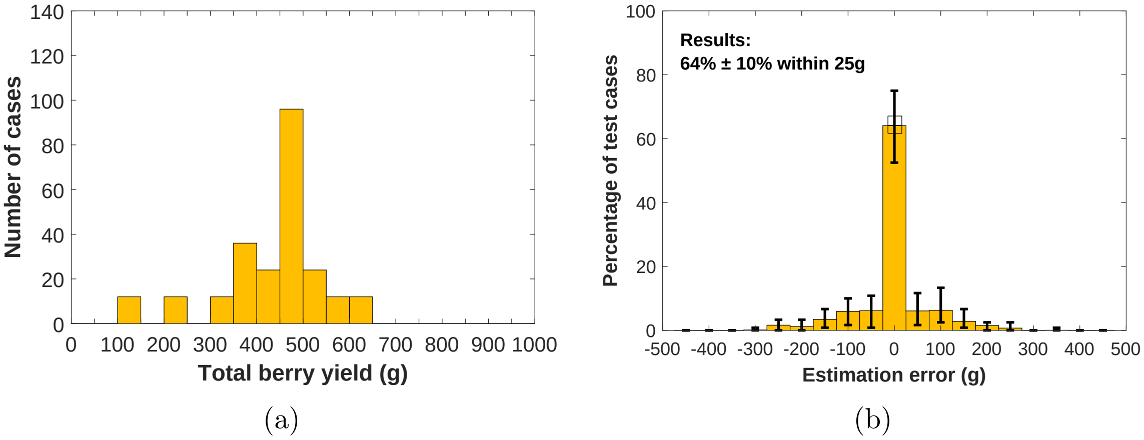
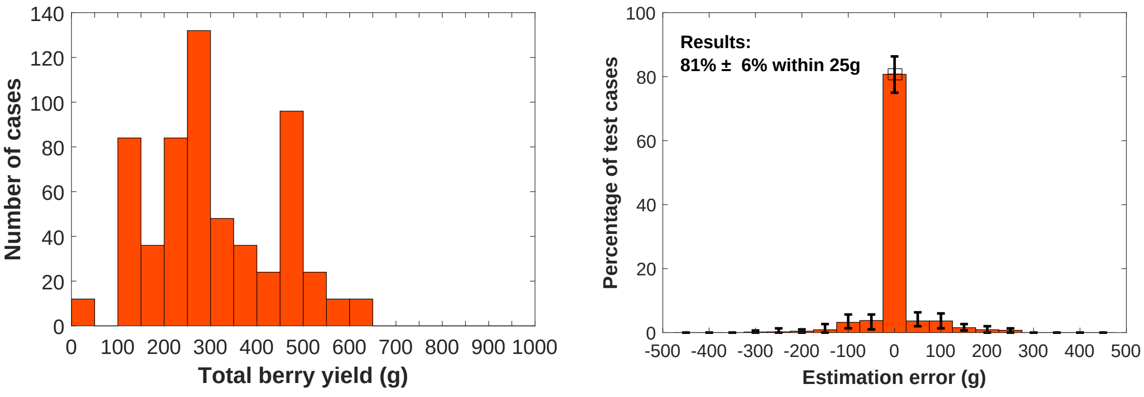


Figure 5.3: (a) Distribution of total-canopy cranberry mass per 1-ft^2 site encompassing 240 field measurements from 20 sites in Testbed 2. (b) Average distribution of estimation error across 40 permutations of training and testing data (120 reflection coefficient signals in each). Error bars represent the maximum and minimum bin-heights across the 40 permutations.

The ground truth yield and prediction error distributions for combined Testbed 1 and 2 are shown in Fig. 5.4. The mean ground truth yield was 317 g, and the standard deviation was 141.6 g. The percentage of test cases with less than 25 g of error was 81% ± 6%. Importantly, the yield values for Testbed 1 did not coincide with any of the yield values for Testbed 2. Therefore, during training there was no conflation of signals or groups of signals from different soil moisture contents. In other words, labels on the signals were unique for a particular soil moisture and yield pair. In this manner soil moisture was accounted for by the classification algorithm with this particular training and testing evaluation.

The statements of predictive accuracy so far correspond to single test cases. So, if we test our trained algorithm on a single reflection coefficient signal from an arbitrary 1-ft^2 site, the yield prediction will be within 25 g of the true yield with ∼80% probability. This performance metric is valuable for yield mapping purposes in that each 1-ft^2 site within a cranberry bed is assigned a yield prediction with error bounds. However, for whole bed



(a)

(b)

Figure 5.4: (a) Distribution of total-canopy cranberry mass per 1-ft^2 site encompassing 600 field measurements from all 50 sites from Testbed 1 and Testbed 2. (b) Average distribution of estimation error across 40 permutations of training and testing data (300 reflection coefficient signals in each). Error bars represent the maximum and minimum bin-heights across the 40 permutations.

yield predictions it's important for the prediction error to be unbiased across 1-ft^2 sites. For example, if the trained algorithm overestimates yield on a 1-ft^2 site basis at the same rate that it underestimates yield on a 1-ft^2 site basis, then across many 1-ft^2 sites (a whole bed) there will be zero error. We calculated the mean prediction error as the average of all prediction errors generated from the testing procedure applied to each testbed. The mean error in yield prediction was less than 3 g for each testbed. For reference, 3 g is approximately 1.3% of the average ground truth yield of 227 g for Testbed 1, and 0.7% of the average ground truth yield of 452 g for Testbed 2. This provides evidence that when applied to a whole bed, this system has the potential for much less uncertainty than the rough estimate of ∼15% uncertainty [2] reported for the current handpicking extrapolation technique.

5.2.4 Summary and Conclusion

We presented results of a preliminary study investigating cranberry crop yield estimation using statistical learning with active microwave sensing. We tested a machine learning approach that included Principal Component Analysis, Linear Discriminant Analysis, and error-correcting output codes. The trained algorithm predicted yield within 25 g of the ground truth in 64% and 82% of the test cases, when the mean ground truths were 452 g and 227 g respectively. These error bounds correspond to single 1-ft^2 site predictions. Most noteworthy however, the average yield prediction error across the 1-ft^2 sites in each testbed was less than 1.3% with respect to the mean testbed yield. Including reflection coefficient phase might improve prediction accuracy, but that was beyond the scope of this study.

Ultimately, this pilot study demonstrates the potential for statistical learning to accurately predict cranberry yield from signals obtained with our prototype microwave sensor.

5.3 Statistical Learning Algorithm Comparison

5.3.1 Introduction

We demonstrated the feasibility of predicting yield from reflection coefficient measurements obtained with our prototype sensor using a preliminary statistical learning approach. Now, we investigate the yield predictive accuracy of five additional well-developed statistical learning algorithms using microwave data from Fall 2018 and Fall 2019.

5.3.2 Methods for Training and Testing Six Statistical Learning Algorithms

We evaluate the yield predictive accuracy of six statistical learning algorithms in terms of the generalization error. The generalization error is defined as the expected prediction error or expected loss of any method for classification or regression. The generalization error is evaluated using testing data, which is ideally independent to the data used in training the prediction method. This is important for measuring how well the method 'generalizes' to novel data. We evaluated the generalization error in section 5.2 using six microwave signals from each 1-ft^2 site in the Testbed. However, the prediction algorithm was trained on the other six microwave signals from each 1-ft^2 site. Importantly, signals obtained from the same 1-ft^2 site are highly correlated. For example, in some cases we only shifted the aperture of the prototype sensor by small fractions of a wavelength to obtain the 12 microwave measurements at each 1-ft^2 site. Also, evaluating the generalization error as in section 5.2 assumes we encounter new 1-ft^2 sites that are exactly the same as those found in training in terms of yield, canopy structure, and soil moisture. We aim to evaluate the generalization error in terms of yield prediction error for a completely new 1-ft^2 site in terms of yield, soil moisture, and canopy structure after training an algorithm on all data available.

We train six statistical learning algorithms using the procedure illustrated in Fig. 5.5. There are 161 1-ft^2 sites across the Fall 2018 and Fall 2019 field data collections. For each 1-ft^2 site, we collected 12 microwave measurements with the prototype sensor as described previously and illustrated with squiggly lines in Fig. 5.5. Each microwave signal consisted of the magnitude of the reflection coefficient at each 1 MHz increment in frequency from 606 MHz to 1300 MHz, constituting a 695 length vector of real valued elements between [0, 1].

We evaluate the generalization error with what's known as Leave-One-Out Cross-Validation (LOO-CV). We train each algorithm with all 12 microwave signals from each of the 160 1-ft^2 sites shown in red in Fig. 5.5. Then, we evaluate the predictive accuracy of the trained algorithm on the 12 microwave signals from the 1-ft^2 testing site 'left out' of training, shown in yellow in Fig. 5.5 and Fig. 5.6. Fig. 5.5 and Fig. 5.6 illustrate a single testing evaluation.

We repeat this training and testing procedure 161 times for each separate 1-ft^2 site in the full dataset. Consequently, our training and testing procedure evaluates an algorithm's error in predicting yield from microwave signals obtained with our prototype sensor, trained with 12 microwave signals from 160 1-ft^2 sites in central Wisconsin. Importantly, the prediction error calculated with LOO-CV reflect the ground truth distribution of yields and soil moistures. In other words, the calculated prediction error is conditioned on the empirical distribution of yields, soil moistures, and canopy structures across the 161 1-ft^2 sites comprising the data.

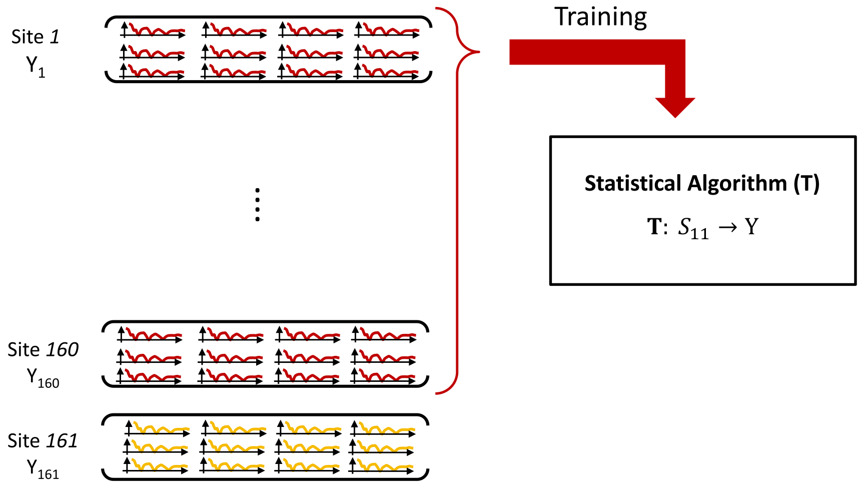

Site 1
Y_1
Site 160
Y_{160}
Site 161
Y_{161}
Training
Statistical Algorithm (T)
$\mathbf{T}$: $S_{11} \rightarrow$ Y

Figure 5.5: Training stage in evaluating generalization error using microwave data for a single testing case in leave-one-out cross validation (LOO-CV). The squiggly lines represent 12 reflection coefficient signals collected within each 1-ft^2 site. Each reflection coefficient signal is the magnitude of the reflection coefficient at frequencies from 606 MHz to 1300 MHz spaced by 1 MHz intervals. There are 161 1-ft^2 sites in total from the Fall 2018 and Fall 2019 field data collections used to evaluate the yield predictive accuracy of each statistical learning algorithm.

We optimized each statistical learning algorithm during the training step. The procedure for optimizing the statistical learning algorithm is illustrated in Fig. 5.7. We used 32 fold cross validation to evaluate the training-generalization error across 160 1-ft^2 sites. We trained each algorithm with the 12 microwave signals from 155 1-ft^2 sites and tested on the 12 microwave signals from the five 1-ft^2 sites reserved for testing in a single training, testing iteration. We repeat this training and testing step 32 times with five 1-ft^2 sites randomly chosen without replacement for testing. We minimize the mean squared prediction error across the 32 testing sets with respect to all parameters of each statistical learning algorithm.

For example, in linear regression with l^2 norm regularization there are 697 parameters to optimize. There are 695 parameters β representing regression coefficients on each variable in the input vector i.e. $|S_{11}(f_i)|$, a bias term, and a regularization parameter λ in the regularization term $\lambda\beta^T\beta$. In training, we supply an initial guess for each of the 697 parameters,

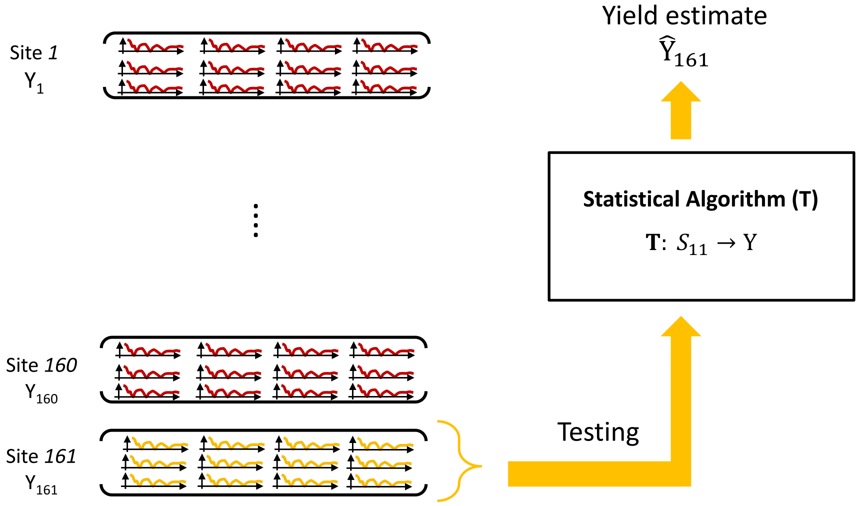


Figure 5.6: Testing stage in evaluating generalization error using microwave data for a single testing case in leave-one-out cross validation (LOO-CV). The prediction error across the 161 test cases is representative of the prediction error for an algorithm trained on 12 microwave signals from 160 1-ft^2 sites and tested using 12 microwave signals from a single new 1-ft^2 site.

evaluate the mean squared prediction error with 32 fold cross validation applied to the 160 1-ft^2 sites, and update the parameter values iteratively until the mean squared prediction error is minimized. The optimal parameter values are then used in the testing evaluation on the 161st 1-ft^2 site left out of training.

The six statistical learning algorithms we evaluated are well known, fully developed theoretically, and have been successfully demonstrated in practice. Details on each algorithm can be found in [46]. The first algorithm we evaluate in terms of yield predictive accuracy is the Linear Discriminant Analysis algorithm described in section 5.2. We also evaluate linear regression with regularization on the regression coefficients. We chose to apply linear regression with two different types of regularization in order to evaluate the predictive accuracy of a linear method using all microwave features versus a subset of microwave features to predict yield. We refer to features as elements of the input vector i.e. the magnitude of the reflection coefficient at each sampled frequency between 606 MHz to 1300 MHz $|S_{11}(f_i)|$, which the algorithm maps to a yield prediction output.

Ridge regression denotes linear regression with regularization on the sum of squared regression coefficients. This type of regularization results in nonzero regression coefficients for all 695 features, albeit smaller coefficients compared to linear regression without regularization. On the other hand, LASSO regression is regularization on the sum of the absolute values of the regression coefficients. LASSO regression tends towards few nonzero regression coefficients that emphasize a subset of 'important' features.

The three algorithms just discussed are based off of linear methods and are therefore

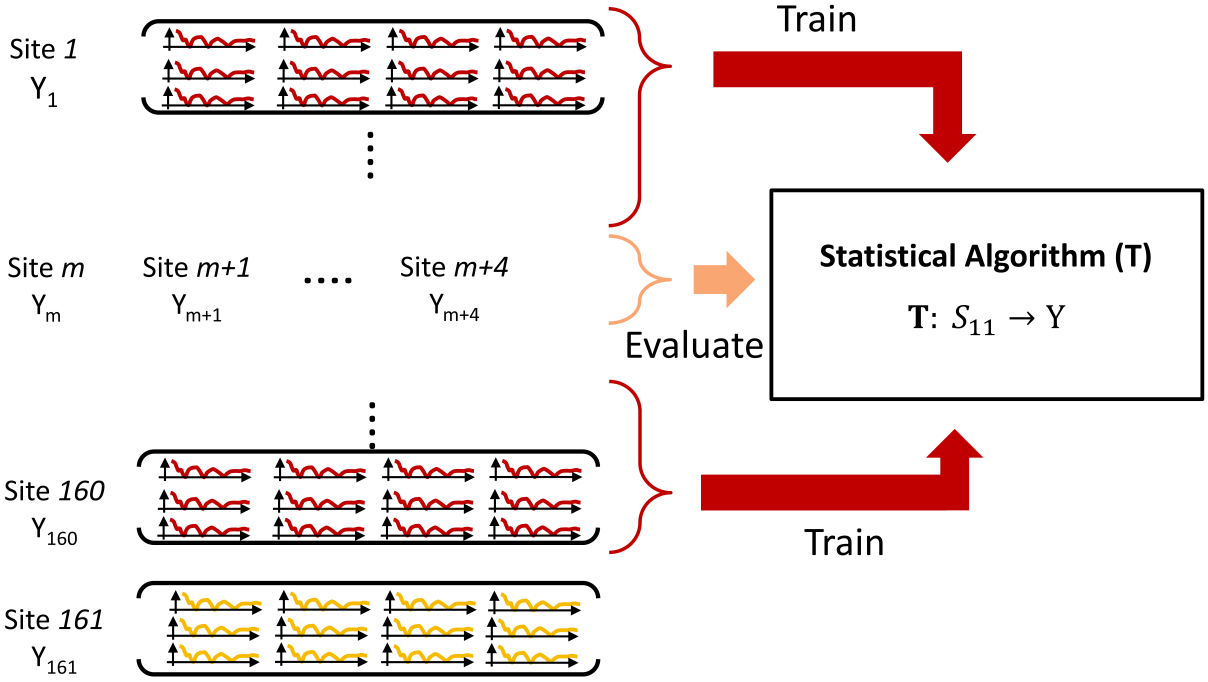


Figure 5.7: Illustration of 32-fold cross validation used during training to optimize the parameters of each statistical learning algorithm. In 32-fold cross validation, there are 32 sets of five unique 1-ft^2 sites used to evaluate the prediction error of the statistical learning algorithm. In each of the 32 evaluations, 155 1-ft^2 sites are used to train the algorithm, and the prediction error is evaluated using the remaining five 1-ft^2 sites. The prediction error across the 32 test evaluations is the loss (function) that is minimized with respect to the algorithm parameters during training.

limited in describing the mapping from $\overrightarrow{S_{11}}$ to yield. The three remaining algorithms are nonlinear and are thus flexible enough to describe a much wider array of mappings from $\overrightarrow{S_{11}}$ to yield. Random forest (RF) regression involves training an ensemble of regression trees, which are piecewise constant fits to the training data. Support Vector Machine regression (SVM) with a Gaussian kernel consists of a nonlinear function fitted to the training data with regularization to avoid over-fitting. Finally, Gaussian Process (GP) regression involves transforming input vectors into a new feature space, fitting a linear function to the transformed input vectors, and adding a term described by a Gaussian Process calculated from the input vector. Since the fit involves a probabilistic Gaussian Process term, the outputs or yield predictions are probabilistic as well. Overall, the six statistical learning algorithms comprise a wide range of regression functions. It is possible that the best possible regression function is contained in the set of functions available across the six algorithms, in which case we would obtain the best possible yield prediction using the microwave signals collected with our prototype sensor. However, given the wide range of possible regression functions we will achieve near-optimal yield prediction even if the optimal regression function is not obtained.

5.3.3 Yield Prediction and Generalization Error Results

To illustrate the generalization error, we plot the prediction results for the Linear Discriminant Analysis algorithm from section 5.2 in Fig. 5.8. As shown on the left-hand plot in Fig. 5.8, there are 161 1-ft^2 test sites corresponding to the LOO-CV where 160 1-ft^2 sites are used in training, the remaining 1-ft^2 site is used in testing, and each 1-ft^2 site is cycled through as the testing site. Each dot in the left-hand plot of Fig. 5.8 represents the mean absolute error in yield prediction across the 12 microwave signals contained in the 1-ft^2 testing site. We calculate statistics on these site-wise mean absolute prediction errors as percentiles on 161 mean absolute errors, illustrated on the right hand side of Fig. 5.8. The minimum and maximum prediction errors are shown for reference but are not necessarily indicators of the performance a prediction algorithm since they could be the result of outliers.

We plot statistics on the mean absolute prediction errors across the 161 1-ft^2 testing sites for six algorithms in Fig. 5.9. The black horizontal lines represent the 25th and 75th percentiles, and the horizontal red line is the 50th percentile. LDA from section 5.2 had the worst predictive accuracy with a median error of 56.7 g. Ridge and LASSO regression had similar median errors of 44.8 g and 42.3 g respectively, however there was a wider spread in error distribution for the LASSO than for the Ridge. This suggests that applying linear regression to a subset of features with the LASSO increases variance in prediction error compared to using the full set of features as in Ridge regression. Random Forest (RF) regression performed comparably to Ridge regression. This is somewhat surprising since the physical relationship between reflection coefficient as measured by the prototype and yield is nonlinear, and a Random Forest can capture nonlinearity whereas Ridge regression can not. Support Vector Machine (SVM) regression had the smallest median error of 39.4 g and the smallest 25th percentile of 24.5 g. The 75th percentile in prediction error for SVM was 67.1 g. This indicates that in 50% of the test sites SVM predicted yield with an absolute

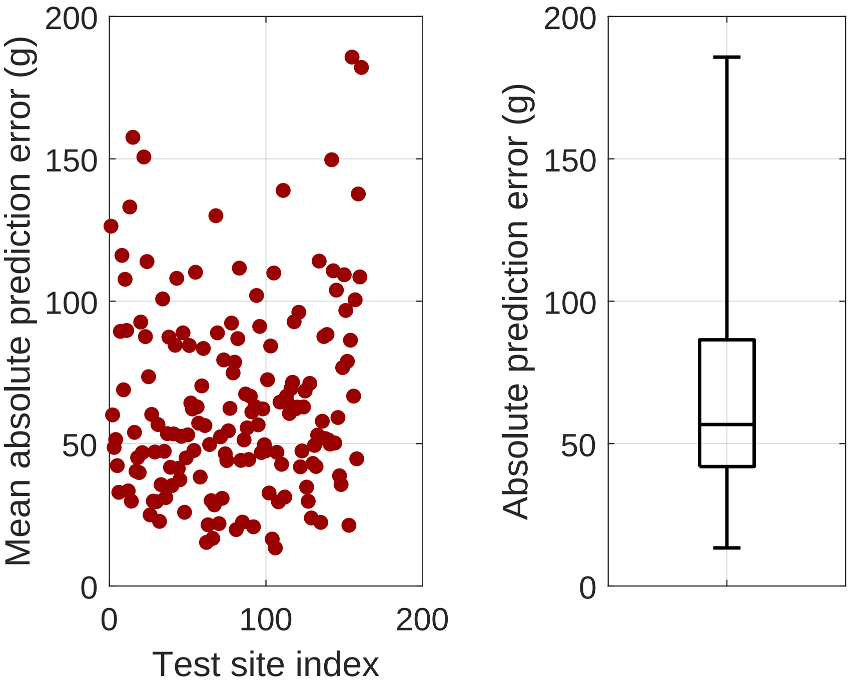


Figure 5.8: Prediction error from the Linear Discriminant Analysis algorithm described in section 5.2. (left) Mean of the absolute prediction errors (MAE) from the 12 signals at each 1-ft^2 test site. (right) Statistics on the MAE across the 161 testing sites in LOO-CV. The five statistics depicted represent the 0th, 25th, 50th, 75th, and 100th percentiles on the MAE across the 161 1-ft^2 testing sites.

prediction error ranging from 25.4 g to 67.1 g. Gaussian Process (GP) regression had a median error of 43.1 g, a 25th percentile of 30.0 g, and a 75th percentile of 63.4 g.

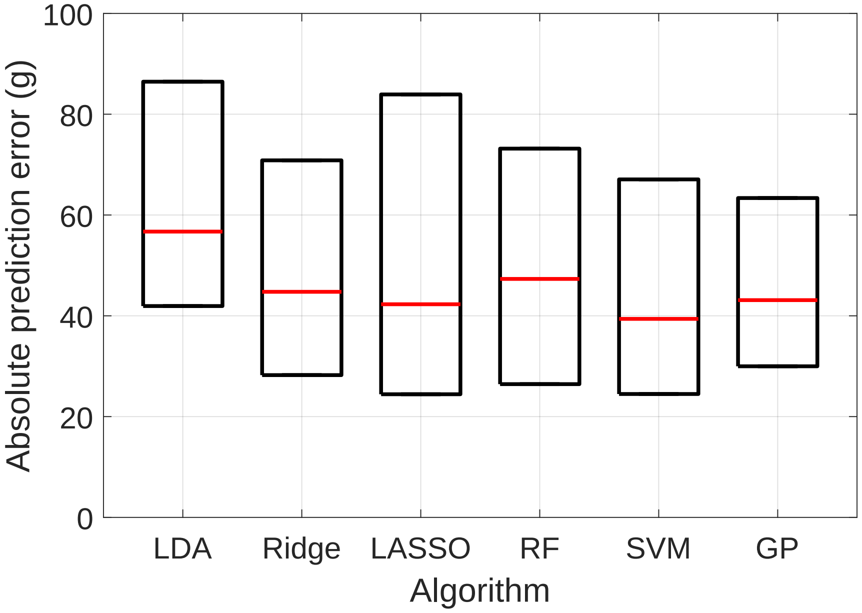


Figure 5.9: Statistics on the generalization error using LOO-CV for six statistical learning algorithms applied to field data from Fall 2018 and Fall 2019. The 25th and 75th percentiles are shown in black, and the 50th percentile is shown in red. 'LDA' represents the Linear Discriminant Analysis algorithm from section 5.2. 'Ridge' and 'LASSO' denote linear regression with l^2 norm and l^1 norm regularization on the regression coefficients respectively. 'RF' represents random forest regression, 'SVM' is support vector machine regression, and 'GP' is Gaussian process regression.

These prediction results indicate that Support Vector Machine regression with a Gaussian kernel has the largest probability of predicting yield with less than 40 g of error within a new 1-ft^2 test site, out of the six algorithms. Another benefit of SVM regression given the size of our dataset is rapid training and testing. Gaussian Process and Random Forest regression required the most time to train and test, followed by LDA. The time required to train and test SVM was slightly longer than for Ridge or LASSO. Also of note, the mean yield across the 161 1-ft^2 sites is 342 g. Assuming that the empirical ground truth yield distribution across the 161 1-ft^2 sites approximates the ground truth yield distribution in all commercial cranberry beds, the expected yield in a 1-ft^2 site within a commercial cranberry bed is 342 g. Less than 10% yield prediction error across a whole bed is required to be of value to growers and supply chain managers. Therefore, on average the yield prediction error needs to be less than 34.2 g in each 1-ft^2 site. We compared the algorithms in terms of absolute prediction error, but in terms of raw prediction error each algorithm produced roughly unbiased predictions across the 161 1-ft^2 sites. In other words, each algorithm produced an error distribution with a mean of roughly zero g. A zero g mean prediction error corresponds to applying this approach across many 1-ft^2 sites. However, cranberry researchers require less than 10% error across their much smaller plots which are 9-ft^2 in terms of canopy surface area. Therefore, having an algorithm produce accurate yield predictions on a 1-ft^2 site basis is important. Hence, SVM will produce the best average yield prediction out of the six algorithms on a 1-ft^2 site basis, which will be applicable to small research plots and whole beds alike.

5.3.4 Summary

We evaluated the generalization error of six well developed statistical learning algorithms using the microwave reflection coefficient data collected with our prototype sensor in Fall 2018 and Fall 2019. All five regression methods had better yield predictive accuracy than the classification based LDA approach from section 5.2. Support Vector Machine regression had the best median absolute prediction error of 39.4 g per 1-ft^2 site and was among the fastest algorithm to train and test. The generalization error i.e. expected absolute prediction error using SVM regression was $39.4/342 = 11.5\%$ when applied to a single 1-ft^2 testing site. We found that the error distribution for SVM regression was roughly unbiased across 161 testing sites, suggesting much better than 10% error across a whole bed.

5.4 The Impact of Soil Moisture and Yield Diversity on Predictive Accuracy

5.4.1 Introduction

We aim to evaluate the impact of soil moisture and yield heterogeneity on the yield predictive accuracy of a statistical learning algorithm applied to field data collected in Fall 2018 and Fall 2019 with our prototype sensor. As shown in chapter 4, soil moisture content can contribute to a high degree of similarity for the microwave signals measured in 1-ft^2 sites with significantly different yields. A statistical learning algorithm relies on dissimilarity between inputs to produce dissimilar outputs, in our case yield predictions. Therefore, our hypothesis is that soil moisture confounds yield prediction from the microwave signals measured with our prototype sensor.

5.4.2 Materials and Methods

We have microwave reflection coefficient, soil moisture, and ground truth yield data for 161 1-ft^2 sites across the Fall 2018 and Fall 2019 field data collections. We plot the soil moisture content and ground truth yield for the 161 1-ft^2 sites in Fig. 5.10, separated by testbed. Each testbed corresponds to a separate day of data collection within a single cranberry bed. Therefore, the soil moisture content has small variation within each testbed as seen in Fig.

5.10, where there is at most a 15% variation in soil moisture content within each testbed. The number of 1-ft^2 sites contained in a testbed ranges from 10 to 38. Within each 1-ft^2 site we have 12 microwave measurements obtained with the prototype sensor as described in chapter 4.

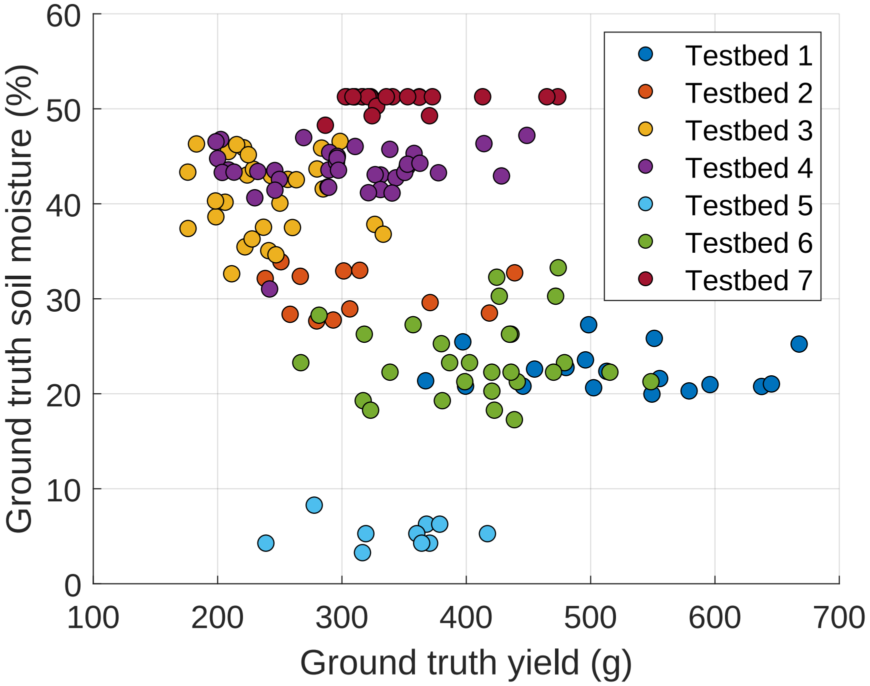
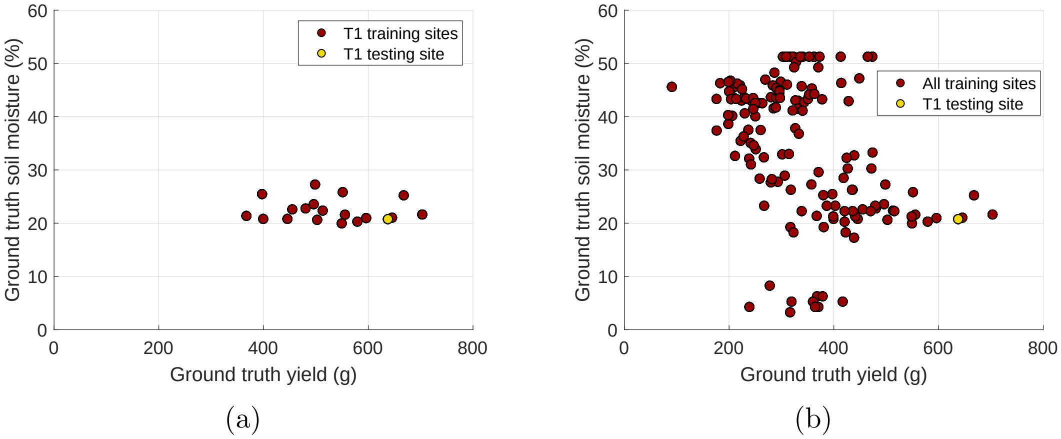


Figure 5.10: Ground truth yield and soil moisture for each testbed from Fall 2018 and Fall 2019. Each dot represents a unique 1-ft^2 site with 12 microwave measurements. Each testbed corresponds to a single day of data collection within a single 2-acre cranberry bed. The number of 1-ft^2 sites per testbed ranges from 10 to 38 across the seven testbeds.

We evaluate the impact of soil moisture heterogeneity through the generalization error calculated using Support Vector Machine regression with a Gaussian kernel. We calculate the generalization error through site-wise LOO-CV in three scenarios. First, in Scenario 1 we evaluate the LOO-CV prediction error using the microwave data from a single testbed. This is illustrated in Fig. 5.11a for Testbed 1 (T1). There are microwave data for 20 1-ft^2 sites in Testbed 1. Therefore, LOO-CV amounts to training on the 12 microwave signals from 19 1-ft^2 sites, testing on the 12 microwave signals from the remaining 1-ft^2 testing site, and iterating through all 20 1-ft^2 sites selected for testing. A single training-testing iteration is shown in Fig. 5.11a, where the red dots represent the 12 microwave signals for each of the 19 1-ft^2 training sites, and the yellow dot represents the 12 microwave signals for the single 1-ft^2 testing site. There are 20 training-testing iterations where each dot in Fig. 5.11a is yellow just once. We calculate the mean absolute error in yield prediction across the 12 microwave signals in each 1-ft^2 testing site. Representative statistics are calculated and reported using the 20 mean absolute errors in total from the 20 1-ft^2 testing sites (for Testbed 1). If there are N 1-ft^2 sites in a testbed, then statistics are calculated and reported using the N mean absolute errors in total from the N 1-ft^2 testing sites. This scenario constitutes a relatively homogeneous soil moisture environment.



(a) (b)

Figure 5.11: Training and testing scenarios for evaluating the impact of heterogeneity of soil moisture content on statistical learning based yield prediction. (a) Ground truth yield and soil moisture for Testbed 1 (T1). (b) Ground truth yield and soil moisture across all testbeds. Prediction error is evaluated using LOO-CV. A single testing evaluation is depicted in which red dots represent the 1-ft^2 sites used in training and the yellow dot represents the single 1-ft^2 site left out of training and used for prediction error evaluation.

The second scenario (Scenario 2) involves a heterogeneous soil moisture environment. Scenario 2 is visualized in Fig. 5.11b, in which all 161 1-ft^2 sites are used in training and testing. As seen in Fig. 5.11b, 160 1-ft^2 sites are used to train the SVM. We test the algorithm in this scenario on the same 1-ft^2 testing sites as in Scenario 1. For example, in Scenario 1 with data from Testbed 1, there are 19 1-ft^2 sites used in training. In Scenario 2 with data from Testbed 1, there are the same 19 1-ft^2 sites used in training as in Scenario 1, but we include 141 additional 1-ft^2 sites in training from the other testbeds. We test the trained algorithms in Scenario 1 and 2 on the same 1-ft^2 testing site, and we do this 20 times in total for Testbed 1 by iterating through each site used in testing. Consequently, the algorithm in Scenario 2 is exposed to a wider range of yields and soil moistures than in Scenario 1. Scenario 1 and 2 are implemented for each of the seven Testbeds spanning data collection from Fall 2018 and Fall 2019.

We also evaluate the yield predictive accuracy in a third scenario. Scenario 3 is identical to Scenario 2, except in terms of the inputs to the statistical learning algorithm. In both Scenario 1 and 2, the inputs are 695 length vectors where each element corresponds to the magnitude of the reflection coefficient at a specific frequency between 606 MHz and 1300 MHz. In Scenario 3, we append the soil moisture content for each 1-ft^2 site to the input vectors, resulting in 696 length input vectors.

5.4.3 Prediction Results

In Fig. 5.12 we compare the predictive accuracy of testbed specific yield prediction i.e. separate testbeds in Scenario 1 with data from all testbeds available in Scenario 2, as well as with with data from all testbeds available and soil moisture content (smc) in Scenario 3. The blue bars represent Scenario 1, the orange bars Scenario 2, and the yellow bars Scenario 3. Error bars correspond to the 25th and 75th percentiles, and the height of the bars represents the mean site-wise prediction error. Again, these statistics are calculated from the mean absolute prediction errors across testing sites (LOO-CV).

Fig. 5.12 illustrates comparable prediction error distributions for all three scenarios. This indicates that soil moisture does not have a confounding effect on yield prediction for testing applied to Testbed 1. In other words, the microwave signals from Testbeds 2-7 are sufficiently different from the signals in Testbed 1 that the algorithm does not mistake the signals from Testbed 1 for signals from Testbeds 2-7. Interestingly, signals from Testbed 6 are similar to the signals from Testbed 1 due to similar soil moisture contents and ground truth yields as seen in Fig. 5.11. However, even though the signals in Testbed 6 have similar yields compared to Testbed 1 the algorithm does not make large errors in yield prediction. In SVM, the predicted yield is a weighted average of microwave signals in the transformed feature space. As long as the microwave signals that are similar in the transformed feature space have similar yields there will be small deviations in predicted yield as additional data is included in training.

From Fig. 5.12 it is apparent that Scenario 1 produced a superior prediction error distribution compared to Scenario 2 and Scenario 3 for Testbed 5. The mean prediction error is 21.7 g for Scenario 1, 37.2 g for Scenario 2, and 36.8 g for Scenario 3. From Fig. 5.11 it's apparent that the data for Testbed 5 is relatively isolated from the other testbeds in terms of soil moisture. The soil moisture for Testbed 5 is roughly 5% whereas all other testbeds have soil moisture contents of 20% or greater. As seen in chapter 4, the confounding effect of soil moisture occurs for significantly different soil moistures and yields. In the case of Testbed 5, we included microwave signals from drastically different soil moistures and yields in training for Scenario 2 and 3. The significantly different soil moistures allowed for similarity in microwave signal despite large differences in ground truth yield, resulting in degraded yield predictive accuracy. The prediction results from Testbed 7 corroborate this result.

Interestingly, the prediction results from Testbed 4 indicate improved yield prediction in Scenario 2 and 3 compared to Scenario 1. The mean prediction error is 49.6 g for Scenario 1, 40.4 g for Scenario 2, and 39.8 g for Scenario 3. The soil moisture in Testbed 4 is almost as high as in Testbed 7, and there is a range of yields comparable to Testbed 7. However, prediction error is not degraded by including microwave signals from Testbed 5 which has significantly different yields and soil moistures. One potential reason for non-degraded prediction error is the denser distribution of microwave signals arising from environments with similar yield and soil moisture i.e. Testbed 3 and 4, similar to Testbed 1 and 6. Ultimately, is seems that prediction error is only negatively impacted by variation in soil moisture content

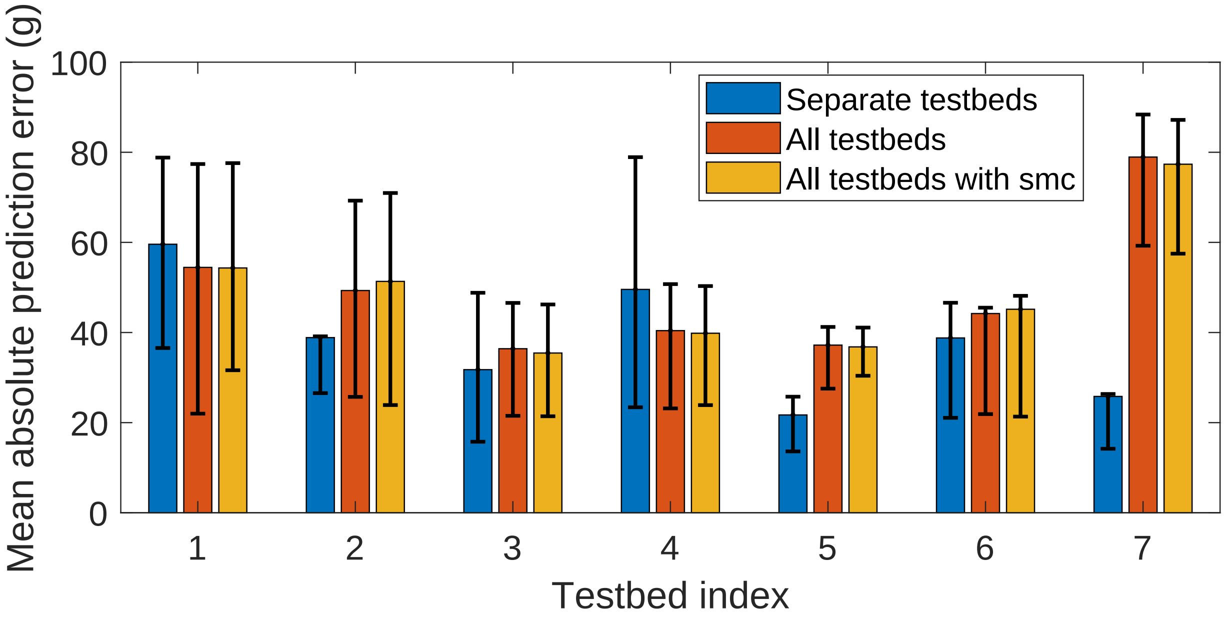


Figure 5.12: Statistics on the mean absolute prediction error generated by LOO-CV. From the 12 microwave signals at each test site there are 12 yield predictions. The mean absolute error across the 12 predictions represents the site-specific prediction error. The statistics displayed are calculated across site-specific testing errors. Error bars represent the 25th and 75th percentiles. The height of each bar represents the mean of the site-specific testing errors.

when testing in 1-ft^2 sites that have significantly different soil moisture contents than the the 1-ft^2 sites comprising the training set.

Across all seven Testbeds, the prediction error is roughly the same between Scenario 2 and Scenario 3. Thus, it appears that appending soil moisture to the inputs is not an effective way to account for soil moisture using SVM regression. This could be due to the fact that adding a single dimension onto a 695 dimensional object does not significantly impact the relative positions of the microwave signals with respect to one another. In other words, the signals in Scenario 3 are effectively unchanged from those in Scenario 2.

5.4.4 Summary

We evaluated the impact of soil moisture and yield heterogeneity on yield predictive accuracy of a SVM regression algorithm applied to field data collected in Fall 2018 and Fall 2019 with our prototype sensor. Yield prediction error was only significantly impacted by variation in soil moisture content when the soil moistures in the testing site differed significantly from the soil moistures in the training set. Otherwise, yield prediction was not adversely affected by heterogeneity in soil moisture across the training and testing sites. This suggests that soil moisture variation will not degrade yield predictive accuracy for a densely sampled distribution of ground truth yields and soil moistures present in the training set.

5.5 Averaging for Enhanced Yield Prediction

5.5.1 Introduction

We investigate enhanced yield prediction through averaging microwave data collected from within individual 1-ft^2 sites. Our investigation includes 12 microwave measurements obtained with our prototype sensor at each of 161 1-ft^2 sites from the Fall 2018 and Fall 2019 field data collections. Specifically, we investigate the impact of averaging microwave signal inputs as well as yield prediction outputs arising from single 1-ft^2 sites.

5.5.2 Methods for Averaging

We evaluate prediction error using LOO-CV as in section 5.2. Thus, we select 160 1-ft^2 sites for training, evaluate prediction error on the remaining 1-ft^2 testing site, and iterate through the 161 1-ft^2 sites used in testing. Training is the same as in section 5.2 where 32-fold cross validation is implemented to optimize the SVM. This analysis only differs from section 5.2 in regards to testing.

We illustrate an example of input averaging in Fig. 5.13 applied to the 1-ft^2 testing site for a single iteration of training-testing. As shown in Fig. 5.13, the testing site is denoted by 12 microwave signals and the corresponding yield for the fixed site m. Fig. 5.13 illustrates the case where we average six inputs. However, we average anywhere from 2-12 signals in this study. For the six input average case, six signals are selected randomly from the 12 in the testing site and averaged. The average is an ensemble average such that the resulting signal is the same length as each input signal. The input signals are 695 length vectors whose elements are the magnitudes of the reflection coefficient at each 1 MHz increment in frequency from 606 MHz to 1300 MHz. Therefore, each element of the resulting average vector is the mean of six reflection coefficient magnitudes at the corresponding frequency. The input-average signal is then mapped to a single yield prediction $\widehat{Y_m}$, and this procedure is repeated 100 times starting at the step where 6 input signals are randomly selected for averaging. For a single testing site and a fixed number of signals averaged (e.g. six signal average) there are 100 yield predictions. Thus, across the 161 testing sites there are a total of 16,100 yield predictions for a fixed number of signals averaged. Statistics on the prediction error are calculated from all 16,100 yield predictions.

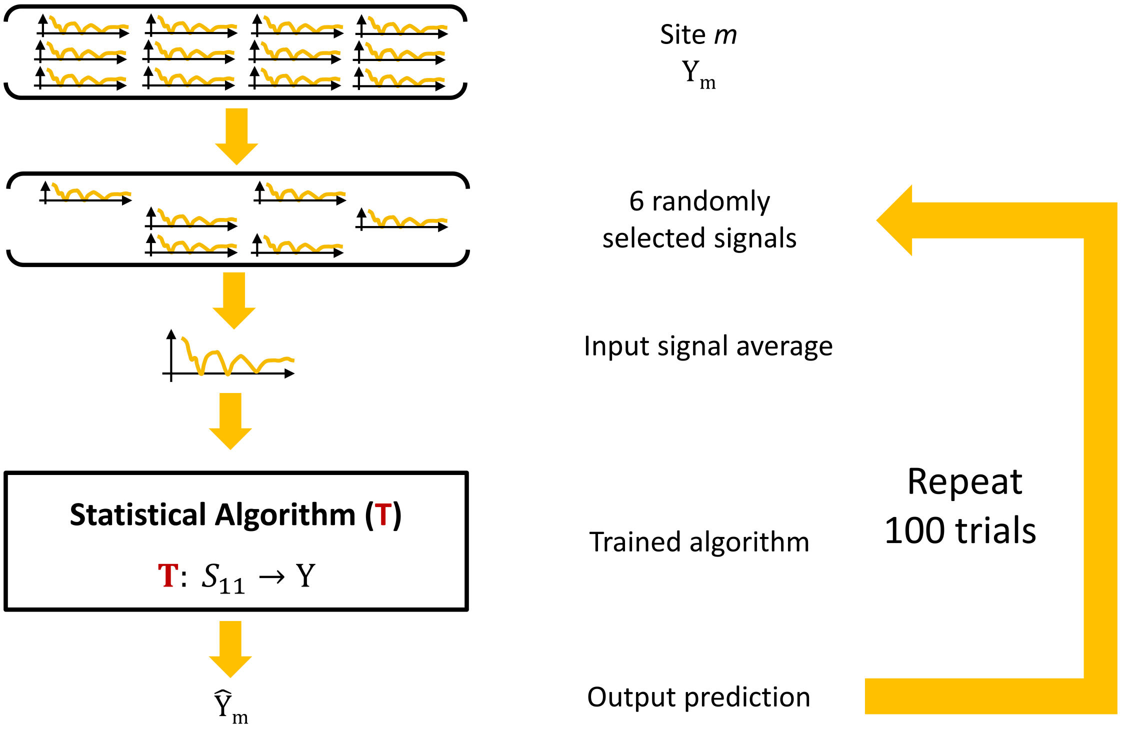


Figure 5.13: Illustration of input averaging with six microwave reflection coefficient signals during testing. Squiggly lines represent the microwave reflection coefficient signals, which consist of reflection coefficient magnitudes at frequencies spaced by 1 MHz intervals from 606 MHz to 1300 MHz (695 discrete frequencies). Input averaging is done in an ensemble fashion for the 695 dimensional input vectors. The trained statistical learning algorithm maps these 695-dimensional input vectors to real-valued yield outputs Y.

We illustrate an example of output averaging in Fig. 5.14 applied to the testing site. As shown in Fig. 5.14, the testing site is denoted by 12 microwave signals and the corresponding yield for the fixed site m. Fig. 5.14 illustrates the case where we average six output predictions. However, we average anywhere from 2-12 outputs in this study. For the six output average case, six S_{11} signals are selected randomly from the 12 total in the testing site. Each signal is mapped to an independent yield prediction $\widehat{Y_i}$. The final yield prediction $\widehat{Y_m}$ is the average of the six independent yield predictions. This procedure is repeated 100 times starting at the step where 6 input signals are randomly selected. Across the 161 testing sites there are a total of 16,100 yield predictions for a fixed number of output predictions averaged. Statistics on the prediction error are calculated from all 16,100 yield predictions.

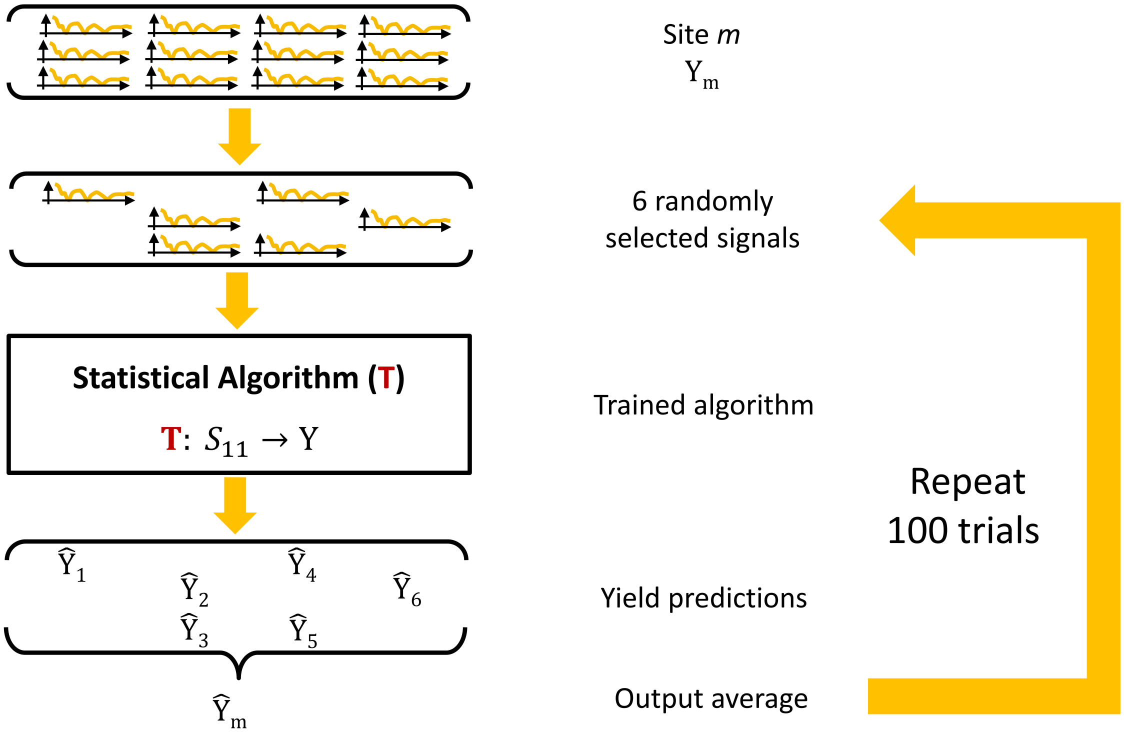


Figure 5.14: Illustration of output averaging with six microwave reflection coefficient signals during testing. The squiggly lines represent microwave input signals, which consist of reflection coefficient magnitudes at frequencies spaced by 1 MHz intervals from 606 MHz to 1300 MHz. The trained statistical learning algorithm maps these 695-dimensional input vectors to yields Y. Output averaging is applied to the independent output predictions from some fixed number of input microwave signals, in this case six.

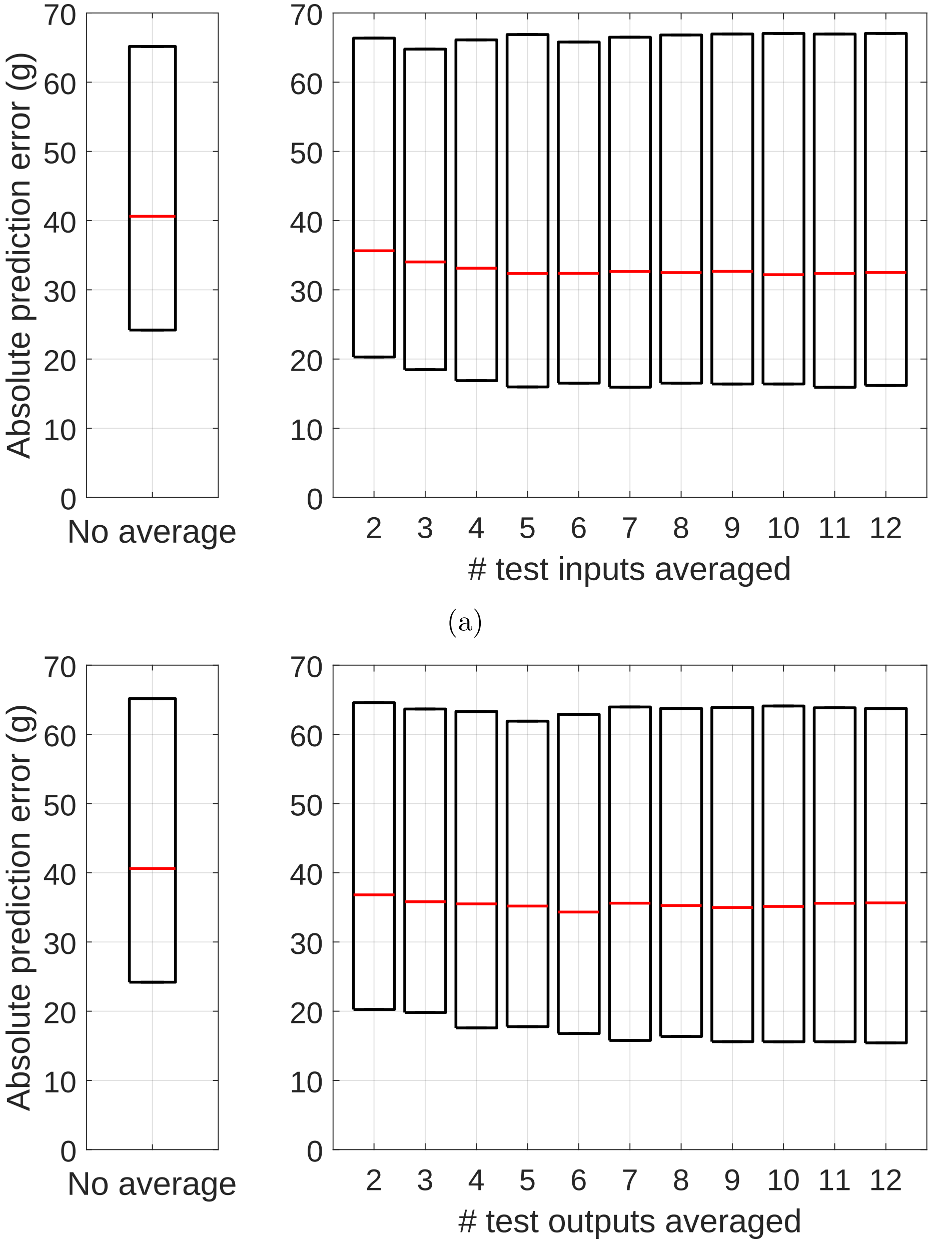
5.5.3 Averaging Results

We plot the 25th, 50th, and 75th percentiles of the absolute prediction error for each number-average in Fig. 5.15 obtained trough LOO-CV. As seen in Fig. 5.15a, the 25th and 50th percentiles decrease as we average an increasing number of test inputs. The 25th and 50th percentiles for the no-average case are 24 g and 41 g respectively. The 25th and 50th percentiles for the 5-input average case are 16 g and 32 g respectively. The 32 g median prediction error is 9.4% of the mean yield of 342 g across the 161 1-ft^2 sites. Thus, averaging five or more input signals coupled with SVM regression achieves the desired accuracy of $<$ 10% error on a 1-ft^2 basis. This automatically entails $<$ 10% error across a cranberry bed. Averaging more than 5 inputs does not seem to improve the prediction error significantly.

As seen in Fig. 5.15b, the 25th and 50th percentile decrease as we average an increasing number of output yield predictions up to six. The 7-12 output average scenarios produce roughly the same prediction error as a 6-output average. The 25th and 50th percentiles for the no-average case are 24 g and 41 g respectively. The 25th and 50th percentiles for the 6-output average case are 17 g and 34 g respectively. The 34 g median prediction error is 9.9% of the mean yield of 342 g across the 161 1-ft^2 sites. Thus, averaging six independent output predictions coupled with SVM regression achieves the desired accuracy of $<$ 10% error on a 1-ft^2 basis. This automatically entails $<$ 10% error across a cranberry bed. Averaging input signals appears to be more effective at reducing prediction error than averaging output predictions. This could be due to the nonlinear transformation of signals to the feature space in which SVM based prediction occurs.

5.5.4 Summary

We have shown that averaging microwave input signals obtained with the prototype sensor improves yield prediction. Averaging five or more input signals from within a single 1-ft^2 testing site coupled with SVM regression achieves the desired accuracy of $<$ 10% error on a 1-ft^2 basis, which automatically entails $<$ 10% error across a cranberry bed. Averaging six independent yield prediction outputs from within a single 1-ft^2 testing site coupled with SVM regression also achieves the desired accuracy of $<$ 10% error on a 1-ft^2 basis. Averaging more than five or six inputs or outputs does not improve the yield predictive accuracy significantly.



(a)

(b)

Figure 5.15: Absolute error in yield prediction versus the number of (a) input signals averaged or (b) output predictions averaged. Prediction error is calculated with LOO-CV and SVM regression applied to 161 1-ft^2 sites from Fall 2018 and Fall 2019. Black horizontal lines represent the 25th and 75th percentiles, and the red horizontal line represents the median absolute prediction error across 16,100 testing evaluations for a fixed number-average.

5.6 Yield Prediction vs. Size of the Training Set

5.6.1 Introduction

We aim to evaluate the rate at which yield prediction error decreases as we increase the number of 1-ft^2 sites used to train a statistical learning algorithm. This analysis will provide an estimate of how many 1-ft^2 sites are required to achieve a desired level of predictive accuracy.

5.6.2 Methods

Across the Fall 2017, Fall 2018, and Fall 2019 field data collections there are 213 1-ft^2 sites. As previously discussed, there are 12 microwave reflection coefficient measurements obtained within each 1-ft^2 site using the prototype sensor, represented by squiggly lines in Fig. 5.16. We illustrate an example training procedure in Fig. 5.16. We fixed the training set size at three 1-ft^2 sites in the figure, but in the actual study we varied the training set size from 20 to 200. We use all 12 microwave signals from each 1-ft^2 site in the training set to train the SVM, fixing the Gaussian kernel parameter at 1.4. We then select 13 1-ft^2 sites at random from the remaining data for testing. Each of the 12 microwave signals in the 13 1-ft^2 testing sites is mapped to a yield prediction. For one round of testing there are 13*12 = 156 yield predictions. We repeat random site selection for training and testing 200 times. Thus, for a fixed training set size there are 156*200 = 31,200 yield predictions from which error statistics are calculated and reported.

5.6.3 Results

We plot statistics on the Root-Mean-Square (RMS) prediction error for training set sizes ranging from 20 to 200 in Fig. 5.17. The box plot displays 0th, 25th, 50th, 75th, and 100th percentiles on RMS prediction error. From Fig. 5.17 it is apparent that the median RMS prediction error decreases monotonically as the number of 1-ft^2 sites in training increases. The median RMS prediction error for a training set consisting of 20 1-ft^2 sites is 96.7 g. The median RMS prediction error for a training set consisting of 100 1-ft^2 sites is 61.7 g. The median RMS prediction error for a training set consisting of 180 1-ft^2 sites is 49.0 g. Thus, an initial addition of 80 1-ft^2 sites in training produces a decrease in RMS prediction error of 35 g. Then, an additional 80 1-ft^2 sites in training produces a decrease in RMS prediction error of only 12.7 g. Thus, the rate at which the RMS prediction error decreases with respect to the size of the training set decreases as the training set size increases. In other words, the median RMS prediction error is concave with respect to the number of sites in the training set.

We fit an exponential function to the mean RMS prediction error as shown in Fig. 5.18. The resulting fit is shown in equation 5.1 where n represents the number of 1-ft^2 sites in the training set. The fit is a good match visually with the mean RMS prediction error.

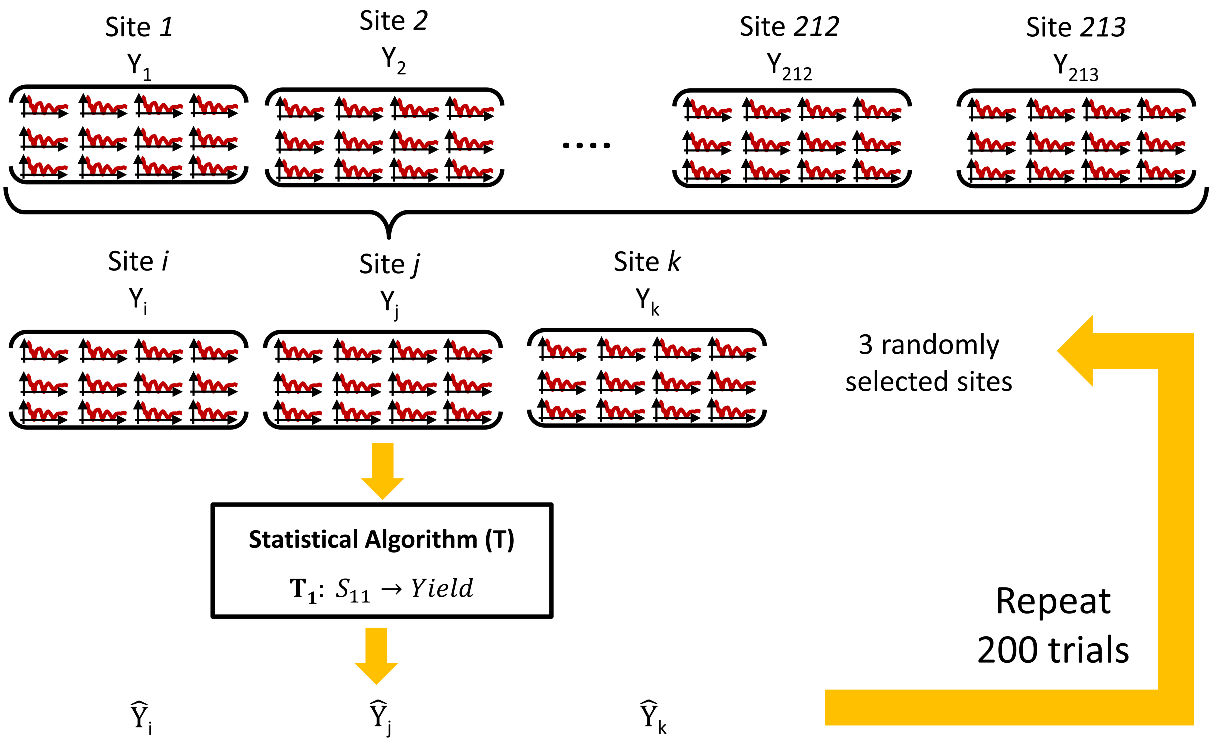


Figure 5.16: Illustration of training set size analysis for a training set size of three 1-ft^2 sites selected randomly from the full set of field data collected in Fall 2017, Fall 2018, and Fall 2019 encompassing 213 1-ft^2 sites. Squiggly lines represent reflection coefficient magnitudes between 606 MHz and 1300 MHz measured with the prototype sensor. There are 12 microwave signals collected within each 1-ft^2 site used to train a Support Vector Machine with a Gaussian kernel. Random selection of training and testing sites is completed 200 times for prediction error evaluation.

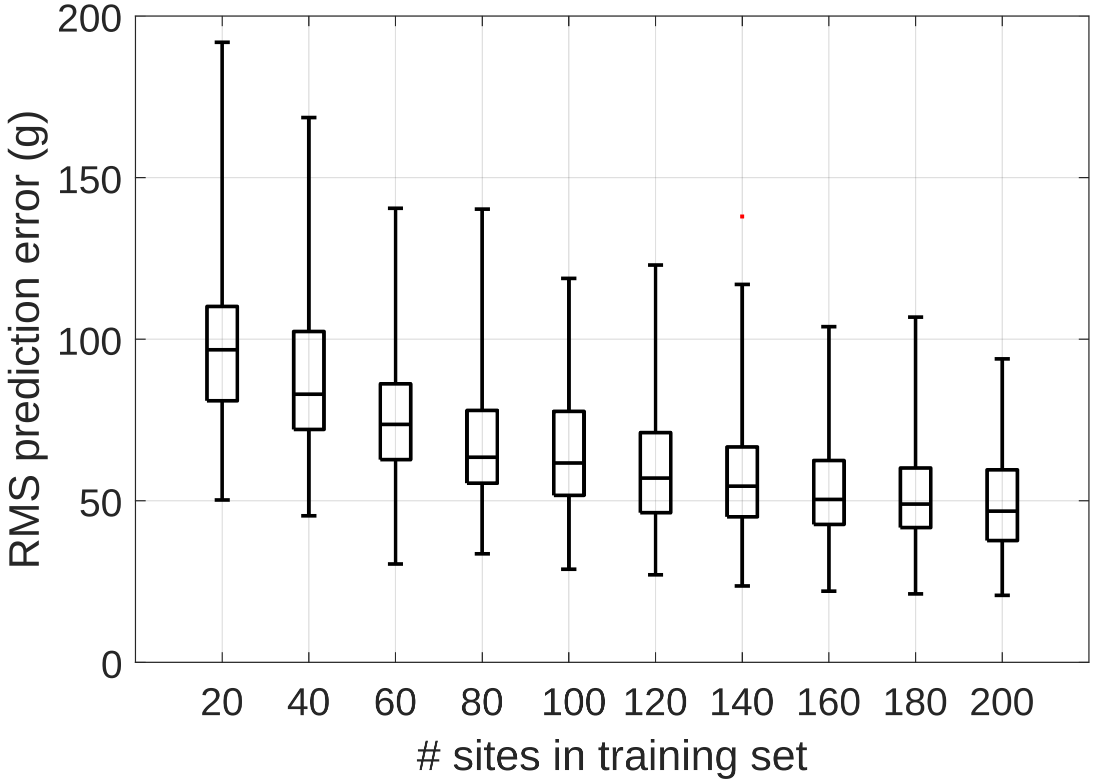


Figure 5.17: RMS prediction error vs. the number of 1-ft^2 sites used in training i.e. the training set size. The boxes correspond to the 0th, 25th, 50th, 75th, and 100th percentiles with respect to the RMS prediction error across 200 random draws of training and testing sites for each fixed training set size.

According to the fit, we would need to collect field data at a total of 361 1-ft^2 sites to achieve a mean RMS prediction error of 35.0 g, which is 10% of the mean yield across the 213 sites. This assumes that the fit perfectly represents the mean RMS prediction error as a function of the number of sites in the training set and that future sites have similar ground truth distributions of yield and soil moisture.

$$194e^{-0.26n^{0.32}} \tag{5.1}$$

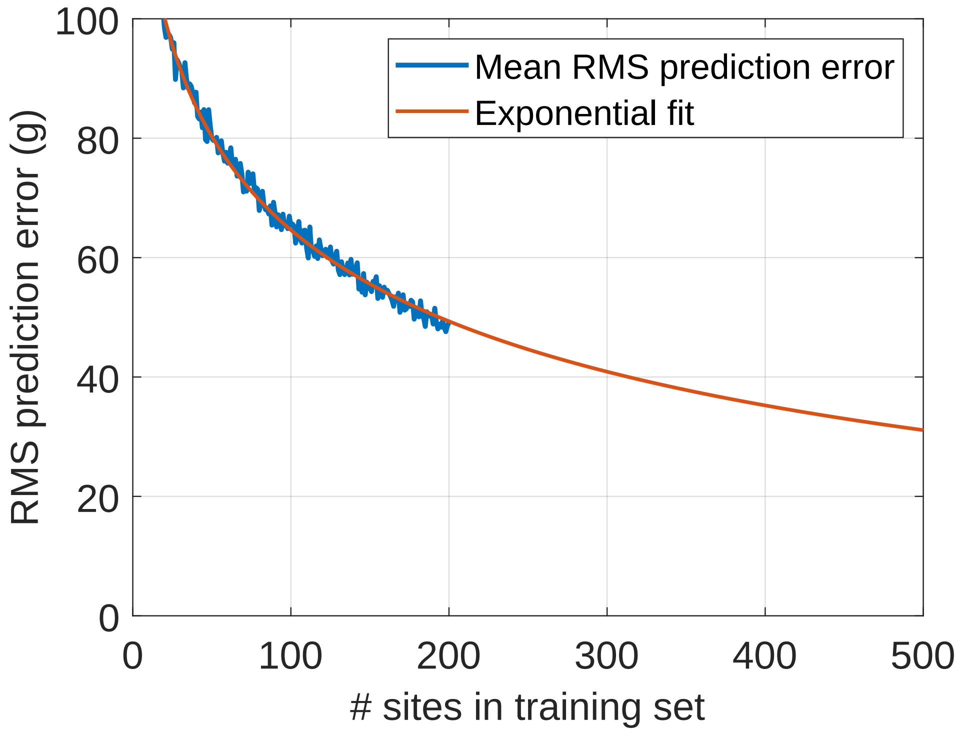


Figure 5.18: Mean of the RMS prediction error across 200 random draws of training and testing sites for each fixed training set size and the exponential fit of Eq. 5.1.

5.6.4 Summary

We evaluated the rate at which yield prediction error decreases as a function the number of 1-ft^2 sites used to train a SVM with a Gaussian kernel parameter of 1.4. We found that the decrease in median RMS prediction error was monotonic as a function of the number of 1-ft^2 sites used in training. We also found that the mean RMS prediction error was described well by a decaying exponential. At least 360 1-ft^2 sites are needed to achieve a mean RMS prediction error of 35 g, which is 10% of the mean yield across the 213 sites.

5.7 Microwave Features for Improved Yield Prediction

5.7.1 Introduction

We aim to improve microwave based yield prediction by generating and selecting important features from existing microwave signals. We generate features based on the slope of the reflected power as a function of frequency. Then, we select important features as yield predictors through Supervised Principal Component Analysis (SPCA). We compare the yield predictive capacity of these microwave features using Support Vector Regression with a Gaussian kernel and LOO-CV.

5.7.2 Methods

In Fall 2018 and Fall 2019 we logged the magnitude of the reflection coefficient at each 1 MHz increment in frequency from 606 MHz to 1300 MHz with our prototype sensor across 161 1-ft^2 sites. We generate additional features based on the slope of the magnitude of the reflection coefficient at each of the sampled frequencies. For the reflection coefficient at 606 MHz we calculate the slope as a forward difference. For the reflection coefficient at 1300 MHz we calculate the slope as a backward difference. For reflection coefficients at 607 MHz, 608 MHz, ..., 1299 MHz we calculated the slope as a centered difference. In total there are 695 reflection coefficient magnitudes spanning 606 MHz to 1300 MHz and from these 695 first derivatives. We scale the first derivatives to fall in the range [0, 1].

We evaluate the predictive accuracy of SVM regression applied to the magnitude of the reflection coefficient, the slope of the reflection coefficient, the magnitude concatenated with the slope, and finally SPCA applied to the concatenated magnitude and slope. Thus, for the magnitude of the reflection coefficient the SVM maps a 695-dimensional vector to yield. Then, for the slope of the reflection coefficient the SVM maps a 695-dimensional vector to yield. For the magnitude concatenated with slope, the SVM maps a 1390-dimensional vector to yield. Finally for SPCA applied to the concatenated magnitude and slope, SVM maps a p-dimensional vector to yield where p is the number of principal components retained in SPCA.

We illustrate Supervised Principal Component Analysis (SPCA) in Fig. 5.19 for the case where the slope vector is concatenated with the magnitude vector. In SPCA, the univariate regression coefficients are calculated for each feature i.e. each element in the 1390 length input vector. Features from the original set are retained if their univariate regression coefficients (Eq. 5.2) exceed a threshold. Principal Component Analysis then applies a linear transformation to the retained features to decorrelate them. Finally, a subset of the principal components with maximum variance are retained and a Support Vector Machine is trained to map them to yield. SPCA has two parameters that are optimized during the 32-fold cross validation training as described in section 5.2. The first parameter is a correlation threshold for the original features, and the other parameter is the number of principal components to retain. SPCA is described in more detail in [46].

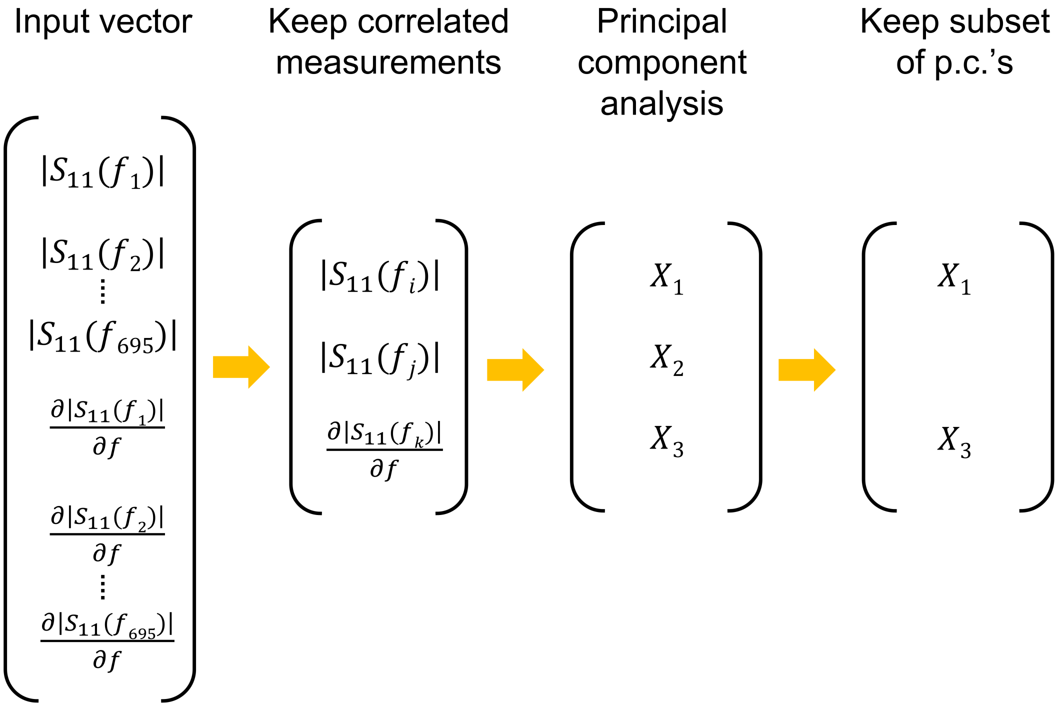

Input vector
Keep correlated measurements
Principal component analysis
Keep subset of p.c.'s
$|S_{11}(f_1)|$
$|S_{11}(f_2)|$
$|S_{11}(f_{695})|$
$\frac{\partial|S_{11}(f_1)|}{\partial f}$
$\frac{\partial|S_{11}(f_2)|}{\partial f}$
$\frac{\partial|S_{11}(f_{695})|}{\partial f}$
$|S_{11}(f_i)|$
$|S_{11}(f_j)|$
$\frac{\partial|S_{11}(f_k)|}{\partial f}$
X_1
X_2
X_3
X_1
X_3

Figure 5.19: Supervised Principal Component Analysis (SPCA) applied to the magnitude of the reflection coefficient concatenated with the slope of the magnitude of the reflection coefficient. The univariate regression coefficient is calculated for each feature i.e. each random variable e.g. $|S_{11}(f_i)|$. Features with regression coefficients larger than a threshold are retained and input to Principal Component Analysis. A subset of principal components with the largest variance are retained and used as inputs to the Support Vector Machine that maps them to yield.

$$C = \frac{\mathbf{x}^T\mathbf{y}}{\sqrt{\mathbf{x}^T\mathbf{x}}} \tag{5.2}$$

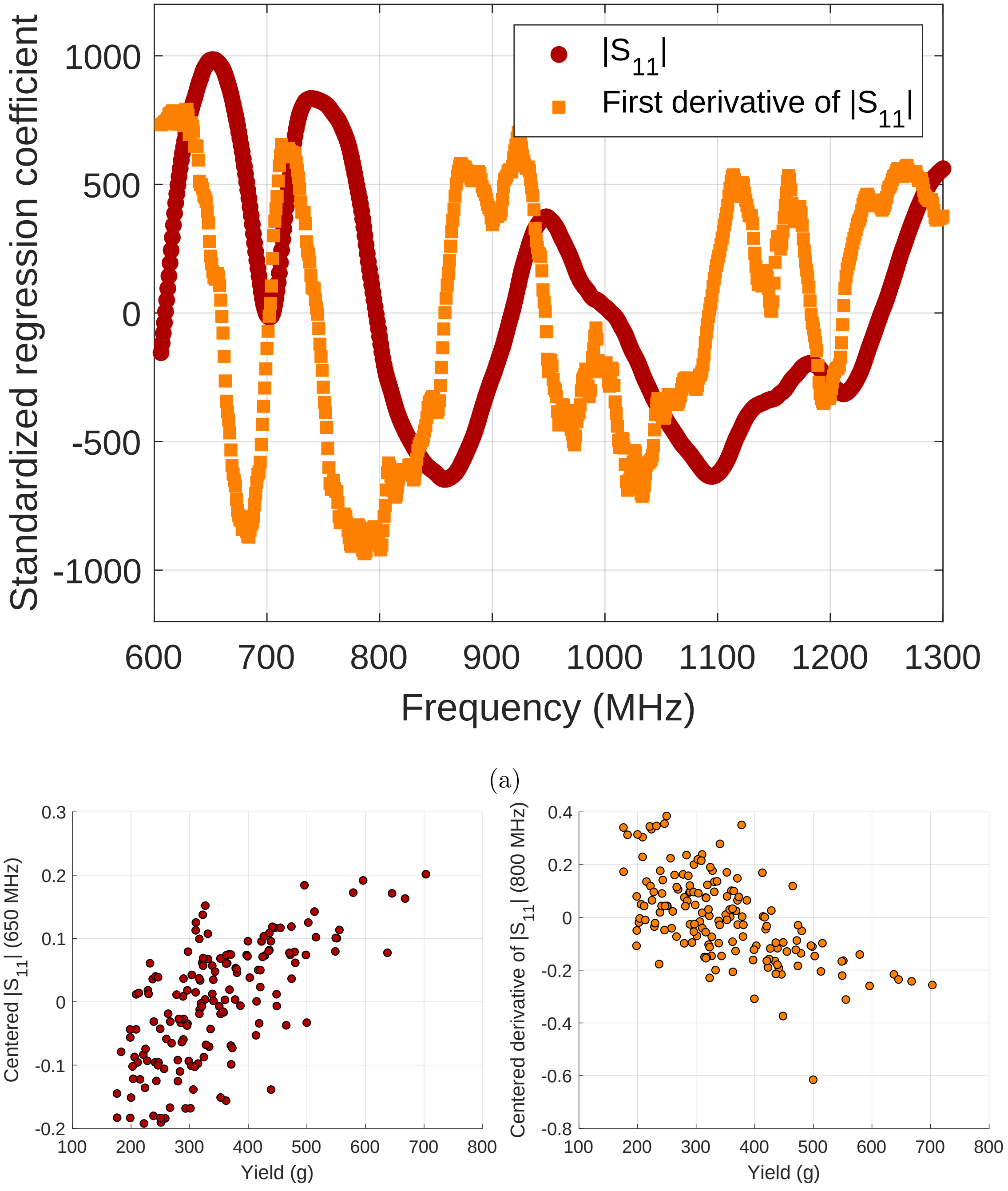
5.7.3 Results

In Fig. 5.20a we plot the standardized regression coefficients with respect to the magnitude of the reflection coefficient and slope as a function of frequency. Standardized regression coefficient is another term for univariate regression coefficient. The univariate regression coefficient describes a normalized correlation between a single feature and the yield. As seen in Fig. 5.20a, both the magnitude of the reflection coefficient and the slope of the magnitude of the reflection coefficient are correlated with yield in certain frequency bands more strongly than others. For reference, the standardized regression coefficient of the magnitude of the reflection coefficient at 650 MHz is 1000, and the data from 161 1-ft^2 sites are plotted in Fig. 5.20b. As seen in Fig. 5.20b, there is a definite positive trend corresponding to a positive standardized regression coefficient of 1000. The standardized regression coefficient of the slope at 800 MHz is -1000, and the data are plotted in Fig. 5.20c. As seen in Fig. 5.20c, there is a definite negative trend corresponding to a negative standardized regression coefficient of -1000.

As seen in Fig. 5.20b, the data show a large variation in terms of yield for a fixed feature value. For example, when $S_{11}(650) = 0$ yield varies from 200-450 g. Therefore, given a measurement of $S_{11}(650) = 0$ the best yield prediction might be 325 g (the mean across the data at that value), in which case there could be up to a ± 125 g error. This example illustrates the necessity of multiple features that are correlated to yield. However, additional features that have no correlation with yield act as noise. Thus, SPCA is an automated method for determining which features are correlated and important for yield prediction.

We plot the prediction results for the magnitude and slope features with and without SPCA in Fig. 5.21. The horizontal black lines represent the 25th and 75th percentiles, and the horizontal red line represents the median. The statistics are calculated from LOO-CV as in section 5.1. Prediction using the magnitude of the reflection coefficients produces a median error of 39.6 g. Prediction using the slope of the magnitude of the reflection coefficients produces a median error of 38.7 g. Prediction using the magnitude concatenated with the slope produces a median error of 37.0 g. Prediction using the magnitude concatenated with the slope with SPCA produces a median error of 37.3 g. The error distributions for the slope and magnitude with and without SPCA are tighter than for magnitude or slope alone.

It appears that the slope provides slightly more accurate yield prediction than the magnitude. However, using both the magnitude and slope provided the best prediction with or without SPCA applied. SPCA did not significantly alter the predictive accuracy when applied to the concatenated magnitude and slope data. We could calculate higher order derivatives in the hopes of finding better yield predictors, but noise in the original $|S_{11}|$ is amplified with each derivative calculation.



(a)

(b)

(c)

Figure 5.20: (a) Standardized or univariate regression coefficients for the magnitude of the reflection coefficient and slope between 606 MHz and 1300 MHz, calculated with Eq. 5.2 from 161 1-ft^2 sites between Fall 2018 and Fall 2019. The regression coefficient illustrates a normalized correlation between each microwave feature and yield. (b) The magnitude of the reflection coefficient at 650 MHz vs. yield. (c) The slope of the magnitude of the reflection coefficient at 800 MHz vs. yield.

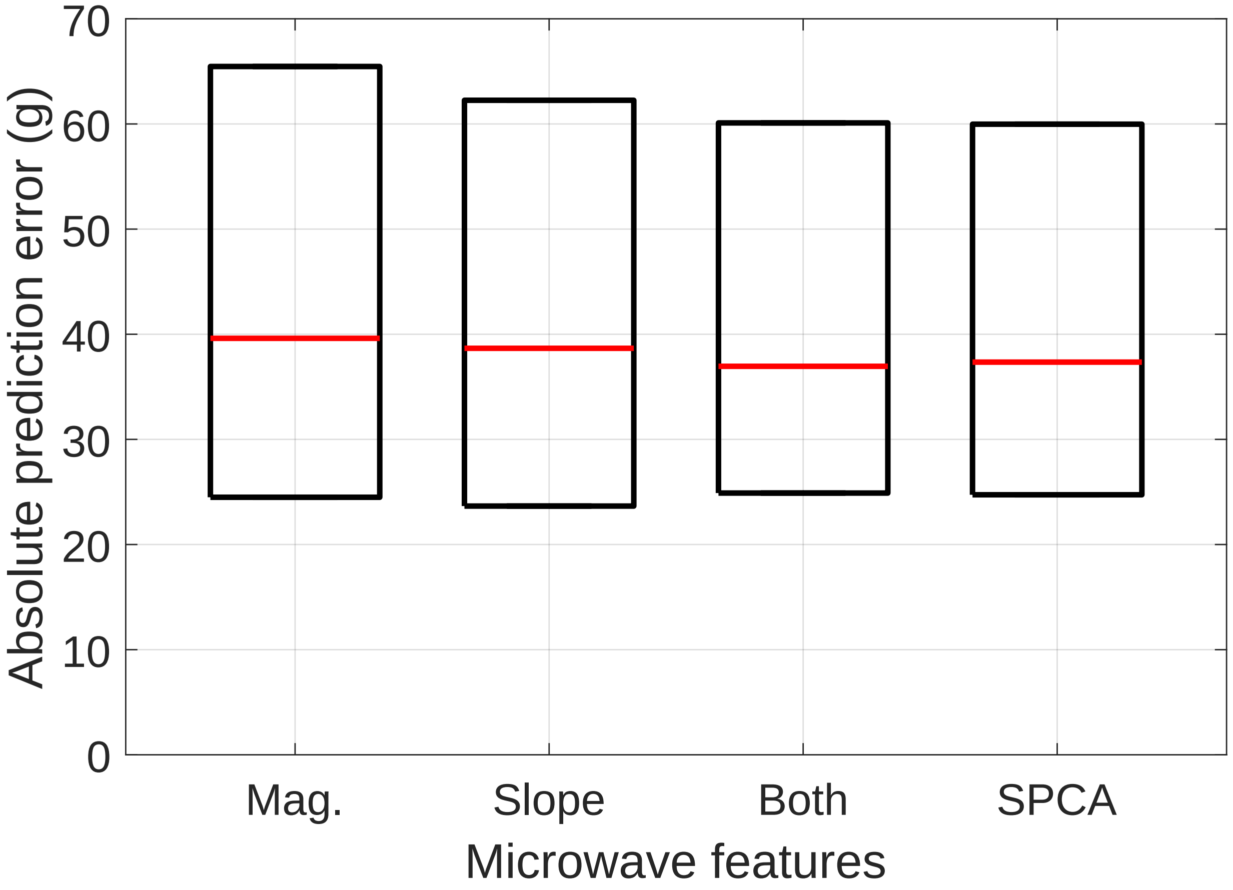


Figure 5.21: Statistics on the prediction error across 161 1-ft^2 sites evaluated with LOO-CV. Black horizontal lines represent the 25th and 75th percentiles, and a horizontal red line represents the median.

5.7.4 Summary

We aimed to improve microwave based yield prediction by generating and selecting important features from existing microwave signals. We generated features based on the slope of the magnitude of the reflection coefficient as a function of frequency. The addition of slope improved the predictive accuracy of the approach. Feature selection via SPCA did not significantly change the predictive accuracy using SVM with a Gaussian kernel to map the magnitude and slope of the reflection coefficients to yield.

Chapter 6

Data Acquisition and Data Fusion Strategies for Enhancing Yield Prediction (Aim 4)

6.1 Introduction

Our aim is to investigate data acquisition and data fusion strategies to enhance statistical learning based cranberry yield prediction, specifically with regard to microwave sensor design and multi-modality sensing. As part of the data acquisition investigation, we evaluate yield prediction with our computational sensor model with and without a calibration step, and compare it to yield prediction using plane wave sensing. As part of the data fusion investigation, we evaluate the utility of canopy photographs in predicting yield and combine the optical data and microwave data for improved yield prediction.

6.2 Simulated Data Acquisition for Improved Yield Prediction

6.2.1 Introduction

In our computational, experimental, and field data studies we evaluated yield prediction using the reflection coefficients measured at the coaxial input to the waveguide transition on the prototype sensor. However, microwaves propagate through the transition and body of the waveguide before exiting the aperture. The waveguide is constructed from a lossy aluminum material which introduces power loss. Therefore, the power transmitted and received at the waveguide aperture is significantly transformed due to internal propagation, reflection, and loss when measured at the coaxial input. We investigate the impact of calibration to remove these effects on yield prediction.

We developed the prototype microwave sensor to transmit and receive microwave power at frequencies below 2 GHz as explained in chapter 4. However, the prototype sensor operates in the near-field. This isn't an inherent disadvantage for yield sensing. However, we showed in Chapter 4 that the reflection coefficients measured with our computational sensor model were more sensitive to a 0.5" change in separation of the aperture from the top of the canopy than to a 20 g change in yield. In this case, the near-field sensing aspect of our sensor is a disadvantage. We investigate the yield predictive accuracy of far-field plane-wave sensing which is not significantly affected by aperture separation.

6.2.2 Methods

Simulated Calibration

We investigate sensor calibration in simulation using our computational sensor model discussed in Chapter 4. The calibration is completed using a S-parameter network that models reflection and lossy propagation within the sensor. Equation 6.1 displays the mathematical formula for the S-parameter network in the frequency domain i.e. each term in the equation is frequency dependent. In Eq. 6.1, S_{11} denotes the reflection coefficient at the coaxial input, S_{12} denotes the transmission coefficient from the coaxial input to the aperture plane of the waveguide, S_{21} denotes the transmission coefficient from the aperture plane of the waveguide to the coaxial input, and S_{22} is the reflection coefficient at the aperture of the waveguide. S_{11} captures the reflection due to an impedance mismatch between the 50 Ohm coaxial input and the coaxial-waveguide transition. S_{12} and S_{21} capture attenuation due to loss during propagation and a corresponding phase shift. S_{22} captures the reflection at the aperture of the waveguide due to a mismatch between the internal waveguide impedance and the impedance of free space. Γ_{in} is the measured reflection coefficient at the coaxial input, and Γ_{ap} is the reflection coefficient at the aperture plane.

$$\Gamma_{in} = S_{11} + \frac{S_{12}S_{21}\Gamma_{ap}}{1 - S_{22}\Gamma_{ap}} \tag{6.1}$$

Walking through Eq. 6.1, if Γ_{ap} is zero then the measured reflection coefficient $\Gamma_{in} = S_{11}$. In this case, the measured reflection coefficient is just the reflection at the coaxial input, which provides no information about the canopy-environment outside of the aperture. Next, assume that $S_{22} = 0$ in which case there is no impedance mismatch at the aperture of the waveguide. In this case, the measured $\Gamma_{in} = S_{11} + S_{12}S_{21}\Gamma_{ap}$. Thus, every Γ_{in} measurement is biased by S_{11}, and the reflection coefficient at the aperture Γ_{ap} is transformed by $S_{12}S_{21}$ whose absolute value is always less than one due to loss. In other words, multiplication by $S_{12}S_{21}$ decreases the difference between independent Γ_{ap} values i.e. decreases signal to noise ratio. Finally, S_{22} is non-zero in practice such that $\frac{S_{12}S_{21}\Gamma_{ap}}{1-S_{22}\Gamma_{ap}}$ creates a nonlinear function in Γ_{ap}. Γ_{ap} is already a nonlinear function of yield. A nonlinear mapping of a nonlinear function is unlikely to produce a less nonlinear function of yield. Thus, the calibration aims to reduce the impact of these sensor effects on the yield predictive capacity of our approach.

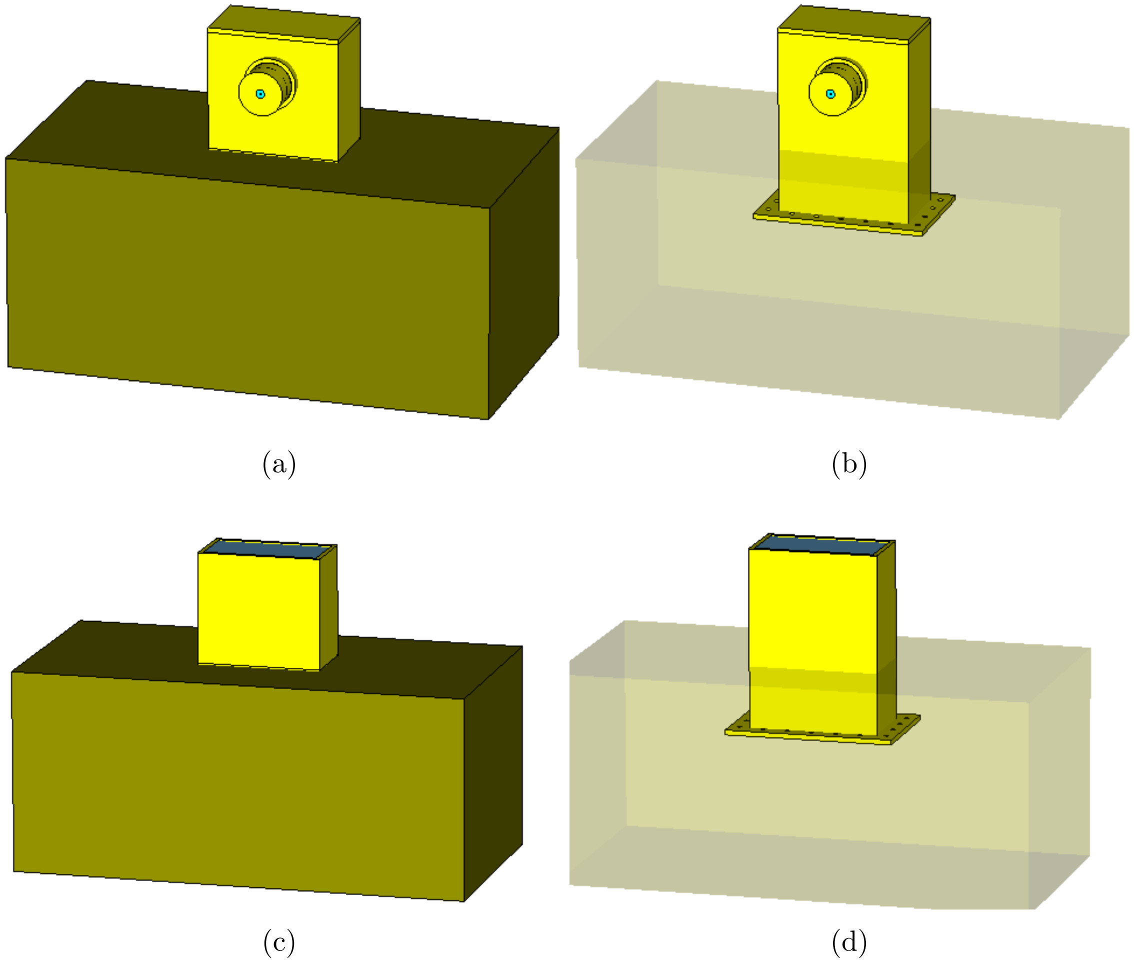
At each frequency, S_{11}, $S_{12}S_{21}$, and S_{22} need to be determined in order to invert Eq. 6.1 and transform any Γ_{in} measurement to Γ_{ap}. Thus, at each frequency we need three sets of independent measurements of Γ_{in} and Γ_{ap} in order to determine the network parameters. We obtain three independent measurements of Γ_{in} and Γ_{ap} by immersing the computational sensor model and a modified computational model in three different dielectric environments in Computer Simulation Technologies (CST) - Microwave Studio. An illustration of the computational and modified sensor are shown in Fig. 6.1. In Fig. 6.1a the computational sensor model is submerged in a dielectric half-space environment. This provides a simulated measurement of Γ_{in}. Figure 6.1b illustrates the modified computational sensor model in which the coaxial-waveguide transition is removed. The modified model consists of a section of WR975 waveguide and the same flange as in the original. The modified model is stimulated with a waveguide port at the top opening (opposite of flange end). Reflection coefficients measured at the waveguide port are referenced to the aperture plane at the flange end using CST's built in S-parameter network model for de-embedding measurements, thereby providing a simulated measurement of Γ_{ap}.

The three dielectric environments consist of Teflon, moist sandy soil, and water. All three materials have a permeability of 1.0. The relative permittivity for Teflon is 2.1. The relative permittivity of moist sandy soil is 13.0 and there is a loss tangent of 0.29 at 10 GHz. Water has a relative permittivity of 74 and an electrical conductivity of 3.53 S/m. These three dielectric materials span a wide range of permittivities which allows for better calculation of the S-parameters. From the three sets of Γ_{in} and Γ_{ap} we calculated S_{11}, $S_{12}S_{21}$, and S_{22} at each frequency. We minimize the squared difference between the model for Γ_{in} in Eq. 6.1 and the simulated measurement of Γ_{in}.

Plane Wave and Computational Sensor Model Simulations

We simulated the computational sensor model and a plane wave model in multiple canopy-soil environments to evaluate the capacity of each modality for accurate yield prediction. The canopy environments are similar to those in Chapter 3 and 4. The canopy is represented by a slab in which cranberry spheres are randomly distributed. The canopy material consists of 90% air dielectric and 10% healthy leaf dielectric as measured in Chapter 3. The berry material is that of cranberry flesh as measured in Chapter 3. The canopy slab is backed by a wet soil half space as described in Chapter 3. The aperture of the computational model waveguide is separated from the top of the canopy slab by 1", 2", 3", and 4". The soil moisture is fixed at 10% or 20%. The thickness of the canopy slab is fixed at 3" or 4". Yield varies from 100 to 600 g in 20 g increments for a total of 26 unique yields. The plane wave simulations are identical to the sensor model simulations except for a plane wave source instead of the sensor model source. The plane wave source is located one meter from the top of the canopy slab and produces a horizontally polarized plane wave normally incident upon the top of the canopy slab. The reflection coefficient is measured at the plane wave source location.

We apply Multi-Dimensional Scaling (MDS) to visualize the reflection coefficient data as



(a)

(b)

(c)

(d)

Figure 6.1: Computational sensor models submerged 12.84 cm in a dielectric half-space. (a) Original WR975 prototype computational sensor model in dielectric half-space. (b) Original WR975 prototype computational sensor model shown embedded within the dielectric material. (c) Modified computational sensor model in dielectric half-space. (d) Modified computational sensor model shown embedded within the dielectric material. The modified model is a flanged section of WR975 waveguide and is stimulated with a waveguide port at the top opening. Both models are simulated at frequencies between 606 MHz and 1300 MHz.

a function of yield. We briefly summarize MDS, but more details can be found in [46]. The MDS algorithm attempts to maps data of any dimension to any given dimension while attempting to preserve pairwise distances between all sinals. We apply MDS to the microwave reflection coefficient signals because they are 695-dimensional, which can not be plotted in a manner that humans can readily understand in terms of geometric distances or curves. Thus, we map the microwave signals onto 2-dimensions using MDS in order to visualize the mapping from microwave signal to yield.

We calculate the generalization error using LOO-CV applied to the simulated reflection coefficients for the 26 yields at fixed soil moisture contents, aperture separations, and canopy depths in order to evaluate the predictive power of the computational sensor and plane wave modalities. We simulated a single random berry distribution for each yield. Thus, we train an SVM with a Gaussian kernel parameter to predict yield using 25 microwave signals, and we test on the remaining microwave signal. We cycle through each of the 26 microwave signals used in testing for a total of 26 prediction error evaluations i.e. generalization errors. We calculate statistics on the generalization errors and report them. During training, we complete 25-fold cross validation to optimize the SVM. In the 25-fold cross validation we train the SVM on 24 signals, evaluate prediction error on the 25th test signal, and cycle through the 25 signals for testing. We minimize the mean squared prediction error across the 25 tests with respect to the SVM parameters.

6.2.3 Results

Calibration S-parameters

In Fig. 6.2, we plot the calculated S-parameters using the reflection coefficients simulated with the computational sensor submerged in Teflon, soil, and water dielectrics. The $|S_{12}S_{21}|$ is $<$ 0.3 between 606 MHz and 630 MHz. This corresponds to a lossy propagation from the coaxial input to the aperture of the waveguide, which is to be expected close to the 606 MHz cutoff frequency of the TE_{10} mode. Above 650 MHz the $|S_{12}S_{21}|$ is $>$0.9 and above 1000 MHz is is close to 1.0. Thus, there is a minimal amount of propagation loss between the aperture and coaxial input. The $|S_{11}|$ and $|S_{22}|$ are $<$0.3 between 700 MHz and 1300 MHz. Thus, there is some power reflected at the coaxial input and the aperture. However, the reflected power is less than 9% in both cases between 700 MHz and 1300 MHz. Thus, propagation, reflection, and attenuation are present in the Γ_{in} measurements collected with the prototype sensor but only mildly.

Plane Wave and Computational Model - Simulated Signals and Predictions

In Fig. 6.3 we plot representative reflection coefficients for the uncalibrated computational model Γ_{in}, the calibrated model Γ_{ap}, and the plane wave model as a function of yield for a 20% soil moisture, 2" aperture separation, and a 4" canopy depth. There are 26 reflection coefficient curves plotted in total which represent yields varying from 100 g to 600 g. In Fig. 6.3a, we plot the reflection coefficients measured at the coaxial input Γ_{in}

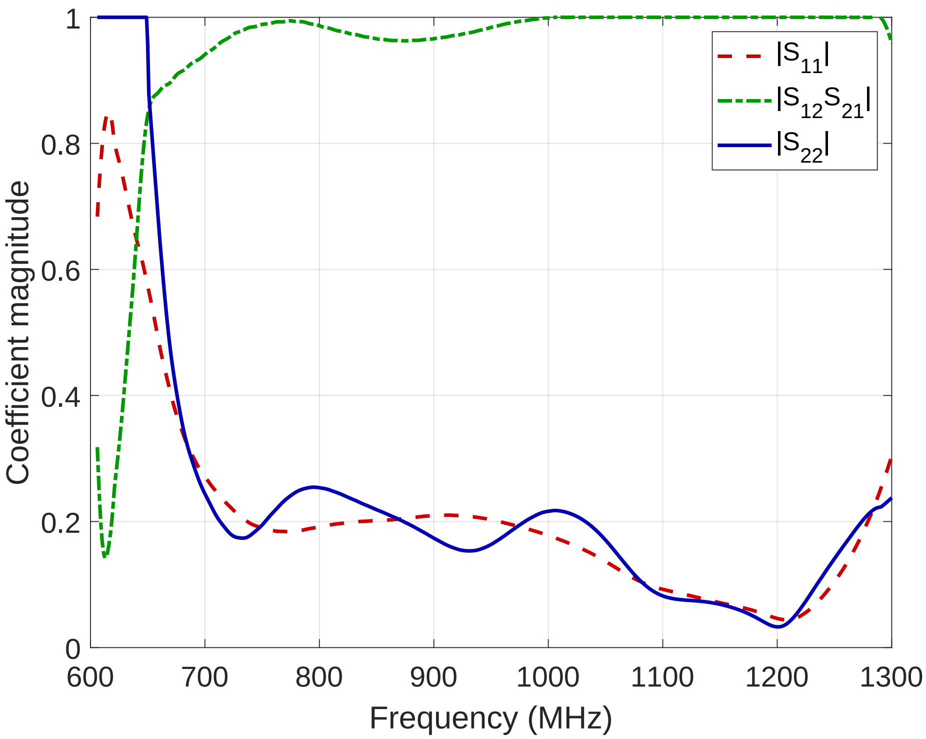


Figure 6.2: Frequency dependent S-parameters in computational sensor model network described by Eq. 6.1. S_{11} denotes the reflection coefficient at the coaxial input, S_{12} denotes the transmission coefficient from the coaxial input to the aperture plane of the waveguide, S_{21} denotes the transmission coefficient from the aperture plane of the waveguide to the coaxial input, and S_{22} is the reflection coefficient at the aperture of the waveguide.

in the computational sensor model. Yields produce visibly different reflection coefficients across the operating bandwidth. It appears that yield has a larger impact on the measured reflection coefficient from 1000-1300 MHz than from 606-1000 MHz. Also, the impact of yield on the reflection coefficient does not appear to be consistently linear, which can be seen at 1000 MHz where higher yields produce more clustering than lower yields.

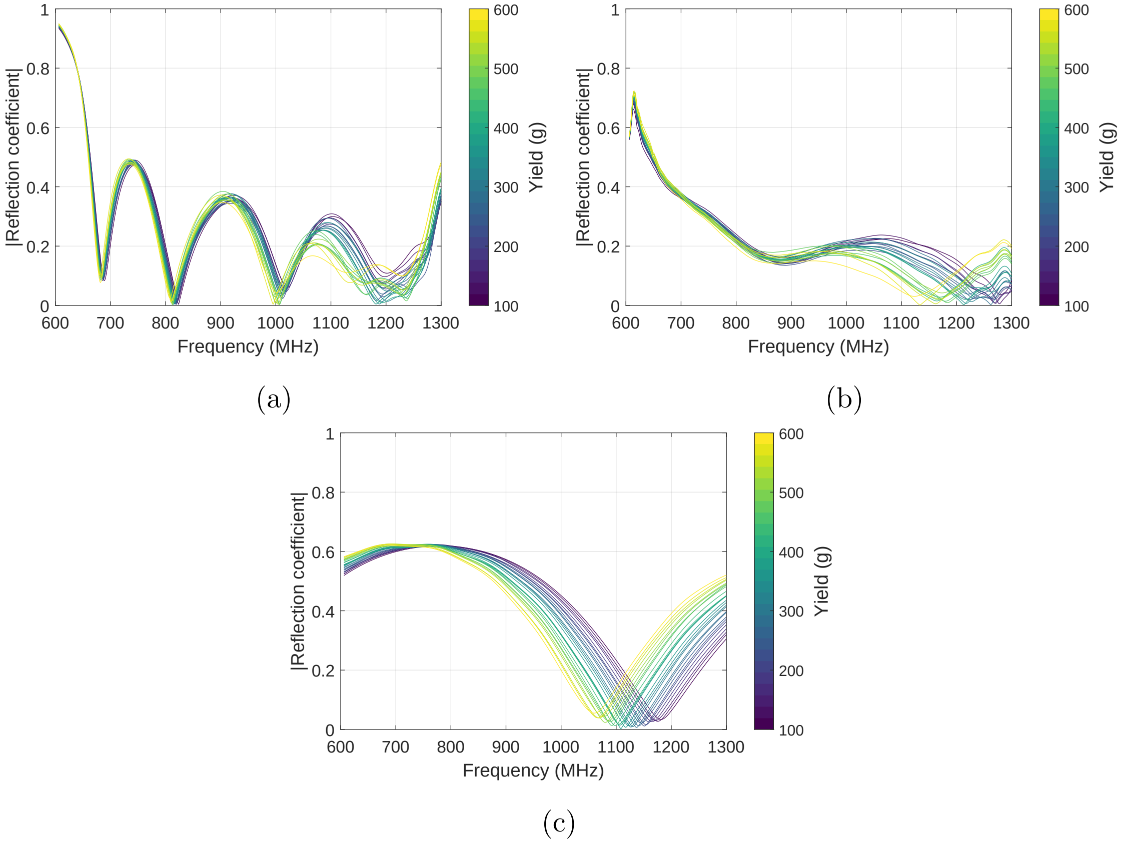
We plot the calibrated reflection coefficients Γ_{ap} in Fig. 6.3b. Similar to the measured reflection coefficients Γ_{in} in Fig. 6.3a, the calibrated reflection coefficients are visibly different across the operating bandwidth. Again, it appears that yield has a larger impact on the calibrated reflection coefficient from 1000-1300 MHz than from 606-1000 MHz. Also, the impact of yield on the calibrated reflection coefficients does not appear to be consistently linear, which can be seen at 1000 MHz where higher yields produce more clustering than lower yields. It is interesting to note the effect of calibration on the reflection coefficents. In Fig. 6.3a, there are peaks and nulls that correspond to internal resonances of the waveguide. In Fig. 6.3a, these peaks and nulls have been removed. Thus, the structure of the calibrated reflection coefficients is significantly different from the structure of the uncalibrated reflection coefficients.

We plot the simulated reflection coefficients for the plane wave model in Fig. 6.3c. Yields produce visibly different reflection coefficients across the operating bandwidth. It appears that yield has a larger impact on the measured reflection coefficient from 1000-1300 MHz than from 606-1000 MHz. Interestingly, the impact of yield on the reflection coefficient appears to be consistently linear across the full bandwidth. The peak and null correspond to a resonance related to the canopy depth. The canopy was modeled with a slab geometry which produces resonant behavior.

In Fig. 6.4, we plot the same reflection coefficients as in Fig. 6.3 with Multi-Dimensional Scaling applied. The 26 reflection coefficient signals are projected onto two coordinates. We verified that the projection preserved the pairwise distances between all 26 signals qualitatively well. In Fig. 6.4a, it appears that the uncalibrated microwave signals have a nonlinear relationship with yield. For example, along Coordinate 1 at a value of 0.5 there is a visually large jump from one blue-green dot to the next blue-green dot at a value of 0.65. Also it seems that higher yields in yellow are not clustered as tightly as light blue dots corresponding to lower yields. It is possible to fit a highly nonlinear function to the data in Fig. 6.4a, but we would be overfitting. The highly nonlinear fit is susceptible to noise, and small variations in input signal can produce large variations in yield predictions once fitting is complete.

In Fig. 6.4b, we plot the calibrated reflection coefficients with Multi-Dimensional Scaling applied. Surprisingly, Fig. 6.4b is almost identical to Fig. 6.4a. This indicates that the calibration procedure does not impact pairwise distances between measured reflection coefficient signals with respect to yield, even though there is attenuation, reflection, and propagation effects associated with the uncalibrated reflection coefficients. This implies that yield prediction from the uncalibrated microwave reflection coefficient signals will be comparable to yield prediction from the calibrated microwave reflection coefficient signals.

In Fig. 6.4c, we plot the reflection coefficients from the plane wave simulations with Multi-Dimensional Scaling applied. The projected microwave signals appear to have a



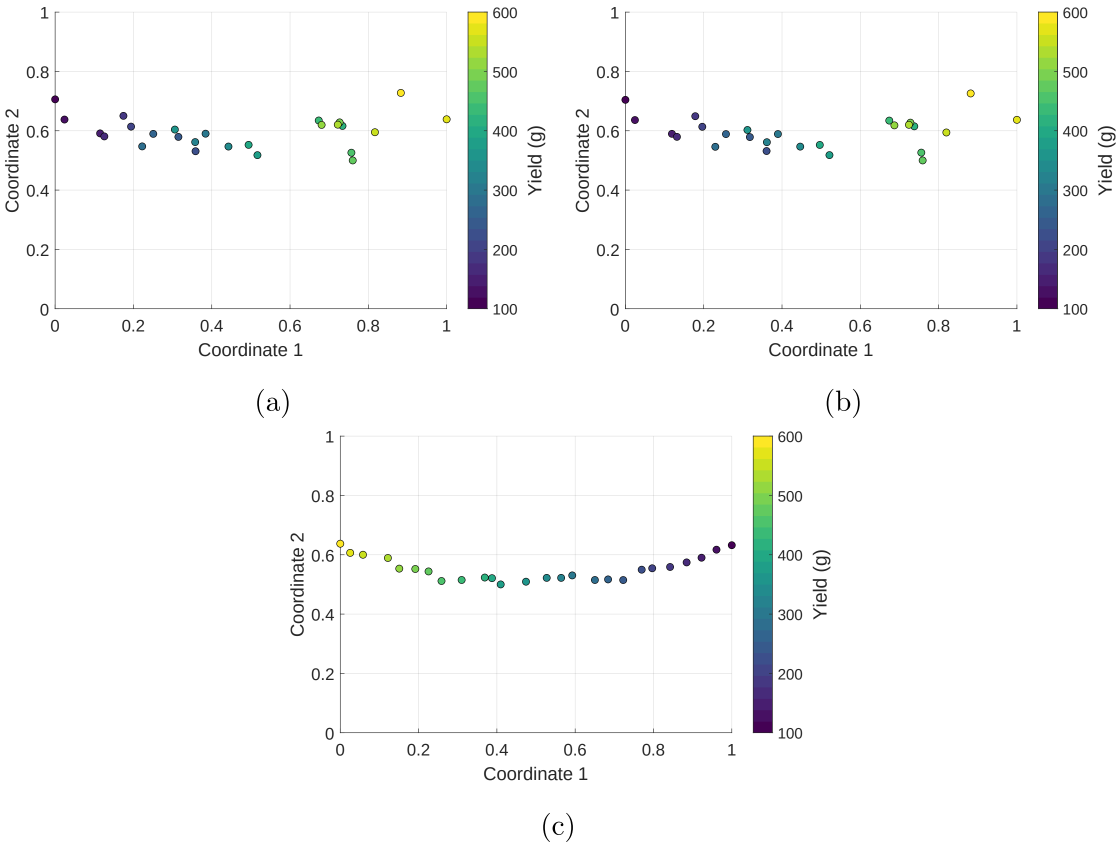
(a)

(b)

(c)

Figure 6.3: Simulated reflection coefficients for the computational sensor and plane wave models in a canopy environment with 20% soil moisture content and 4” canopy depth. (a) Simulated reflection coefficients measured at the coaxial input Γ_{in} with a 2” aperture separation. (b) Simulated reflection coefficients measured at the coaxial input with a 2” aperture separation and calibrated using Eq. 6.1 for Γ_{ap}. (c) Simulated reflection coefficients captured with the computational plane wave model.

monotonic and possibly linear relationship with yield along Coordinate 1. This suggests that the plane wave signals should provide accurate yield prediction due to the simplicity of the relationship between yield and microwave signal.



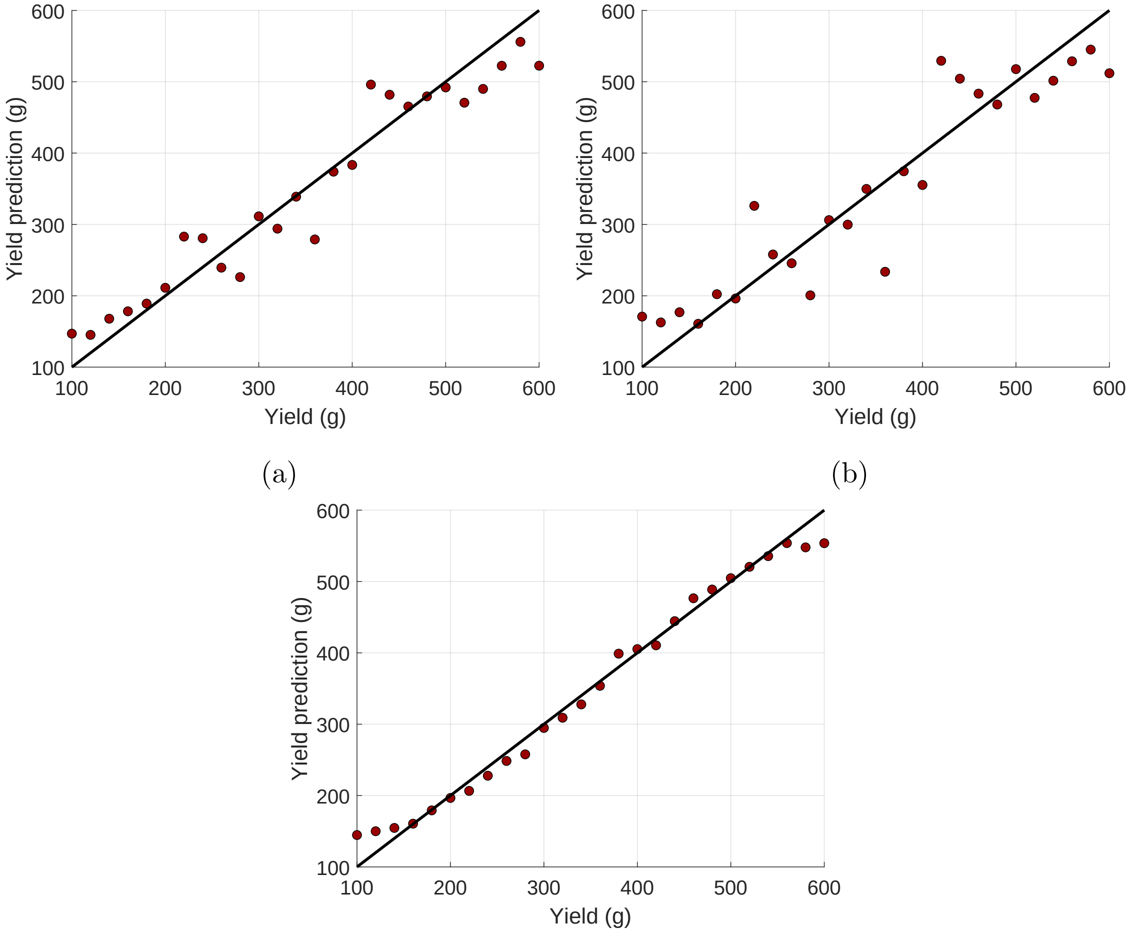
(a)

(b)

(c)

Figure 6.4: Multi-Dimensional Scaling (MDS) applied to simulated reflection coefficients for the computational sensor and plane wave models in a canopy environment with 20% soil moisture content and 4” canopy depth. (a) MDS applied to simulated reflection coefficients measured at the coaxial input Γ_{in} with a 2” aperture separation. (b) MDS applied to simulated reflection coefficients measured at the coaxial input with a 2” aperture separation and calibrated using Eq. 6.1 for Γ_{ap}. (c) MDS applied to simulated reflection coefficients captured with the computational plane wave model. Coordinate 1 and Coordinate 2 correspond to the two coordinates on which the 695-dimensional data were projected.

In Fig. 6.5 we plot the SVM based yield predictions against the ground truth yield for the uncalibrated sensor model, the calibrated model, and the plane wave model for the 20% soil moisture, 2” aperture separation, and a 4” canopy depth simulation environments. There are 26 dots displayed in each plot corresponding to test predictions from each reflection coefficient signal in LOO-CV. In Fig. 6.5a, we plot the prediction results for the reflection coefficients from the uncalibrated sensor model. In Fig. 6.5b, we plot the prediction results for the calibrated reflection coefficients from the sensor model. In Fig. 6.5c, we plot the prediction results for the reflection coefficients from the plane wave model. As suggested from the MDS projections, the plane wave reflection coefficients provided the most accurate yield predictions for this simulation environment. Also, the calibrated and uncalibrated reflection coefficients measured with the sensor model had comparable predictive accuracy as suggested by the MDS projections as well.



(a)

(b)

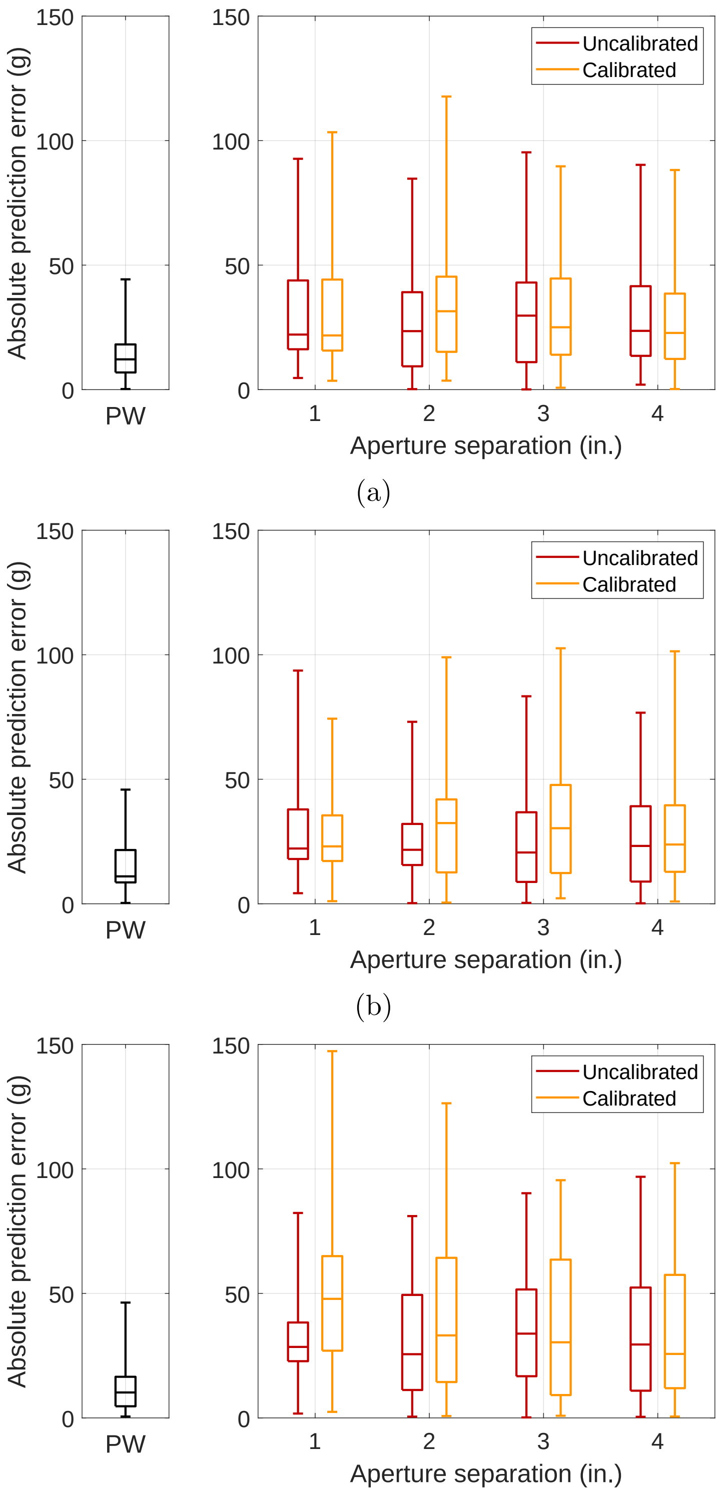
(c)

Figure 6.5: Yield predictions from the simulated reflection coefficients measured by the (a) uncalibrated sensor model, the (b) calibrated sensor model, and the (c) plane wave model. Ground truth yield is plotted on the x-axis. There are 26 dots displayed in each plot corresponding to test predictions from each reflection coefficient signal in LOO-CV.

In Fig. 6.6, we plot the prediction error statistics across the full set of simulation environments. Statistics are calculated from the 26 test prediction errors resulting from LOO-CV for a fixed soil moisture content, aperture separation, and canopy depth. In Fig. 6.6a, we plot the prediction error statistics for a 10% soil moisture content and a 3” canopy depth. In Fig. 6.6b, we plot the prediction error statistics for a 20% soil moisture content and a 3” canopy depth. In Fig. 6.6c, we plot the prediction error statistics for a 20% soil moisture content and a 4” canopy depth. It is visible from Fig. 6.6 that plane wave reflection coefficients provided the most accurate yield prediction using SVM with a Gaussian kernel across all soil moisture contents, aperture separations, and canopy depths investigated. The uncalibrated and calibrated reflection coefficients had comparable predictive accuracy in most cases.

6.2.4 Summary

We investigated the impact of calibration on microwave based yield prediction using our computational prototype sensor model. We found that calibration using an S-parameter network did not significantly alter the yield predictive accuracy of our computational sensor model. Importantly, we simulated far-field plane wave based yield prediction and found that it was superior to our near-field computational model across a range of canopy environments.



(a)

(b)

(c)

Figure 6.6: Prediction error statistics across the full set of simulation environments. Statistics are calculated from the 26 test prediction errors resulting from LOO-CV for a fixed soil moisture content, aperture separation, and canopy depth. (a) Simulation environment with 3" canopy depth and 10% soil moisture content. (b) Simulation environment with 3" canopy depth and 20% soil moisture content. (c) Simulation environment with 4" canopy depth and 20% soil moisture content. Box plots display the 0th, 25th, 50th, 75th, and 100th percentiles on absolute prediction error.

6.3 Experimental Data Fusion

6.3.1 Introduction

Statistical prediction of cranberry yield from optical data has been attempted before. However, the studies involved limited optical features and mainly linear regression algorithms. We evaluate the predictive accuracy of optical based yield prediction using the canopy photographs with their corresponding ground truth yield collected during Fall 2018 and Fall 2019. We compare yield predictive accuracy using the optical data to yield predictive accuracy using the microwave data. Finally, we propose two methods for using both the microwave and optical data to enhance yield prediction.

6.3.2 Methods

We evaluate the predictive accuracy of SVM regression applied to optical data and microwave data using the field data collected in Fall 2018 and Fall 2019. The field data encompasses 12 microwave measurements obtained with the prototype sensor, a canopy photograph, a soil moisture measurement, and the harvested ground truth yield at 161 1-f^2 sites. We evaluate the predictive accuracy of SVM regression with a Gaussian kernel applied to just the microwave data. We calculate the generalization error as in Chapter 5 using LOO-CV. The 12 microwave measurements at each of 160 1-f^2 sites are used in training, the 12 microwave measurements at the remaining 1-f^2 testing site are used to evaluate prediction error, and we cycle through the 1-f^2 site used in testing a total of 161 times. The microwave data are mathematically illustrated in Fig. 6.7. The magnitude of the reflection coefficient at 695 discrete frequencies from 606 MHz to 1300 MHz is concatenated with the first derivative of the reflection coefficient at each of the 695 discrete frequencies. Thus, the SVM is mapping a 1390 length input vector to yield.

We evaluate the predictive accuracy of SVM regression with a Gaussian kernel applied to just the optical data. We calculate the generalization error as in Chapter 5 using LOO-CV. Each optical image at 160 1-f^2 sites are used in training, the optical image at the remaining 1-f^2 testing site is used to evaluate prediction error, and we cycle through the 1-f^2 site used in testing a total of 161 times. The optical data are illustrated in Fig. 6.8. Each canopy photograph is first cropped such that it only captures the 1-ft^2 site. The canopy photograph is represented digitally by an optical image. Each optical image was roughly 2400 pixels by 2400 pixels for a total of 5760000 pixels. Each pixel has an associated Red, Green, and Blue (RGB) intensity, where the intensity ranges in value from 0 to 1. We calculated 25 color indices at each of the 5760000 pixels for each image. The color indices are defined in Fig. 6.9 and cover a range of visual representations of each pixel. For example, there are color indices that emphasize each color in three different representations of color e.g. the HSV colormap and there are color indices that describe differences between color intensities e.g. Excess Green Index. As shown in Fig. 6.9, we calculate the mean and variance of each color index across the 5760000 pixels for each image. We include two additional features derived

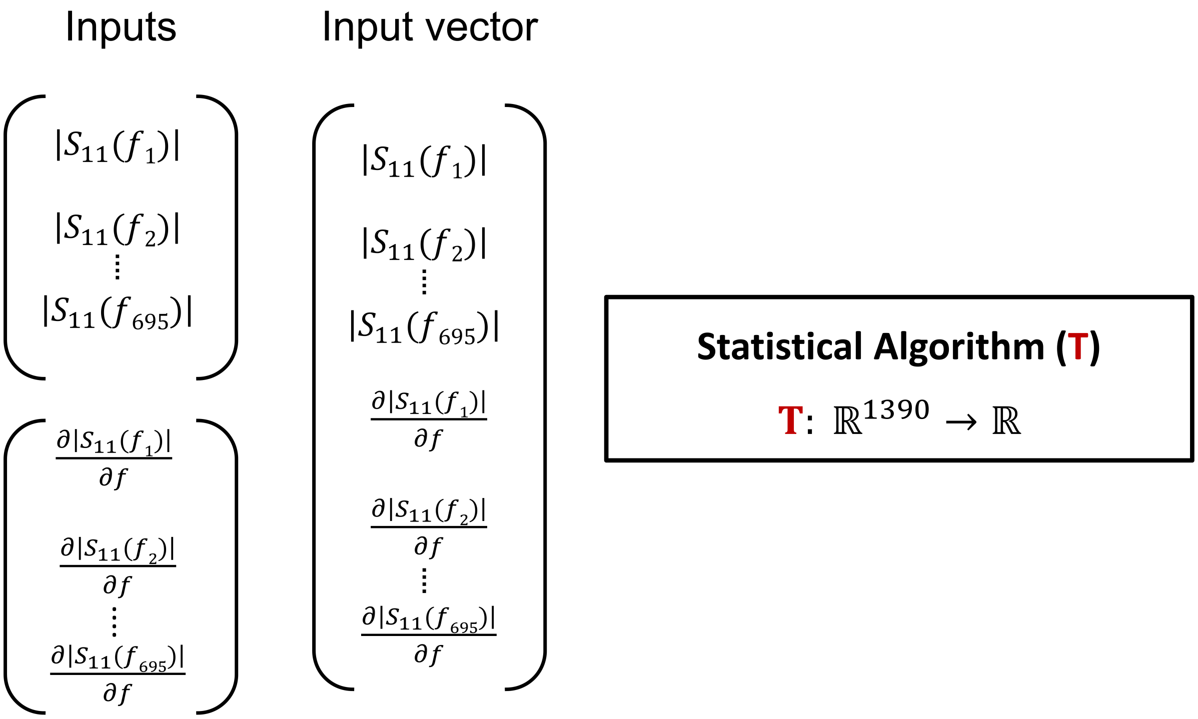

Inputs
Input vector
$|S_{11}(f_1)|$
$|S_{11}(f_2)|$
$|S_{11}(f_{695})|$
$\frac{\partial|S_{11}(f_1)|}{\partial f}$
$\frac{\partial|S_{11}(f_2)|}{\partial f}$
$\frac{\partial|S_{11}(f_{695})|}{\partial f}$
Statistical Algorithm (T)
$T\colon \mathbb{R}^{1390} \rightarrow \mathbb{R}$

Figure 6.7: Mathematical representation of microwave data and the statistical mapping to yield. The magnitude of the reflection coefficient is recorded at 695 discrete frequencies from 606 MHz to 1300 MHz spaced by 1 MHz intervals. The derivative of the magnitude of the reflection coefficient with respect to frequency at each of the 695 frequencies is illustrated. The magnitude and slope constitute 695 length vectors. After concatenating the magnitude and slope vectors the resulting vector has a length of 1390. The statistical learning algorithm (T) maps the real-valued 1390 length vector to a real valued yield.

from a leaf count and a leaf segmentation fraction calculated through the Region-Based Convolutional Neural Network of [53]. In total, for each image there are 52 features i.e. a 52-length input vector which the SVM maps to yield.

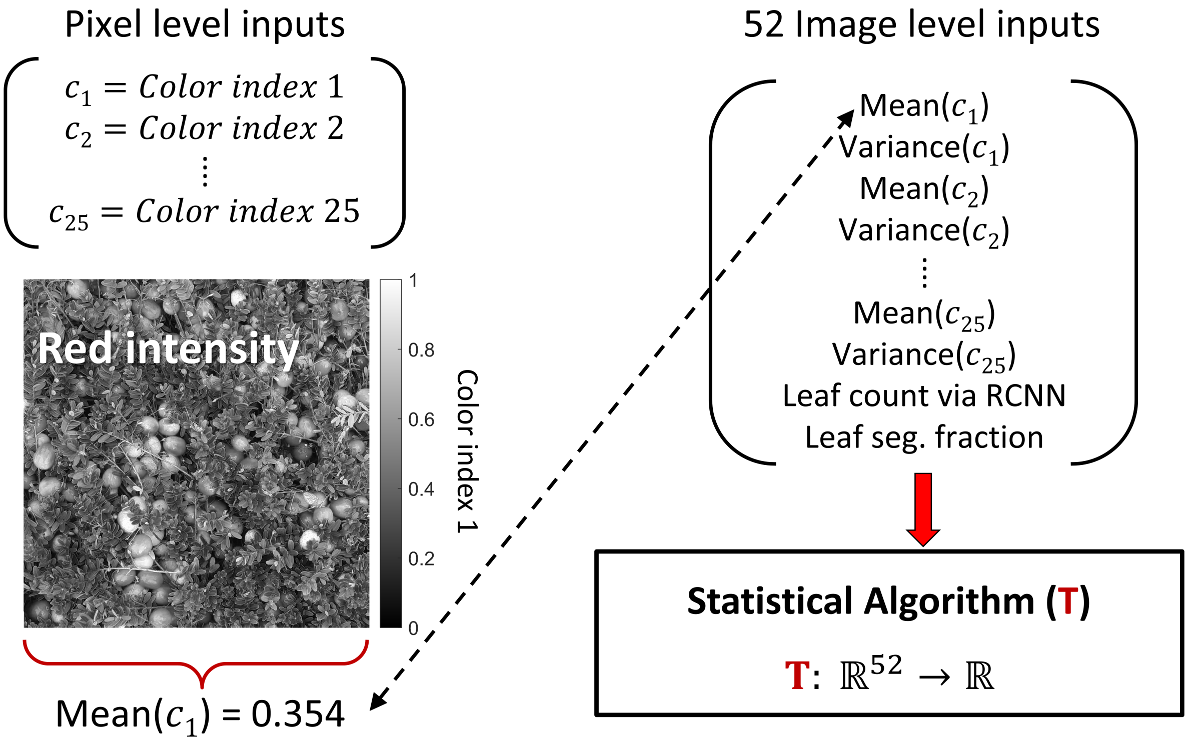
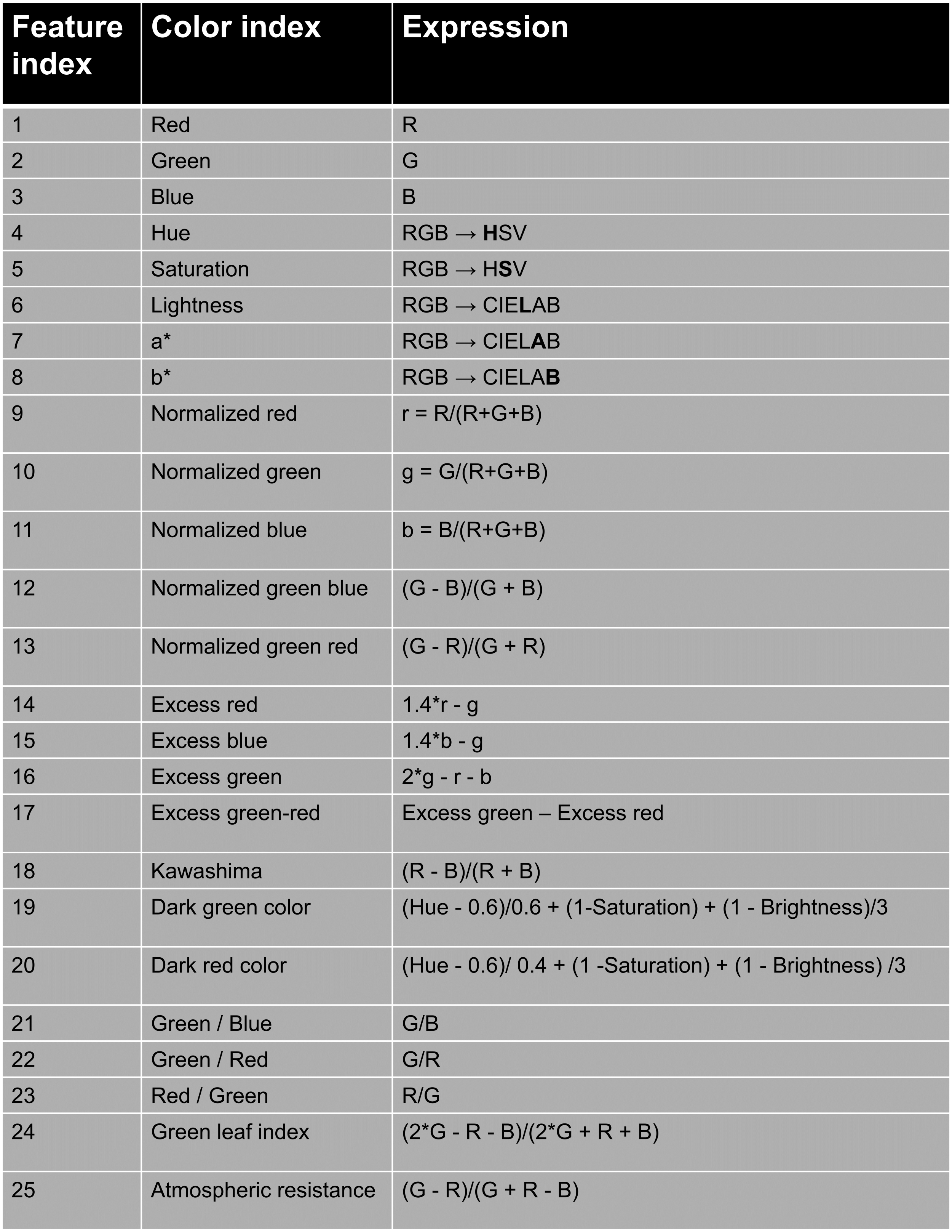


Figure 6.8: Conversion of optical images to optical feature vectors and the statistical mapping to yield. Raw canopy images are represented by Red, Green, and Blue (RGB) intensities at each pixel. At each pixel, 25 color indices are calculated from the RGB values. Optical features are calculated from the mean and variance of each color index across the image. Two additional features are included in the input vectors based on a leaf count and leaf segmentation fraction calculated with a Region-Based Convolutional Neural Network trained on synthetic leaf data [53]. The statistical learning algorithm (T) maps the real-valued 52 length vector to a real valued yield.

We combine the microwave and optical data through a linear regression on separate predictions as shown in Fig. 6.10. In this approach, we have two separate Support Vector Machines. The first SVM maps the 1390 dimensional microwave inputs to yield while the other SVM maps the 52 dimensional optical inputs to yield. Then, the final yield prediction is a linear combination of the prediction from the microwave-SVM and the prediction from the optical-SVM. This is called model stacking and it is discussed in further detail in [46]. We still complete 32-fold cross validation during training as described in Chapter 5. However, we are optimizing more parameters with this hybrid approach. There are parameters associated with the microwave-SVM, parameters associated with the optical-SVM, and



Feature index	Color index	Expression
1	Red	R
2	Green	G
3	Blue	B
4	Hue	RGB → **H**SV
5	Saturation	RGB → H**S**V
6	Lightness	RGB → CIE**L**AB
7	a*	RGB → CIEL**A**B
8	b*	RGB → CIELA**B**
9	Normalized red	r = R/(R+G+B)
10	Normalized green	g = G/(R+G+B)
11	Normalized blue	b = B/(R+G+B)
12	Normalized green blue	(G - B)/(G + B)
13	Normalized green red	(G - R)/(G + R)
14	Excess red	1.4*r - g
15	Excess blue	1.4*b - g
16	Excess green	2*g - r - b
17	Excess green-red	Excess green – Excess red
18	Kawashima	(R - B)/(R + B)
19	Dark green color	(Hue - 0.6)/0.6 + (1-Saturation) + (1 - Brightness)/3
20	Dark red color	(Hue - 0.6)/ 0.4 + (1 -Saturation) + (1 - Brightness) /3
21	Green / Blue	G/B
22	Green / Red	G/R
23	Red / Green	R/G
24	Green leaf index	(2*G - R - B)/(2*G + R + B)
25	Atmospheric resistance	(G - R)/(G + R - B)

Figure 6.9: Optical features calculated from Red (R), Green (G), and Blue (B) intensities at each pixel in canopy image. The RGB intensities fall in the range [0, 1].

finally the weights for the final prediction a, b that are optimized during training.

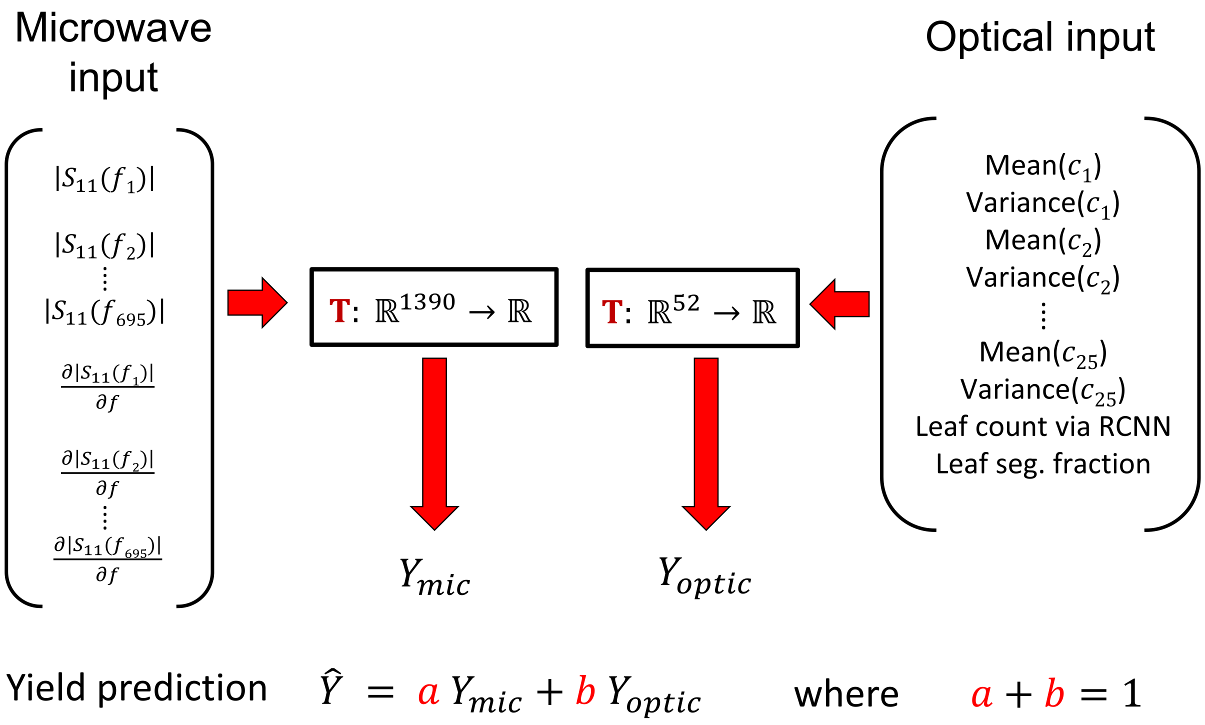

Microwave input
$|S_{11}(f_1)|$
$|S_{11}(f_2)|$
$|S_{11}(f_{695})|$
$\frac{\partial|S_{11}(f_1)|}{\partial f}$
$\frac{\partial|S_{11}(f_2)|}{\partial f}$
$\frac{\partial|S_{11}(f_{695})|}{\partial f}$
$\mathbf{T}$: $\mathbb{R}^{1390} \to \mathbb{R}$
$\mathbf{T}$: $\mathbb{R}^{52} \to \mathbb{R}$
Optical input
Mean(c_1)
Variance(c_1)
Mean(c_2)
Variance(c_2)
Mean(c_{25})
Variance(c_{25})
Leaf count via RCNN
Leaf seg. fraction
Y_{mic}
Y_{optic}
Yield prediction $\hat{Y} = a\,Y_{mic} + b\,Y_{optic}$ where $a + b = 1$

Figure 6.10: Yield prediction using a linear combination of a microwave based yield prediction and an optical based yield prediction. The parameters of each statistical learning algorithm and the final weights a, b are determined during the training step using 32 fold cross validation.

We also evaluate another hybrid microwave-optical approach shown in Fig. 6.11. In the second hybrid approach, we simply concatenate the microwave input vector with the optical input vector. The Microwave input vectors consist of 1390 elements and optical input vectors consist of 52 elements. Concatenating the inputs results in a 1442 length input vector. We map the 1442-dimensional input vectors to yield through a Gaussian kernel SVM. Finally, we concatenate soil moisture to the 1442-dimensional microwave-optical input vector and evaluate the generalization error using a Gaussian kernel SVM.

Finally, we complete a sampling study on the prediction error using the hybrid approach where there are two separate Gaussian kernel Support Vector Machines (SVM) that predict yield from the microwave and optical data separately, whose predictions are combined linearly. There are 12 predictions from each of the 161 1-ft^2 sites used in evaluating this hybrid approach. First, we calculate the mean error across the 12 predictions at each 1-ft^2 site. This is equivalent to averaging the 12 outputs, then calculating the prediction error. Thus, there is a single prediction error for each 1-ft^2 site. Next, we keep 149 out of the 161 average prediction errors corresponding to sites whose yield was in the range [200, 600] g. This set

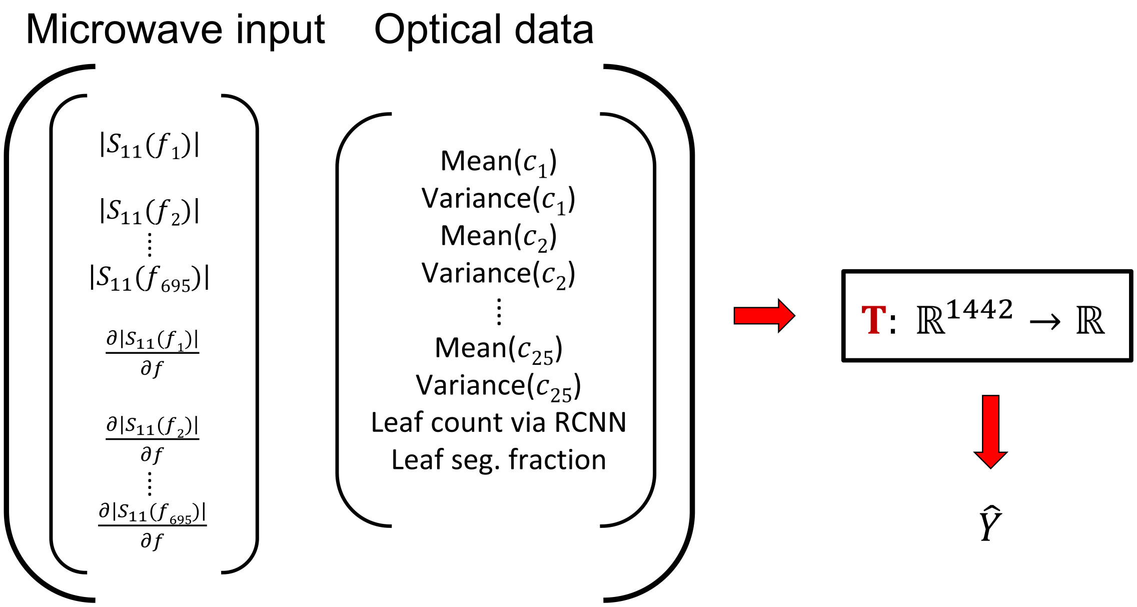


Figure 6.11: Combination of microwave and optical inputs through concatenation. Microwave input vectors consist of 1390 elements and optical input vectors consist of 52 elements. Concatenating the inputs results in a 1442 length input vector which the statistical learning algorithm maps to a real valued yield prediction.

of 149 prediction errors represents the distribution of error we would expect in a commercial cranberry bed when training the hybrid approach with 12 microwave signals and a canopy photograph at 160 1-ft^2 sites, and averaging 12 SMV-output predictions from 12 microwave signals collected in a single 1-ft^2 test site.

We want to evaluate how the average prediction error of our approach behaves as we test it on more than a single 1-ft^2 site. We emulate this through sampling the error distribution. For a fixed number of 1-ft^2 test sites, we sample the error distribution 10000 times and report the absolute value of the average error across the number of 1-ft^2 test sites. For example, in one case we fix the number of 1-ft^2 test sites at 20. We then randomly sample 20 prediction errors from the 149 1-ft^2 sites' prediction error. Then, we average the 20 prediction errors and calculate the absolute value. This represents a single scenario where we test the trained hybrid approach on 20 random 1-ft^2 test sites in a commercial cranberry bed and evaluate the average total prediction error across the 20 1-ft^2 test sites. We repeat sampling 20 prediction errors 10000 times.

6.3.3 Results

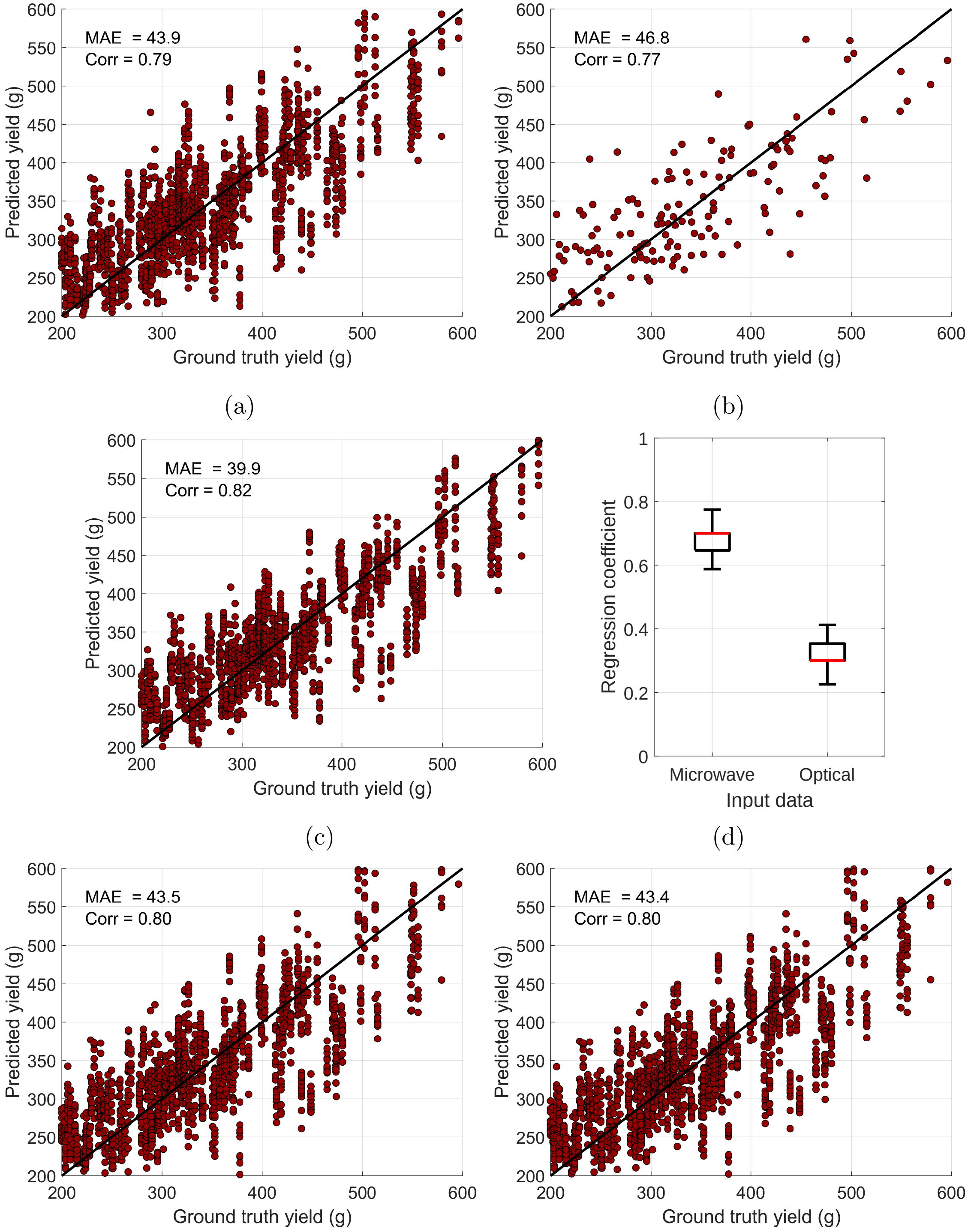
We plot the yield predictions from each approach resulting from LOO-CV in Fig. 6.12. From Fig. 6.12a we see that the microwave based prediction resulted in a mean absolute error (MAE) of 43.9 g and a correlation coefficient between the predictions and ground truth yield of 0.79. It is apparent that the microwave based yield prediction was more accurate than the optical based yield prediction as shown in Fig. 6.12b, where the MAE was 46.8 g and the correlation coefficient was 0.77. The first hybrid approach where microwave and optical predictions were combined linearly resulted in the best prediction accuracy as seen in Fig. 6.12c. The MAE for the linear combination of predictions was 39.9 g and the correlation coefficient was 0.82. The larger correlation between predicted and ground truth yields for the hybrid approach is visually apparent when comparing Fig. 6.12a with Fig. 6.12b.

The statistics on the prediction weights for the hybrid approach across the 161 testing evaluations are displayed in Fig. 6.12d. There is a tight distribution for each weight. The median weight for the microwave prediction was 0.7, whereas the median weight for the optical prediction was 0.3. Thus, the microwave based prediction provided a better estimate of the true yield than the optical based prediction. The second hybrid approach (concatenation of microwave and optical data) with and without soil moisture resulted in predictive accuracy comparable to just microwave based prediction. Consequently, concatenation between fundamentally different measurements does not appear to be an effective method for improving yield prediction.

The prediction error results for the sampling study are shown in Fig. 6.13. Fig. 6.13a displays the empirical prediction error distribution using the hybrid optical-microwave SVM approach, with a 12-output average as the final yield prediction. It appears that the prediction error is roughly unbiased with a large probability of predicting yield with less than 10 g of error. In Fig. 6.13b and Fig. 6.13c, we plot the average prediction error of our approach as a function of the number of 1-ft^2 sites we sample for testing the approach. The mean prediction error when testing on a single 1-ft^2 site is 37 g. The mean prediction error when testing on a five 1-ft^2 sites is 18 g. The mean prediction error when testing on a 9 1-ft^2 sites is 12 g. The mean prediction error when testing on 120 1-ft^2 sites is 2 g. The mean yield per 1-ft^2 site across the 149 1-ft^2 sites is 342 g. Thus, an average error of 37 g represents 10.8% error across a bed. Also, 18 g represents a 5.3% total bed error, 12 g represents a 3.5% total bed error, and 2 g represents a 0.6% total bed error. Thus, it is clear that as we sample more 1-ft^2 sites our approach becomes more accurate on average, and we can achieve extremely accurate yield prediction when applied to a full cranberry bed.

6.3.4 Summary

We evaluated the predictive accuracy of optical based yield prediction, microwave based yield prediction, and combinations of microwave and optical based yield prediction. We found that combining microwave and optical data through a linear combination of separate microwave and optical predictions enhanced yield predictive accuracy. We also found that microwave based predictions were more reliable for predicting yield than optical based predictions. Finally, through our sampling study we found that our hybrid microwave-optical SVM-based prediction approach could achieve total bed-wise error on the order of 1% of the total bed yield or less.



(a)

(b)

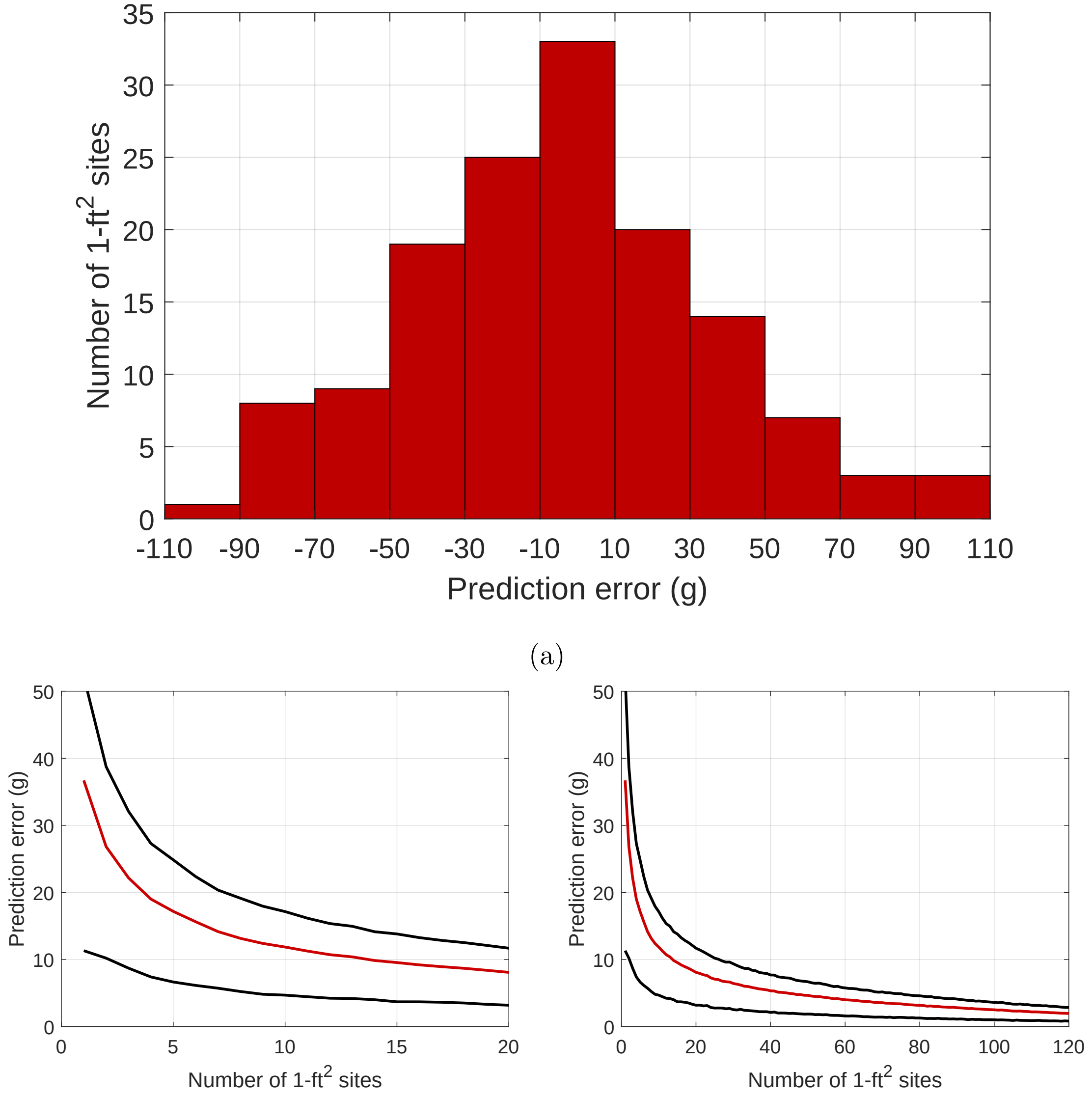
(c)

(d)

(e)

(f)

Figure 6.12: Yield predictions and statistics for microwave and optical based prediction. (a) Microwave based yield predictions. (b) Optical based yield predictions. (c) Linear combination of microwave and optical based yield predictions. (d) Weights for the linear combination of microwave and optical based yield prediction. (e) Concatenated microwave and optical input based predictions. (f) Concatenated microwave, optical, and soil moisture input based predictions. Predictions are the result of LOO-CV applied to the 161 1-ft^2 sites from field data collection in Fall 2018 and Fall 2019. 'MAE' is the mean absolute error in prediction and 'Corr.' is the correlation coefficient between the predicted and ground truth yields.



(a)

(b)

(c)

Figure 6.13: Prediction error statistics across the 149 1-ft^2 sites comprising the field data from Fall 2018 and Fall 2019 with ground truth yields between 200 g and 600 g. LOO-CV was used to evaluate the prediction error using Gaussian kernel SVM regression and the hybrid approach that averages microwave and optical predictions. (a) Mean prediction error across 12 microwave refection coefficient signals at each site. (b) Absolute value of the mean prediction error as a function of the number of sites sampled for testing. (c) Absolute value of the mean prediction error as a function of the number of sites sampled for testing across a wider range of the number of sites sampled for testing. The black lines in (a) and (b) represent the 25th and 75th percentiles of prediction error, and the red line represents the mean prediction error.

Chapter 7

Summary, Conclusions, and Future Work

We investigated cranberry-crop yield prediction using primarily microwave sensing. In Chapter 3, we measured and formulated the dielectric properties of cranberries, leaves, and soil and incorporated them into simulations of microwave scattering from cranberry canopy models. From the simulations, we found that frequencies below 2000 MHz were suitable for yield prediction using an active microwave sensor.

In Chapter 4, we developed and characterized an experimental prototype microwave sensor. We developed a computational model for the prototype sensor and conducted simulations to evaluate its sensing volume and sensitivity to yield, soil moisture, canopy depth, and aperture separation from the canopy. We found that the reflection coefficients measured by the computational sensor model are affected by cranberries extending throughout the full depth of the canopy and extending laterally 12-15" in both transverse directions. We also found that the reflection coefficients measured by the computational sensor model are more sensitive to soil moisture and aperture separation than to yield. With the experimental prototype we measured a total of 2556 microwave signals across 213 1-ft^2 sites from nine days of field data collection in central Wisconsin. We harvested the ground truth yield in each site. In 161 1-ft^2 sites we also measured soil moisture and captured a photograph of the canopy.

In Chapter 5, we investigated accurate yield prediction using statistical learning applied to the microwave signals collected with the prototype sensor in the field. We found that Gaussian kernel Support Vector Machine regression performed better than 5 other well-developed algorithms. We found that collecting multiple microwave signals from each 1-ft^2 site and averaging them enhanced yield predictive accuracy. It appeared that yield prediction was hindered by sparsity in training data, specifically in terms of soil moisture and ground truth yield representation. Training data size was exponentially related to prediction error on a 1-ft^2 site basis. Finally, we found that the slope of the measured reflection coefficients with respect to frequency was a valuable predictor of yield.

In Chapter 6, we simulated a far-field plane wave source and our computational sensor model above cranberry canopy models. We found that plane wave illumination provided microwave data that could be used to predict yield more accurately than our computational sensor model. We also investigated the use of optical data from the canopy photographs in predicting yield. We found that microwave based predictions were more accurate than optical based predictions. However, a hybrid statistical learning approach using optical and microwave data provided more accurate yield prediction than either sensing modality alone. Finally, we demonstrated outstanding predictive accuracy when applying our hybrid microwave-optical approach to many 1-ft^2 sites, suggesting total bed-wise error on the order of 1% of the total bed yield or less.

We have developed and demonstrated an approach for yield prediction based on active microwave sensing and statistical learning. All other methods for yield prediction have produced errors on the order of 15% of the total bed yield or more across a cranberry bed. With our prototype, we were able to achieve errors on the order of less than 1% of the total bed yield for a 120 1-ft^2 testbed. In addition, the error using our prototype was decreasing as a function of the number of 1-ft^2 sites in our testbed, suggesting even smaller error across a full cranberry bed. Our approach has the potential to automate yield prediction in commercial cranberry beds with remarkable accuracy.

Future Directions

There are many exciting avenues for future research despite the success of our approach to yield prediction through the prototype sensor. We obtained our results for a specific set of soil moistures and yields in commercial beds. Regardless of the sensor used for yield prediction, it is important to evaluate predictive performance across all potential yields, soil moistures, and canopy depths. A denser sampling of canopy environments is recommended. A related but different investigation relates to sampling. Our approach or any sensing approach can only predict yield in the areas in which active sensing occurs. An important question is how many sites need to be sampled in order to achieve a certain level of prediction error if for some reason the final approach only allows for a small fraction of each bed to be sampled with the sensor(s).

Ground based antenna arrays pose one approach for sensing a majority of the cranberry bed quickly. They can consist of planar elements that are lightweight, relatively inexpensive, and allow for signal processing via synthetic aperture radar. Thus, the antenna arrays could conceivably extract localized measurements across the width of the bed, while the boom on which they are mounted is driven across the length of the bed. This approach would allow the whole bed to be illuminated rapidly. We would need to investigate this approach thoroughly with regard to a couple of variables. First, it is possible that there is an angular dependence on scattering such that a fixed yield can produce multiple backscattered signals depending on the illumination and viewing angle. Second, we would need to verify that the array approach produces roughly unbiased prediction errors across many testing sites so that full bed yield prediction is accurate. Finally, we would need to design the array such that collecting ground truth yield is feasible.

Our prototype demonstrated accurate yield prediction with a small sensing volume or sensing extent, meaning it would have to be moved many times to cover a cranberry bed. The proposed ground-based antenna array approach could be designed to cover a larger sensing extent either physically or with signal processing. However, such an approach needs to be driven along a bed and trained using microwave data with ground truthed yield (laborious). Thus, I propose another research direction that uses microwave data obtained from active sensors in space to predict yield across cranberry beds. There are antennas mounted on spacecraft continuously monitoring the earth, and a lot of data is either freely available or not restricted. An advantage for the space based sensor is in ground truthing. The antennas in space have spatial resolution on the order of a cranberry bed's surface area. Each bed is harvested and the total bed yield is measured during the harvest. Thus, with little effort it is possible to obtain full cranberry bed microwave measurements with the corresponding ground truth full bed yield every cranberry harvest. I believe this is an opportunity to augment any approach to full bed yield prediction.

Bibliography

[1] L. Pozdnyakova, D. Gimenez, and P. Oudemans, “Spatial Analysis of Cranberry Yield at Three Scales,” *Agronomy Journal*, vol. 97, pp. 49-57, 2005.

[2] L. DeVetter, J. Colquhouna, J. Zalapa, and R. Harbut, “Yield estimation in commercial cranberry systems using physiological, environmental, and genetic variables,” *Scientia Horticulturae*, vol. 190, pp. 83-93, July 2015.

[3] L. Pozdnyakova, P. Oudemans, M. Hughes, and D. Gimenez, “Estimation of spatial and spectral properties of phytophthora root rot and its effects on cranberry yield,” *Computers and Electronics in Agriculture*, vol. 37, no. 1-3, pp. 57-70, December 2002.

[4] R. Kerry, P. Goovaerts, D. Gimenez, P. Oudemans, and E. Muñiz “Investigating geostatistical methods to model within-field yield variability of cranberries for potential management zones,” *Precision Agriculture*, vol. 17, no. 3, pp. 247-273, June 2016.

[5] R. Kerry, P. Goovaerts, D. Gimenez, and P. Oudemans “Investigating temporal and spatial patterns of cranberry yield in New Jersey fields,” *Precision Agriculture*, vol. 18, no. 4, pp. 507-524, August 2017.

[6] A. Haufler, J. Booske, S. C. Hagness, B. Tilberg, L. Wells-Hansen, and R. Serres, “Feasibility of efficient and accurate estimation of cranberry crop yield using microwave sensing,” 2017 IEEE International Symposium on Antennas and Propagation and USNC/URSI National Radio Science Meeting, San Diego, CA, USA, July 2017, pp. 375-376.

[7] A. Haufler, J. Booske, S. C. Hagness, and B. Tilberg, “An experimental pilot study of cranberry crop yield estimation using near-field microwave sensing,” 2018 IEEE International Symposium on Antennas and Propagation and USNC/URSI National Radio Science Meeting, Boston, MA, USA, 2018, pp. 1147-1148.

[8] S. C. Steele-Dunne, H. McNairn, A. Monsivais-Huertero, J. Judge, P.-W. Liu, and K. Papathanassiou, “Radar Remote Sensing of Agricultural Canopies: A Review,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 5, pp. 2249-2273, May 2017.

[9] J. Chen, H. Lin, and Z. Pei, "Application of ENVISAT ASAR Data in Mapping Rice Crop Growth in Southern China," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 3, pp. 431-435, July 2007.

[10] Y. Oh, K. Sarabandi, and F. T. Ulaby, "An Empirical Model and an Inversion Technique for Radar Scattering from Bare Soil Surfaces," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 30, no. 2, pp. 370-381, Mar. 1992.

[11] Y. Oh, K. Sarabandi, and F. T. Ulaby, "Semi-Empirical Model of the Ensemble-Averaged Differential Mueller Matrix for Microwave Backscattering from Bare Soil Surfaces," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 6, pp. 1348-1355, Jun. 2002.

[12] M. Hallikainen, F. Ulaby, M. Dobson, M. El-rayes, and L. Wu, "Microwave Dielectric Behavior of Wet Soil-Part 1: Empirical Models and Experimental Observations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-23, no. 1, pp. 25-34, Jan. 1985.

[13] J. Betbeder, R. Fieuzal, and F. Baup, "Assimilation of LAI and Dry Biomass Data From Optical and SAR Images Into an Agro-Meteorological Model to Estimate Soybean Yield," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 6, pp. 2540-2553, June 2016.

[14] M. Zribi, O. Taconet, S. Le Hegarat-Mascle, D. Vidal-Madjar, C. Emblanch, C. Loumagne, and M. Normand, "Backscattering behavior and simulation comparison over bare soils using SIR-C/X-SAR and ERASME 1994 data over Orgeval," *Remote Sens. Environ.*, vol. 59, no. 2, pp. 256-266, Feb. 1997.

[15] A. Toure, K. P. B. Thomson, G. Edwards, R. J. Brown, and B. G. Brisco, "Adaptation of the MIMICS Backscattering Model to the Agricultural Context-Wheat and Canola at L and C Bands," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 1, pp. 47-61, Jan. 1994.

[16] J. van Zyl, "On the importance of polarization in radar scattering problems," Ph.D. dissertation, Calif. Inst. Technol., Pasadena, CA, Dec. 1985.

[17] S. Bakhtiari and R. Zoughi, "A Model for Backscattering Characteristics of Tall Prairie Grass Canopies at Microwave Frequencies," *Remote Sensing of Environment*, vol. 36, pp. 137-147, 1991.

[18] G. Cookmartin, P. Saich, S. Quegan, R. Cordey, P. Burgess-Allen, and A. Sowter, "Modeling Microwave Interactions with Crops and Comparison with ERS-2 SAR Observations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 2, pp. 658-670, Mar. 2000.

[19] G. Macelloni, S. Paloscia, P. Pampaloni, F. Marliani, and M. Gai, "The Relationship Between the Backscattering Coefficient and the Biomass of Narrow and Broad Leaf Crops," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 4, pp. 873-884, Apr. 2001.

[20] M. Burgin, D. Clewley, R. M. Lucas, and M. Moghaddam, "A Generalized Radar Backscattering Model Based on Wave Theory for Multilayer Multispecies Vegetation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 12, pp. 4832-4845, Dec. 2011.

[21] S. L. Durden, J. J. van Zyl, and H. A. Zebker, "Modeling and Observation of the Radar Polarization Signature of Forested Areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 27, no. 3, pp. 290-301, May 1989.

[22] P. Siqueira and K. Sarabandi, "A numerically derived electromagnetic scattering model for grass grain heads," 1996 IEEE International Geoscience and Remote Sensing Symposium, Lincoln, NE, USA, 1996, vol. 2, pp. 1337-1339.

[23] A. N. Arslan, J. Koskinen, J. Pulliainen, and M. Hallikainen, "A semi empirical backscattering model of forest canopy covered by snow using SAR data," 2000 IEEE International Geoscience and Remote Sensing Symposium, Honolulu, HI, USA, 2000, vol. 5, pp. 1904-1906.

[24] S. Paloscia, E. Santi, G. Fontanelli, F. Montomoli, M. Brogioni, G. Macelloni, P. Pampaloni, and S. Pettinato, "The Sensitivity of Cosmo-SkyMed Backscatter to Agricultural Crop Type and Vegetation Parameters," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 7, pp. 2856-2868, Jul. 2014.

[25] M. Pardé, J.-P. Wigneron, A. Chanzy, Y.H. Kerr, J.-C. Calvet, P. Waldteufel, S Schmidl Søbjærg and N. Skou, "N-Parameter Retrievals from L-band Microwave Observations Acquired over a Variety of Crop Fields," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1168-1178, Jun. 2004.

[26] P. F. Polatin, K. Sarabandi, and F. T. Ulaby, "An Iterative Inversion Algorithm with Application to the Polarimetric Radar Response of Vegetation Canopies," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 1, pp. 62-71, Jan. 1994.

[27] S.-B. Kim, M. Moghaddam, L. Tsang, M. Burgin, X. Xu, and E. G. Njoku, "Models of L-Band Radar Backscattering Coefficients Over Global Terrain for Soil Moisture Retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 2, pp. 1381-1396, Feb. 2014.

[28] A. Tabatabaeenejad and M. Moghaddam, "Inversion of Subsurface Properties of Layered Dielectric Structures With Random Slightly Rough Interfaces Using the Method

of Simulated Annealing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2035-2046, Jul. 2009.

[29] T. Le Toan, F. Ribbes, L. Wang, N. Floury, K. Ding, J. Kong, M. Fujita, and T Kurosu, "Rice Crop Mapping and Monitoring using ERS-1 Data Based on Experiment and Modeling Results," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 1, pp. 41-56, Jan. 1997.

[30] S. Park, S. Kweon, and Y. Oh, "Validity Regions of Soil Moisture Retrieval on the LAI $-\ \theta$ Plane for Agricultural Fields at L-, C-, and X-Bands," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 6, pp. 1195-1198, Jun. 2015.

[31] R. D. de Roo, Y. Du, F. T. Ulaby, and M. C. Dobson, "A Semi-Empirical Backscattering Model at L-band and C-band for a Soybean Canopy with Soil Moisture Inversion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 4, pp. 864-872, Apr. 2001.

[32] S. S. Saatchi and M. Moghaddam, "Estimation of Crown and Stem Water Content and Biomass of Boreal Forest using Polarimetric SAR Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 2, pp. 697-709, Mar. 2000.

[33] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced Spectral Classifiers for Hyperspectral Images: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 1, pp. 8-32, Mar. 2017.

[34] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep Learning-Based Classification of Hyperspectral Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094-2107, Jun. 2014.

[35] A. Plaza, J. A. Benediktsson, J. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, J. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent Advances in Techniques for Hyperspectral Image Processing," *Remote Sensing of Environment*, vol. 113, pp. S110-S122, Sep. 2009.

[36] L. E. Pierce, K. Sarabandi, and F. T. Ulaby, "Application of an artificial neural network in canopy scattering inversion," 1992 IEEE International Geoscience and Remote Sensing Symposium, Houston, TX, 1992, vol. 2, pp. 1067-1069.

[37] C. Notarnicola, F. Posa, and M. Angiulli, "Soil parameters retrieval from remotely sensed data: efficiency of neural network and Bayesian approaches," 2004 IEEE International Geoscience and Remote Sensing Symposium, Anchorage, AK, USA, 2004, vol. 7, pp. 4682-4685.

[38] S. Fukuda, R. Katagiri, and H. Hirosawa, "Unsupervised approach for polarimetric SAR image classification using support vector machines," 2002 IEEE International Geoscience and Remote Sensing Symposium, Toronto, Ont., Canada, 2002, vol. 5, pp. 2599-2601.

[39] S. Paloscia, P. Pampaloni, S. Pettinato, and E. Santi, “A Comparison of Algorithms for Retrieving Soil Moisture from ENVISAT/ASAR Images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 10, pp. 3274-3284, Oct. 2008.

[40] K. J. Bois, A. D. Benally, P. S. Nowak, and R. Zoughi, “Cure-State Monitoring and Water-to-Cement Ratio Determination of Fresh Portland Cement-Based Materials using Near-Field Microwave Techniques,” *IEEE Transactions on Instrumentation and Measurement*, vol. 47, no. 3, pp. 628-637, Jun. 1998.

[41] F. Ulaby and M. El-rayes, “Microwave Dielectric Spectrum of Vegetation - Part II: Dual-Dispersion Model,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-25, no. 5, pp. 550-557, Sep. 1987.

[42] F. Ulaby and E. Wilson, “Microwave Attenuation Properties of Vegetation Canopies,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-23, no. 5, pp. 746–753, Sep. 1985.

[43] N. R. Peplinski, F. T. Ulaby, and M. C. Dobson, “Dielectric Properties of Soils in the 0.3-1.3-GHz Range,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, no. 3, pp. 803-807, May 1995.

[44] Microwave Radar and Radiometric Remote Sensing. Ann Arbor: The University of Michigan Press, 2014.

[45] J. Caron, S. Bonin, S. Pepin, L. Kummer, C. Vanderleest, and W. L. Bland, “Determination of Irrigation Set Points for Cranberries from Soil- and Plant-Based Measurements,” *Canadian Journal of Soil Science*, vol. 96, no. 1, pp. 37-50, Mar. 2016.

[46] T. Hastie, R. Tibshirani, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, 2nd ed. New York, NY: Springer, 2009.

[47] E. Allwein, R. E. Schapire, Y. Singer, “Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers,” *Journal of Machine Learning Research*, vol. 1, pp. 113-141, 2000.

[48] Fürnkranz, Johannes, “Round Robin Classification,” *Journal of Machine Learning Research*, Vol. 2, 2002, pp. 721-747.

[49] H. Skriver, F. Mattia, G. Satalino, A. Balenzano, V. Pauwels, N. Verhoest, and M. Davidson, “Crop classification using short-revisit multitemporal SAR data,” IEEE J. Sel. Topics Earth Obs. Remote Sens., vol. 4, no. 2, pp. 423–431, Jun. 2011.

[50] T. L. Toan, F. Ribbes, L.-F. Wang, N. Floury, K.-H. Ding, J. A. Kong, M. Fujita, and T. Kurosu, “Rice crop mapping and monitoring using ERS- 1 data based on experiment and modeling results,” IEEE Trans. Geosci. Remote Sens., vol. 35, no. 1, pp. 41–56, Jan. 1997.

[51] S. Paloscia, E. Santi, G. Fontanelli, F. Montomoli, M. Brogioni, G. Macelloni, P. Pampaloni, and S. Pettinato, “The sensitivity of CosmoSkyMed backscatter to agricultural crop type and vegetation parameters,” IEEE Journal of Selected Topics In Applied Earth Observations and Remote Sensing, vol. 7, no. 7, pp. 2856-2868, July 2014.

[52] R. Prasad, “Retrieval of crop variables with field-based X-band microwave remote sensing of ladyfinger.” Adv. Space Res., vol. 43, pp. 1356–1363, 2009.

[53] D. Ward and P. Moghadam, “Synthetic Arabidopsis Dataset. v4, CSIRO. Data Collection.” https://doi.org/10.25919/5c36957c0af41, 2018.