UNIVERSITY OF CALIFORNIA
SANTA CRUZ

THREE CASE STUDIES IN QUANTITATIVE APPROACHES TO AGROECOSYSTEM MANAGEMENT

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ENVIRONMENTAL STUDIES

by

Graeme Joel Baird

March 2019

The Dissertation of Graeme Joel Baird is
approved:

_____
Professor Carol Shennan, chair

_____
Professor Madeleine Fairbairn

_____
Professor Kai Zhu

_____
Lori Kletzer
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures and Tables

Three Case Studies in Quantitative Approaches to Agroecosystem Management

Graeme Joel Baird

Abstract

Effective ecological management of agroecosystems for both productivity and sustainability is by design a messy and complex task, producing problems which benefit from highly data-focused analyses. Here, three case studies using quantitative approaches to these problems are presented. First, using results from soil nitrogen monitoring in a long-term organic vegetable/strawberry cropping dataset, an ensemble machine learning model and process model are contrasted and used to reveal key drivers of soil mineral nitrogen asynchrony and loss potentials. Environmental factors, nutrient inputs, and management practices interact to determine the magnitude of nitrogen mineralization, and key combinations of these factors, such as early- or late-season disturbance and irrigation, may increase the risk of generating loss-vulnerable pools. Second, a Bayesian network modeling approach is used to synthesize data across multiple lab and field experiments, using cross-experimental data in an integrative manner, furthering our understanding of treatment dynamics in anaerobic soil disinfestation, an ecological soil pathogen control method, and providing a step forward in recommendations to strawberry growers seeking to optimize implementation in their systems. A strong relationship between carbon inputs and soil temperatures suggests that growers may be able to ease environmental restraints with additional inputs during treatment applications. Finally, an

unsupervised cluster analysis is applied to a broad survey of on-farm management practices and approaches to disease control in walnut production, detecting two primary groups of divergent management practices. These groups, broadly characterized by moderate versus high levels of data and technology use, utilize markedly different approaches towards the integration of information and technology into on-farm management decisions.

This work could not have been without my mentors, colleagues, family, and friends, to whom I am indebted.

Learning ecological nitrogen management: model approaches to predicting nitrate asynchrony in organic vegetable/strawberry cropping systems

Introduction

As agricultural systems continue to develop in the 21st century, a variety of projections and critical objectives have been proposed to help guide researchers and land managers towards fruitful avenues of development. Of specific concern is agricultural intensification, the concomitant demands placed on system-level productivity, and how to best meet projected yield requirements while maintaining the integrity of peri-agricultural environmental, social, and economic systems. In particular, there is a continued need to tighten nutrient cycles within agroecosystems, with particular emphasis on the movement, use efficiencies, and losses of nitrogen (Gruber and Galloway 2008) and phosphorous (E. M. Bennett, Carpenter, and Caraco 2001).

Despite extensive documentation of impacts (Gruber and Galloway 2008, Howarth 2008, Robertson and Vitousek 2009) and contemporary research on improved management practices (Follett 2012), nitrogen (N) losses from agricultural fields remain one of the primary and intractable sources of N pollution to global ecosystems (Robertson and Vitousek 2009). Modes of off-farm transportation of agricultural N, very frequently in the same season as it was brought to the farm as synthetic N fertilizer, persist at relatively high levels with modern estimates of nitrogen use

efficiency (on-farm and harvest N retention versus total applied) remaining below 50% (Cassman, Dobermann, and Walters 2002). Many of these losses result from asynchrony between soil N availability, microbial N uptake, and plant N requirements, producing a concomitant accumulation of loss-vulnerable pools of soil inorganic N (Drinkwater and Snapp 2007). While these pools are particularly prominent in systems that maintain soil inorganic N saturation through applications of synthetic N fertilizers, excess N can exist in any soils where fertilizer inputs and net N mineralization outpace real-time crop N requirements.

Some success in reducing N asynchrony has been found in the use of model-calibrated fertilizer recommendations, which use estimations of underlying N transformations and plant uptake to specify fertilization rates, application timings, and irrigation management. However, these models tend to be targeted towards large-scale conventionally-managed cereal systems (Follett 2012) and their results may be poorly extendable to intensively managed vegetable crops with markedly different climactic conditions, crop phenologies, and management practices (Kersebaum 2007).

Regardless, while the effects of N losses on local and regional ecological processes are particularly prominent in areas with surface water flows and industrialized cereal cropping systems (e.g. domestically, the Mississippi River Basin (McIsaac et al. 2001), globally (Diaz and Rosenberg 2008)), vegetable cropping systems in the Central California Coast region (San Mateo, Santa Cruz, Monterey counties, hereafter "CCA") can lose a considerable amount of N to the environment, primarily through

leaching to shallow groundwater (leading to surface water pollution) and to deeper aquifers (Fogs, LaBolle, and Weissmann 1999), and as N2O emissions, a potent greenhouse gas (Harter et al. 2014).

Manipulation of management and input-timing is relatively understudied in systems which use ecological nitrogen management ("ENM") strategies. Agroecosystems utilizing ENM frameworks, most commonly organically managed systems, avoid the input of synthetic N fertilizers, instead relying on in-season mineralization of soil organic N, biomass fertilizers, and crop residues. Further, long-term management of soil microbial biomass and recalcitrant N pools are seen as a fundamental component of fertility management (Drinkwater and Snapp 2007).

Applications of predictive modeling to complex ecological management systems is appealing - with the caveat that existing N-management crop models are often calibrated towards managing N dynamics in systems reliant on synthetic fertilizers as the primary source of plant available N, where N transformations are largely subsumed by the magnitude of mineral N inputs. Organic vegetable cropping systems reliant on ecological nitrogen management (ENM) strategies depend on a more diverse and unstable set of immobilization-mineralization reactions.

While the underlying microbial transformations and soil-plant interactions which drive mineralization patterns from soil organic nitrogen (SON) pools are well-studied (Benbi and Richter 2002), predicting the relative contributions of management events and environmental factors in driving net N mineralization and immobilization

remains a daunting task, leaving CCA growers who wish to use ecological practices little scientific support on how best to manage system N.

Model approaches to ENM

One approach involves the use of process models, or mechanistic models, which seek to directly represent the underlying biogeochemistry of the modeled systems. These are typically continuous simulation models, generally with hard-coded parameters, that use time-step simulations of soil processes such as water transport and nutrient transformations, alongside models of plant growth. In theory, by simulating these processes to a certain degree of specificity, emergent and measurable phenomena can be adequately captured, even in systems which are dominated by biological N turnover processes (Kersebaum 2007, Giltrap, Li, and Saggar 2010).

Conversely, instead of emphasis on underlying biological processes an alternative approach is the use of data-driven prediction models, i.e. those derived from statistical and machine learning paradigms. These models consider the data features themselves to be the system under consideration. In some cases, domain-specific expert knowledge may be incorporated into the modeling procedures, for example in the construction and specification of a deep hierarchical Bayesian model, and in other cases the modeling specifications may be driven only by improving model performance on key evaluative measures (out-of-sample loss, specificity/sensitivity equilibria, etc.). In the latter category, modern nonparametric machine learning methods contain useful features such as robustness to parameter collinearity, high

degrees of predictive performance, and feature selection capacity during the model fitting process. A summary of the high-level differences between these two analyses paradigms are provided in Table 1-1.

Statistical/ML models (data-focused)	Mechanistic models (process-focused)
Foundation - Model or algorithm performance and robustness - Verification of distributional / inferential assumptions in data **Basis** - Fitted model or learned parameters describing data **Training** - Typically via an objective function, such as error or loss, which may be problem-specific according to outcome goals **Outcome** - Direct model interpretation, if possible, via learned structure - Forecasting and prediction from combination of model structure and training data **Dependencies** - Adequate vetting of model structure, assumptions, and bias - Capture or acknowledgement of systematic error in data-generating process - Domain knowledge generally required for inference - Protection against overfitting for generalizable results	**Foundation** - Fundamental characteristics of data-generating process - Verification of mechanistic assumptions of data **Basis** - Learned / parameterized equations describing process **Training** - Starting conditions, outcome boundaries, and parameters set on a system-specific basis when known, or best guess **Outcome** - Direct model interpretation sometimes possible via simulation - Forecasting and prediction using model structure and set parameters **Dependencies** - Adequate construction of mechanistic structure and assumptions - Process-based capture of systematic error in data-generating process - System-specific knowledge and empirical inputs for learning model parameters - Upstream models for imputing missing data

Table 1-1. General characteristics and differences between data-focused and process-focused modeling approaches.

DNDC

One process model is ``DeNitrification DeComposition" (DNDC), a simulation model which builds crop growth and nutrient movement patterns from underlying carbon and nitrogen biogeochemical modeled processes (Giltrap, Li, and Saggar 2010). DNDC has been parameterized for a wide variety of agroecosystems, including animal agriculture and perennial systems, contains simulation modules specifically calibrated to soil N immobilization-mineralization processes, and shows promise for adaptation to the intensive vegetable systems of the CCA. Importantly for this application, the DNDC model also provides functionality to directly simulate soil NO3 values.

The code for DNDC is closed-source, but the overall structure of hard-coded pathways and user-defined parameters is as described in Li 2009. Via a complex biogeochemical simulation framework, the DNDC model is capable of estimating daily pools of soil nitrate and is capable of linking these simulated nitrate pools to unique environmental and management scenarios.

Random forests

A data-driven machine learning model considered here is the random forest model (RF model). Random forests leverage a combination of bagging, random parameter

sampling, and an underlying 'base learner' model called "classification and regression trees" (CART), to generate outputs which, via weighted combinations, leverage the weak predictions from the base CART learners into a robust overall prediction.

CART models are non-parametric prediction models which use a method called binary recursive partitioning to structure data; by using binary subdivision to structure data into a series of nested forks and nodes, data series can be broken down into smaller series of binary prediction tasks. The algorithm is generally accomplished by searching through data features, testing a variety of binary split points for each feature, and choosing the feature-split combination that minimizes variance, impurity, or some other loss function. This procedure is then repeated until the loss function can no longer be minimized with additional splits, or when user-specified constraints are encountered.

The exact form of the underlying CART model can be calibrated to the prediction task at hand (i.e. by manipulation of the loss functions, manipulation of hyperparameters, modification of how the tree is built) and then used to build the overall ensemble model. (The 'depth' of each tree, or the number of levels the tree is allowed to grow until a terminal node, is often manipulated as a hyperparameter which controls overfitting.)

In particular, improvements can be made to the CART algorithm, and by extension the RF model predictions, by changing the methodology of how splits are generated.

The original formulation of CART (Breiman et al. 1984) expresses a general tendency to over-utilize covariates with many possible splits, biasing the model against fully utilizing the potential feature space, and a general tendency to overfit by generating splits with very small improvements over the loss function (although this behavior can be somewhat controlled with hyperparameter tuning).

The formulation of conditional inference trees, first proposed by (Hothorn, Hornik, and Zeileis 2006), posits a method to address both of these concerns, by modifying the binary split procedure and introducing a distributional test to evaluate whether a split is 'worth' making. By extension, conditional random forests leverage these improvements on the underlying CART model to create a more performant ensemble model, using the same strategy of bagging and boosting to maximize data utilization, and with conditional inference trees as the base learners.

Interpreting ML models

A major challenge in the use of machine learning prediction models in inferential tasks is translating their generalized predictive powers to more interpretive modes of use. This can sometimes be done via direct examination of how individual marginal effects or simple interactions of effects influence final predictive outcomes via examination of posterior predictions given simulated inputs, but such 'black box' models generally lack the interpretable ease of, say, the coefficients of a linear model.

This has overall led to calls for additional development of 'interpretable machine learning' methods (Doshi-Velez and Kim 2017) to facilitate both interpretation of

how and why predictions are made from machine learning models. The latter question of why predictions are structured in particular ways given particular datasets is essentially a reframing of an inferential problem, and answers to this question can be leveraged here to interpret the underlying mechanisms in our machine learning approach.

Shapley values

One method for interpretable models is an approach borrowed from coalitional game theory, the Shapley value, a model-agnostic explanatory value adapted from a 'player-payout' structure in theoretical games. The full treatment of how this value can be formulated is available here: (Molnar 2018); some key aspects of the methodology presented there are repeated below for clarity.

Given some arbitrary prediction function $\hat{f}(x_{i.})$, such as a linear model:

$$\hat{f}(x_{i1}, \dots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Where each x_{ij} is a sample at observation i and feature j, and each β are feature weights, we can consider the feature effect ϕ_{ij} of x_{ij} to be:

$$\phi_{ij}(\hat{f}) = \beta_j x_{ij} - \beta_j E(X_j)$$

This is the difference between the mean effect of this feature on all data and the effect of this feature on data point i. By iterating through all datapoints in the target dataset, a distribution of observation-specific feature effects can be assembled (unique to that

dataset), and by iterating through all features in the same way, each feature can be assessed for its relative contribution to the dataset predictions.

This general principle, here applied to a linear model, can be transferred to a model-agnostic form by the use of Shapley values, which consider each feature as a player in an overall game of predicting outcomes, and the contribution of each feature to that outcome is their 'payout', or Shapley value. By computing the distributions of values for each observation-feature set, we can obtain average and marginal values appropriate for analysis in a form similar to $\phi_{ij}(\hat{f})$ (Štrumbelj and Kononenko 2014). Importantly, the above form illustrates the interpretive meaning of ϕ_{ij} - for every observation i, ϕ_{ij} represents the amount that feature j 'pulls' the prediction away from the global mean (whether in large or small magnitude), and thus can decompose a complex non-linear model into intepretable, observation-specific chunks.

The utility of this calculation can be explored via plotting simulated effects and Shapley values (Figure 1-1). Four scenarios are plotted: a positive linear relationship, negative linear relationship, nonlinear relationship, and no consistent effect relationship. For interpretability in visualization, the observed values are scaled to range between -1 and 1.

The linear relationships are straightforward to interpret (Figure 1-1, A B) - each simulated pair of Shapley value and observed value are plotted via the x-axis position (Shapley value) and point color (scaled observed value). The observed value at each point pushes the model prediction in the direction and magnitude indicated by the

corresponding Shapley value. The nonlinear relationship follows a similar pattern (Figure 1-1, C), but switches direction in the middle range of observed values. Finding this relationship would indicate that observed values at either high/low extremity did not influence model predictions, but values in the middle of the observed distribution did. Finally, the distribution of Shapley and observed values if there is no clear effect manifests as a generally mixed distribution around 0 (Figure 1-1 D).

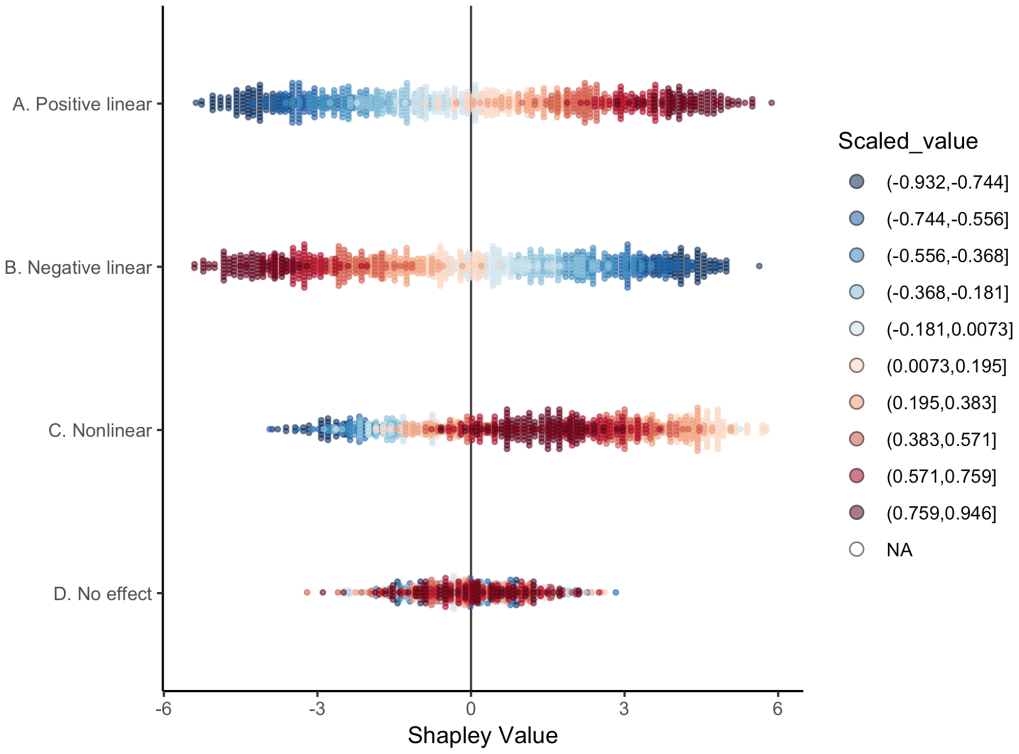


Figure 1-1. Example relationships represented by a Shapley value plot. Simulated Shapley values (x-axis) for each relationship (y-axis) are plotted with simulated underlying scaled parameter values (point color).

Study goals

This work approaches the problem of understanding N asynchrony in mixed vegetable cropping systems using these two linked methods. First, using a multi-year, multi-crop ENM dataset, we evaluate the outcomes of the DNDC model to determine 1) which components of the N cycle are estimated as most important to overall dynamics, with specific attention to processes generally considered as important to ENM (namely, N release from OM pools), and 2) how well process-based estimates match observed soil mineral N levels. This serves as the "bottom-up" approach to N dynamics in ENM, building from N cycle processes to outcomes.

Second, using the same dataset, we apply a purely machine learning approach via a conditional random forest model. Model structure analysis and simulation provides inference into the primary drivers of soil N dynamics in this regionally-specific dataset. This serves as the "top-down" counterpart approach, taking a process-naïve analysis structure to infer environmental and management features critical to N asynchrony in ENM systems.

Methods

Field experimental design - treatment and spatial layout

The Mother-baby trials are a set of field experiments composed of two linked components: the Mother trial, a replicated field experiment conducted on the UCSC farm, and the Baby trials, a set of smaller unreplicated treatments exported to on-farm

trials. Both trials are designed to simulate actual cohorts of on-farm practices, combined in ways that constitute overall management strategies. Only the data from the Mother trial is analyzed here.

The mother trial is a split-split plot design experiment which varies treatments by three levels: (1) length of rotation: 2 years vs 4 years, (2) crops in rotation: a broccoli-dominant sequence vs a lettuce-dominant sequence, (3) 4 levels of combined disease/fertility treatments: no-fertilizer, no-fertilizer, mustard seed meal, and compost/feather-meal additions. Paired with these treatments are winter-cropping regimes that vary which species are planted as cover crop over winter, respectively: bare fallow, mixed-legume cover crop (a 45%/45%/10% mixture of bell bean *Vicia faba*, hairy vetch *Vicia villosa*, and cereal rye *Secale cereale*), cereal rye cover crop (100% *Secale cereale*), and mixed-legume cover crop (same mixture proportions). The overall layout of the management and treatment schedule is illustrated in Table 1-2 and 1-3, as well as the correspondence between schedules and treatment letterings, which are used later in the paper for comparisons of error.

	Year 1, 5			**Year 2, 6**		
Treatment	Fall 2011	Winter	Summer 2012	Fall 2012	Winter	Summer 2013
1a	cc	cc	Broccoli	cc	cc	Lettuce / Cauliflower
2a	cc	cc+c+f	Broccoli	cc	cc+c+f	Lettuce / Cauliflower
3a	rcc	rcc+mc	Broccoli	rcc	rcc+mc	Lettuce / Cauliflower
4a	bf	bf	Broccoli	bf	bf	Lettuce / Cauliflower
5a	cc	cc	Lettuce	cc	cc	Broccoli
6a	cc	cc+c+f	Lettuce	cc	cc+c+f	Broccoli
7a	rcc	cc+mc	Lettuce	cc	cc+mc	Broccoli
8a	bf	bf	Lettuce	bf	bf	Broccoli
1b	cc	cc	Broccoli	asd	Strawb +	Strawb +
2b	cc	cc+c+f	Broccoli	asd + c	Strawb +	Strawb +
3b	rcc	cc+mc	Broccoli	mc	Strawb +	Strawb +
4b	bf	bf	Broccoli		Strawb	Strawb
5b	cc	cc	Lettuce	asd	Strawb +	Strawb +
6b	cc	cc+c+f	Lettuce	asd + c	Strawb +	Strawb +
7b	rcc	cc+mc	Lettuce	mc	Strawb +	Strawb +
8b	bf	bf	Lettuce		Strawb	Strawb

Table 1-2. Crop rotation and management treatments. cc=legume/cereal cover crop, rcc=rye covercrop, bf=bare fallow, mc=mustard seed meal, f=fertilizer, Strawb=strawberry, Strawb+= strawberry+ fertigation. Treatments 1a to 8a are 4 year rotations, and 1b to 8b are 2 year rotations.

	Year 3			Year 4		
Treatment	Fall 2013	Winter	Summer 2014	Fall 2014	Winter	Summer 2015
1a	cc	cc	Broccoli	asd	Strawb +	Strawb +
2a	cc	cc+c+f	Broccoli	asd + c	Strawb +	Strawb +
3a	rcc	rcc+mc	Broccoli	mc	Strawb +	Strawb +
4a	bf	bf	Broccoli		Strawb	Strawb
5a	cc	cc	Lettuce	asd	Strawb +	Strawb +
6a	cc	cc+c+f	Lettuce	asd + c	Strawb +	Strawb +
7a	rcc	cc+mc	Lettuce	mc	Strawb +	Strawb +
8a	bf	bf	Lettuce		Strawb	Strawb
1b	cc	cc	Broccoli	asd	Strawb +	Strawb +
2b	cc	cc+c+f	Broccoli	asd + c	Strawb +	Strawb +
3b	rcc	rcc+mc	Broccoli	mc	Strawb +	Strawb +
4b	bf	bf	Broccoli		Strawb	Strawb
5b	cc	cc	Lettuce	asd	Strawb +	Strawb +
6b	cc	cc+c+f	Lettuce	asd + c	Strawb +	Strawb +
7b	rcc	rcc+mc	Lettuce	mc	Strawb +	Strawb +
8b	bf	bf	Lettuce		Strawb	Strawb

Table 1-3. rop rotation and management treatments. cc=legume/cereal cover crop, rcc=rye covercrop, bf=bare fallow, mc=mustard seed meal, f=fertilizer, asd+c= asd+compost, Strawb=strawberry, Strawb += strawberry + fertigation. Treatments 1a to 8a are 4 year rotations, and 1b to 8b are 2 year rotations.

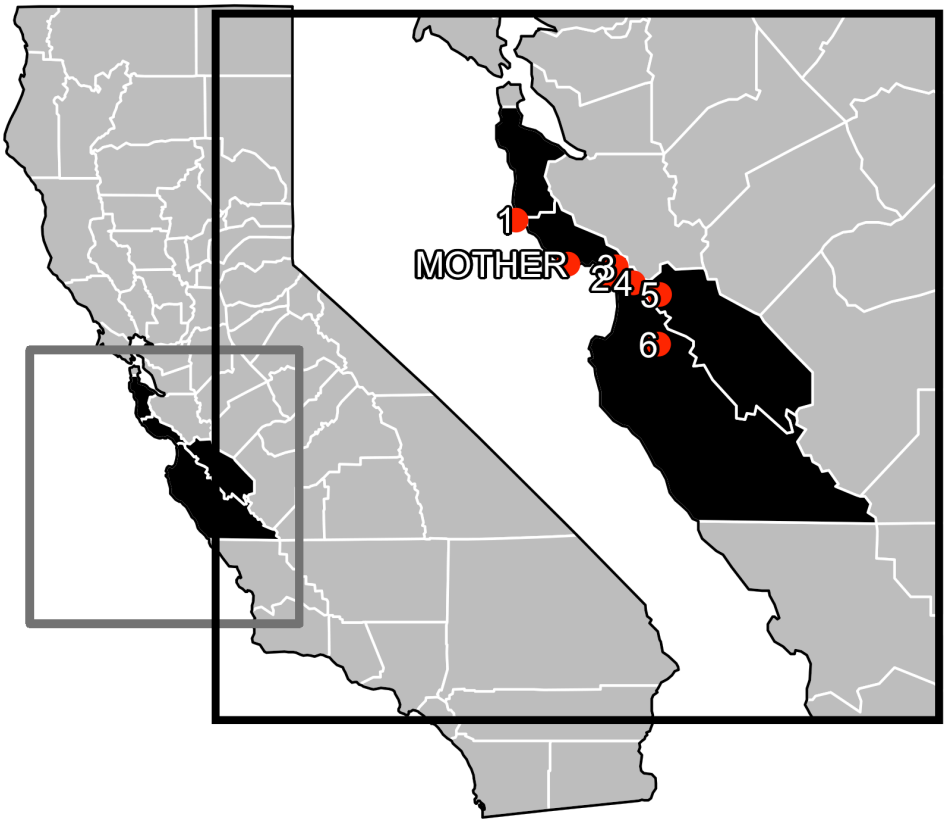


Figure 1-2. Illustration of Mother-Baby Trial field sites. Within the state of California, USA, four counties, San Mateo, Santa Cruz, San Benito, and Monterey counties are highlighted, and approximate field trial locations within these counties are provided by red points and white labels.

The soil, environmental, and management data from the Mother trial is not analyzed as a traditional treatment framework, but instead is approached as a dataset of observations with decomposed individual parameters (precipitation, fertilization, tillage, etc.) associated with each data point. When appropriate, model predictions

from the decomposed dataset are pooled together back into assemblages of management systems via treatment labels.

Data collection

Soil inorganic N

Soil inorganic N levels (NO3, NH4) were monitored via direct sampling (see Appendix 2 for sampling dates). Soil samples were taken from each plot at two soil depths (0-15cm, 15-30cm) using an 3cm internal diameter soil probe, and field extracted or stored under refrigeration until lab extraction. Extraction was accomplished by placing approximately 5 grams of soil into 25ml of 2M KCl, followed by 30 minutes of agitation on a table shaker, filtration through Whatman Grade SA 720 type filter papers, and the resulting solute stored in a freezer until analysis. Process blank samples were taken for each sampling date to monitor sources of contamination during the sampling and filtration process. KCl extracted samples were analyzed for NO3- and NH3+ using an automated colorimetric flow injection analysis technique (Lachat FIA 8500). At each sampling date, moisture samples were obtained for gravimetric moisture analysis and back-conversion of inorganic-N levels to a dry soil basis (mg NO3 / kg dry soil). Gravimetric moisture analysis was conducted by weighing samples before and after 24 hours in an oven set to 105C.

Fertility input rates and CN

Fertility inputs were evaluated for total carbon and nitrogen content via combustion analysis. Samples were taken from all fertility inputs prior to application. Samples were then oven dried to remove all moisture, ground and homogenized to pass a .5mm sieve, and analyzed in-house using a combustion-based gas analysis CNS protocol. Dry-weight amendment rates and CN values were then used to calculate the absolute contribution of N g / kg amendment and C g / kg amendment. These values form the basis of estimating kg/ha carbon and nitrogen input rates.

Cover crop biomass rates and CN

Mineralization effects from decomposing crop residues and cover crop biomass were estimated by direct measurement of biomass and CN content. Crop residues were extrapolated from measured wet weights during crop harvest to generate estimations of kg / ha residue biomass remaining in the field after harvest. Sub-samples were retained for laboratory CN analysis. Cover crop biomass rates were extrapolated from biomass sampling using two .25 m^2 quadrats per plot, separation of biomass by plant type, and subsequent wet weighing. Sub-samples from each plant type were retained for laboratory CN analysis. Both the crop residue CN samples and cover crop CN samples were then oven dried to remove all moisture, ground and homogenized to pass a .5mm sieve, and analyzed in-house using a combustion-based gas analysis CNS protocol. Dry-weight biomass and CN values were then used to calculate the

absolute contribution of N g / kg biomass and C g / kg biomass, and subsequently extrapolated to a kg / ha basis.

Yield biomass rates and CN

Yields from all plots were measured via direct sampling of marketable yield rates, averaged over two harvest events for vegetable crops, and cumulatively sampled over season-wide harvest events for berry crops. Sub-samples of biomass from all yield events were retained for moisture and CN analysis. Yield values were then used to parameterize biomass production processes in the DNDC model.

Environmental monitoring

Environmental conditions such as precipitation, temperature, soil temperature, and ETo were collected from a combination of data streams from an on-farm weather station and nearby CIMIS weather stations (https://cimis.water.ca.gov/). Data for all parameters were collected on an hourly basis and missing values were imputed using nearest-neighbor averaging.

Management events

Management events such as planting dates, incorporation, tillage, and irrigation were tracked via maintenance of field records by field managers. For irrigation data, date of application, application type (sprinkler/drip) and irrigation amount (cm) were retained. Incorporation and tillage data were tracked by date, type, and depth of soil disturbance.

Windowing functions

Environmental parameters temperature, soil temperature, and ETo were transformed with mean-value windowing functions from 0 to 20 days prior to the date of observation, i.e. “soil temperature 0” as the parameter value on the day of observation, and “soil temperature -20” as the average parameter value over the 20 days prior to observation. Precipitation and irrigation were transformed using cumulative windowing functions from 0 to 20 days prior to the date of observation, using a sum of all observations in the window period. Fertilization C and N rates were transformed via cumulative windowing functions 90 and 180 days before observation.

Final data matrix

After collection of all associated data and transformed parameters, the final data matrix of 3117 date-plot unique observations and 120 parameters was used for analysis. All but one parameter (crop type, factorial) were treated as numeric or integer types.

Model fitting - DNDC

As the DNDC model operates on daily simulations, the underlying dataset was first converted into a compliant format. Weather data (precipitation, maximum/minimum daily temperatures), fertilizer data (kg/ha C and N), tillage data (date, depth, type), planting/harvest data (crop type, planting/harvest dates), irrigation data (date, depth, type) were used as direct inputs to the model.

Crop growth parameters were adjusted to reflect observed crop biomass characteristics and biomass distribution ratios (sampled via "yield biomass rates"), as well as adjustments to nitrogen fixation, nutrient uptake, and water uptake parameters.

The DNDC simulations were conducted using DNDC version 9.5, published by the University of New Hampshire and available for download at http://www.dndc.sr.unh.edu/. Simulations were run in 'site' mode, using input .dnd and .txt files that are stored in this repository and which contain the entirety of parameters and data necessary to reproduce these simulations.

After simulation over the entire observation window in the input data, simulated NO3 values were extracted and matched with observation dates from the original dataset for further analysis.

Model fitting - conditional random forest

An RF model was fit using the cforest function from the partykit package in R Statistical Software (Team 2013). The randomly preselected input features for each underlying CART learner model was set at 11, the approximate square root of the input parameter count, a standard baseline. The number of trees was evaluated by searching over the hyperparameter space between 250 and 1500 trees until RMSE stabilized, and the lowest number of trees within this space was kept (1000). Underlying CART learners were left unpruned and allowed to grow to maximum depth. Model fit results are presented in (Appendix 6).

Following the initial model fit, variable importance scores were calculated to estimate the relative contributions of individual parameters to overall predictive capacity. These scores were calculated with the conditional permutation importance framework outlined by (Strobl et al. 2008), proposed as an extension of Breiman's original measure of variable importance (Breiman 2001) as more robust to detecting variable importance within correlated parameters.

To calculate conditional permutation importance, out-of-bag error was first calculated for the entire ensemble of parameters, such that the prediction error from the non-bootstrapped data for each tree is used to generate an overall calculation of error. Rather than permuting the values of each parameter independently and sequentially, a grid of permuted parameter values with conditional dependencies is used to re-calculate OOB error, and the difference between the original intact OOB error and the permuted OOB error is used to assign a variable importance score to each parameter (Strobl et al. 2008). These scores were calculated using the varimp function from the partykit package in R.

Following the initial model fit and variable importance score calculation, prior knowledge on the collinearity of windowed environmental parameters was used to exclude all but the best-performing parameter from each category. This procedure was used to improve final model parsimony and interpretability.

RF posterior predictions

From the RF model, predictions on new data can be generated fairly simply via a combination of input data x, fitted weighting functions w, and original data Y_i (Meinshausen 2006), i.e.:

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

Posterior predictive checks and simulated management scenarios from the final RF model were generated via the formula above, via iteration through original data or simulation data (Appendix 7) using the party_predict method from the partykit package in R.

Shapley values

Shapley values were constructed from the selected conditional random forest model using the Shapley function in the shapleyR package. This package implements calculation of Shapley values using the algorithm as described in (Štrumbelj and Kononenko 2014) and partially detailed above.

Model comparison

To facilitate comparison between models, data were clustered based on treatments and predictive outcomes evaluated on a treatment-date basis, as the DNDC model did not incorporate plot-specific effects. To implement this, prediction error scoring was calculating using root mean square error within each treatment-date cluster, where the

loss for each replicated observation $i,\ i = \{1,2,3,4\}$ given model $f(.)$, data x_i, and real outcome y_i is calculated as:

$$\mathcal{L}_{RMSE}(x, y) = \sqrt{\frac{1}{4}\sum_{i=1}^{\{1,2,3,4\}}(f(x_i) - y_i)^2}$$

Calculation of RMSE for each cluster provides an estimate of date-treatment-specific error, which is then used to evaluate changes in error between models, management regimes (treatments), and seasonal heteroskedasticity.

Results

DNDC model results

While the DNDC model did not effective capture the volatility or overall seasonal patterns in the NO3 data, synchrony between estimated local hotspots of NO3 availability suggests that the underlying mineralization simulations tying specific management events, such as soil disturbance, fertilization, or changes in soil moisture, are at least somewhat grounded in reality (Figure 1-3).

Large, persistent simulated peaks in soil NO3 were predicted by the DNDC model but not supported by data, and appear to be primarily driven by simulated fertilization events in the mid- and high- fertility treatment plots. In contrast, mineralization patterns from the unfertilized treatments were almost entirely uncaptured.

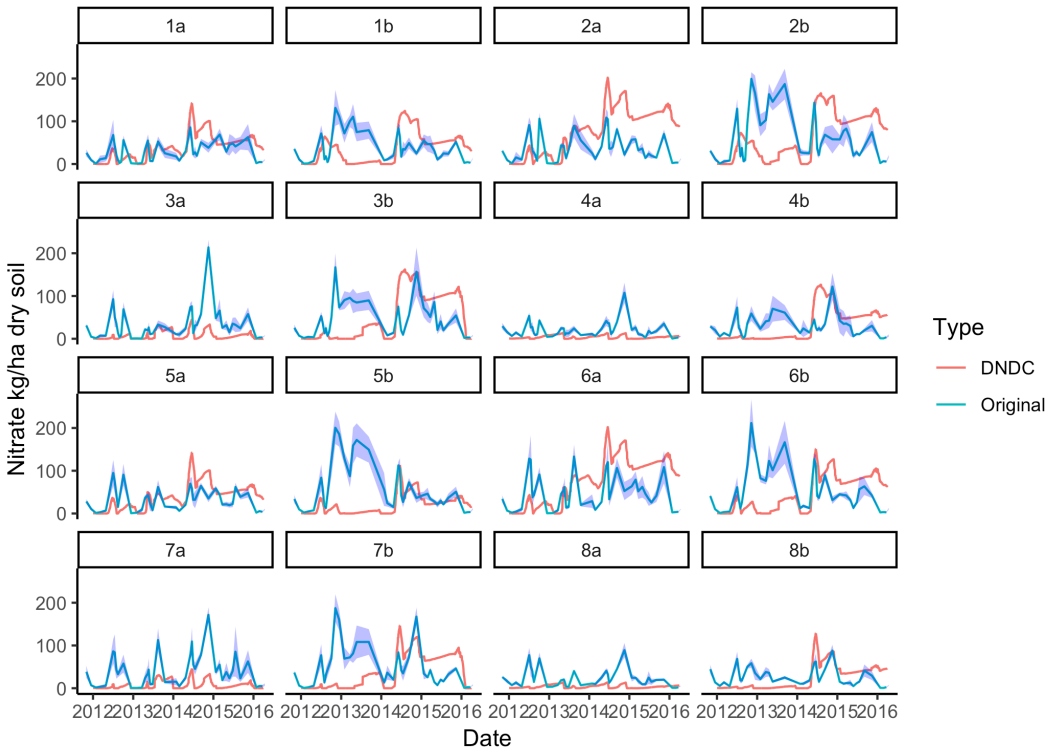


Figure 1-3. Soil nitrate levels, measured in NO3 kg /ha. Observed mean values across replicates per date-observation are plotted as blue lines, with blue shading representing standard error around the mean. DNDC values are directly simulated using treatment-specific data and date-treatment point outcomes are plotted as red lines.

RF model variable importance

In the initial model fit to all possible parameters, variable importance via permutation indicates an overwhelming dominance of environmental factors as important predictors, second only to crop type (Appendix 4). Dominant environmental predictors include soil moisture, baseline soil organic N, air temperature (20 day average), soil temperature (20 day average), ETo (4 day average), precipitation (4 day

cumulative), and irrigation (10 day cumulative). Days since planting, cultivation, and incorporation rank highly, as well as average nitrogen inputs, cover crop nitrogen inputs, and total contributions of carbon from fertilizers over a window of 90 and 180 days (Appendix 4). Ranking of relative performance remained stable after re-fitting the RF model with the selected subset of parameters and re-calculation of the variable importance score (Figure 1-4).

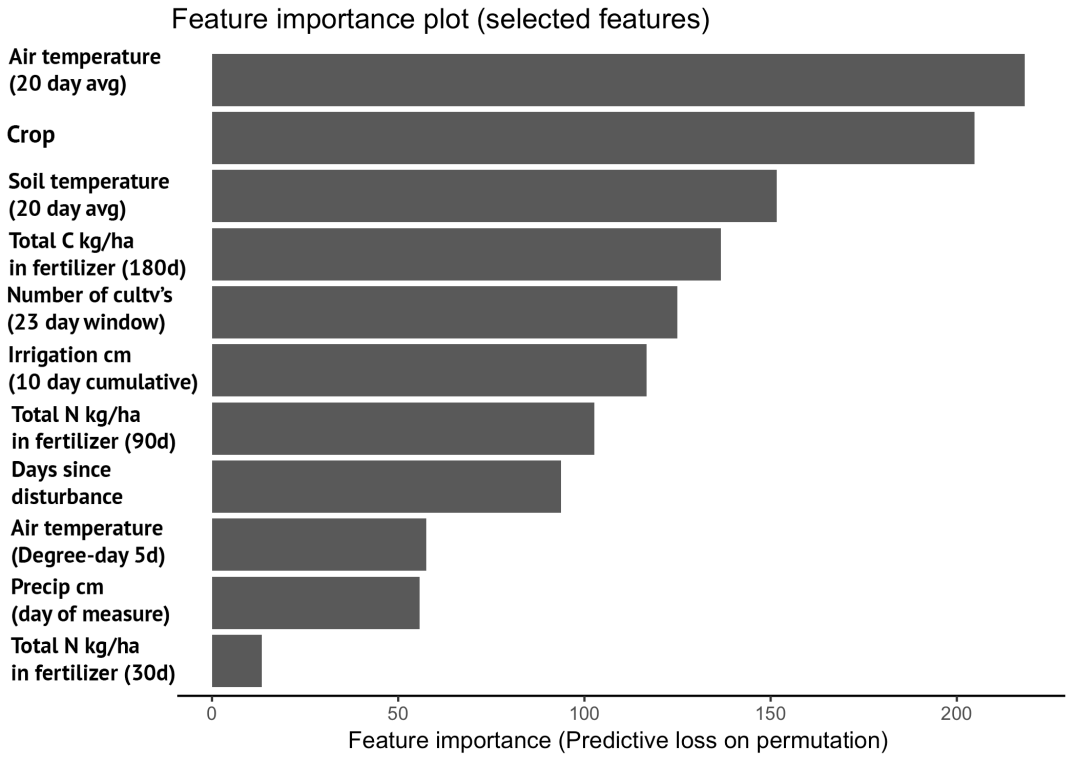


Figure 1-4. Feature importances, measured as error loss on permutation, for the restricted set of parameters after elimination of lower-performing transformed parameters from the tested feature set. Higher values indicate a greater importance of the parameters to the overall model.

RF model residuals and posterior predictive

Posterior predictions from the RF model display a good fidelity to original data, and generally capture critical peaks. Overall error is concentrated around NO3 peaks, with posterior predictions underestimating the magnitude of mineralization peaks (Figure 1-5).

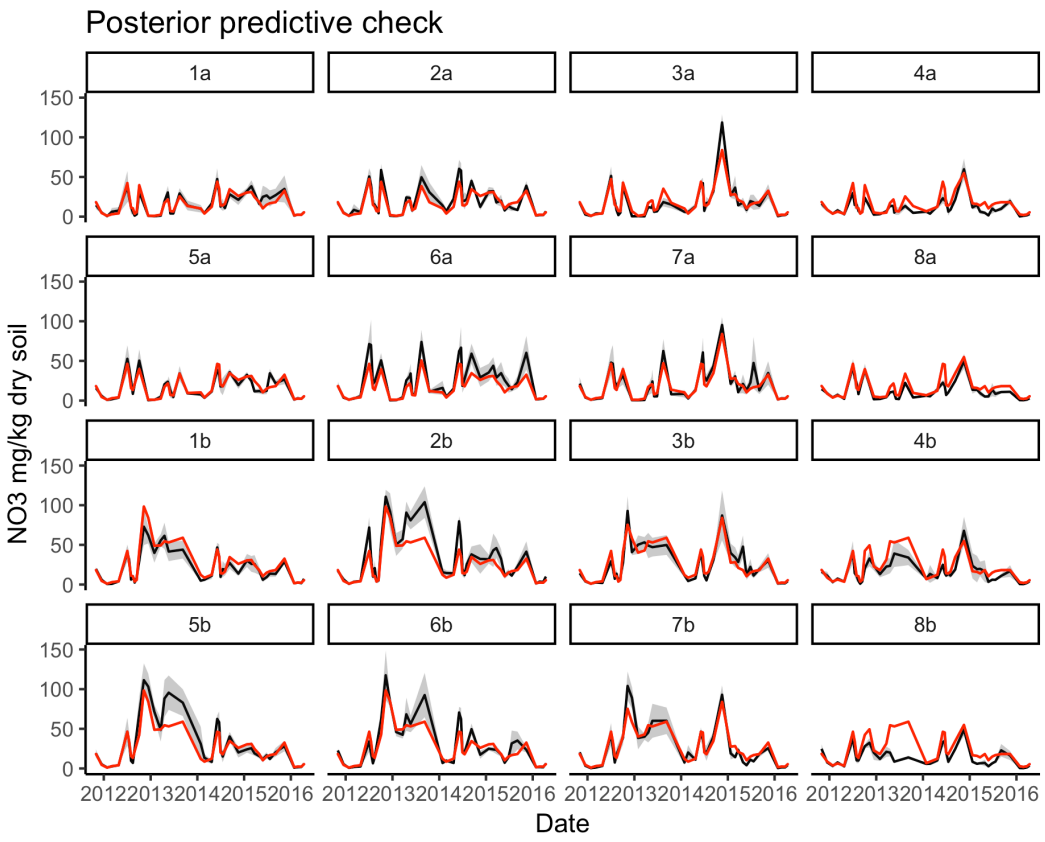


Figure 1-5. Soil nitrate levels and RF model projections, measured in NO3 mg / kg dry soil. Observed mean values are across replicates per date-observation are plotted as black lines with black shading representing standard error around the mean. RF values are simulated on a plot-specific basis, averaged across replicates per date-

observation, and plotted as red lines with red shading representing standard error around the mean.

Shapley values

The overall distributions of Shapley values are, with the exception of crop type, heavily distributed around 0, indicating the extreme nonlinearity of variable interactions in the modeled system - on an observation-by-observation basis, the effect of individual parameters can vary widely from being strongly positive to strongly negative. This finding, in combination with the overall good fit of the RF model to observed data, suggests that successful observation-specific predictions of soil NO3 is possible by complex combinations of underlying covariates (Figures 2-6, 2-7).

In particular, the importance of the only factor covariate in the model (crop type), both via the variable importance measure and the observed extremities of Shapley values, suggests there may by carry-forward effects within the CART models within crop types. This postulate is confirmed by individual examination of crop-specific Shapley values (Figures 2-6, 2-7), where differentiation in values across the 0-axis emerges, indicating crop-specific segregation of effects.

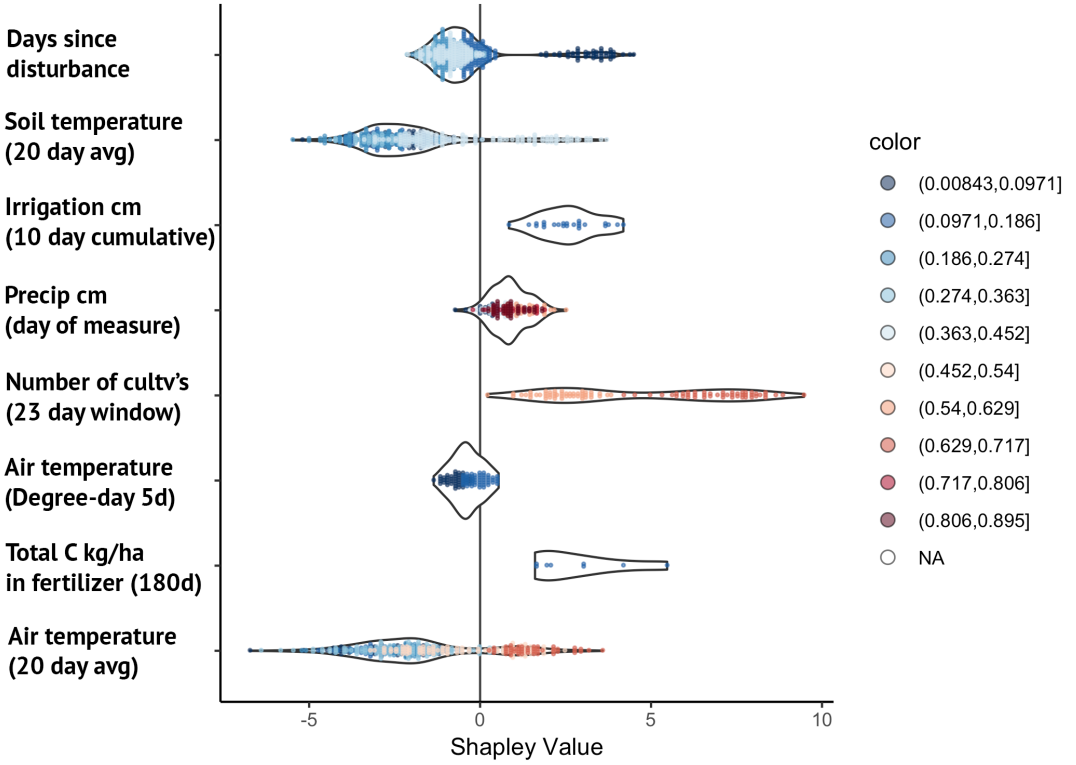


Figure 1-6. Shapley values for parameters in the final RF model (cover crop and fallow).

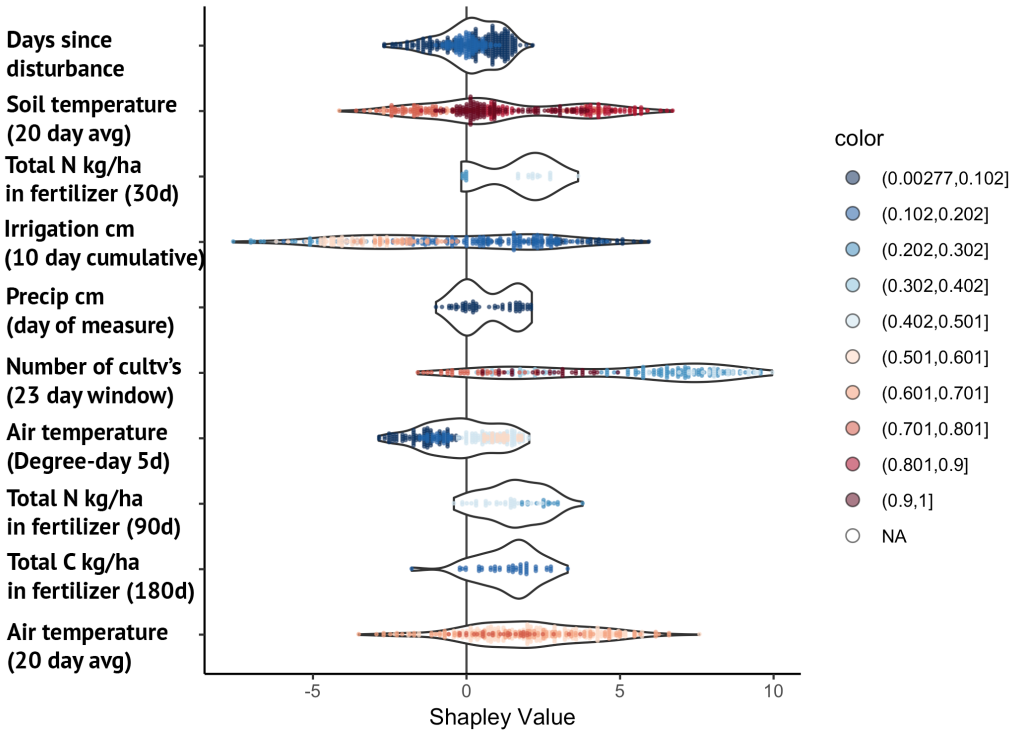


Figure 1-7. Shapley values for parameters in the final RF model (lettuce and broccoli)

Comparing DNDC and RF models

The random forest model far outperforms the DNDC model in predicting date-specific levels of soil nitrate pools, despite a considerably weaker assembly of underlying features and autoregressive properties (Figure 1-8).

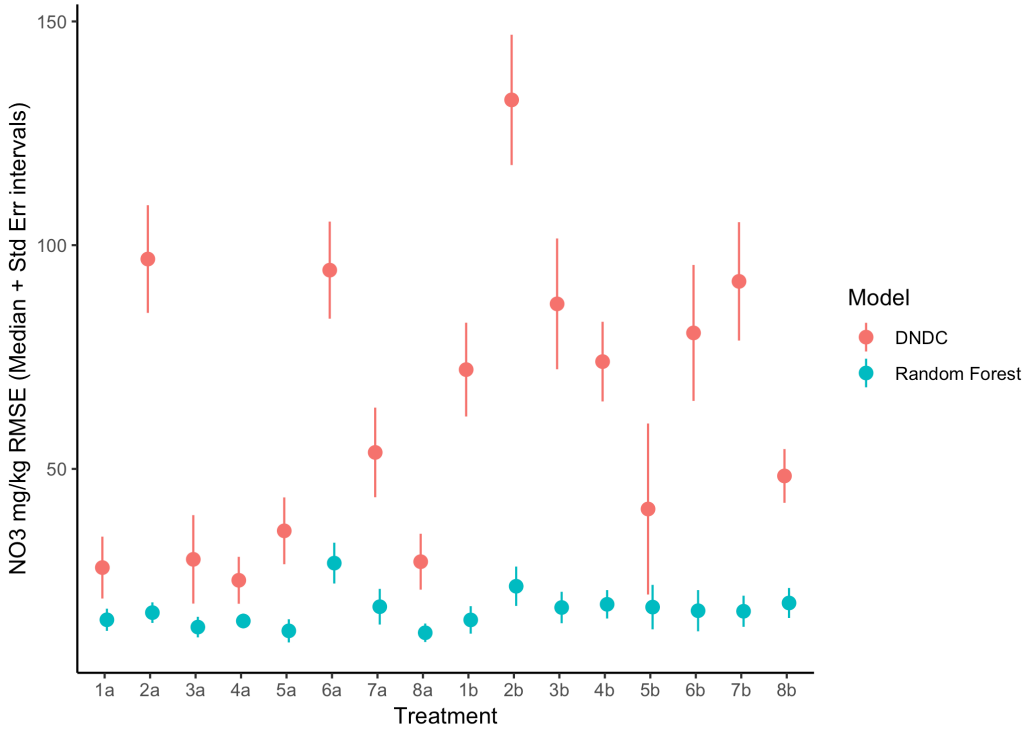


Figure 1-8. Date-observation RMSE values calculated for DNDC and RF models, clustered by treatment. RMSE values from DNDC are back-transformed to be on a mg NO3 / kg dry soil basis.

Period-specific predictions

Calculation of Shapley values for period-specific data provides an insight into the drivers of asynchrony in post-incorporation data. Subsetting all data to be bounded within 30 days of crop incorporation and no more than 15 days after planting provides a restricted set of values (Figure 1-9).

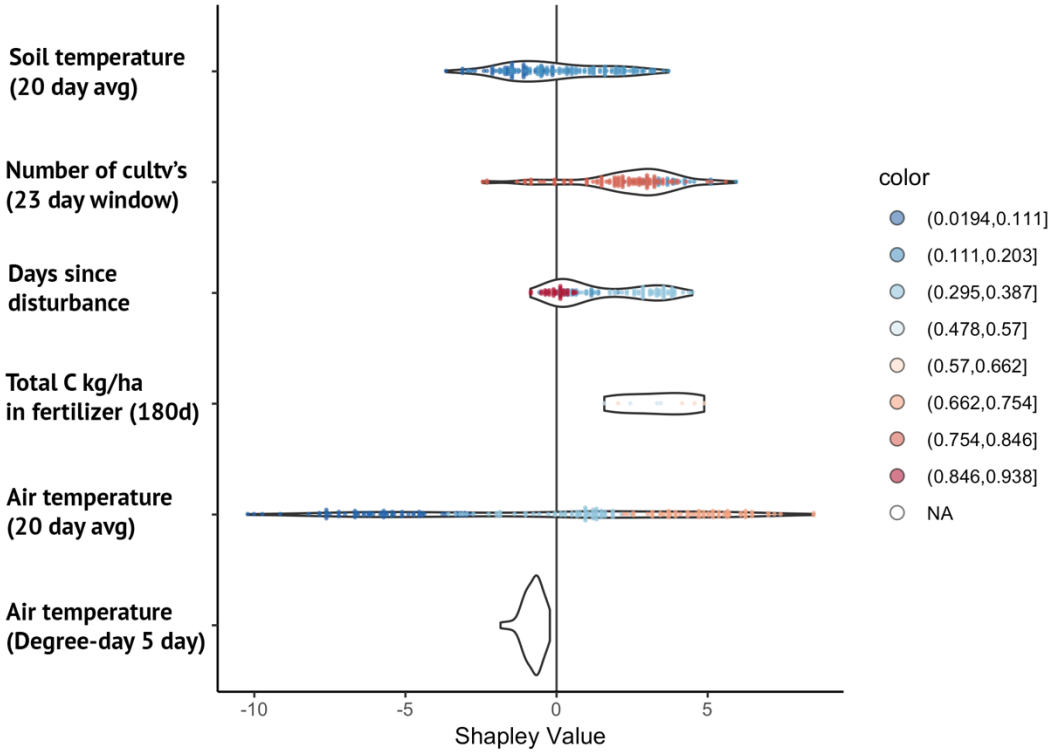


Figure 1-9. Shapley values calculated for observations within 30 days of crop incorporation.

Discussion

DNDC vs RF

A great deal of the error in the DNDC model appears to be attributable to a general underestimation of two factors: first, the magnitude of volatility in the real system is far greater than model estimates. Second, the year-over-year losses of nitrate from the system appear to be grossly underestimated, with almost no instances of movement from high nitrate levels to undetectable levels, even under winter precipitation regimes that would almost certainly leach any free nitrates from the 0-10cm profile

under consideration. In fallow, unfertilized no-cover-crop treatments (4a,8a,4b,8b), where N dynamics are presumably entirely driven by mineralization-immobilization patterns from pools of soil organic N, nearly all peaks of NO3 availability were completely absent from simulated data (Figure 1-10). A notable exception is the N peaks produced during the strawberry crop in treatments 4b and 8b (Figure 1-10, panels 4b/8b), which were generally predicted by the DNDC simulation, but subsequent uptake/losses of NO3 were not adequately modeled.

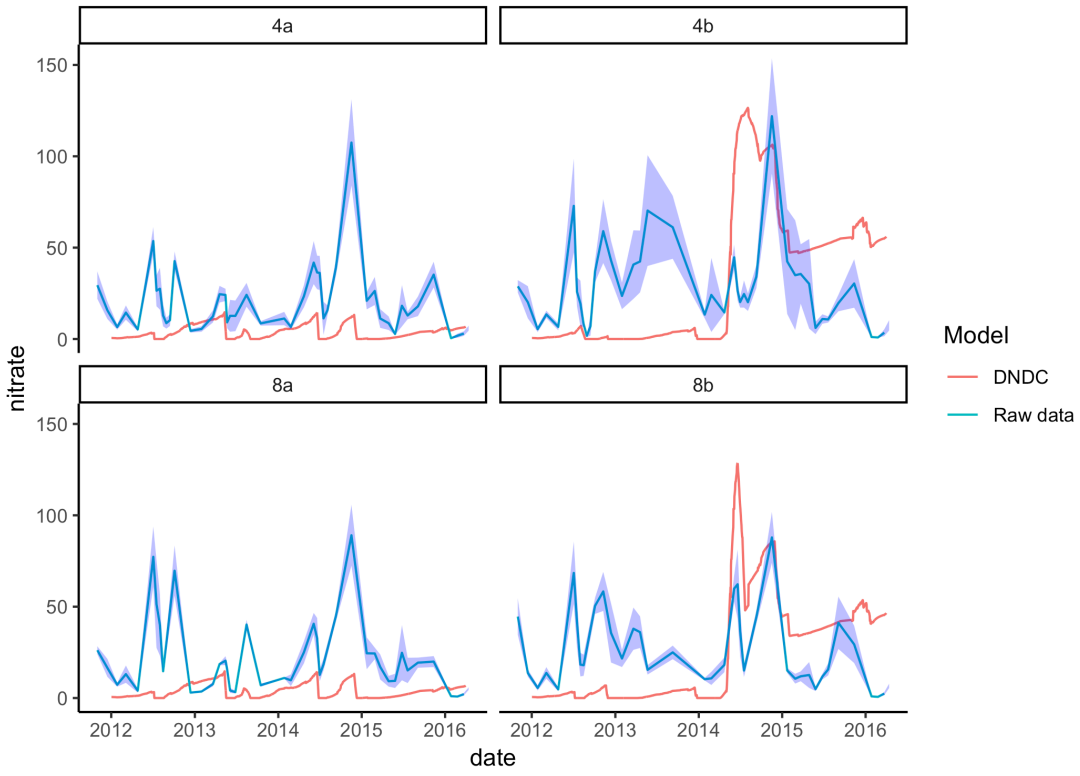


Figure 1-10. Soil nitrate levels and DNDC model projections, measured in NO3 mg / kg dry soil, for treatments without winter crops or any applied fertilizers. Observed mean values across replicates per date-observation are plotted as blue lines, with blue shading representing standard error around the mean. DNDC values are directly

simulated using treatment-specific data and date-treatment point outcomes are plotted as red lines.

While this could be an issue with insufficient or incorrect parameterization of the underlying process model, it remains notable that despite extensive tuning using data relatively unusual in specificity, DNDC predictions remain somewhat inaccurate. It may be the case that for prediction of soil nitrate pools themselves, underlying model processes are relatively underdeveloped, as the primary development goals of this model were for estimations of gaseous losses from agroecosystems.

In contrast, posterior predictive estimations from the random forest model indicate good fit to the real data, both in general fit to the time-series pattern (Figure 1-10) and in overall RMSE (Figure 1-8). This, despite the markedly aperiodic form of the nitrate data, suggests that the time-series can be successfully decomposed into functions associated covariates at each individual date, with little loss of predictive power from an absence of an autoregressive term. This is in agreement with (Finney, Eckert, and Kaye 2015), who successfully used a similar approach of parameter transformation, CART models, and RF models to analyze longitudinal soil nitrate data.

While autoregressive features may be important to effective simulation in process models such as DNDC, their absence is a benefit to the overall interpretability and potential applications of this model to ENM. With no prior knowledge of the nitrate levels in a soil at some arbitrary time in the past, environmental covariates can be

assembled to provide a reasonable prediction of the magnitude of NO3 concentration in the soil, and a decision theoretic framework can then take over.

Interpreting RF results

In the RF model, the pronounced role of environmental factors such as temperature, precipitation, and evapotranspiration in producing a close fit to real-world data illustrate the importance of key periods in the cropping cycle to the development of microbial activity and resultant mineralization patterns, and agree with prior work describing these factors as critical drivers in ENM systems (Drinkwater and Snapp 2007).

The strongly positive effect size and overall frequent inclusion of cultivation timing lends further support to suggestions that mineralization patterns can be deliberately stimulated via soil disturbance processes, whether via breakdown of soil aggregates, introduction of oxygen into soil environments, or mixing of decomposing materials (Booth, Stark, and Rastetter 2005).

Management recommendations

This finding, together with the evidently powerful role of soil moisture and irrigation on N pools, further suggests that a combination of soil temperature manipulation, tillage practices, and irrigation, can be used to provoke transformations of soil organic N reserves when plant N uptake may be expected, especially in situations where may

wish to reserve N contributions from incorporated cover crops until late-season plantings, as has been previously noted (Kaye and Quemada 2017).

In complement, these findings suggest that inappropriately timed tillage and irrigation may pronounce N mineralization beyond levels that early-season plants, with low N requirements and underdeveloped root systems, may be able to uptake. In this scenario, or in situations where post-incorporation soils are simply fallow, these standing pools, while also capable of immobilization transformations back into organic N, are the canonical sources of N pollution via water and gas losses (Schimel and Bennett 2004).

Model-agnostic forms of interpreting machine learning models such as Shapley values provide a novel path for researchers interested in both inference and prediction from their datasets to produce actionable insights, and the potential to approach datasets that were previously considered intractable or necessitated excessive simplification before they could be analyzed.

In particular, an increasing number of tools are available for growers to use in their farm nitrogen management protocols, with offerings from both the public and private sphere. Both the calibration of these tools with grower-provided information and the production of mineralization or N pathway predictions rely on appropriate model structures. The results from this study suggest both that (1) some combination of process and machine learning models may provide the best site- and data-specific but biologically-grounded results, and (2) if these services are provided to fields, growers,

or management systems with a particular emphasis on the provision of nitrogen via ENM or organic matter more generally, careful attention should be paid to the inclusion and calibration of OM / SOM mineralization pathways.

Cross-experimental synthesis to determine environmental and management drivers in Anaerobic Soil Disinfestation, an ecological pathogen management technique

Introduction

In California's strawberry production systems, soil-borne fungal pathogens are a critical pest that can cause substantial crop losses or crop failures, making pre-plant soil fumigation a functional necessity. However, the long-term sustainability of this keystone technology is doubtful. In the peri-urban coastal growing regions of strawberry production, increasingly restrictive regulatory policy, mounting environmental and human health concerns, and residential development have limited the rates of fumigants that growers are permitted to apply.

Viable alternatives to chemical fumigation, especially biological control methods, are limited, and growers are often economically restricted from cultural practices that would allow for system-inherent pathogen control. The limitations of these practices are especially true for organic strawberry producers, who cannot use pre-plant fumigation and instead rely almost exclusively on cultural management in order to limit disease losses.

A method for pathogen biocontrol, Anaerobic Soil Disinfestation (here called "ASD", also called "BSD" and "biosolarization"), was recently adapted for use in California strawberry systems (Shennan et al. 2013, Rosskopf et al. 2015, Shennan et al. 2018). Early work on ASD has indicated a promising capacity to effectively suppress

pathogens and provide similar crop outcomes as fumigation, but has not been intensively monitored after grower-directed applications (Rosskopf et al. 2015).

The basic application method of ASD is to incorporate a carbon source into the soil, saturate soil pore spaces with water, and maintain saturation during the treatment period. The biotic and abiotic changes that occur during maintenance of anaerobic conditions appear to be responsible for the disease-suppressive properties of this treatment, but the exact mechanisms of action and optimal application parameters are the subject of continued research (Shennan et al. 2018).

Years of ASD development across multiple research groups have produced an extensive amount of work on the various controlled and uncontrolled parameters that influence treatment outcomes of ASD in both laboratory and field settings. During the ASD treatment, a variety of biological and chemical shifts occur (Momma et al. 2013). The initial period of treatment sees a rapid drop in soil redox potential, presumably from biological activity (Momma et al. 2013), which is maintained via periodic irrigation or maintenance of a gas-impermeable barrier. The cumulative exposure to extreme reductive conditions, sometimes calculated as Eh hours below -200mV, has been proposed as an important predictor of treatment success (Shennan et al., 2014). Additionally, chemical species with anti-pathogen activity such as Mn 4+, Fe 3+, NH4+, and various fatty acids are generated during treatment, which may have some anti-pathogenic activity (Momma et al. 2013).

The primary driver in this process appears to be a complex relationship between anaerobicity, temperature, carbon-source type, and carbon-source application rate.

Anaerobicity is usually maintained via saturation of soil pore spaces with management of irrigation water and where plastic mulch application is feasible. However, there is some evidence that different pathogens may express different sensitivities to combinations of these factors, such as Fusarium's noted survival capacity under low-temperature treatment conditions, leading to the formation of critical soil temperature thresholds which ASD applications must exceed to achieve pathogen-specific suppression (Shennan et al. 2018).

Manipulation of soil temperature is a relatively difficult task, only feasibly accomplished by moving the treatment dates towards warmer seasons or by selection of specialized mulch material. In cases where these options are insufficient or unavailable, equivalent outcomes may sometimes be achieved through increased application rates of C-sources (D. M. Butler et al. 2012). With this said, while a variety of carbon sources, such as ethanol, cereal bran, molasses, manure, standing cover crops, or other locally available carbon-rich materials have been successfully tested in ASD applications at a variety of rates (Rosskopf et al. 2015), there is substantial evidence that the microbial shifts linked to pathogen control may express a substantial dependence on C-source type (Mazzola, Muramoto, and Shennan 2018).

With these and other factors that could influence the operational outcome of ASD disease control in production systems, it is evident that the biological complexity in real-world application is significant. One way of approaching this complexity is via a global-level analysis to infer critical thresholds, cumulative requirements,

associations, or other parameter features that could produce a generalized estimate of the treatment and environmental conditions most important in operational settings.

Synthesizing datasets with networks

An ideal candidate model type for integration across experimental data in a complex, multivariate system are probabilistic graphical models. These models provide a useful structure for combining prior knowledge, both in the form of system structure and prior distributions, with inference gleaned from data evidence.

Probabilistic graphical models utilize system-wide knowledge to represent relationships and uncertainties in a system using probabilistic structures via encoding probability distributions in a structure called Directed Acyclic Graphs (DAGs). DAGs are defined by two object types: nodes, representing observed instances of random variables, and edges, representing dependencies between nodes. Encoding data structures as DAGs provides a direct method for decomposing complex multivariate systems into tractable structures via the specification of joint probability distributions.

Bayesian networks

Bayesian Networks, a specialized case of probabilistic graphical models, accomplish this specification by factorizing the joint probability distribution into a set of marginal and joint likelihoods, viz. the likelihood-prior factor of Bayes Equation, encoding conditional dependence structures in the joint likelihoods to represent structural dependencies. Prior work has found modeling using Bayesian networks to be robust

in difficult conditions such as noisy or highly sparse data (Tsamardinos, Brown, and Aliferis 2006), providing further support to the utility that a network-based approach provides in the analysis of complex, noisy multivariate data.

Determining the joint probability likelihoods is, however, not a trivial task. Even in smaller graphs (typically defined as below 30 nodes), a graph could contain as many as 2^{32} probabilities (fully-connected) requiring consideration, a computationally infeasible task. In this sense the process of analyzing data in a Bayesian network framework (and in PGNs more generally) is generally split into two components: structural discovery, where the pre-defined nodes are linked via directed edges, and parameter inference, where the edge relationship between any two nodes is assigned a parametric relationship. This parametric relationship is typically Gaussian for continuous variables or encoded as conditional probability tables for discrete variables but can also leverage link functions to accommodate nonlinear relationships.

Once the joint probability distributions are defined over the graphical model as a Bayesian network, probability maximization approaches can then be applied to search over the parameter space by whatever means is desired or appropriate for the task at hand, such as likelihood or posterior maximization, direct sampling, or other algorithms.

Expert knowledge

Bayesian networks are sometimes called expert knowledge networks in reference to the clearly defined role domain-specific expertise can play in informing the modeling process. While modern algorithmic approaches to structure discovery can be extremely effective in discovering network structures within data - and in some cases this is the experimental goal - domain knowledge can still play an important role in defining node and edge characteristics. This can be accomplished by a-priori definition of dependency strengths and variances, blacklisting (complete elimination) or whitelisting (complete inclusion) of edge presences from inferred graphs, or a-priori establishment of edge directions.

These characteristics may, for example, be particularly useful in the analysis of data from experimental data where treatments or manipulations have been established, or in observational data where prior domain knowledge has clearly defined causal relationships. In these cases, edge directions where background knowledge is available can be explicitly set, allowing for inference to be conducted on the edge parameters solely.

Additional information may be encoded in systems via the use of ‘levelness’ in data, where graph structures are encoded in directed levels. These directed levels cluster nodes in groups of cascading effects, as instanced in systems containing environmental, mediating, and outcome variables. In such a system, edge

dependencies moving from outcome to environmental are physically impossible and can be excluded from structural inference a priori.

Study goals

This study seeks to leverage the analytical ability of probabilistic graphical network modeling to bring together several datasets across multiple years, sites, and contexts (in both field and incubator studies), to produce an integrated analysis of the relationship between management, environmental, and treatment factors in ASD implementation, with a specific focus on inference of key outcomes and discovery of factors that can be leveraged to maximize the treatment success.

In particular, applying Bayesian networks as synthesis models provides a methodology and foundation for future work to combine disparate ASD datasets containing mixes of treatments and observed variables, into a coherent framework of analysis that can produce inference using all available experimental and observational data.

Methods

Source datasets

The datasets used for this analysis are extracted from datafiles collected via laboratory and field experiments conducted at University of California Santa Cruz, or from on-farm trials at sites nearby the UC Santa Cruz campus. Individual experiments

covered diverse subsets of total system parameters, but no trial sites contained observations from all parameters.

Year conducted	Location	Variables tracked
2015-2016	UC Santa Cruz	Diameter, yield, wilt, Verticillium, soil temperature, soil moisture, Eh, carbon source, carbon rate, soil EC
2016-2017	UC Santa Cruz	Diameter, yield, wilt, Verticillium, soil temperature, soil moisture, Eh, carbon source, carbon rate
2013-2014	Watsonville, CA	Yield, wilt, soil temperature, carbon source, carbon rate
2016-2017	Watsonville, CA	Eh, carbon source, carbon rate
2011-2012	UC Santa Cruz	Yield, wilt, Verticillium, carbon source, carbon rate, prior crop
2011-2012	Salinas, CA	Yield, wilt, soil temperature, carbon source, carbon rate, prior crop
2014-2015	UC Santa Cruz	Yield, wilt, diameter, pH, EC, Verticillium soil count, Verticillium

		plant infection, carbon source, carbon rate
2012-2013	UC Santa Cruz	Yield, wilt, carbon source, carbon rate, soil temperature
2010-2011	Castroville, CA	Yield, soil temperature, Eh, carbon source, carbon rate
2013-2014	La Selva Beach, CA	Wilt, EC, pH, carbon source, carbon rate, soil temperature
2014-2015	La Selva Beach, CA	Wilt, Eh, carbon source, carbon rate, soil temperature
2012-2013	Watsonville, CA	Carbon source, carbon rate, yield
2013-2014	Watsonville, CA	Carbon source, carbon rate, yield
2013	UC Santa Cruz	Verticillium, carbon source, carbon rate, soil temperature, soil moisture, Eh, pH

Table 2-1. Datasets used.

Yield measurements, wilt and plant diameter measurements

Yield measurements were determined by end-of-season cumulative yield, measured in units of [lbs. marketable fruit weight per acre], generally sampled twice a week,

and summed over the entire growing season. Yields were then standardized by conversion to a yield ratio using the basic equation

$$Y_{ratio} = Y_{trt}/Y_{control}$$

where the yield ration R is the comparison of each ASD treatment to the within-experiment control, typically a completely untreated or only fertilized treatment. This conversion serves to standardize yield measurements to account for methodological differences between experiments, including differences in observation window, variety- or environment- specific differences in yield not related to ASD or pathogen damage, and any measurement biases.

Wilt measurements were determined using a standard plant pathology wilt scoring system based on visual determination of disease presence and severity. Similar to yield, wilt measurements were standardized via a ratio transformation in a manner identical to yield ratio calculation, providing partial control for experimental differences in methodology or overall plant health.

$$W_{ratio} = W_{trt}/W_{control}$$

In all experiments, plant diameter was calculated via visual observation of the strawberry plant crown and measurement of the top-down longest dimension of the plant canopy (measured in cm). Diameter measurements were then converted to ratio measurements in a manner identical to yield ratio calculations.

$$D_{ratio} = D_{trt}/D_{control}$$

For a single field site (OREI Mother Trial), $Y_{control}$, $W_{control}$, and $D_{control}$ were set in these equations to a different non-ASD treatment, as the control treatment was compromised by additional fertility treatments that excessively lowered yields in those plots.

Anaerobic soil conditions

Soil anaerobicity was measured via determination of the relative potential, or Eh of the soil environment. Included datasets universally measured Eh by the use of ORP probes (manufacturer information), which use a platinum electrode sensor and installation-based monitoring to provide hourly measures of Eh. Prior work has indicated that the level or duration of anaerobicity, measured as detected Eh values below a threshold of 200mV, is an indicator of the intensity of ASD treatment (Wang et al. 1993).

$$Eh_{sum} = \sum(Eh_d - Eh_{THRESH}) \; \forall \; (Eh_d - Eh_{THRESH} > 0)$$

where Eh_{THRESH} for this study is 200mV.

Carbon sources and rates

Carbon sources and rates were documented as meta-data alongside experimental data for all sampled datasets. Application rates were treated as fixed [tons/acre] amounts whether applied as pre-plant dry materials or via an injection method (i.e. for molasses treatments). Carbon sources were converted to binary presence/absence

variables in a [0,1] interval, and rates were combined into a single unscaled "carbon source" variable.

Soil temperatures and threshold calculations

Soil temperature data were obtained on a daily basis from on-site weather stations or soil temperature sensors, if available, or from soil temperature estimates generated by the nearest CIMIS weather station dataset. Temperatures were sampled from a 3-week window capturing the ASD application process, and converted to three parameters: maximum soil temperature observed, average soil temperature, and a sum-thresholded area, calculated by a function which calculates a summary value T_{sum}, for any daily mean soil temperature observation t_d above some threshold temperature t_{THRESH}:

$$T_{sum} = \sum(t_d - t_{THRESH}) \; \forall \; ((t_d - t_{THRESH}) > 0)$$

The temperature threshold value was evaluated via a model-fitting framework during the final modeling process and set at 18C.

Soil EC and pH

Soil EC (electrical conductivity) and pH values were obtained via direct probe measurements in-situ or via bench analysis on collected soil samples.

Soil Verticillium counts and plant Verticillium counts

Soil *Verticillium dahliae* levels were evaluated in all datasets by the same methodology. First, representative samples were taken from experimental units using a soil probe to 15cm depth - incubator containers, if an incubator experiment, or experimental plots, if a field experiment. Samples were then air-dried, ground with mortar and pestle to homogenize and mix, and passed through a .5mm sieve. Approximately 2 grams of the homogenized sample was passed through an Anderson sampler onto a set of 5 petri-dishes with semi-selective media. After 1 month of growth, individual colonies were tallied on a colony-forming unit (CFU) and back-converted to a CFU / kg dry soil basis.

Plant-infected *Verticillium* counts were evaluated in all datasets via direct sampling of potentially-infected root hairs from a random sample of plants in each experimental unit. Root hair segments were then placed onto a set of 5 petri-dishes with semi-selective media. After 1 month of growth, colonies were evaluated for presence or absence of *Verticillium*.

Temperature projections for treatment suitability

Temperature projections were generated by using historical weather data from 2017. First, estimated daily soil maximum temperature values were obtained from the CIMIS system (cimis.water.ca.gov, Figure 2-1), an irrigation management system that provides high-quality weather data via a distributed network of weather stations. Weather stations are placed in agricultural zones and provide a variety of

instrumentation-based data - in the case of soil temperatures, via direct measurement using a thermistor (cimis.water.ca.gov). An evenly spaced coordinate sampling grid was then placed within the general boundaries of the major strawberry growing regions of California (Figure 2-2). Daily soil temperatures were imputed to these coordinate points by 3-nearest-neighbor averaging of daily soil temperatures from nearest CIMIS stations, yielding a coordinate sampling grid with associated temperatures. For four months - August, September, October, and November, temperatures were then summarized using the same averaging and sum-thresholding function applied to experimental data, yielding two measurements (threshold-sum, mean) over four months for each coordinate point.

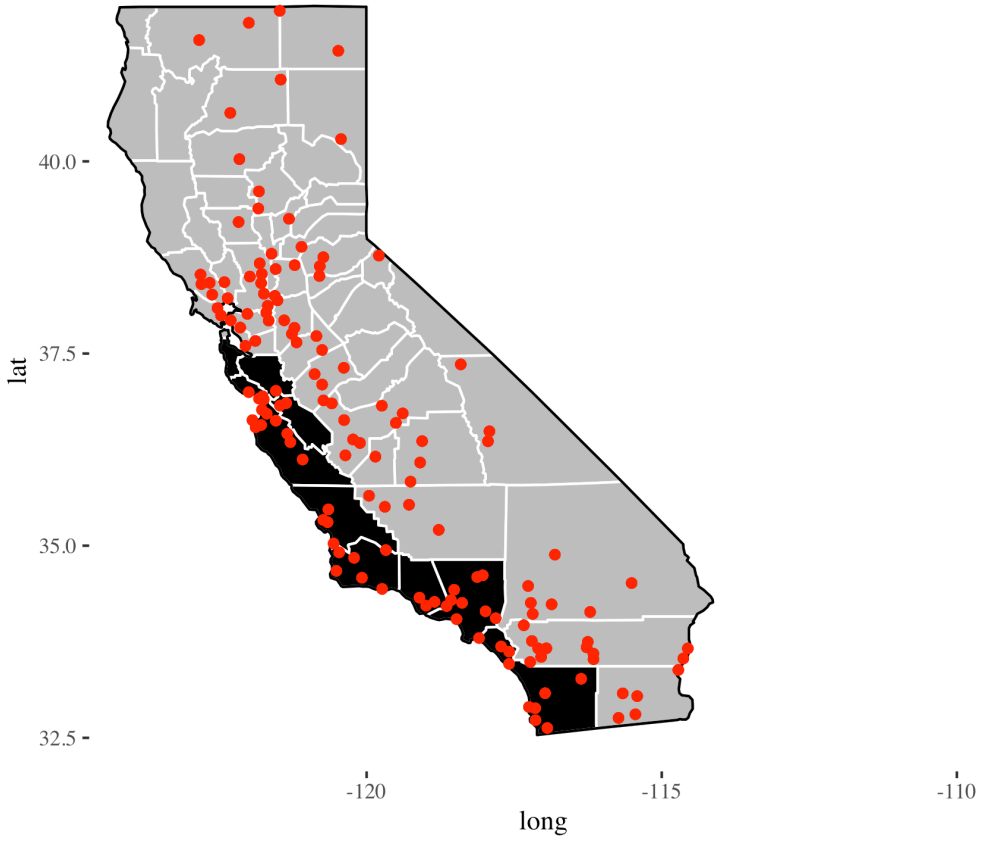


Figure 2-1. A map of California, with counties outlined. Red points indicate coordinates of CIMIS weather stations, which provide daily estimates of maximum soil temperatures. Counties filled with black color indicate coverage of the majority of strawberry-growing areas, and the boundary for temperature simulations. CIMIS stations provide underlying data for projecting treatment success based on historical temperatures.

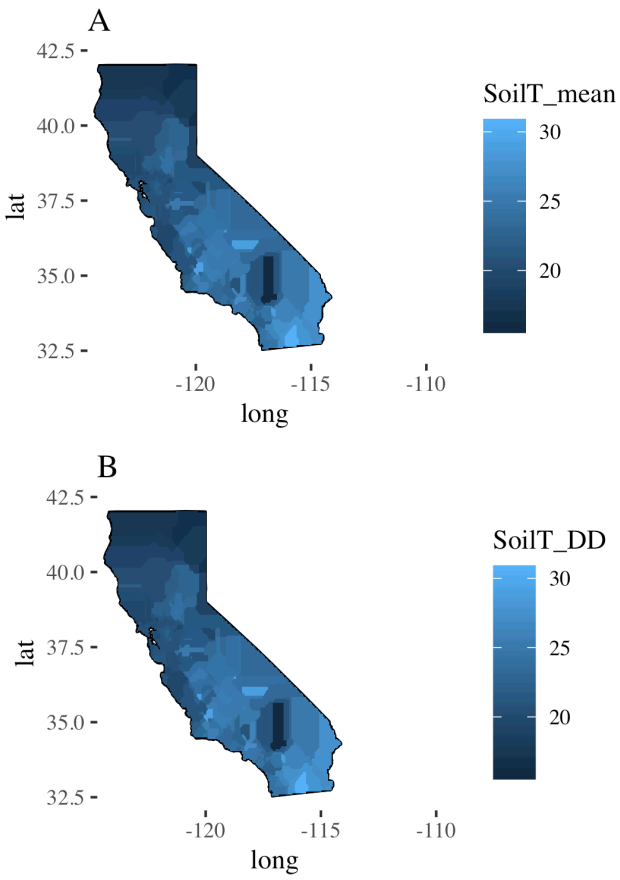


Figure 2-2. A map of California, with mean maximum soil temperatures (A) and accumulated soil degree days (B) as fill cells, from data collected in September 2017. Data sourced from daily estimates via CIMIS weather stations.

Data preparation

All available explanatory variables and outcomes from each dataset were targeted for inclusion in this study and transformed using the above described methodologies. Further manipulation of factorial treatment data was used to isolate rate from type, with equivalent C tons / ac serving as a continuous variable of **rate**, while binary index variables were generated for each C input **type** to provide a [0, 1] parameter of

C input presence. Interaction effects were manually generated by creation of derived interaction variables for all environmental and treatment inputs, but no interaction effects were generated for outcome variables, as these would not be interesting or interpretable.

Network structure discovery

Structure discovery was conducted by first encoding the underlying mixed continuous/discrete data as binned sets of observations, which allows for model analysis to proceed using continuous probability tables as per (Friedman, Goldszmidt, and others 1996). To learn the network structure, expressed here as an n × n matrix where any cell i, j takes a value of 1 when node i is connected to node j, a max-min hill-climbing (MMHC) algorithm described in (Tsamardinos, Brown, and Aliferis 2006) was applied. The MMHC algorithm uses repeated applications of a sub-protocol, the max-min parent-child algorithm, to iteratively search through a range of possible edges, searching and scoring a series of edge and node combinations until the highest-scoring DAG is found.

Expert knowledge was incorporated into structure discovery via a-priori enforcement of directed levels in the network structure, both in order to facilitate faster model convergence and to limit the presence of causally impossible edge directions, i.e. for some defined variable groupings A and B, the dataset D is subset into two parameter sets X and Y so that all variables $X \in A$ and $Y \in B$ can only be linked by the edge $X \rightarrow Y$, not $Y \rightarrow X$. Pre-defined directed groups were established for four layers. (1)

Environmental: pre-treatment vert soil CFU (continuous +), soil temperature threshold (continuous +), soil temperature (continuous +), soil moisture accumulated (continuous +), soil moisture (continuous +), rice bran (binary), cover crop (binary), molasses (binary), mustard meal (binary), carbon rate (continuous +), soil temperature X carbon rate (continuous +), soil temperature threshold X C (continuous +). (2) Intermediate: Eh-h < 200mV (continuous +), soil temperature X Eh-h < 200mV (continuous +), post-treatment Verticillium strawberry infection (proportion), pre- vs post-treatment Verticillium suppression (proportion). (3) Plant health: strawberry crown diameter (ratio vs control), strawberry plant wilt score (ratio vs control). (4) Yield (ratio vs control). No other structuring, such as mandatory groups or white/blacklisting were enforced on structure discovery.

Weighted network structure

While the MMHC approach provides a robust methodology of discovering the highest-scoring network structure, the stability of the discovered network to perturbations in the data is uncaptured, as the DAG generation minimizes a loss function over the entire dataset for selection of a single network. To provide confidence measures on the existence or robustness of any edge in the discovered network, MMHC was conducted within a bootstrapping framework, allowing for generation of confidence measures for each edge, and the creation of a "weighted DAG". The wDAG is generated by the following algorithm from (Friedman, Goldszmidt, and Wyner 1999):

For $i = 1,2,\ldots m$, sample with replacement N instances from D to create a new dataset D_i.

Apply structural learning to D_i to learn network structure $\hat{G}_i = \hat{G}\ (D_i)$.

For each feature, define the weight as:

$$p_N^{*,n}(f) = \frac{1}{m}\sum_{i=1}^{m} f\ (\hat{G}_i)$$

i.e. the total number of occurrences of any edge in all networks learned from all m bootstrapped samples. This algorithm was applied here using $m = 100$ bootstrapped samples with a sampling rate of $N = 276$, the original data size, using the bnstruct package in R (Franzin, Sambo, and Di Camillo 2016, R Core Team 2013).

Parameter learning

From all observed $\hat{G}_i$ the final graphical structure was extracted via manual examination. Isolated variables (nodes with no edges) and variables with lower than 80% confidence in the wDAG were excluded from the final network structure prior to parameter learning. The final model $\mathcal{G} = (\mathbf{V}, A)$, where $\mathbf{V}$ is the global probability distribution $\mathbf{V} = \{X_1, X_2, \ldots X_j\}$ for all j variables and A is the set of discovered edges, provides the corresponding joint probability distribution as a factorized form of $\mathbf{V}$ and conditional independence is learned via the presence and direction of edges in A.

The determined conditional independence structure was then used to construct a Gaussian Bayesian Network, where independence is maintained by treating every random variable **X** as multivariate-normal distributed, so that for all $\{X_1, X_2, \ldots X_j\}$,

$$p(X_1, X_2, \ldots X_j) = N(\mu, \Sigma)$$

where μ is a vector of means corresponding to each variable X and Σ is the variance-covariance matrix, with conditional independence relationships are encoded by a priori sparsity, i.e. the covariance $\sigma_i \sigma_j$ is set to 0 where conditional independence between X_i, X_j is assumed. Parameters from this probability distribution were then learned using maximum likelihood estimation, generating MLE estimators $\hat{\mu}$ and $\hat{\Sigma}$ (Scutari 2009). Where learned covariance $\sigma_i \sigma_j$ was zero, indicating poor support for direct effects, the edge was manually removed, and any resulting orphan nodes were discarded. This algorithm was applied here using the bnlearn package in R (Scutari 2009, R Core Team 2013).

Prediction

Parameter estimates over the entire probability distribution $p(X_1, X_2, \ldots X_j)$ provide a fully generative model. Prediction is then accomplished by evaluating the probability of a fixed outcome X_i in the system generally, $p(X_i = x|\mathbf{X}, \mu, \Sigma)$, single maximum likelihood estimates given a set of observations $p(\mu_i|\mathbf{X}, \mu, \Sigma)$, or any combination of maximum likelihood estimates , single maximum likelihood estimates given a set of observations $p(\mu_i, \mu_j|\mathbf{X}, \mu, \Sigma)$. As conditional independence within the network

allows marginalization of node likelihoods to direct parents, predictions on any subset of variables $p(X_1, X_2, \ldots X_j)$ only requires consideration of parents $\Pi_{X_{1..j}}$, i.e. $p(X|\Pi_X)$ (Scutari 2009).

Results

Graph structure

Structure discovery found strong support for nearly all environmental variables, including evidence for mediation of effects via plant health characteristics (diameter, wilt) on yield, as well as direct effects on yield. There was no support for differentiation of carbon inputs by type. Instead, direct carbon rates were found to influence wilt and yield both as individual and as interaction effects with soil temperature and soil temperature threshold (Figure 2-3).

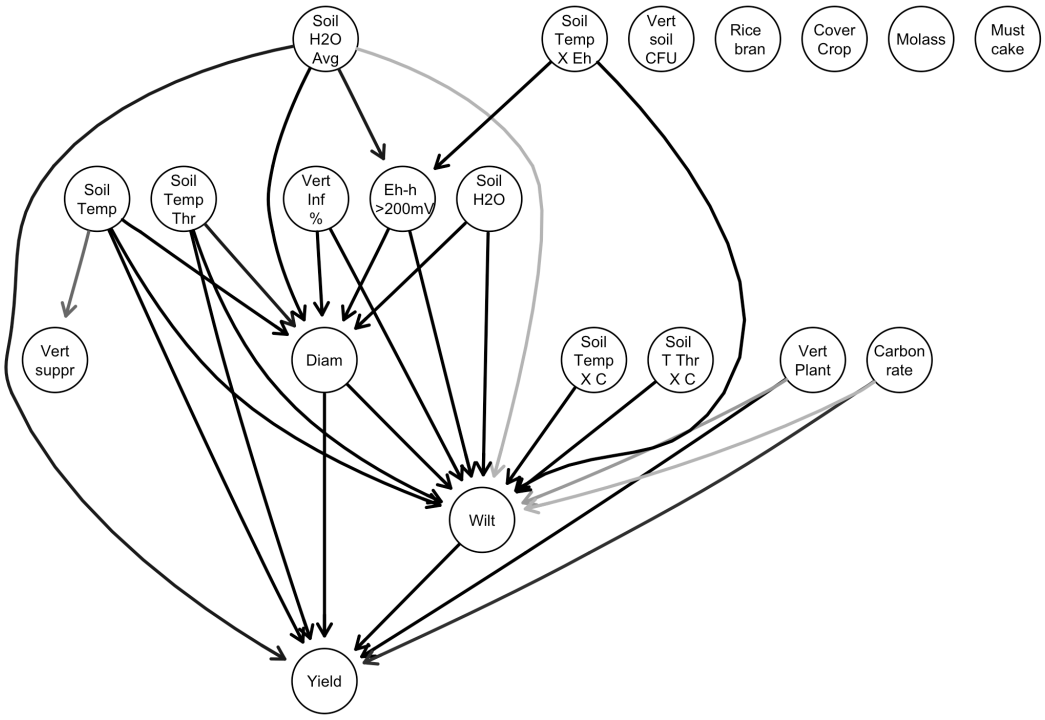


Figure 2-3. Weighted directed acyclic graph structure. Structure was learned using 100 iterations of bootstrap sampling over the entire dataset. Arrows indicate the presence and direction of edges, and shading indicates the strength of the relationship, measured in the number of times the edge was learned over all iterations. The adjacency matrix for this wDAG is available in Appendix 2.

Pruned graph structure and parameters

After parameter fitting and removal of null edges and orphaned nodes, the final graph structure simplified into a set of three primary input drivers (temperature, carbon rate, and Eh), influencing the three primary outcomes (plant diameter, plant wilt, and yield) (Figure 2-4). Evidence for both direct effects to outcome variables and

mediated effects are apparent, the latter effect being primarily driven by plant diameter and plant wilt serving as mediating variables.

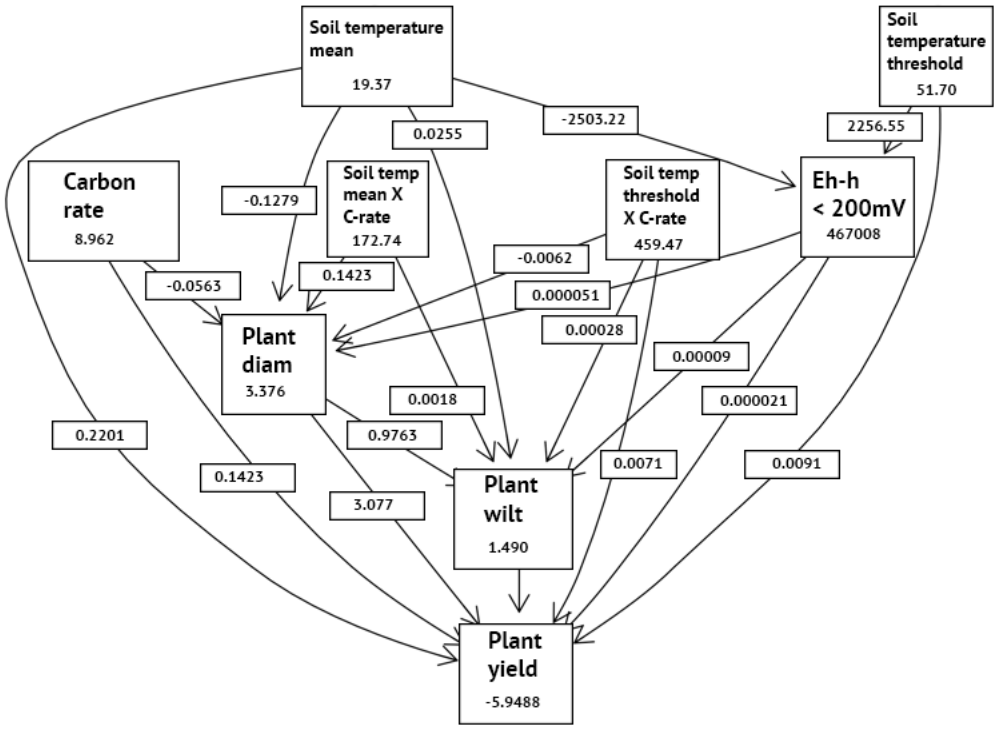


Figure 2-4. Gaussian Bayesian network structure, after removal of orphan nodes and null edges. Nodes and edges represent means and covariance parameters as factorized in the multivariate normal representation of the joint probability distribution.

Predictive outcomes - marginal

Posterior predictive outcomes indicate a strong interaction effect between soil temperatures and carbon inputs. Figure 2-5 illustrates this relationship by evaluating the probabilistic outcome of yield exceeding a 20% boost over the control, via the likelihood $p(X_{yield} > 1.2|\mathbf{X}, \mu, \Sigma)$, and values of carbon source and mean maximum daily temperature simulated over realistic ranges.

A bivariate inflection ridge is apparent in the projection separating the predictive values into low- and high-likelihood regions. A similar but slightly moderated ridge is apparent in Figure 2-5, which evaluates the same likelihood but with carbon source and mean degree days daily temperature as the simulation variables.

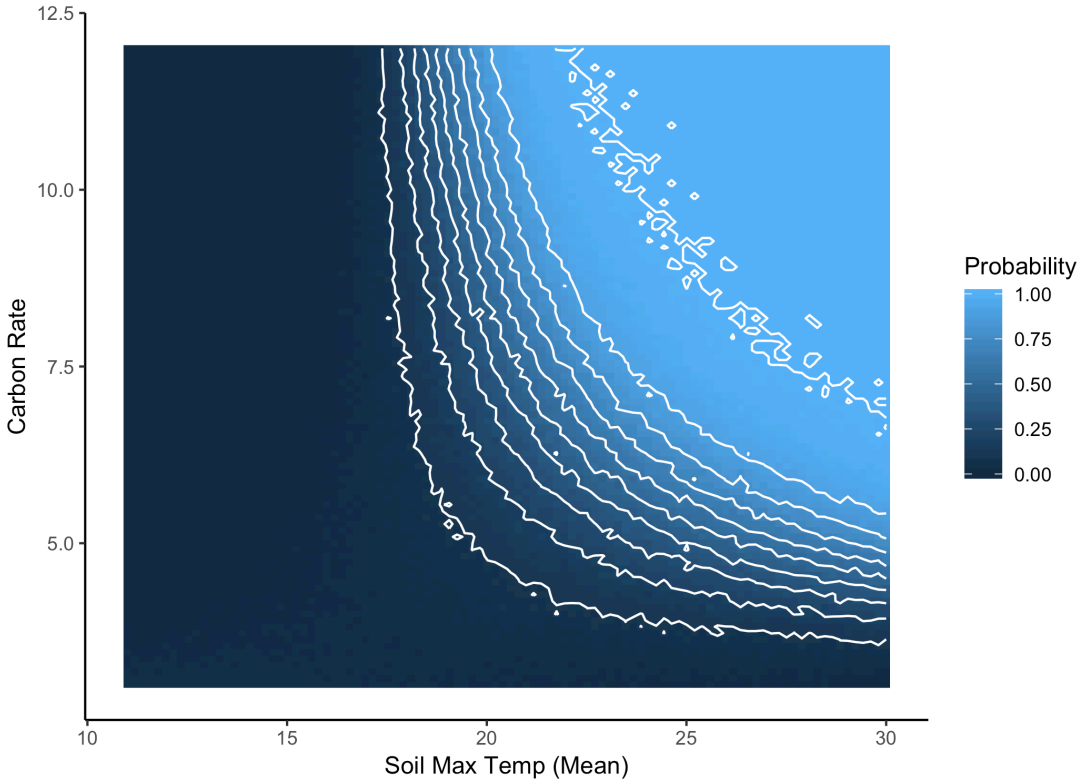


Figure 2-5. Predictive surface estimating the probability of a 20% yield boost given varying soil maximum temperatures and carbon rates. Probabilities are derived from evaluating the model over a range of simulated C rates and soil temperatures. Contour lines mark every .1 increase in posterior probability estimation.

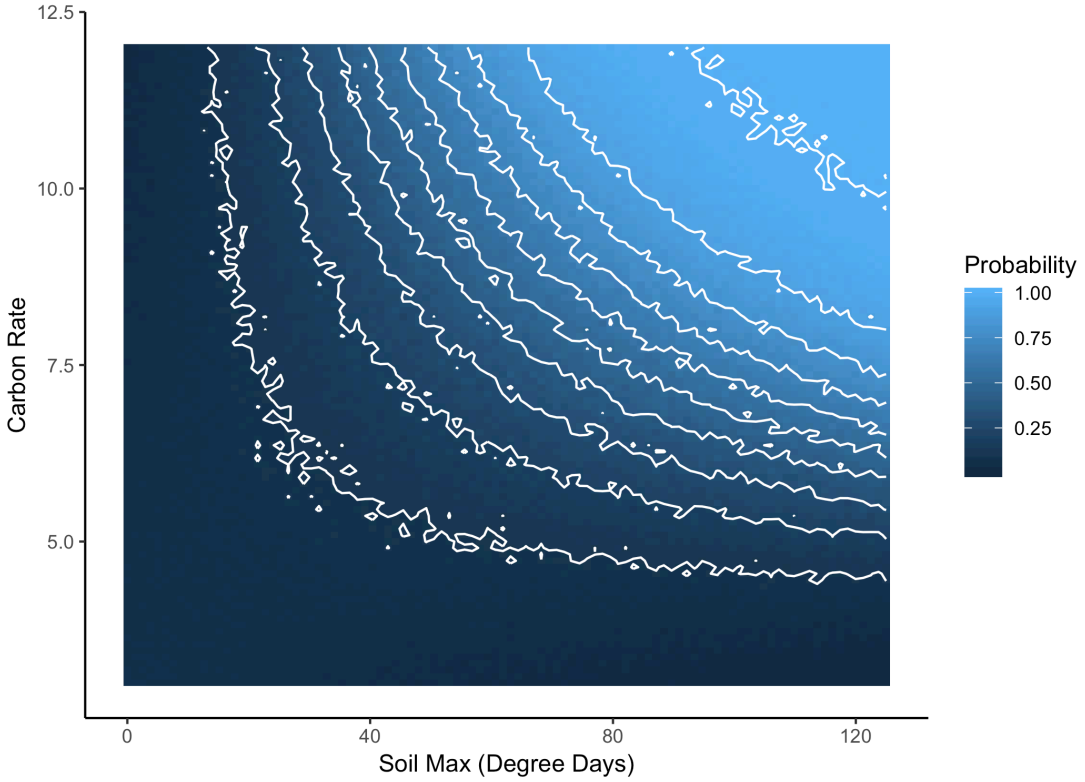


Figure 2-6. Predictive surface estimating the probability of a 20% yield boost given varying soil degree-day temperatures and carbon rates. Probabilities are derived from evaluating the model over a range of simulated C rates and soil temperatures. Contour lines mark every .1 increase in posterior probability estimation.

Predictive outcomes - geographical

These same findings are repeated when applied to the geographic data derived from rasterized CIMIS soil temperature data extracted and interpolated between weather stations for four months (August, September, October, November). Given three carbon rates (11 tons / ac, 9 tons / ac, 7 tons / ac), the probability of a 20% yield boost

closely tracks soil temperatures, generally along a north-south gradient but with some notable regional variation, presumably due to topographic variation.

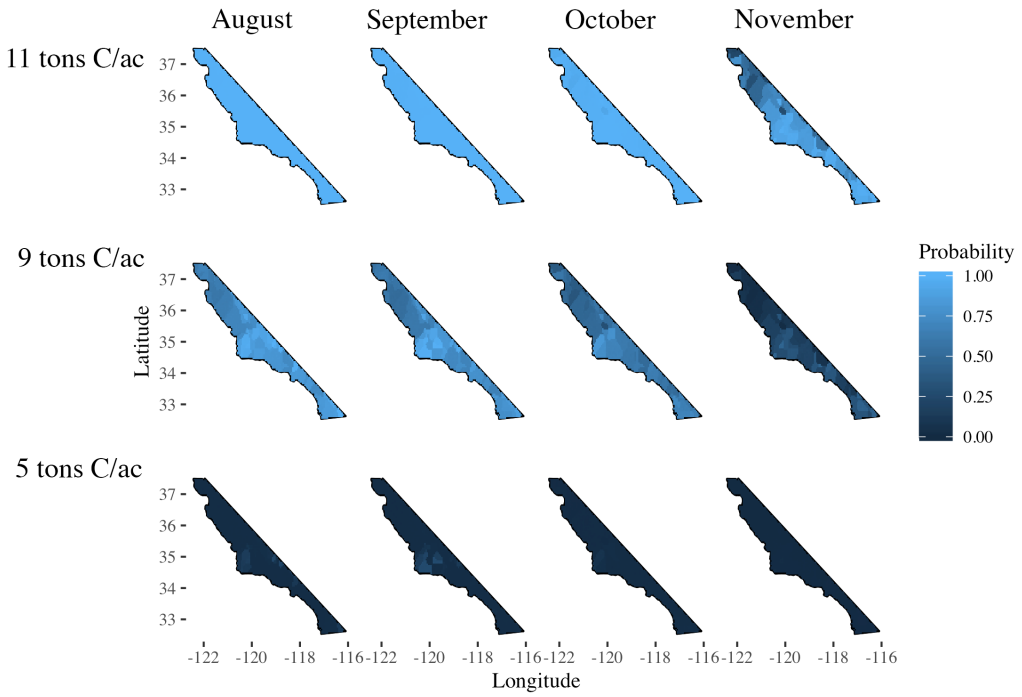


Figure 2-7. Projections of the probability of a 20% yield boost given varying carbon rates, months of year, and geographic location. Month of year and geographic location are used to index rasterized temperature data, which is sampled from 2017 historical CIMIS data.

Discussion

Key factors to successful treatment

Results from both the structural and parameter learning process strongly confirm the critical importance that environmental parameters play in determining the outcomes

of ASD applications, reinforcing the existing recommendations that applications of this method in on-farm scenarios requires careful consideration of the application parameters and environmental conditions that will influence outcome successes.

The greedy nature of the MMHC algorithm used in structural learning is likely responsible for the sharp differences between the larger network derived initially and the smaller network that was retained for parameter learning; by leveraging iterative evaluations of conditional independence and bootstrapping, and by fitting on conditional probability tables, MMHC is far more sensitive to the presence or absence of edges than the gaussian network's MLE algorithm used for parameter discovery (Scutari 2009).

The presence of edges and nodes in the MMHC-derived network structure may suggest important avenues for future research; while soil moisture parameters appeared to influence both plant diameter and wilt, their parameters could not be estimated effectively. Prior work supports the hypothesis that maintenance of soil moisture is important for both stimulation of microbial activity and maintenance of anaerobic conditions (also supported by the presence of an edge between soil moisture average and Eh-h) (Shennan et al. 2018). Future work may consider evaluating the role of irrigation and soil moisture maintenance under a manipulative experimental framework to generate additional data on this relationship.

In the gaussian Bayesian network, the dominant role of carbon rate and soil temperatures is also in agreement with prior work (Hewavitharana, Ruddell, and

Mazzola 2014) and an underlying hypothesis of ASD treatment mechanisms - as ASD relies on the stimulation of soil microbial activity and respiration, and on the facilitation of metabolic product generation, there should exist a strong relationship between carbon resource additions and temperatures on outcomes.

Surprisingly, effects the type of carbon source were not detected in either MMHC or MLE learning, despite prior work indicating that lability and other qualities may make certain organic amendments more suitable for ASD than others (Butler et al. 2012). Further work or expansion of the present model to accommodate additional data sources should explore these relationships.

Both the marginal plots of carbon rate X temperature and geographic projections of treatment success probabilities indicate the potential for adaptive use of carbon rates in response to projected temperature regimes. For growers in southern growing areas, or with the flexibility to apply ASD on ground in early-season, model projections suggest the possibility of carbon rate reductions with little impact on the projected ability to generate improved yield / plant vigor.

Conversely, the presence of compensatory effects may allow growers in northern growing areas or late-season applications who still wish to apply ASD to "make up for" the temperature penalty by increasing their carbon application rates. While the present analysis was not able to capture disease pressure-specific parametric relationships, the natural extension of these findings is to evaluate whether soils with higher disease burdens, either through CFU or pathogen species, could be adaptively treated with higher temperature or carbon rates on a sub-field basis.

Limitations

The use of yield ratios in this study provides a convenient standardization of results across experiments, especially useful in strawberry systems where absolute yield and plant characteristics may vary widely from small changes to cultivar selection or plant spacing. However, ratios also require more considered interpretation, as they simultaneously represent the potential yield change from ASD applications and the baseline performance of the reference plot.

In fields with very little disease pressure and overall high yields, this may suppress the apparent effect of even a well-applied ASD treatment, and in fields with very strong, this may pronounce the same effect (assuming the ASD application worked). In theory, this can result in paradoxical findings, such as a high baseline disease pressure having an overall positive effect on the yield outcome. While disease-count parameters were not estimable in this study, this type of challenge of interpretation remains salient.

Further, the strong majority of experiments included in this synthesis (with the exception of two fields) were only documented as containing infestation of *Verticillium dahliae* as the primary pathogen. Prior work on ASD applications in fields infested with *Macrophomina* and *Fusarium* indicates that the required treatment conditions in these fields may be quite different (Yonemoto et al. 2006, Muramoto et al. 2016, Ebihara and Uematsu 2014); most notably, the soil

temperatures required to achieve suppression of species other than *Verticillium* are notably higher.

From this perspective, the Bayesian network learned here is best characterized as *Verticillium*-specific. Future work integrating additional datasets into a synthetic analysis should take careful consideration on model structures that accommodate pathogen-specific system differences, either through completely unpoled model fitting (separate structures and parameters for each pathogen system), or via more complicated hierarchical structures.

Future work

The risk aversion of many growers is profound even within the consideration of switching between fumigant types, not to mention non-fumigant control options (Asci et al., n.d.), and the high capitalization and extreme consequences of fungal pathogens in strawberry production (Carter et al. 2005, Koike and Gordon 2015) likely exacerbate this aversion.

The outcomes of this project provide both a methodology and results that may provide a path forward for reducing the uncertainty about optimized ASD treatment recommendations and their predicted outcomes. Integration of additional data from experiments and on-farm would improve the specificity and reliability of predictions and may provide additional avenues for farm-specific management recommendations, such as nesting of meteorological projections, mulch selection for increasing soil

temperatures if model projections indicate insufficient temperatures, or other management interventions to improve success estimates.

Further, both structure discovery and parameter learning are Bayesian processes which allow for integration of new data on existing model structures without re-learning the entire dataset. Via belief propagation methodologies, the existing MMHC-derived model structure can be updated and evaluated with additional data, facilitating continuous inclusion of additional data.

Unsupervised clustering of farmer approaches to information use, land management, and pathogen control in walnut production systems in Chile

Introduction

In the last 10 years, walnut production in Chile has undergone a significant expansion in land under cultivation, from an estimated 5,000 hectares in 2007 to 40,000 hectares in 2018 (Guajardo et al. 2019). This production area represents an increasingly economically important industry, and most projections of future demand for walnuts and walnut products indicate that the market will continue to be supply-limited, encouraging future growth (Guajardo et al. 2019).

Within this acreage, a significant portion of the walnut-producing area appears to present evidence of root- and stem-rot diseases associated with infection via pathogenic Phytophthora species (Guajardo et al. 2017). While the damage is often non-lethal and controllable, Phytophthora infection can cause complete loss of productivity from individuals or progress to full mortality, representing a major threat to the productivity of affected orchards (Mircetich, Matheron, and others 1983).

Further, the geographic distribution of walnut production systems in Chile appears to be shifting, at least in part due to changes in geographic suitability attributable to global climate change (Guajardo et al. 2019). In areas previously unexposed to walnut production, soil conditions may affect the type and behavior of Phytophthora spp and produce root-pathogen interactions more severe than traditional production geographies.

Although the factors affecting Phytophthora infection in Chilean walnut production are multivariate, a few key features of the production systems are salient. First, the susceptibility of an orchard to attack by Phytophthora spp is well-described in agronomic literature as primarily management-based (Browne et al. 2006), in that with appropriate irrigation, sanitation, and varietal practices an orchard may generally be inured to significant damage from Phytophthora-related die-off. Additionally, the majority of plantings utilize Juglans regia rootstock (Guajardo et al. 2017). While this rootstock is admired for its high-performing yield characteristics, it is also notoriously susceptible to attack by Phytophthora species (Browne et al. 2006).

Several gaps of knowledge in this system are present. First, while several species and sub-species of Phytophthora have been identified as parasitic/pathogenic to walnuts, the prevalence and distribution of these species within Chile is unknown, nor whether their geographic distribution is indeed a source of novel severity in disease outbreaks. Second, while alternative rootstocks are available, it is as yet unknown whether introduction of these rootstocks will actually reduce morbidity.

Finally, there is a considerable absence of information on the actual on-farm management practices in Chilean walnut production, and how integrated agronomic programs can best serve these operations. It is to this final point that the present work is directed.

Extension work and technology transfer

Contemporary research in agricultural extension and outreach has seen an increasing call for further incorporation of stakeholder-oriented methodologies into the design and implementation of agricultural research, with particular emphasis towards the role socioeconomic stratification may play in the incorporation and relative benefit of technological innovation in agroecosystems (MacMillan and Benton 2014, Levidow, Pimbert, and Vanloqueren 2014).

While traditional models of technological development by researchers and extensions have primarily focused on methodological and epistemic challenges as the primary motivators to inquiry, proponents of a more integrated approach to extension work encourage a conceptualization of agronomic improvements as components of an overall integrated agroecosystem, with a concomitant responsibility to down-stream effects of released technologies (Hauser et al. 2016).

Further, it has been well-noted that agronomic work may serve to benefit a restricted strata of land managers - in particular, farms of greater capitalization and technological adoption may be more prepared to invest in, acquire, and implement further technological improvements produced by scientific research, further cementing the potential role of research in directly shaping the agricultural landscape (Leeuwis 2013, for an overview of the subject).

With this in mind, manager-oriented research, such as participatory action research, establishes a clearly defined role of farmers and land managers within the scientific

inquiry process, both as collaborators to the research process and recipients of the outcomes (Hauser et al. 2016). Such projects frequently incorporate expert knowledge and feedback from farmer-stakeholders into several or all of the steps in a research pipeline, from initial ideation, development of project outcomes, and implementation and data collection, sometimes even involving farmer-stakeholders as active participants in data collection and monitoring (ibid.).

In the absence of direct involvement of farmers within the research cycle, conscientious understanding of the agroecological landscape is recommended as a first step towards directing research and technology transfer activities (Liu, Bruins, and Heberling 2018, Hauser et al. 2016). While often overlooked, the diversity, frequency, and sufficiency of actively implemented farm practices can provide a valuable source of knowledge and direction for both future research and effective provisioning of extension services.

Further, it may be the case that managers in a study system present a diversity of characteristics that is neither unified nor random, but instead can be understood within broad groups of management approaches. This phenomenon has been observed in multiple study systems previously, from productions as diverse as dairy (Cook et al. 2016), rice (Savary et al. 1994), and mixed vegetable production (Hillger et al. 2006). In this case, understanding the underlying distribution of clustering is a critical task prior to implementing a technology transfer program, as the information on disparate needs or approaches may provide insights as to how a potential intervention may be structured as to maximize (or minimize) particular effects.

Methods

Methods: survey dataset

This work is based on a 2017 dataset surveying management practices and technological attitudes of walnut farmers in Chile. To adequately capture a representative range of socioeconomic and management characteristics in the production landscape, this survey used a stratified sampling design with pre-binned categories by geography and farm size. In each administrative region where walnut production has a meaningful presence (Regions 4, 5, 6, 7, 8, and RM/13), a target of a minimum 16 growers to survey was established; within that target, a sub-target of 4 growers from four farm sizes (<1ha, 1-5ha, 5-25ha, 50ha+) was established, though obtained numbers vary (Appendix 4). Questions addressed socioeconomic characteristics of the farm operations, management behaviors, irrigation management, specific management approaches to the control of Phytophthora and disease generally, and agricultural information management (see Appendix 1 for a full list).

Farm survey geography

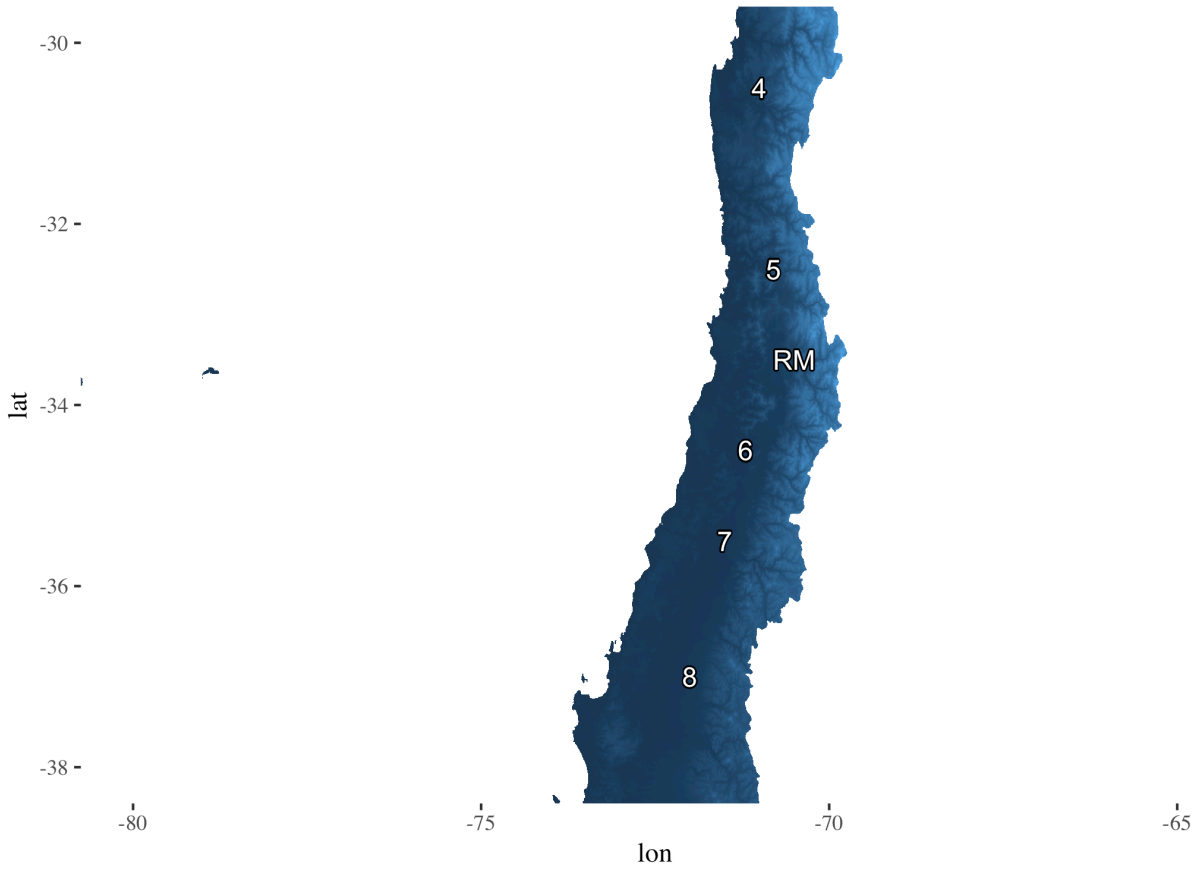


Figure 3-1. Surveyed regions within Chile, numbered by region name.

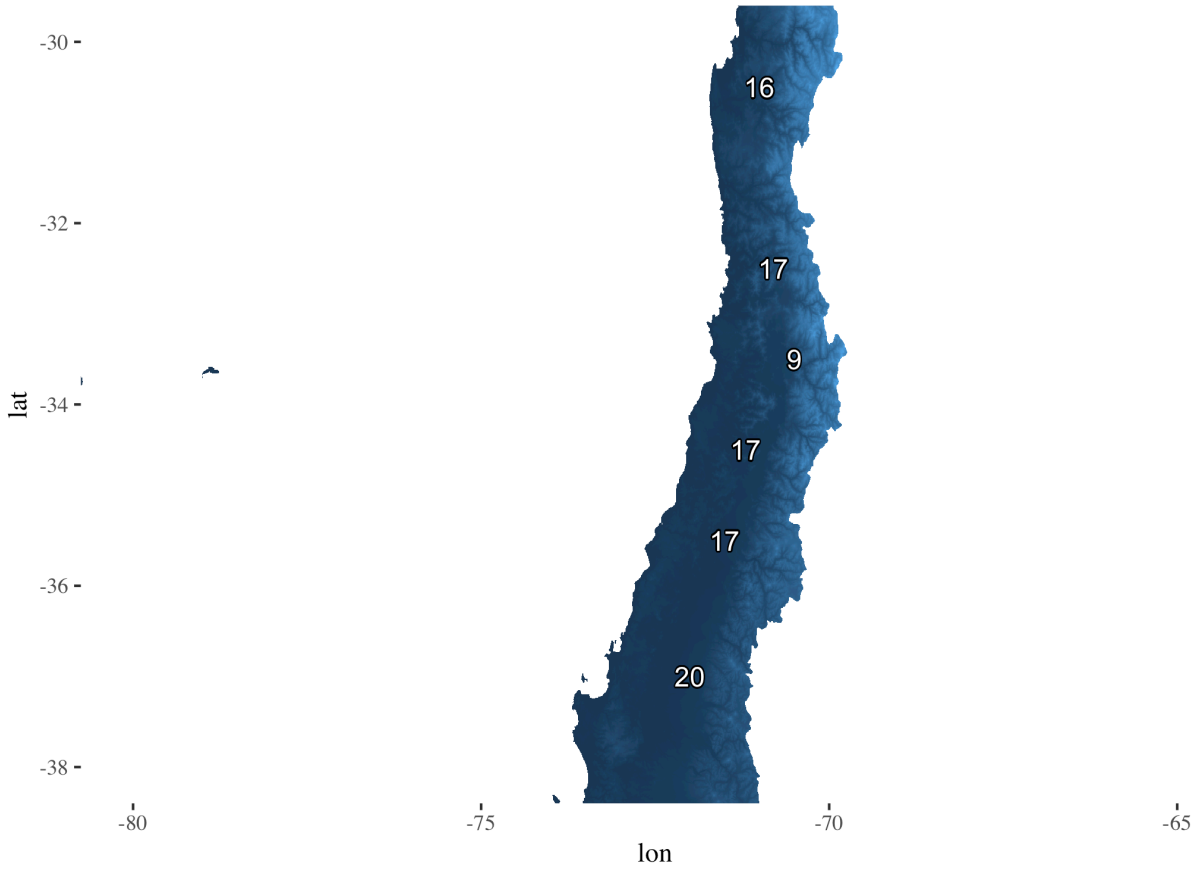


Figure 3-2. Surveyed regions within Chile, numbered by respondent counts.

Converting and subsetting responses

The final dataset is composed of 96 unique survey points with 101 unique questions (Appendix 2). Of these questions, 37 are based on factor-response answer types. Almost all data analysis methods implicitly or explicitly convert factor responses to sets of binary response variables, e.g. “dummy variable encoding” or “1-hot encoding”. In this dataset there are an additional 117 binary responses from possible choices within factor responses, bringing the total parameter count to 218.

Responses were then further subset to reflect the desired parameter pool for clustering. Instead of using the entire feature space, the features were restricted to include only information on demographics, farm-level characteristics, and farm management responses, so that clustering would be conducted on a dataset determined by actual management behaviors rather than underlying information-attitude responses.

Partition around medoids algorithm

The clustering method applied here uses an unsupervised learning algorithm called K-medoids. Similar to the more commonly used K-means algorithm, K-medoids eschews the use of synthesized centers ('means') and instead considers data observations as candidate pivots for generation of cluster centers. A simplified version of the algorithm operates as follows, for some arbitrary number of clusters K:

K-medoid partitioning (Reynolds et al. 2006)

Take an initial guess for centers $c_1, \ldots, c_K$ via random sampling of the actual dataset.

1. Minimize over C: for each $i = 1, \ldots n$, find the cluster center c_k closest to X_i, and let $C(i) = k$
2. Minimize over $c_1, \ldots, c_K$: for each $k = 1, \ldots K$, let $c_k = X_k^*$, the medoid of points in cluster k, i.e., the point X_i in cluster k that minimizes $\sum_{C(j)=k} \| X_j - X_i \|_2^2$

When within-cluster variation doesn't change with a new medoid, the process is stopped. In English, this algorithm assigns an arbitrary label to some random point as an initialization step, searches through the rest of the data and assigns labels based on their similarity to any of the initialized values, finds a medoid within each generated cluster, and repeats the label-assigning and medoid-finding process until no further improvements can be made.

This process generates a few key measurements: clusters derived from the algorithmic iteration, representative medoids for those derived clusters, and the overall fit of derived clusters to the dataset ('silhouette width'). Each are helpful for different aspects of the modeling process: clusters assigned to each data point separate observations into discrete units for later dissection, representative medoids provide real-world examples of a 'canonical' example most-exemplary of each derived cluster, and fit measures provide a measure by which we can search over the hyperparameter space k and find the optimal cluster size.

PAM implementation

Implementation of the PAM algorithm on this dataset was conducted using R Statistical Software (R Core Team 2013), precise software details at end of document). First, selected categorical response variables were decomposed into constituent binary variables using a general 1-hot-encoding technique. Numerical and ordinal response variables were not modified.

Of the 116 original parameters (Appendix 2), a subselection of all questions related to management practices were used as inputs to a clustering method (63 total). The objective was to identify clusters of farmer behaviors within the dataset based on similarities within observed parameters. Some parameters were removed due to covariance (e.g. region ID and coordinates), irrelevance (e.g. file source or email), irregularity (e.g. irrigation frequency information), or insufficient processing (e.g. varying response types to open-ended questions). Information network questions were excluded from clustering as they were not directly related to management practices but are included as profile variables linked to derived clusters.

First, a dissimilarity matrix was constructed by using Euclidean distance (root sum-of-squares difference) for numeric parameters or Gower's distance method (Kaufman and Rousseeuw 2009) for nominal, ordinal, and binary data. Gower's distance is a function which attributes a dissimilarity value $d_i j$ to any two observations i and j based on a weighted mean calculation. These steps were accomplished with the daisy function from the cluster library in R (R Core Team 2013).

This dissimilarity matrix was then used as an input into an implementation of pam algorithm following the original build/swap steps as described above, using the pam function from the cluster library in R (R Core Team 2013)..

Silhouette width

The pam function provides a model fit of generated clusters and overall "silhouette width", a generalized measure which evaluates variance explained within clusters

versus variance unexplained (Rousseeuw 1987). The general approach to this calculation relies on determination of within-group dissimilarity and between-group dissimilarity, via estimation of the function $s(.)$ for every observation i, where the function $a(i)$ determines the average dissimilarity of observation i to all other observations in the final cluster, as determined by the fitting process detailed above, and function $b(i)$ determines the *minimum* average dissimilarity of observation i when iteratively compared to all other clusters. From these underlying functions, we define $s(.)$ as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Which provides a [-1,1] bounded interval of "silhouette width" (Rousseeuw 1987).

Represented by the sil_width class attribute generated by the pam function in the R cluster library, this process was repeated for a series of candidate K values [2,…,20] and the final model fit was selected from the model of value k which maximizes silhouette width, and by extension explanatory power.

Projection onto two-dimensional space

After generation of clusters, examination of the clustering distribution by visualization is generally suggested as a best practice to ensure proposed clusters are robust. Projection of higher-dimensional datasets onto a two-dimensional space for visualization is an important task in many applications and has been a frequent task in machine learning research. Most recently, a popular algorithm called t-distributed

stochastic neighbor embedding, or "t-SNE", has seen widespread acceptance into dimension-reduction and visualization tasks (Maaten and Hinton 2008). Briefly, the t-SNE achieves dimensionality reduction by generating two sets of pairwise dissimilarities, one for the original dataset and one for a set of low-dimensional "map point" counterparts (for original data, dissimilarity is a normalized Euclidean distance, for low-dimensional map points, Euclidean distance). Both sets of dissimilarity measures are then subjected to minimization of a cost function based on Shannon entropy in order to find the map point configuration that contains the highest fidelity to dissimilarities observed in the data. Some additional tweaks are required to adequately calibrate the normalization parameters, but this is left to the original work (Maaten and Hinton 2008).

This procedure is implemented with the dissimilarity matrix previously described, using the `Rtsne` package from the `Rtsne` library in R.

Feature importance scores

Feature importances were inferred by calculating a ratio of cosine similarity for each feature via evaluation of *within-dataset* similarity versus *within-group* similarity. Because clustering produces groups that are likely of different lengths, to calculate dot-products for feature vectors a random sampling approach is used, so that for each feature x the following algorithm is used:

Cosine similarity ratio for feature importance

1. Subset the feature vector **x** into two random samples **a** and **b**.
2. Calculate cosine similarity 1:I times with

$$cos(\theta_i) = \frac{a \cdot b}{\| a \| \| b \|}$$

3. The average cosine similarity is then

$$cos(\overline{\theta}_{all}) = \sum_{i=1}^{I} c\,os(\theta_i)/I$$

4. Repeat steps 1-3 for a subset of the data within each group J = {1,2} for all s_{ij}, so that

$$cos(\overline{\theta}_{group}) = \sum_{i=1}^{I} c\,os(\theta_{i1})/2I + \sum_{i=1}^{I} c\,os(\theta_{i2})/2I$$

5. Calculate the similarity ratio

$$r(x) = \frac{cos(\overline{\theta}_{group})}{cos(\overline{\theta}_{all})}$$

6. Repeat for all features.

This approach provides a ratio centered at 1 (zero difference in sampled similarities between all-data or group-data), and ratios above 1 provide a convenient measure of features that are particularly clustered within groups versus the random sample, akin

to similar methods using cosine similarity as a distance measure for document clustering (Huang 2008).

Results

Optimal cluster size

Silhouette width calculations indicate highest explanatory power at k=2 and an immediate loss on addition of further clusters, with gradual improvement in fit as clusters are added. Overall silhouette width is low-moderate in explanatory power by general interpretive guidelines (Rousseeuw 1987).

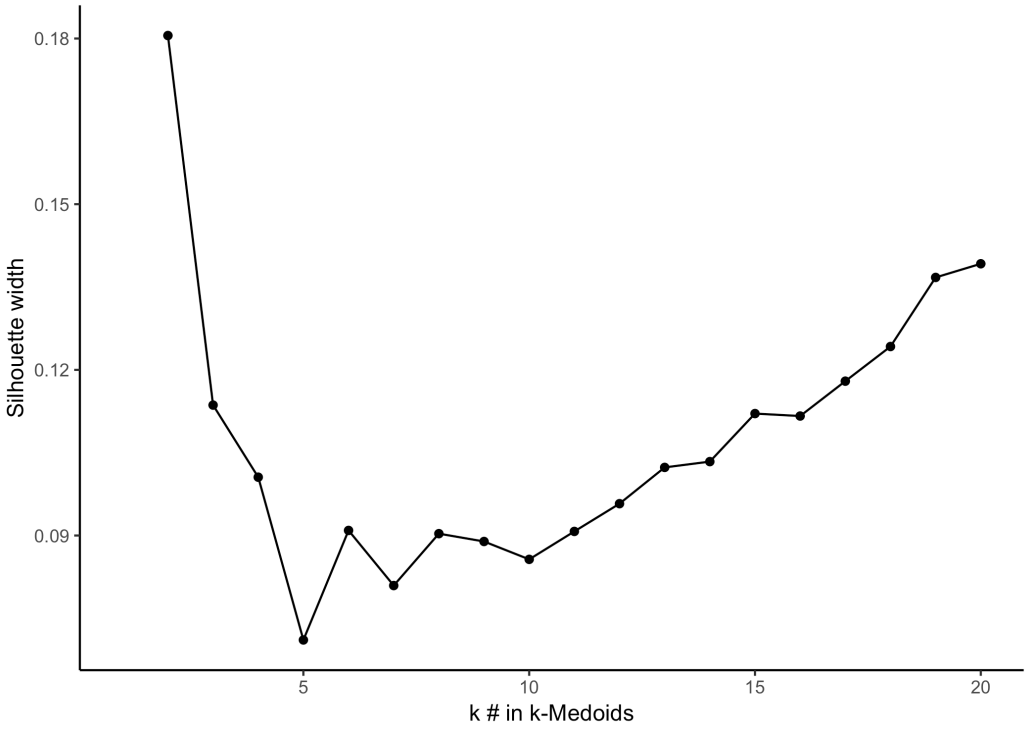


Figure 3-3. Silhouette width versus # of medoids evaluated.

Reduced parameter space projection

Parameter reduction using t-SNE indicates an efficient capture of clustering from the PAM algorithm, an expected result given the similar methodologies of Euclidean distance minimization. Clusters are well-defined with only small evidence of mixing across boundaries.

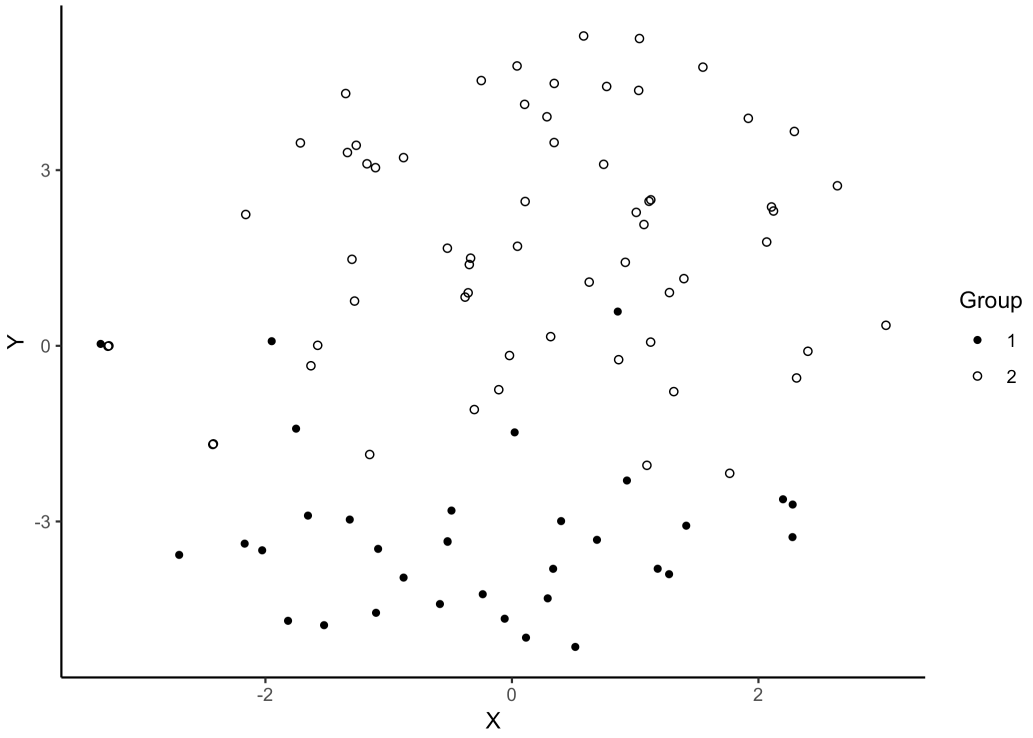


Figure 3-4. t-SNE projection of survey data into a 2-dimensional space. Projection methodology is described in methods above. X and Y axes are synthetic mappings of original dataset. Clusters defined by partition-around-medoid methods were used to color map points post-hoc and were not used in the t-SNE process.

Distributions between clusters

Numerous features of the dataset are immediately evident upon examination of how 2-medoid clustering separated the survey data into group 1 (hereafter, 'G1') and group 2 (hereafter, 'G2'). Results indicate good support for the hypothesis that a primary division of respondents into two categories exist: larger-scale, higher-capital growers, typically conducting operations of greater than 25ha, and smaller, less-capitalized farms, typically conducting operations of smaller than 12 ha.

Demographics and farm characteristics

First, respondents from G1 tend to be concentrated in central regions and overall manage larger farms and almost all of the largest farms (25-50 ha, 50+ha), whereas G2 growers are primarily found in the north and south reaches of the surveyed area and are primarily smaller farms (0-5ha, 5-12ha) (Figure 3-5).

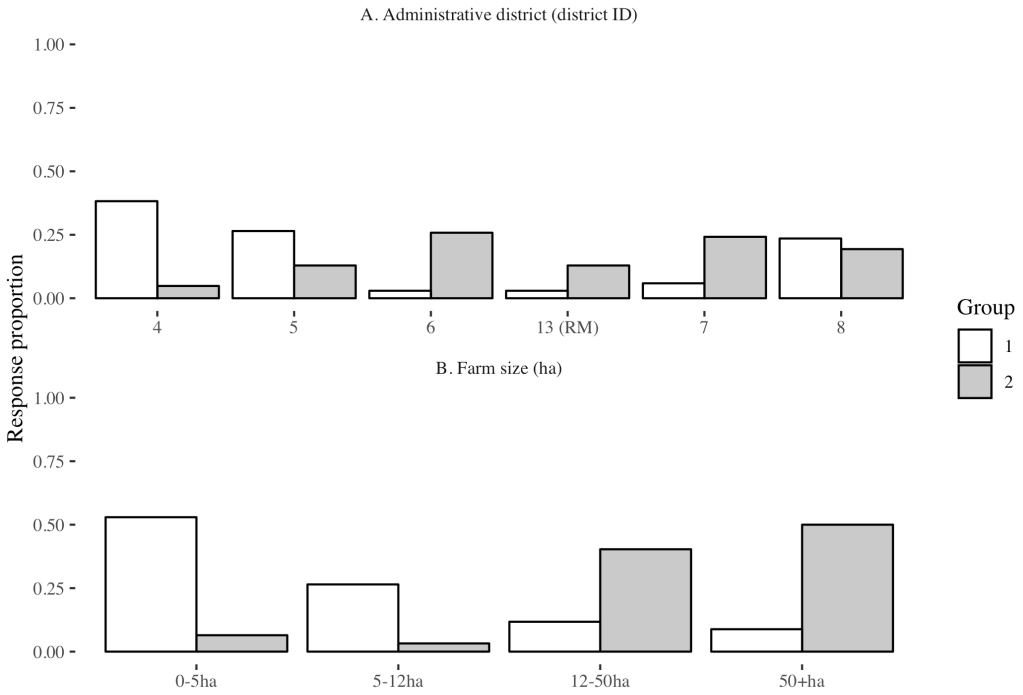


Figure 3-5. Survey response histogram densities. Facet A plots counts of respondents in each group by location in administrative districts. Facet B plots counts of respondents in each group by farm size category. Response counts are presented as empty bars for group 1 and full bars for group 2.

As expected, *Juglans regia* is the near-universal choice for rootstock in both groups, and only a few instances of respondents from G1 who use *Juglans nigra* or Vlach rootstocks (although, conversationally, several growers expressed an interest in obtaining Vlach rootstocks as potentially resistant to Phytophthora). A differentiation in preferred or established scion wood cultivars is present in the data, as G1 growers responded with a higher proportion of "Chandler" cultivars versus G2 growers, who more frequently responded as using "Serr".

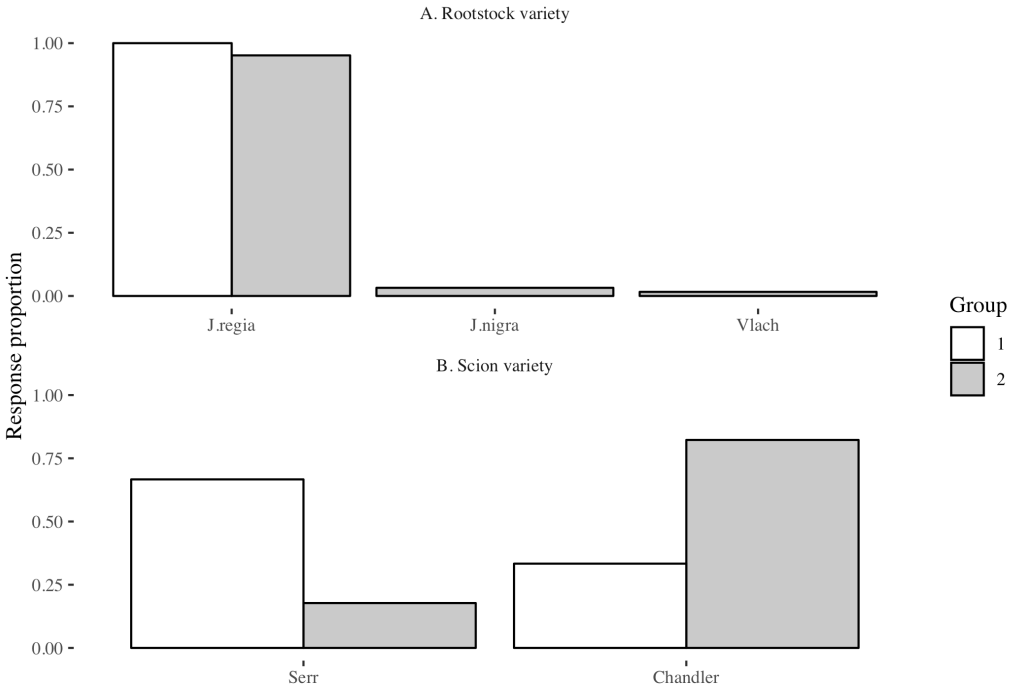


Figure 3-6. Survey response histogram densities. Facet A plots proportions of respondents in each group by their reported plantings of rootstock varieties. Facet B plots proportions of respondents in each group by their reported use of scion varieties. Response proportions are presented as empty bars for group 1 and full bars for group 2.

Examination of density plots indicates differences between groups in terms of productivity, and an overall trend of higher yield in G1 versus G2 respondents.

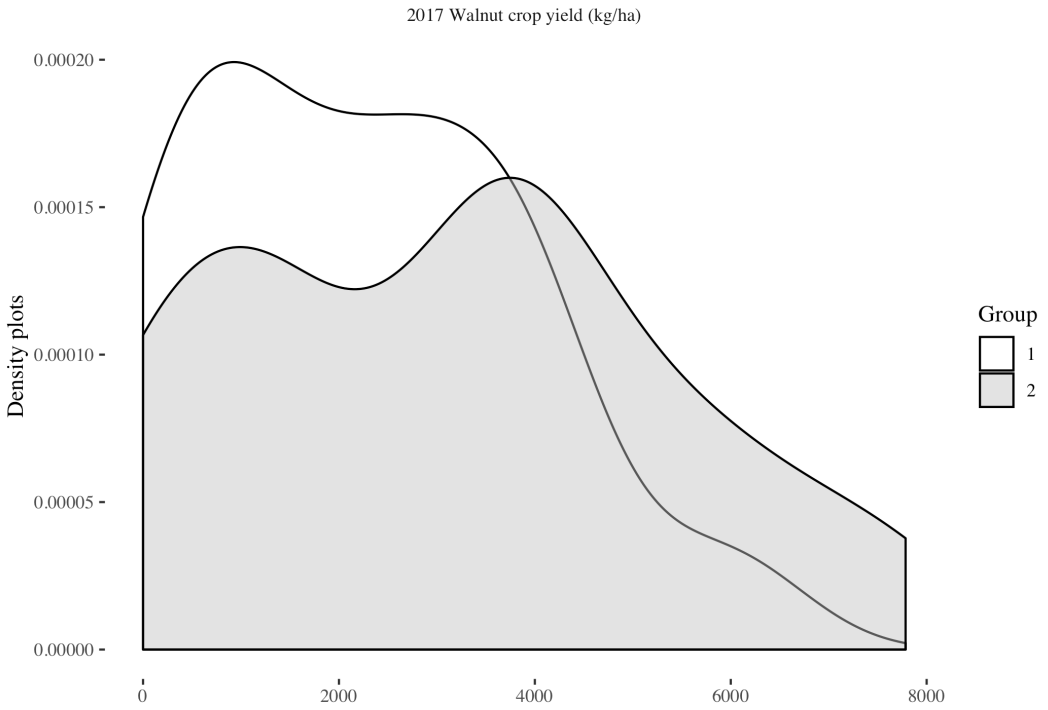


Figure 3-7. Survey response density for crop yield, measured as the estimated kg/ha yield of marketable crop in the 2017 season. Response densities are presented as empty for group 1 and full for group 2.

Disease attitudes and practices

There is not a very strong trend in differences between groups in attitudes towards the severity of Phytophthora as an on-farm issue or the relative efficacy of management solutions to Phytophthora control, although a somewhat reduced belief in Phytophthora as a severe issue and an increased belief in the effectiveness of management solutions may be present in G1 growers (Figure 3-8, facets A and B). A stark difference emerges in the responses addressing preparedness to implement management and technological solutions, as respondents in G2 present as

considerably more pessimistic about their own abilities. Additional differences are evident in the actual strategies respondents take towards control of Phytophthora on their farms (Figure 3-8, facet D), with a majority of G1 growers responding as using chemical strategies or both chemical and management strategies but rarely management on their own, while the majority of G2 growers report using management alone or both strategies, but rarely chemical strategies on their own.

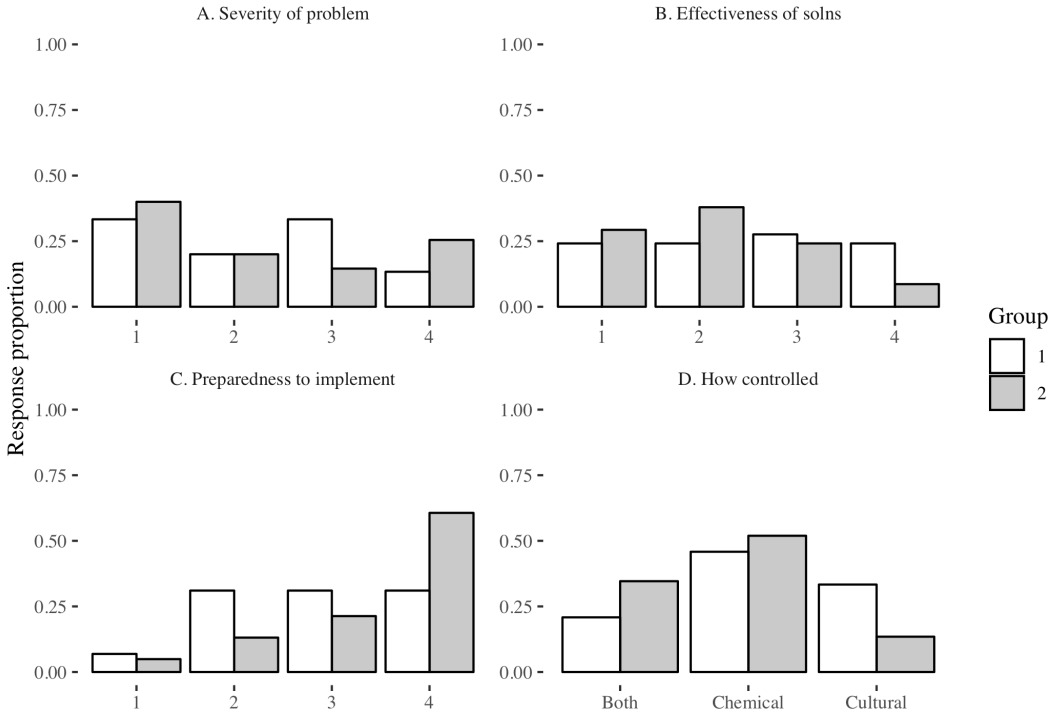


Figure 3-8. Survey response histogram densities. Facets A, B, and C plot proportions of respondents in each group by their response to the questions: A - “How severe do you think the problem of Phytophthora in walnut production is?” (1 Low, 2 Moderate, 3 High, 4 Severe), B - “Do you believe that the solutions that exist today are effective for controlling Phytophthora?” (1 Not effective, 2 Somewhat effective, 3 Moderately

effective, 4 Very effective), C - “How prepared do you feel for implementing these solutions?” (1 Not prepared, 2 Somewhat prepared, 3 Moderately prepared, 4 Very prepared), D - “What kinds of treatment do you use for Phytophthora control?” (1 Cultural, 2 Chemical, 3 Both). Response proportions are presented as empty bars for group 1 and full bars for group 2.

Only slight differences are evident in a series of questions addressing the progression of disease control via a replanting cycle (Figure 3-9), with both groups reporting fairly strong evidence that Phytophthora disease is widespread, growers frequently address tree death via tree removal and replanting in the same soil, only sometimes use disease mitigation strategies before replanting, and these strategies, when used, sometimes fail.

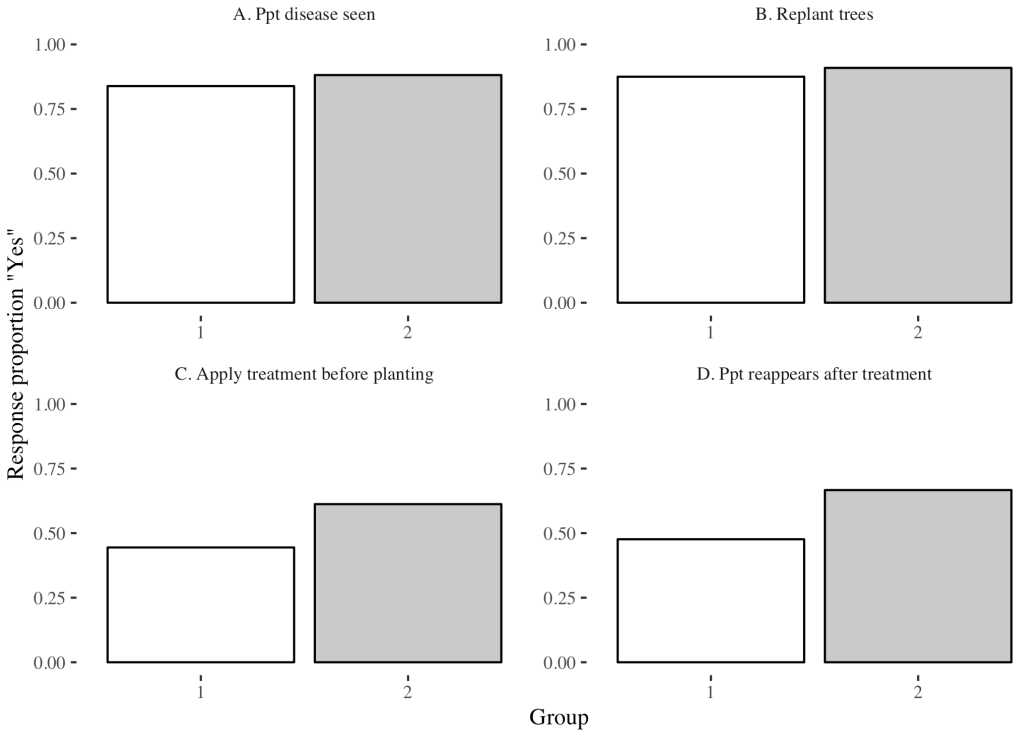


Figure 3-9. Survey binary response histograms, presented as proportions of respondents indicating “Yes” to the following questions: A - “Do you have plants damaged by Phytophthora species?”, B - “Have you had to replant trees after disease-related death in recent years?”, C - “If you replant, do you apply a treatment before planting again?”, D - “If you apply a treatment, does the problem reappear?”. Response proportions are presented as empty bars for group 1 and full bars for group 2.

Irrigation management

Stark differences between groups in their utilization of information for farm management begin to fully emerge in the irrigation data, especially with regards to the use of information systems to monitor and control irrigation (Figure 3-10).

Respondents in G1 overwhelmingly use information-gathering technologies to obtain information about the irrigation states of their farms, and actively use this information to guide decision making on when and how much to irrigate, while fewer than half of respondents in G2 reported any use of information system use.

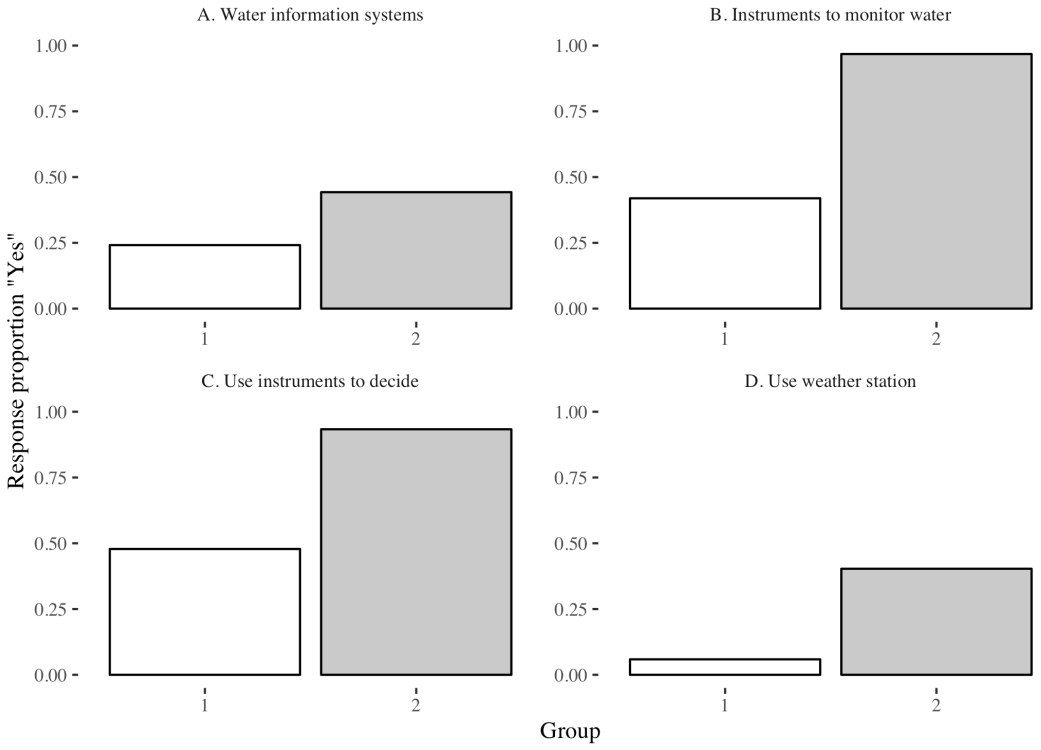


Figure 3-10. Survey binary response histograms, presented as proportions of respondents indicating “Yes” to the following questions: A - “Do you use a system to provide evapotranspiration rates?”, B - “Do you use instruments to monitor irrigation?”, C - “Do you use them to decide when to irrigate?”, D - “Do you use weather station forecasts?”. Response proportions are presented as empty bars for group 1 and full bars for group 2.

Soil and fertility management

Divisions in information-collection patterns continue in responses to soil management questions, as a majority of G1 and minority of G2 respondents use chemical, physical, organic matter, and soil pH testing as information foundations in their management systems (Figure 3-11).

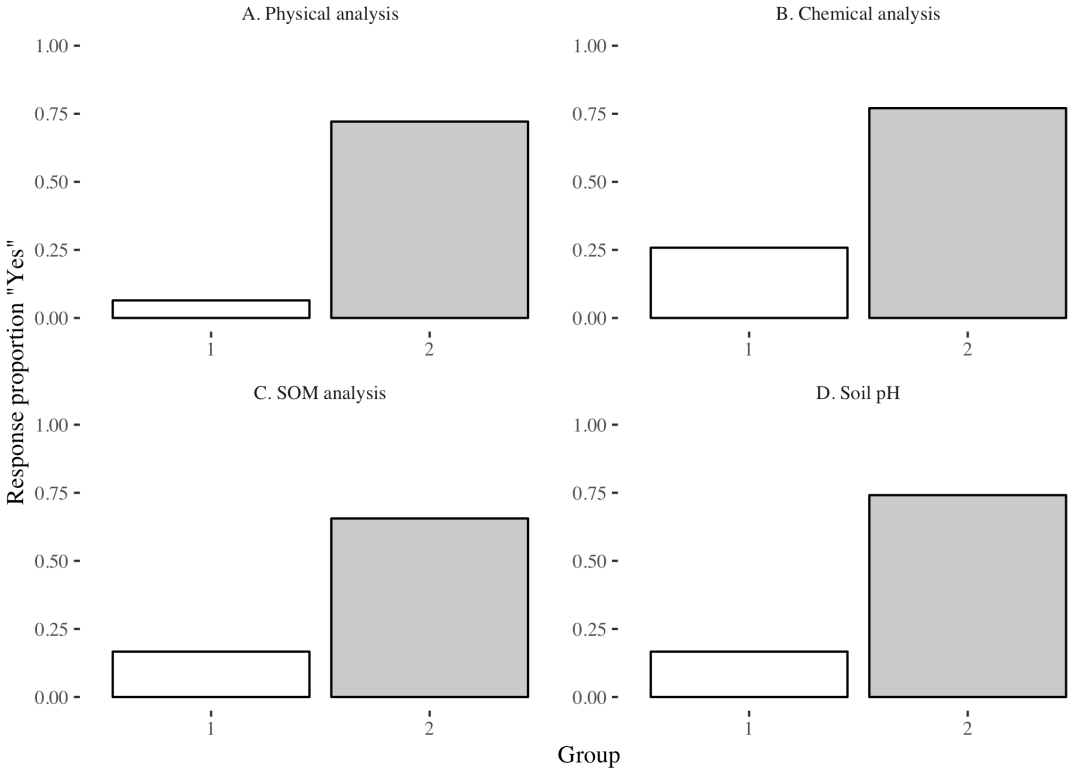


Figure 3-11. Survey binary response histograms, presented as proportions of respondents indicating “Yes” to the following questions: A - “Do you obtain physical analysis information for your soils?”, B - “Do you obtain chemical analysis information for your soils?”, C - “Do you obtain soil organic matter level information

for your soils?", D - "Do you obtain pH level information for your soils?". Response proportions are presented as empty bars for group 1 and full bars for group 2.

While a relatively equal proportion of responses reported using Urea as a nitrogen source, respondents from G1 appear to more frequently use a more diverse source of fertilizers, both as nitrogen inputs and as general organic amendments for soil management (Figure 3-12).

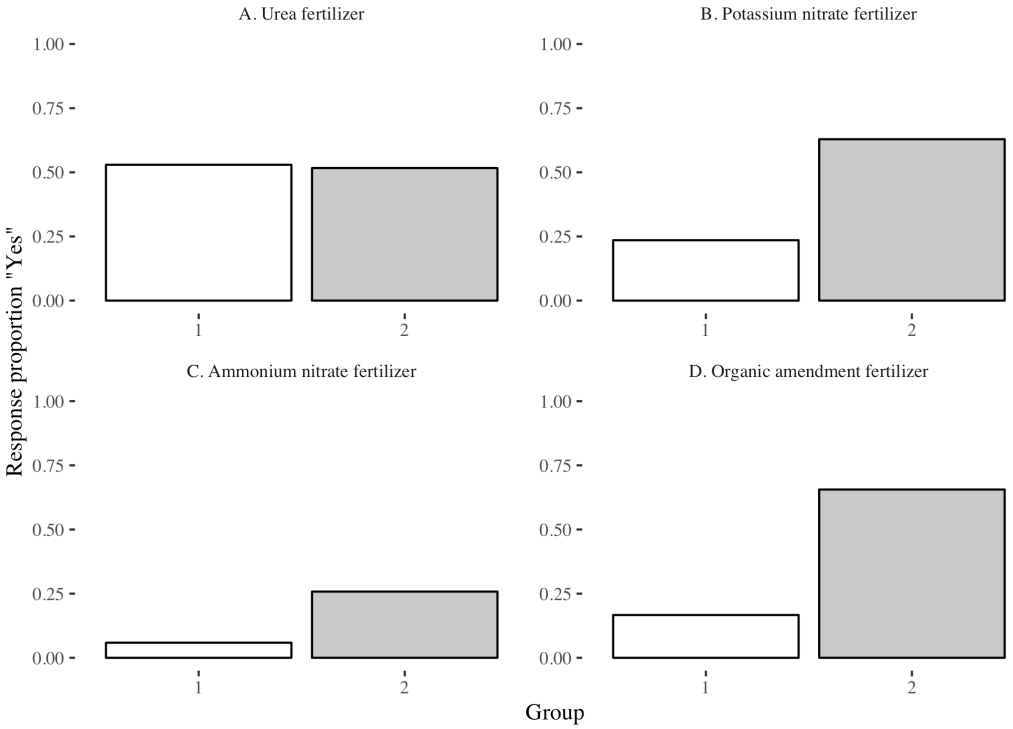


Figure 3-12. Survey binary response histograms, presented as proportions of respondents indicating a positive response to the following questions (responses A, B, and C, were obtained from a multiple-response question) : A, B, C - "Which nitrogen fertilizers do you use as nitrogen inputs?", D - "Do you amend your soils with organic

amendments?". Response proportions are presented as empty bars for group 1 and full bars for group 2.

Information preferences

Across all categories of information sources, respondents in G1 more often reported any answer other than "Never" to questions regarding the frequency of their use of different information sources to guide their decision-making processes. The most frequent information-source across all groups was the use of smartphones or the internet, while some from G2 reported use of journals from information.

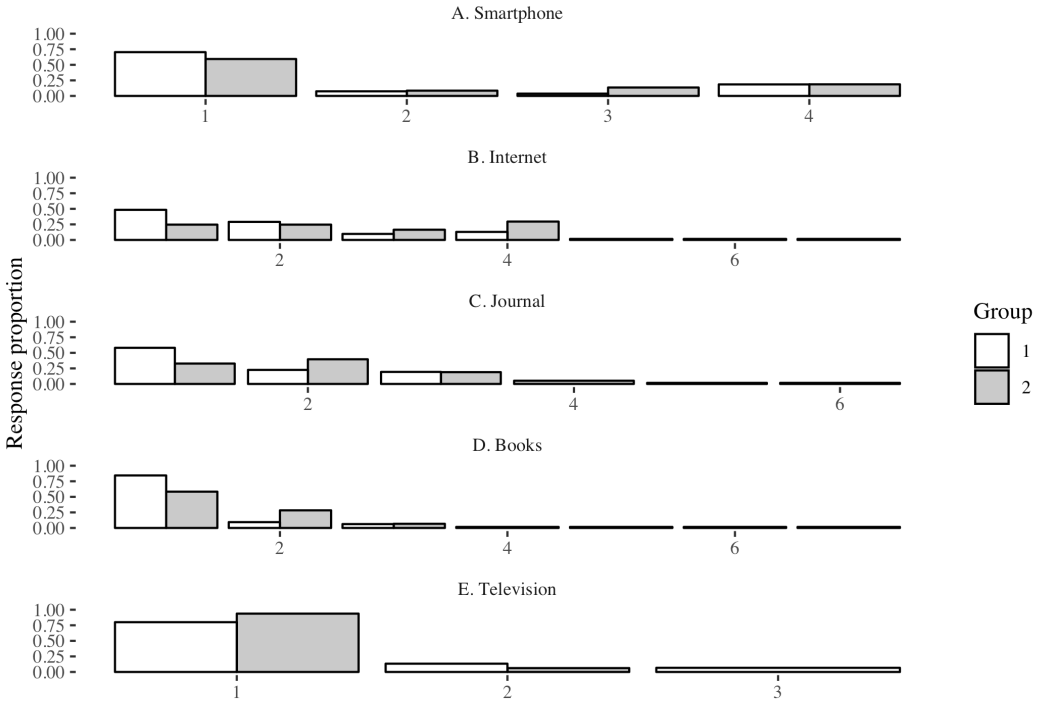


Figure 3-13. Survey histogram densities of the proportion of respondents indicating responses to the question, "How often does your operation use the following source

of information for making decisions about walnut management?": A. Smartphone, B. Internet, C. Journal, D. Books, E. Television. Responses were given on the scale: 1 Never, 2 Less than one time a month, 3 One to two times a month, 4 More than once a week. Response proportions are presented as empty bars for group 1 and full bars for group 2.

In particular, via follow-up questioning, the specific journals "RED Agricola" and "Revista del Campo", extension-funded agronomist publications, were favored by G1 respondents as sources of management information, whereas G2 respondents rarely used either information source.

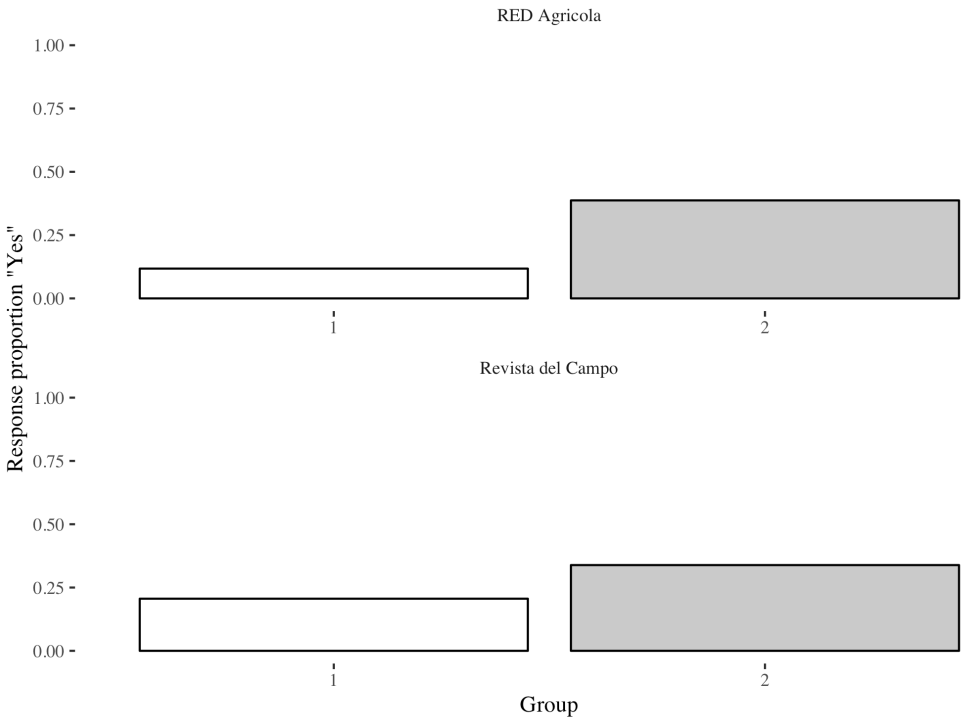


Figure 3-14. Survey binary response histograms indicating responses to the question, "Does your operation use the following journal for making decisions on your farm

about Walnut management?" (decomposed from a multiple-response questions). Response proportions are presented as empty bars for group 1 and full bars for group 2.

Information obtained from direct communication with other persons was most often obtained from consultants by G1 respondents, while government agents / extensionists were most frequently consulted by G2 (although at a lower rate than the utilization of consultants by G1).

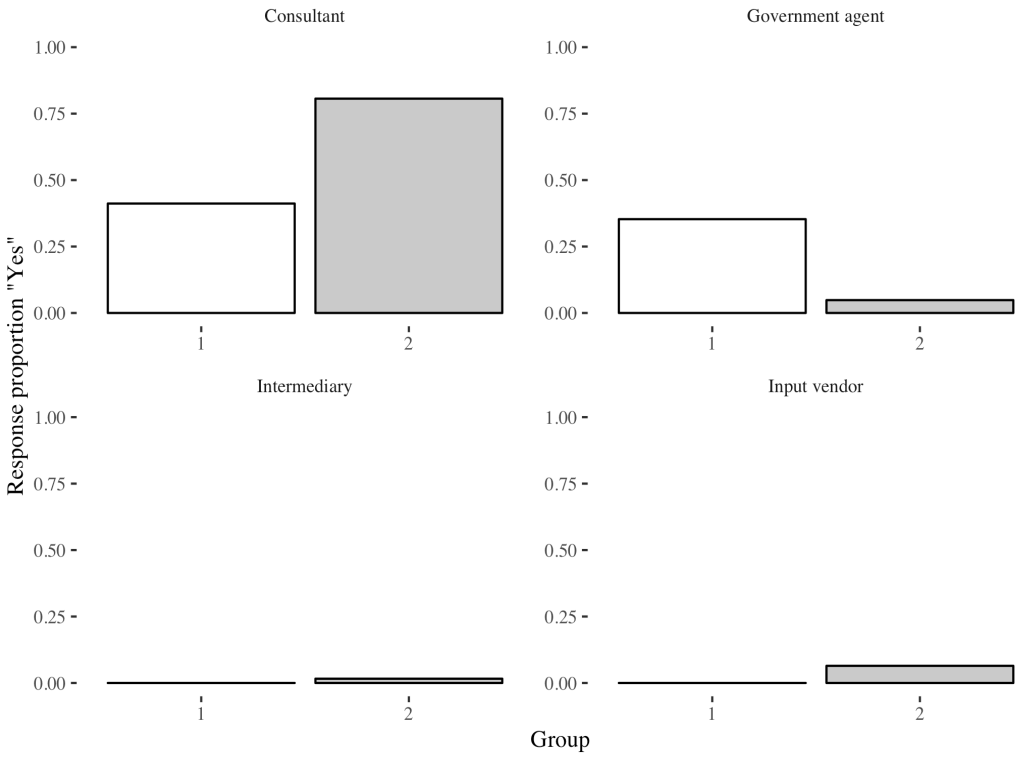


Figure 3-15. Survey binary response histograms, presented as proportions of respondents indicating a positive response to the following questions (responses A, B, C, D, were obtained from a multiple-response question): "If you have a question

about pathogen control, who do you consult?" Response proportions are presented as empty bars for group 1 and full bars for group 2.

Feature importance rankings

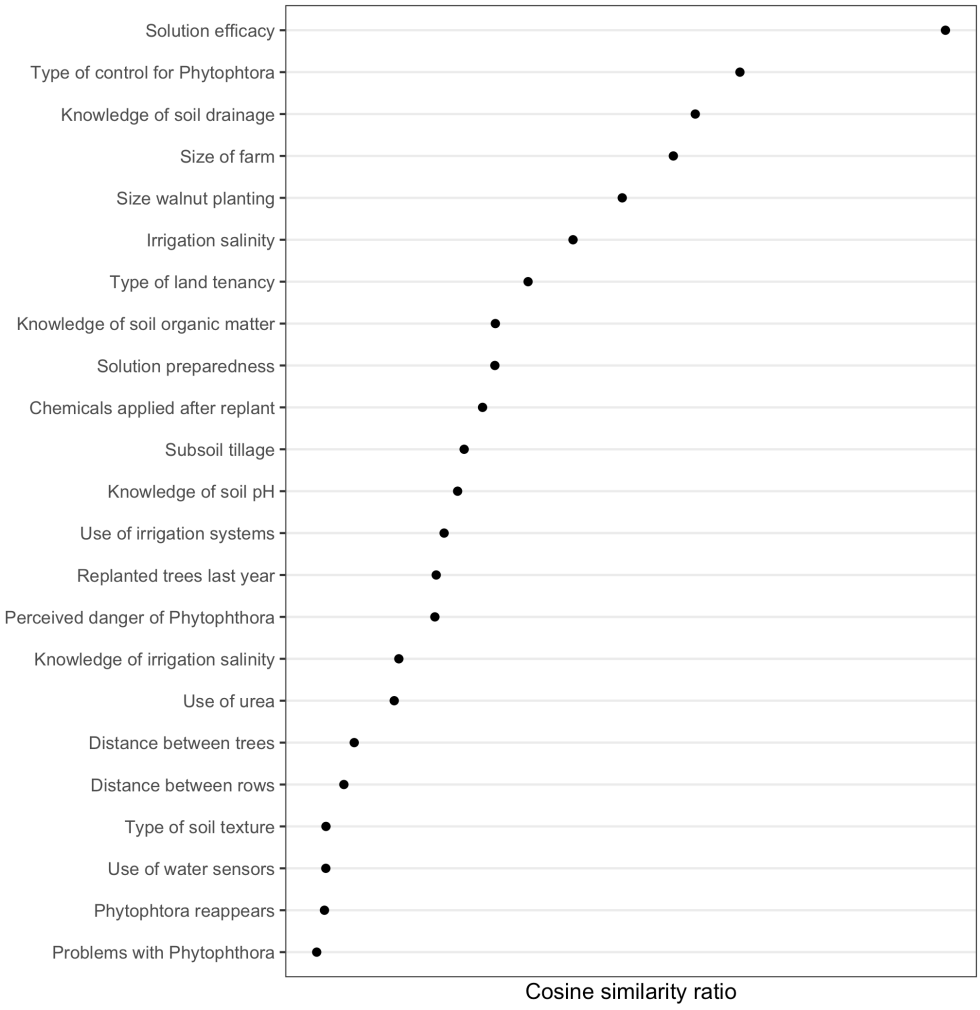


Figure 3-16. Feature importance ranks. Features (y-axis) are plotted against their relative importance (x-axis), as calculated by a cosine similarity ratio. Variables above a cosine similarity ratio of 1 are presented.

Cosine similarity ratios calculated for each feature indicate a cluster of around 20 strongly-defined features, primarily related to management of irrigation, disease control, and use of instrumentation for informing management choices. Use of driplines for irrigation, noted above as a management choice favored by G2 growers, application of Aliette, a counter-Phytophthora prophylaxis, reappearance of Phytophthora issues after replanting, irrigation water source (well versus surface), and the perceived scale of Phytophthora as a pressing issue were highly discriminating factors (Figure 3-16).

Discussion

Segregation of operational characteristics

This strong demographic divide illustrates a point made many times elsewhere in literature studying the modern evolutions of agricultural demographics - not just that there are clear strata in persons and capitalization within farming systems, but that increasingly these strata are being pushed towards opposite ends of small-large farm size and low-high capitalization axes. In consequence, it's particularly evident that, independent of any further data regarding the respondents' agricultural practices or attitudes themselves, the foundational management abilities of each group are likely divergent, whether via access to capital, equipment, or inputs.

Differing attitudes towards information and technology

The clear divisions in information use patterns between groups supports prior work finding that the access to and use of agronomic information can be widely heterogeneous, not just from the point of view of providing information, but from attitudes to whether the information is inherently useful (Garforth et al. 2003). In particular, agronomic systems which most benefit from the input of expert knowledge, whether via inherent system-based knowledge, knowledge obtained from static information sources, or external consultants, may be most vulnerable to stratification based on underlying divisions of information 'ability', as observed in other contexts (Birner and Anderson 2007, Hall et al. 2001).

While both groups sought out the advice of external consultants, G1 tended to utilize private consulting services and G2 growers tended to utilize government extensionists, a subsidized service, suggesting that there may be additional network-level effects on information dissemination and capture by these groups, as prior work has noted differences in behavior and information sourcing in public extensionists versus private consultants (Kidd et al. 2000).

Capitalization and risk

Underlying currents of capitalization-driven separation can be found in many aspects of this study system. At the outset, the generally lower land size, lower yields, and higher use of external supplemental income reasonably places respondents in the G2 cluster in less-capitalized systems. Unsurprisingly, growers in this system tended to

report a lower and less diverse use of fertility inputs, chemical inputs, irrigation monitoring technology, advanced irrigation application systems, and fewer analytical measures to evaluate soil conditions.

All of these features are well-established as behavioral characteristics of growers with access to more capital, whereby land managers with lower access to resources and land-improvement capital are more likely to make management choices that reduce or fail to increase their land productivity or stimulate negative feedback loops in their system's biologies (Daberkow and McBride 2003).

Potential consequences for extension activity

These findings bring strong focus to the imperative on extensionists and researchers to consider the socioeconomic conditions of agricultural systems targeted for technology or information transfer, and agricultural research more generally. Clustering patterns in this system definitively outline divisions in attitudes, resource access, and information channels between the two farm types.

A broader recognition is underway, especially among those in the agricultural sciences, of the responsibility of the scientific community to consider their role as participants within socio-environmental systems, and especially of the role scientific research and outreach activities can play in shaping these systems in a dynamic way (Ashby and Sperling 1995, MacMillan and Benton 2014). Case studies demonstrating the inadvertent shaping of economic systems by land-grant universities in California illustrate this point in agricultural contexts, where the relationship between scientific

communities and land management practitioners can be particularly strong (George and Clawson 2014, Chatterjee, Dinar, and González-Rivera 2016).

The defined groups discovered by survey in this work provide an opportunity to consider how technology transfer activities should proceed. While the problems faced by growers in finding effective and long-term solutions to Phytophthora management in their systems are significant, and basic scientific work should be conducted on potential strategies for control and mitigation of pathogens, it is evident from this work that a segment of the industry is better-prepared to integrate new solutions and technologies into their management systems than another.

Further, unconsidered release of new technologies via means or channels most convenient to research institutions may be implicitly biased towards supporting more-technologized growers over others, as evidenced by the divisions in preferential use of both technology and information for management decisions among growers in this system. If current issues with diseases such as Phytophthora worsen, perhaps along with other environ- and climate-dependent pathogens, these divisions may be exacerbated. Synergistic interactions between shifting climactic factors and pathogen-pathogen interactions or use of opportunistic infection opportunities may pose continuing issues to growers' abilities to manage their orchards, especially under conditions that limit capital or infrastructural investments.

Appendices – Chapter 1

Appendix 1: Original computing environment

This document and all figures, tables, and supporting analyses were generated using R Statistical Software and RMarkdown. The generative RMarkdown file and associated R scripts for data manipulation and analysis are available at https://github.com/graemebaird/orei_ucsc. As reproducibility best practice, the computing environment used to generate this document is detailed below.

R version 3.5.1 (2018-07-02)

Platform: x86_64-apple-darwin15.6.0 (64-bit)

attached base packages: *stats4*, *grid*, *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *bindrcpp(v.0.2.2)*, *ggcorrplot(v.0.1.2)*, *DALEX(v.0.2.4)*, *shadowtext(v.0.0.4)*, *partykit(v.1.2-2)*, *libcoin(v.1.0-1)*, *party(v.1.3-1)*, *strucchange(v.1.5-1)*, *sandwich(v.2.5-0)*, *zoo(v.1.8-4)*, *modeltools(v.0.2-22)*, *mvtnorm(v.1.0-8)*, *forcats(v.0.3.0)*, *stringr(v.1.3.1)*, *purrr(v.0.2.5)*, *readr(v.1.1.1)*, *tidyr(v.0.8.1)*, *tibble(v.1.4.2)*, *tidyverse(v.1.2.1)*, *dplyr(v.0.7.6)*, *ggthemes(v.4.0.1)*, *mapdata(v.2.3.0)*, *maps(v.3.3.0)*, *ggmap(v.2.7.900)*, *pander(v.0.6.3)*, *shapleyR(v.0.1)*, *reshape2(v.1.4.3)*, *ggplot2(v.3.1.0)*, *combinat(v.0.0-8)*, *checkmate(v.1.8.5)*, *mlr(v.2.13)*, *ParamHelpers(v.1.11)*, *magrittr(v.1.5)* and *openxlsx(v.4.1.0)*

loaded via a namespace (and not attached): *TH.data(v.1.0-9)*, *colorspace(v.1.3-2)*, *deldir(v.0.1-15)*, *rjson(v.0.2.20)*, *rprojroot(v.1.3-2)*, *proxy(v.0.4-22)*, *yaImpute(v.1.0-30)*, *rstudioapi(v.0.7)*, *ggpubr(v.0.1.8)*, *lubridate(v.1.7.4)*, *coin(v.1.2-2)*, *xml2(v.1.2.0)*, *codetools(v.0.2-15)*, *splines(v.3.5.1)*, *knitr(v.1.20)*, *Formula(v.1.2-3)*, *jsonlite(v.1.5)*, *broom(v.0.5.0)*, *cluster(v.2.0.7-1)*, *png(v.0.1-7)*, *shiny(v.1.1.0)*, *compiler(v.3.5.1)*, *httr(v.1.3.1)*, *backports(v.1.1.2)*, *assertthat(v.0.2.0)*, *Matrix(v.1.2-14)*, *lazyeval(v.0.2.1)*, *cli(v.1.0.0)*, *later(v.0.7.5)*, *htmltools(v.0.3.6)*, *tools(v.3.5.1)*, *coda(v.0.19-2)*, *gtable(v.0.2.0)*, *agricolae(v.1.2-8)*, *glue(v.1.3.0)*, *gmodels(v.2.18.1)*, *fastmatch(v.1.1-0)*, *Rcpp(v.1.0.0)*, *parallelMap(v.1.3)*, *cellranger(v.1.1.0)*, *spdep(v.0.8-1)*, *gdata(v.2.18.0)*, *nlme(v.3.1-137)*, *inum(v.1.0-0)*, *rvest(v.0.3.2)*, *mime(v.0.5)*, *miniUI(v.0.1.1.1)*, *breakDown(v.0.1.6)*, *gtools(v.3.8.1)*, *XML(v.3.98-1.16)*, *LearnBayes(v.2.15.1)*, *MASS(v.7.3-51.1)*, *scales(v.1.0.0)*, *promises(v.1.0.1)*, *hms(v.0.4.2)*, *parallel(v.3.5.1)*, *expm(v.0.999-3)*, *RColorBrewer(v.1.1-2)*, *BBmisc(v.1.11)*, *yaml(v.2.2.0)*, *gridExtra(v.2.3)*, *rpart(v.4.1-13)*, *stringi(v.1.2.4)*, *AlgDesign(v.1.1-7.3)*, *highr(v.0.7)*, *klaR(v.0.6-14)*, *boot(v.1.3-20)*, *zip(v.1.0.0)*, *spData(v.0.2.9.4)*, *RgoogleMaps(v.1.4.2)*, *rlang(v.0.3.0.1)*, *pkgconfig(v.2.0.2)*, *bitops(v.1.0-6)*, *evaluate(v.0.11)*, *lattice(v.0.20-35)*, *bindr(v.0.1.1)*, *labeling(v.0.3)*, *tidyselect(v.0.2.4)*, *factorMerger(v.0.3.6)*, *plyr(v.1.8.4)*, *R6(v.2.2.2)*, *multcomp(v.1.4-8)*, *ALEPlot(v.1.1)*, *pillar(v.1.3.0)*, *haven(v.2.0.0)*, *withr(v.2.1.2)*, *sp(v.1.3-1)*, *survival(v.2.43-1)*, *modelr(v.0.1.2)*, *crayon(v.1.3.4)*, *questionr(v.0.7.0)*, *rmarkdown(v.1.10)*, *jpeg(v.0.1-8)*, *readxl(v.1.1.0)*, *data.table(v.1.11.4)*, *pdp(v.0.7.0)*, *digest(v.0.6.17)*, *xtable(v.1.8-3)*, *httpuv(v.1.4.5)* and *munsell(v.0.5.0)*

Appendix 2: Sampling dates

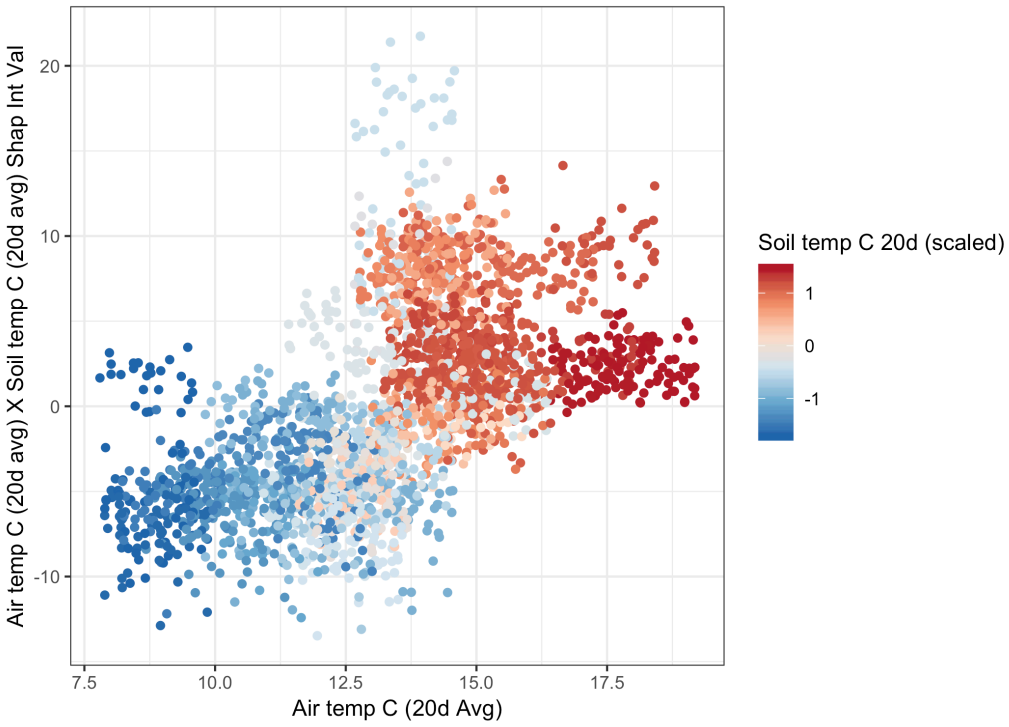
	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8a	8b
2011-11-01	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2011-12-15	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2012-01-27	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2012-03-05	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2012-04-26	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2012-07-03	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2012-07-17	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2012-08-01	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2012-08-15	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2012-08-29	2	4	4	3	2	4	4	4	0	0	0	0	0	0	0	0
2012-09-14	4	4	4	4	4	4	4	4	0	0	0	0	0	0	0	0
2012-10-04	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2012-11-09	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4
2012-12-13	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2013-01-30	4	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4
2013-03-21	4	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4
2013-04-19	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2013-05-15	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0

2013-05-22	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2013-06-04	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0
2013-06-27	3	0	4	0	3	0	4	0	4	0	4	0	4	0	3	0
2013-08-15	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0
2013-09-09	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4
2013-10-16	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0
2014-01-27	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2014-02-25	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2014-04-23	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2014-06-05	4	4	4	2	4	4	4	4	4	3	4	4	4	4	4	4
2014-06-20	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2014-07-01	4	3	4	4	4	4	4	3	4	4	3	3	4	4	4	4
2014-07-18	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4
2014-08-04	4	4	4	4	4	4	4	4	0	0	0	0	0	0	0	0
2014-09-10	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4	4
2014-11-17	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2015-01-23	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2015-02-27	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2015-03-23	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4

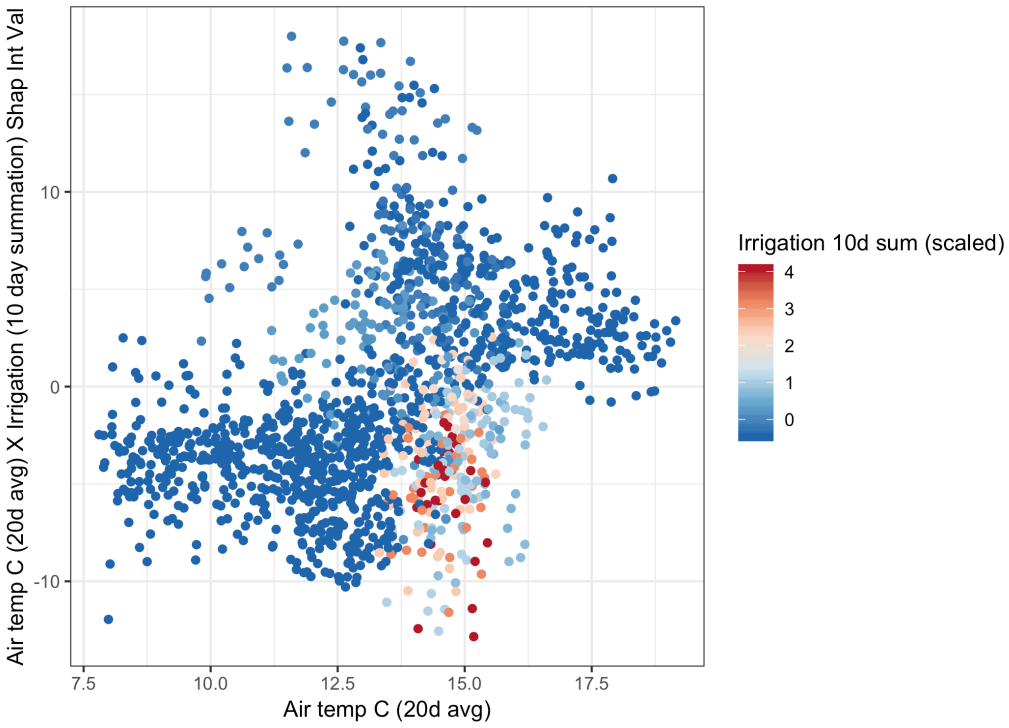
2015-04-29	4	4	4	4	4	4	4	4	4	4	4	4	4	3	4	4
2015-05-27	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2015-06-26	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4
2015-07-21	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2015-09-04	3	4	4	4	4	4	4	4	4	3	4	4	4	4	4	4
2015-11-12	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2016-01-27	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2016-02-24	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4
2016-03-23	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2016-04-13	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4

Table provides observations per date, per treatment.

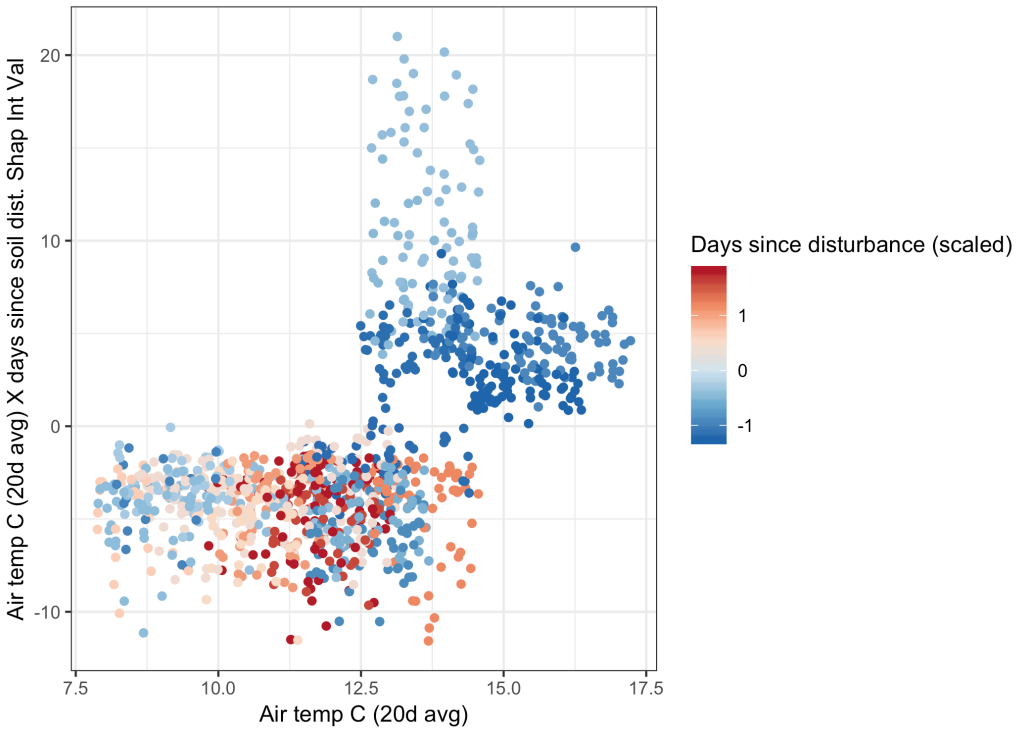
Appendix 3: Additional Shapley value visualizations



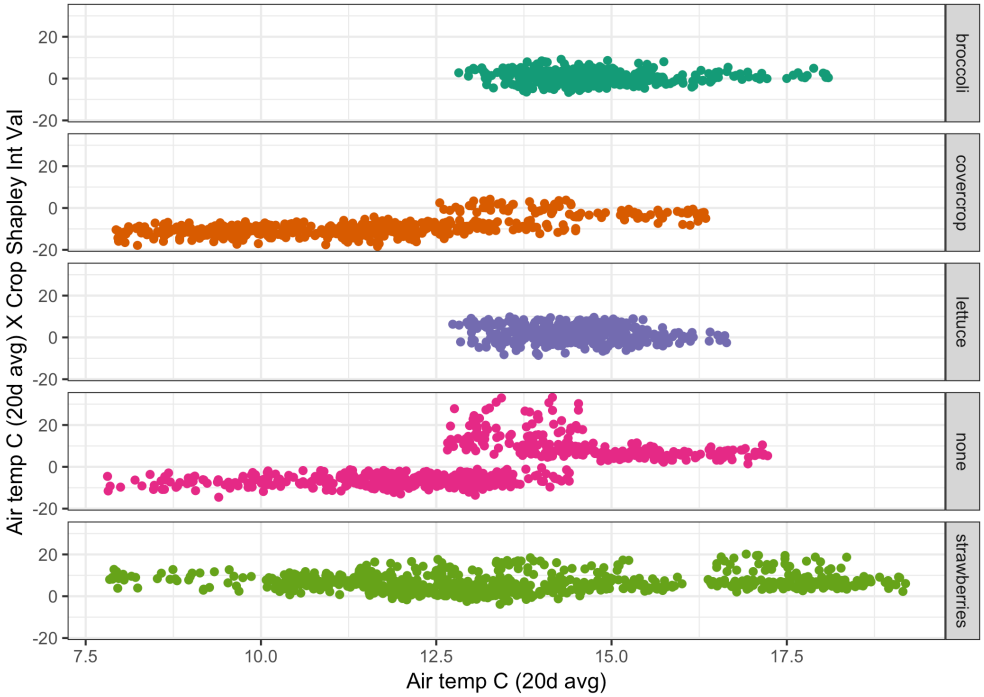
Shapley values (x axis, shading) and interaction Shapley values (the sum of two values, y-axis) plotted.



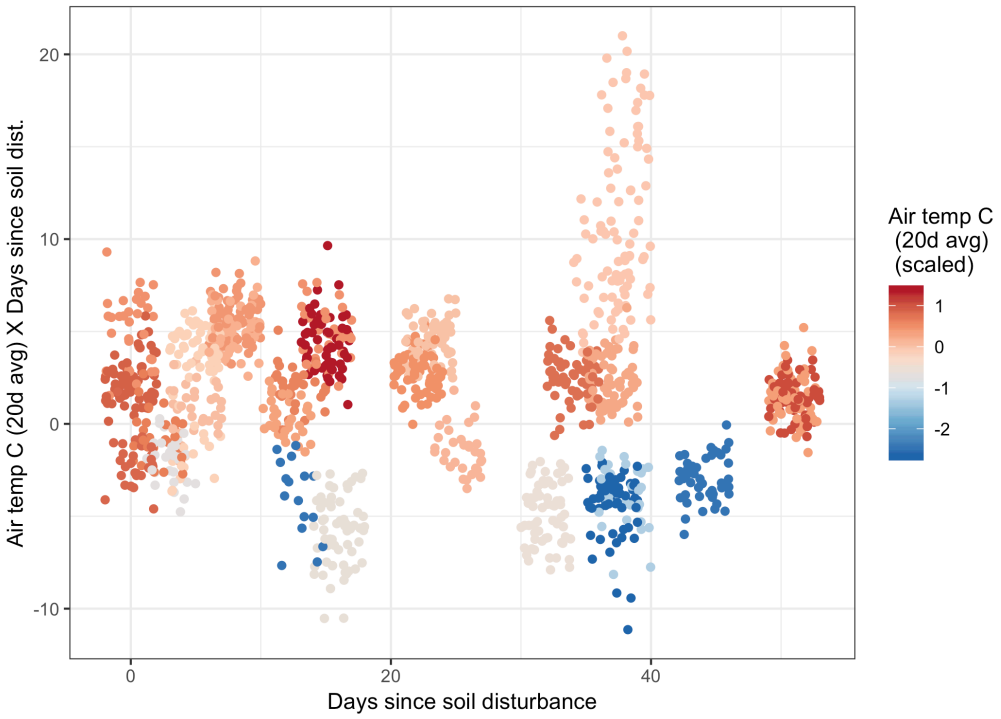
Shapley values (x axis, shading) and interaction Shapley values (the sum of two values, y-axis) plotted.



Shapley values (x axis, shading) and interaction Shapley values (the sum of two values, y-axis) plotted.

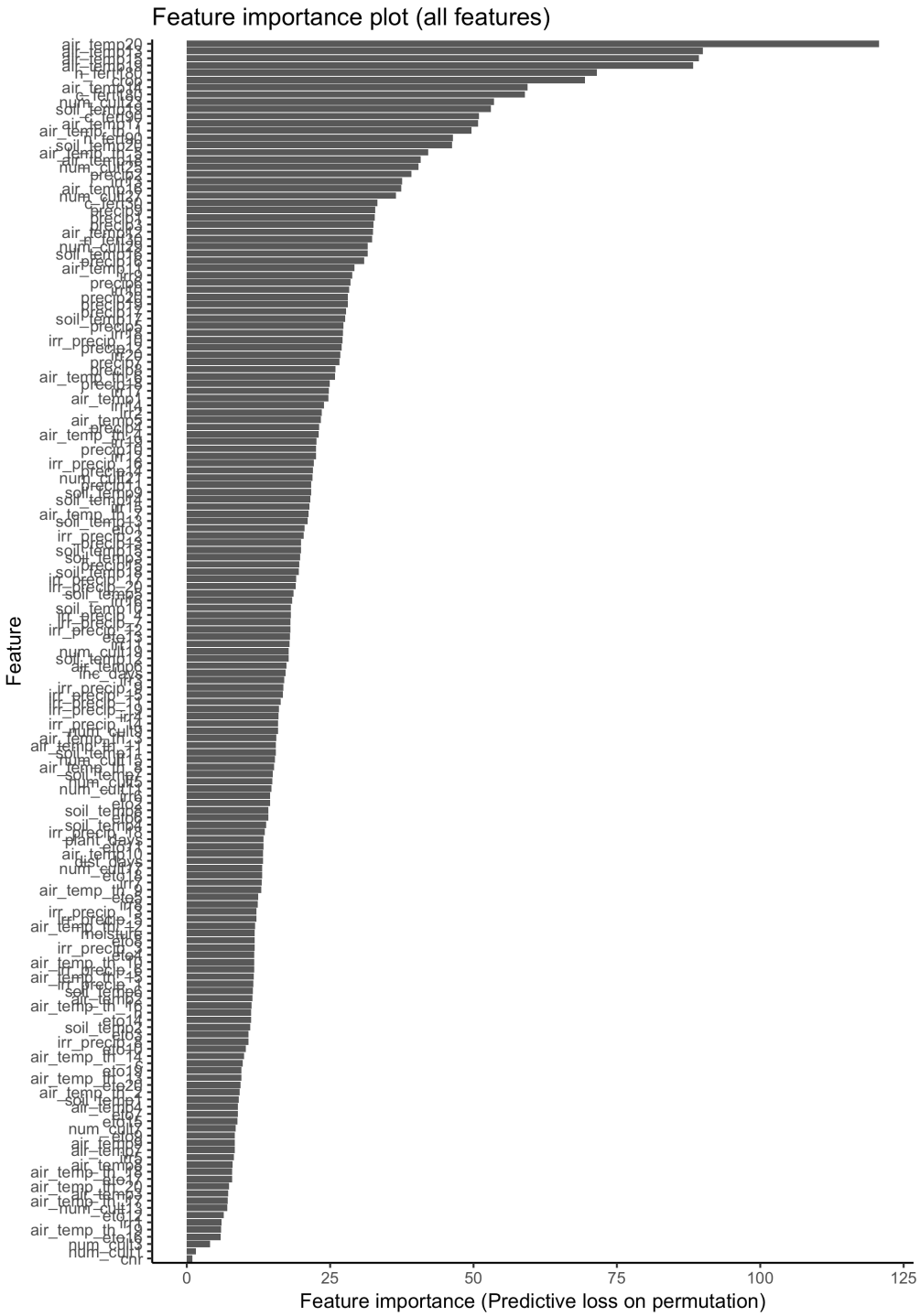


Shapley values (x axis) and interaction Shapley values (the sum of two values, y-axis) plotted, faceted by crop.



Shapley values (x axis, shading) and interaction Shapley values (the sum of two values, y-axis) plotted.

Appendix 4: Full RF feature importance



Feature importances, measured as error loss on permutation, for the entire suite of

original and transformed parameters evaluated in the RF model fit. Higher values indicate a greater importance of the parameters to the overall model.

Appendix 5: RF residuals

The majority of RF model residuals reside within a region of <10 mg NO3 / kg dry soil, and 90% of predictions capture a range of <25 mg NO3 / kg dry soil. The largest residuals, in excess of 100 mg NO3 / kg dry soil, occur when large mineralization peaks in data are entirely uncaptured by smoother estimates produced by RF predictions.

Appendices - Chapter 2

Appendix 1: Original computing environment

This document and all figures, tables, and supporting analyses were generated using R Statistical Software and RMarkdown. The generative RMarkdown file and associated R scripts for data manipulation and analysis are available at https://github.com/graemebaird/asd_syn. As reproducibility best practice, the computing environment used to generate this document is detailed below.

R version 3.5.1 (2018-07-02)

Platform: x86_64-apple-darwin15.6.0 (64-bit)

attached base packages: *grid*, *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *bindrcpp(v.0.2.2)*, *VIM(v.4.7.0)*, *data.table(v.1.11.4)*, *colorspace(v.1.3-2)*, *readr(v.1.1.1)*, *bnlearn(v.4.4)*, *bnstruct(v.1.0.4)*, *igraph(v.1.2.2)*, *Matrix(v.1.2-14)*, *bitops(v.1.0-6)*, *rgdal(v.1.3-4)*, *sp(v.1.3-1)*, *mice(v.3.3.0)*, *lattice(v.0.20-35)*, *reshape(v.0.8.8)*, *janitor(v.1.1.1)*, *dplyr(v.0.7.6)*, *ggthemes(v.4.0.1)*, *mapdata(v.2.3.0)*, *maps(v.3.3.0)*, *ggmap(v.2.7.900)*, *pander(v.0.6.3)*, *shapleyR(v.0.1)*, *reshape2(v.1.4.3)*, *ggplot2(v.3.1.0)*, *combinat(v.0.0-8)*, *checkmate(v.1.8.5)*, *mlr(v.2.13)*, *ParamHelpers(v.1.11)*, *magrittr(v.1.5)* and *openxlsx(v.4.1.0)*

loaded via a namespace (and not attached): *minqa(v.1.2.4)*, *rjson(v.0.2.20)*, *class(v.7.3-14)*, *rio(v.0.5.10)*, *rprojroot(v.1.3-2)*, *codetools(v.0.2-15)*, *splines(v.3.5.1)*,

robustbase(v.0.93-3), *knitr(v.1.20)*, *nloptr(v.1.0.4)*, *broom(v.0.5.0)*, *png(v.0.1-7)*, *rgeos(v.0.3-28)*, *graph(v.1.60.0)*, *compiler(v.3.5.1)*, *backports(v.1.1.2)*, *assertthat(v.0.2.0)*, *lazyeval(v.0.2.1)*, *htmltools(v.0.3.6)*, *tools(v.3.5.1)*, *gtable(v.0.2.0)*, *glue(v.1.3.0)*, *fastmatch(v.1.1-0)*, *Rcpp(v.1.0.0)*, *parallelMap(v.1.3)*, *carData(v.3.0-2)*, *cellranger(v.1.1.0)*, *nlme(v.3.1-137)*, *lmtest(v.0.9-36)*, *laeken(v.0.5.0)*, *stringr(v.1.3.1)*, *lme4(v.1.1-18-1)*, *pan(v.1.6)*, *DEoptimR(v.1.0-8)*, *MASS(v.7.3-51.1)*, *zoo(v.1.8-4)*, *scales(v.1.0.0)*, *hms(v.0.4.2)*, *parallel(v.3.5.1)*, *BBmisc(v.1.11)*, *yaml(v.2.2.0)*, *curl(v.3.2)*, *gridExtra(v.2.3)*, *rpart(v.4.1-13)*, *stringi(v.1.2.4)*, *randomForest(v.4.6-14)*, *e1071(v.1.7-0)*, *BiocGenerics(v.0.28.0)*, *boot(v.1.3-20)*, *zip(v.1.0.0)*, *RgoogleMaps(v.1.4.2)*, *rlang(v.0.3.0.1)*, *pkgconfig(v.2.0.2)*, *evaluate(v.0.11)*, *purrr(v.0.2.5)*, *bindr(v.0.1.1)*, *labeling(v.0.3)*, *tidyselect(v.0.2.4)*, *plyr(v.1.8.4)*, *R6(v.2.2.2)*, *mitml(v.0.3-6)*, *pillar(v.1.3.0)*, *haven(v.2.0.0)*, *foreign(v.0.8-70)*, *withr(v.2.1.2)*, *survival(v.2.43-1)*, *abind(v.1.4-5)*, *nnet(v.7.3-12)*, *tibble(v.1.4.2)*, *crayon(v.1.3.4)*, *car(v.3.0-2)*, *jomo(v.2.6-4)*, *rmarkdown(v.1.10)*, *jpeg(v.0.1-8)*, *readxl(v.1.1.0)*, *Rgraphviz(v.2.26.0)*, *forcats(v.0.3.0)*, *vcd(v.1.4-4)*, *digest(v.0.6.17)*, *tidyr(v.0.8.1)*, *stats4(v.3.5.1)* and *munsell(v.0.5.0)*

Appendix 2: Structure: weighted partially directed acyclic graph

Table continues below

	Diam	Yield	Wilt	Vert soil CFU	Soil Temp Thr	Soil Temp
Diam	0	100	100	2	0	0

Yield	0	0	0	0	0	0
Wilt	0	100	0	1	0	0
Vert soil CFU	0	7	3	0	0	0
Soil Temp Thr	89	100	100	0	0	0
Soil Temp	100	100	100	0	0	0
Soil H2O Avg	100	88	28	2	0	0
Soil H2O	100	11	100	2	0	0
Eh-h >200mV	100	0	100	3	0	0
Rice bran	0	0	0	0	0	0
Cover Crop	0	0	0	0	0	0
Molass	0	0	0	0	0	0
Vert Plant	0	100	40	0	0	0
Must cake	0	0	0	0	0	0
Vert Inf %	100	12	100	0	0	0
Vert suppr	0	0	0	0	0	0
Carbon rate	11	82	28	0	0	0
Soil Temp X C	0	0	100	3	0	0
Soil Temp X Eh	0	0	100	1	0	0

Soil T Thr X C	0	0	100	0	0	0

Table continues below

	Soil H2O Avg	Soil H2O	Eh-h >200mV	Rice bran	Cover Crop	Molass
Diam	0	0	0	0	0	0
Yield	0	0	0	0	0	0
Wilt	0	0	0	0	0	0
Vert soil CFU	0	0	0	0	0	0
Soil Temp Thr	0	0	0	0	0	0
Soil Temp	0	0	1	0	0	0
Soil H2O Avg	0	0	89	0	0	0
Soil H2O	0	0	15	0	0	0
Eh-h >200mV	0	0	0	0	0	0
Rice bran	0	0	0	0	0	0
Cover Crop	0	0	0	0	0	0
Molass	0	0	0	0	0	0

Vert Plant	0	0	0	0	0	0
Must cake	0	0	0	0	0	0
Vert Inf %	0	0	0	0	0	0
Vert suppr	0	0	3	0	0	0
Carbon rate	0	0	4	0	0	0
Soil Temp X C	0	0	4	0	0	0
Soil Temp X Eh	0	0	99	0	0	0
Soil T Thr X C	0	0	0	0	0	0

Table continues below

	Vert Plant	Must cake	Vert Inf %	Vert suppr	Carbon rate
Diam	0	0	0	0	0
Yield	0	0	0	0	0
Wilt	0	0	0	0	0
Vert soil CFU	0	0	0	0	0

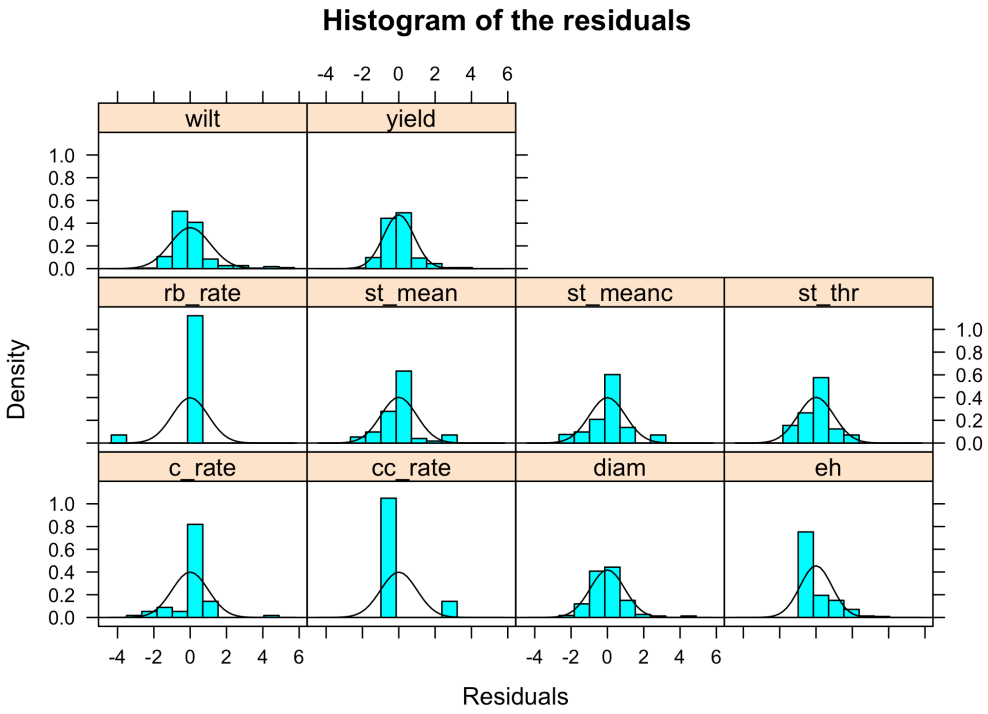
Soil Temp Thr	0	0	0	0	0
Soil Temp	0	0	0	58	0
Soil H2O Avg	0	0	0	0	0
Soil H2O	0	0	0	0	0
Eh-h >200mV	0	0	0	0	0
Rice bran	0	0	0	0	0
Cover Crop	0	0	0	0	0
Molass	0	0	0	0	0
Vert Plant	0	0	0	0	0
Must cake	0	0	0	0	0
Vert Inf %	0	0	0	0	0
Vert suppr	0	0	0	0	0
Carbon rate	0	0	0	0	0
Soil Temp X C	0	0	0	0	0
Soil Temp X Eh	0	0	0	14	0

Soil T Thr X C	0	0	0	0	0

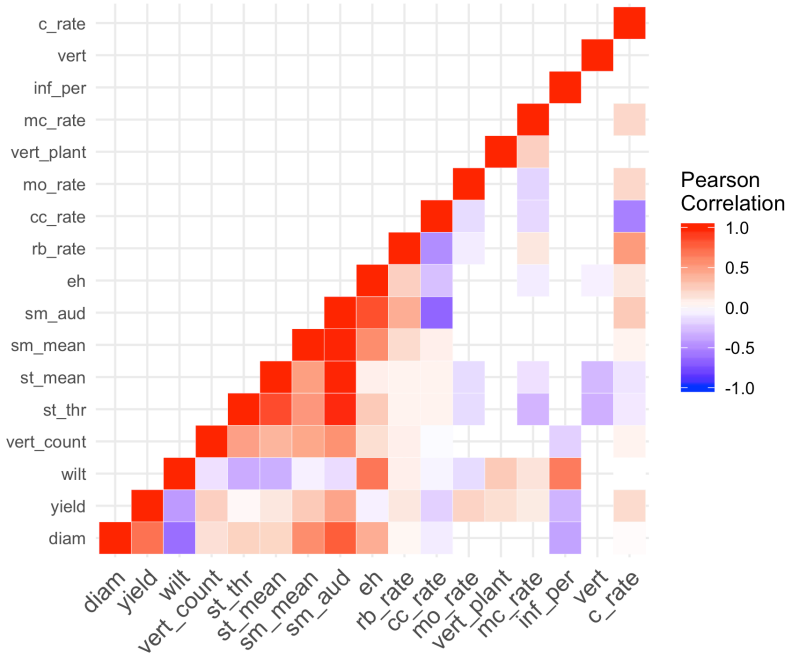
	Soil Temp X C	Soil Temp X Eh	Soil T Thr X C
Diam	0	0	0
Yield	0	0	0
Wilt	0	0	0
Vert soil CFU	0	0	0
Soil Temp Thr	0	0	0
Soil Temp	0	0	0
Soil H2O Avg	0	0	0
Soil H2O	0	0	0
Eh-h >200mV	0	0	0
Rice bran	0	0	0
Cover Crop	0	0	0
Molass	0	0	0
Vert Plant	0	0	0

Must cake	0	0	0
Vert Inf %	0	0	0
Vert suppr	0	0	0
Carbon rate	0	0	0
Soil Temp X C	0	0	0
Soil Temp X Eh	0	0	0
Soil T Thr X C	0	0	0

Appendix 3: Gaussian Bayesian diagnostics

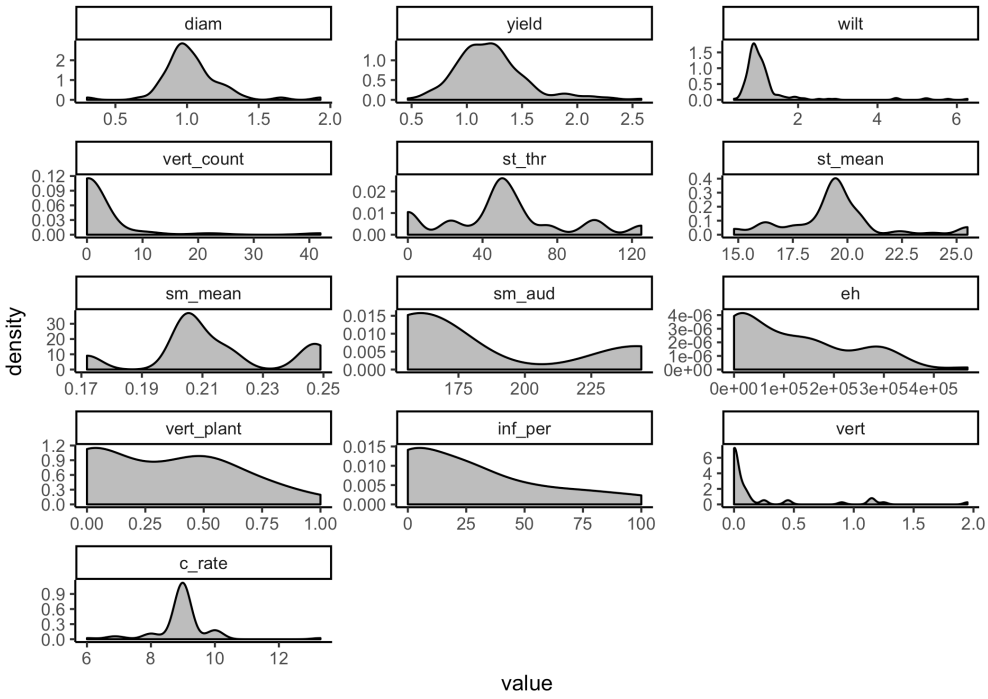


Appendix 4: Raw data correlations



Variable correlations.

Appendix 5



Variable densities.

Appendix Ch3

Appendix 1: Original computing environment

This document and all figures, tables, and supporting analyses were generated using R Statistical Software and RMarkdown. The generative RMarkdown file and associated R scripts for data manipulation and analysis are available at https://github.com/graemebaird/pucv_nogal (Ed note: repository not public yet). As reproducibility best practice, the computing environment used to generate this document is detailed below.

R version 3.5.1 (2018-07-02)

Platform: x86_64-apple-darwin15.6.0 (64-bit)

attached base packages: *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *bindrcpp(v.0.2.2)*, *shadowtext(v.0.0.4)*, *tidyr(v.0.8.1)*, *openxlsx(v.4.1.0)*, *rgeos(v.0.3-28)*, *scales(v.1.0.0)*, *ggthemes(v.4.0.1)*, *pander(v.0.6.3)*, *mice(v.3.3.0)*, *lattice(v.0.20-35)*, *factoextra(v.1.0.5)*, *NbClust(v.3.0)*, *reshape(v.0.8.8)*, *dplyr(v.0.7.6)*, *magrittr(v.1.5)*, *ggplot2(v.3.1.0)*, *Rtsne(v.0.13)* and *cluster(v.2.0.7-1)*

loaded via a namespace (and not attached): *tidyselect(v.0.2.4)*, *purrr(v.0.2.5)*, *splines(v.3.5.1)*, *colorspace(v.1.3-2)*, *htmltools(v.0.3.6)*, *yaml(v.2.2.0)*, *pan(v.1.6)*, *survival(v.2.43-1)*, *rlang(v.0.3.0.1)*, *jomo(v.2.6-4)*, *pillar(v.1.3.0)*, *nloptr(v.1.0.4)*, *glue(v.1.3.0)*, *withr(v.2.1.2)*, *sp(v.1.3-1)*, *bindr(v.0.1.1)*, *plyr(v.1.8.4)*, *stringr(v.1.3.1)*,

munsell(v.0.5.0), *gtable(v.0.2.0)*, *zip(v.1.0.0)*, *codetools(v.0.2-15)*, *evaluate(v.0.11)*, *forcats(v.0.3.0)*, *labeling(v.0.3)*, *knitr(v.1.20)*, *parallel(v.3.5.1)*, *broom(v.0.5.0)*, *Rcpp(v.1.0.0)*, *backports(v.1.1.2)*, *lme4(v.1.1-18-1)*, *digest(v.0.6.17)*, *stringi(v.1.2.4)*, *ggrepel(v.0.8.0)*, *grid(v.3.5.1)*, *rprojroot(v.1.3-2)*, *tools(v.3.5.1)*, *lazyeval(v.0.2.1)*, *tibble(v.1.4.2)*, *crayon(v.1.3.4)*, *pkgconfig(v.2.0.2)*, *MASS(v.7.3-51.1)*, *Matrix(v.1.2-14)*, *assertthat(v.0.2.0)*, *minqa(v.1.2.4)*, *rmarkdown(v.1.10)*, *rstudioapi(v.0.7)*, *rpart(v.4.1-13)*, *mitml(v.0.3-6)*, *R6(v.2.2.2)*, *nnet(v.7.3-12)*, *nlme(v.3.1-137)* and *compiler(v.3.5.1)*

Appendix 2: Survey table

Question	Type	Response range	Responses
Region	Factor	4,5,6,7,8,13	4. Coquimbo, 5. Valparaíso, 6. O´Higgins, 7. Maule, 8. Biobio, 13. Metropolitana
Farm size category	Ordinal	1,2,3,4	1 - Less than 5, 2 - Between 5 and 12, 3 - Between 12 and 50, 4 - More than 50
Date	Date	Date	NA
Locality	Factor	Name	NA

Total farm size	Numeric	Ha	NA
Area of walnut production	Numeric	Ha	NA
Area of management unit	Numeric	Ha	NA
Year of planting	Integer	#	NA
Land use before planting to walnuts	Factor	1,2,3,4,5,6,7,8	1 Grassland, 2 Forest, 3 Cereals, 4 Deciduous trees, 5 Non-deciduous trees, 6 Avocado, 7 Horticultural crops, 8 Other
Distance between trees within rows	Numeric	m	NA
Distance between trees between rows	Numeric	m	NA
Yield in 2016 harvest	Numeric	kg	NA
Rootstock	Factor	1,2,3,4,5	1 Franco, 2 Juglans nigra, 3 Paradox, 4 Vlach, 5 Other

Scion variety	Factor	1,2,3	1 Serr, 2 Chandler, 3 Other
Dose of nitrogen applied	Numeric	units N / Ha	NA
Sources of nitrogen, rates applied, dates applied	Open ended	Open ended	NA
Source of water	Factor	1,2,3	1 Surface, 2 Subterranean, 3 Both
Name of canal (if surface)	Character	Name	NA
Depth of well (if aquifer)	Numeric	#	NA
Use of instruments to monitor irrigation?	Binary	Yes / no	NA
Instruments	Factor	1,2,3,4,5,6	1 Calicata, 2 Flow meter, 3 Moisture sensor, 4 Bomb presiometer, 5 Meteorological station, 6 Other

If another instrument, which?	Character	Name	NA
Do you use them to decide when to irrigate	Binary	Yes / no	NA
Have your instruments been calibrated?	Binary	Yes / no	NA
Do you use a system to provide evapotranspiration rates?	Binary	Yes / no	NA
How do you make the decision of when and how much to irrigate?	Open ended	Open ended	NA
System of irrigation used	Factor	1,2,3,4,5,6	1 Flood, 2 Furrow, 3 Sprinkler, 4 Drip, 5 Microsprinklers, 6 Californian
If it's drip, indicate number of lines	Numeric	#	NA

Do you measure the conductivity of your water?	Binary	Yes / no	NA
If yes, what is the EC of your water	Numeric	#	NA
If no, is your water saline?	Factor	Yes / No / Don't know	NA
If you don't know, an estimated range	Factor	1,2,3	1 Very saline, 2 Somewhat saline, 3 Not saline
Do you cultivate plants between hills?	Binary	Yes / no	NA
Which?	Character	Open ended	NA
Do you know when the flush of root growth is on your farm?	Binary	Yes / no	NA
When (period in year)	Date	Time of year	NA
How many days did you irrigate in September?	Integer	#	NA

Volume irrigated	Numeric	#	NA
How many days did you irrigate in October?	Integer	#	NA
Volume irrigated	Numeric	#	NA
How many days did you irrigate in November?	Integer	#	NA
Volume irrigated	Numeric	#	NA
How many days did you irrigate in the summer?	Integer	#	NA
Volume irrigated	Numeric	#	NA
How many days did you irrigate in the fall?	Integer	#	NA
Volume irrigated	Numeric	#	NA
How many passes with heavy equipment per season?	Integer	#	NA
Do you subsoil till before planting?	Factor	Yes / No / Don't know	NA

If yes, with what do you till?	Factor	1,2,3,4	1 Backhoe, 2 Ripper, 3 Bulldozer, 4 Other
Depth of subsoil tillage	Numeric	cm	NA
Do you use hills?	Binary	Yes / no	NA
Height of hills	Numeric	cm	NA
Chemical analysis of soil	Binary	Yes / no	NA
Physical analysis of soil	Binary	Yes / no	NA
What texture does your soil have?	Factor	1,2,3,4,5	1 Sandy, 2 Sandy loam, 3 Loam, 4 Clay loam, 5 Clay
Do you know what the organic matter % is in your soil?	Binary	Yes / no	NA
If yes, how much?	Numeric	#	NA
Do you add organic amendments to your soil?	Binary	Yes / no	NA
How many times a year?	Integer	Open ended	NA

Type of organic amendment	Factor	1,2,3,4,5,6	1. Humus, 2.. Compost, 3 Bird manure, 4 Livestock manure, 5 Small livestock manure, 6 Residues
Do you know the pH of your soil?	Binary	Yes / no	NA
What range do you have?	Factor	1,2,3,4,5	1 Very acid 4.5 - 5.5, 2 Acid 5.5 - 6.5, 3 Neutral 6.5 - 7.5, 4 Basic 7.5 - 8.5, 5 Alkaline 8.5 - 10
Do you know what the drainage is like on your farm?	Binary	Yes / no	NA
How is it?	Factor	1,2,3	1. Good, 2 Normal, 3 Poor
Do you have plants damaged by Phytophthora species?	Factor	Yes / No / Don't know	NA

Have you perceive the problems of Phytophthora on your farm?	Factor	1,2,3,4	1 Low, 2 Moderate, 3 High, 4 Severe
What kinds of treatment do you use for Phytophthora control?	Factor	1,2,3	1 Cultural, 2 Chemical, 3 Both
If it's chemical, what product do you use?	Character	Open ended	NA
If it's cultural, what method do you use?	Open ended	Open ended	NA
Have you had to replant trees after disease-related death in recent years?	Binary	Yes / no	NA
If yes, how many hectares were affected?	Numeric	#	NA
If yes, after taking out trees do you apply some kind of treatment?	Binary	Yes / no	NA
Which	Character	Open ended	NA

If you apply a treatment, does the problem reappear?	Binary	Yes / no	NA
If yes, after how many years?	Numeric	#	NA
Are the solutions that exist today (chemical products, rootstock, irrigation monitoring) are effective for controlling Phytophthora?	Ordinal	1,2,3,4	1 Not effective, 2 Somewhat effective, 3 Moderately effective, 4 Very effective
Are you prepared to implement these solutions?	Ordinal	1,2,3,4	1 Not prepared, 2 Somewhat prepared, 3 Moderately prepared, 4 Very prepared
If you have a question about pathogen control, who do you consult?	Factor	1,2,3,4,5,6,7	1 Input vendor, 2 Agent of government institution, 3 NGO, 4 University or researcher, 5 Family or

			neighbors, 6 Purchaser, 7 Independent consultants
Other than the persons mentioned, are there any other sources of information?	Factor	1,2,3,4,5,6,7	1 Input vendor, 2 Agent of government institution, 3 NGO, 4 University or researcher, 5 Family or neighbors, 6 Purchaser, 7 Independent consultants
Who do you consider the most reliable sources of information?	Factor	1,2,3,4,5,6,7	1 Input vendor, 2 Agent of government institution, 3 NGO, 4 University or researcher, 5 Family or neighbors, 6 Purchaser, 7 Independent consultants
Do you participate in any program that	Factor	1,2,3,4,5,6,7,8,9,10	1 SAT, 2 PRODESAL, 3 ChileNut, 4 Chilean

provides information or solutions for producers?			Walnut Commission, 5 INIA group of technology transfer, 6 INDAP, 7 Corfo program
Do you use any other source of information for managing your walnuts?	Factor	1,2,3,4,5,6,7,8,9,10	1 Books, 2 Journals, 3 Internet, 4 Television programs, 5 National agronomy magazine, 6 Radio, 7 Smartphone
How often do you use books?	Factor	1,2,3,4	1 Never, 2 Less than one time a month, 3 One to two times a month, 4 More than once a week
Books most used	Open ended	Open ended	NA
How often do you use journals?	Factor	1,2,3,4	1 Never, 2 Less than one time a month, 3 One to two times a

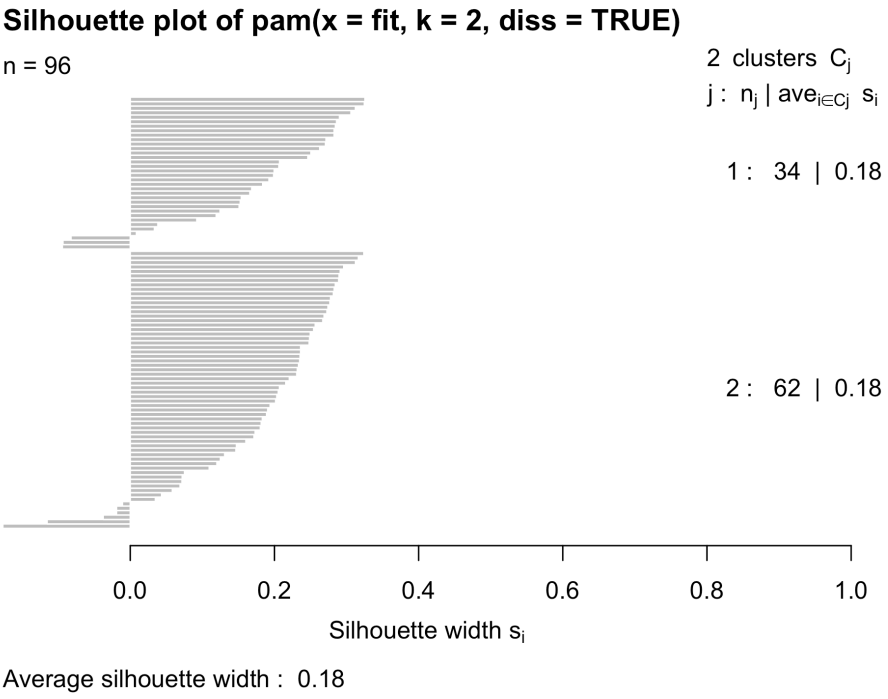
			month, 4 More than once a week
Journals most used	Open ended	Open ended	NA
How often do you use the internet?	Factor	1,2,3,4	1 Never, 2 Less than one time a month, 3 One to two times a month, 4 More than once a week
Sites most visited	Open ended	Open ended	NA
How often do you use television programs?	Factor	1,2,3,4	1 Never, 2 Less than one time a month, 3 One to two times a month, 4 More than once a week
Programs watched most often	Open ended	Open ended	NA
Radio	Factor	1,2,3,4	1 Never, 2 Less than one time a month, 3 One to two times a

			month, 4 More than once a week
Radio shows most used	Open ended	Open ended	NA
Smartphone applications	Factor	1,2,3,4	1 Never, 2 Less than one time a month, 3 One to two times a month, 4 More than once a week
Application most used	Open ended	Open ended	NA

Appendix 3: Survey numbers

	1	2	3	4
4	9	4	2	1
5	7	2	3	5
6	0	1	5	11
7	1	0	8	8
8	4	4	7	5
13	1	0	4	4

Appendix 4: Silhouette of clusters



References

References – Chapter 1

Benbi, Dinesh K, and Jörg Richter. 2002. "A Critical Review of Some Approaches to Modelling Nitrogen Mineralization." *Biology and Fertility of Soils* 35 (3). Springer: 168–83.

Bennett, Elena M, Stephen R Carpenter, and Nina F Caraco. 2001. "Human Impact on Erodable Phosphorus and Eutrophication: A Global Perspective: Increasing Accumulation of Phosphorus in Soil Threatens Rivers, Lakes, and Coastal Oceans with Eutrophication." *AIBS Bulletin* 51 (3). American Institute of Biological Sciences: 227–34.

Booth, Mary S, John M Stark, and Edward Rastetter. 2005. "Controls on Nitrogen Cycling in Terrestrial Ecosystems: A Synthetic Analysis of Literature Data." *Ecological Monographs* 75 (2). Wiley Online Library: 139–57.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1). Springer: 5–32.

Breiman, Leo, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 1984. "Classification and Regression Trees, the Wadsworth Statistics and Probability Series, Wadsworth International Group, Belmont California (Pp. 356)."

Cassman, Kenneth G, Achim Dobermann, and Daniel T Walters. 2002. "Agroecosystems, Nitrogen-Use Efficiency, and Nitrogen Management." *AMBIO: A Journal of the Human Environment* 31 (2). BioOne: 132–40.

Diaz, Robert J, and Rutger Rosenberg. 2008. “Spreading Dead Zones and Consequences for Marine Ecosystems.” *Science* 321 (5891). American Association for the Advancement of Science: 926–29.

Doshi-Velez, Finale, and Been Kim. 2017. “Towards a Rigorous Science of Interpretable Machine Learning.” *arXiv Preprint arXiv:1702.08608*.

Drinkwater, Laurie E, and SS Snapp. 2007. “Nutrients in Agroecosystems: Rethinking the Management Paradigm.” *Advances in Agronomy* 92. Elsevier: 163–86.

Finney, Denise M, Sara E Eckert, and Jason P Kaye. 2015. “Drivers of Nitrogen Dynamics in Ecologically Based Agriculture Revealed by Long-Term, High-Frequency Field Measurements.” *Ecological Applications* 25 (8). Wiley Online Library: 2210–27.

Fogs, GE, Eric M LaBolle, and Gary S Weissmann. 1999. “Groundwater Vulnerability Assessment: Hydrogeologic Perspective and Example from Salinas Valley, California.” *GEOPHYSICAL MONOGRAPH-AMERICAN GEOPHYSICAL UNION* 108. AGU AMERICAN GEOPHYSICAL UNION: 45–62.

Follett, Ronald F. 2012. *Nitrogen Management and Ground Water Protection*. Vol. 21. Elsevier.

Giltrap, Donna L, Changsheng Li, and Surinder Saggar. 2010. “DNDC: A Process-Based Model of Greenhouse Gas Fluxes from Agricultural Soils.” *Agriculture, Ecosystems & Environment* 136 (3-4). Elsevier: 292–300.

Gruber, Nicolas, and James N Galloway. 2008. “An Earth-System Perspective of the Global Nitrogen Cycle.” *Nature* 451 (7176). Nature Publishing Group: 293.

Harter, Johannes, Hans-Martin Krause, Stefanie Schuettler, Reiner Ruser, Markus Fromme, Thomas Scholten, Andreas Kappler, and Sebastian Behrens. 2014. “Linking N 2 O Emissions from Biochar-Amended Soil to the Structure and Function of the N-Cycling Microbial Community.” *The ISME Journal* 8 (3). Nature Publishing Group: 660.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. “Unbiased Recursive Partitioning: A Conditional Inference Framework.” *Journal of Computational and Graphical Statistics* 15 (3). Taylor & Francis: 651–74.

Howarth, Robert W. 2008. “Coastal Nitrogen Pollution: A Review of Sources and Trends Globally and Regionally.” *Harmful Algae* 8 (1). Elsevier: 14–20.

Kaye, Jason P, and Miguel Quemada. 2017. “Using Cover Crops to Mitigate and Adapt to Climate Change. a Review.” *Agronomy for Sustainable Development* 37 (1). Springer: 4.

Kersebaum, Kurt Christian. 2007. “Modelling Nitrogen Dynamics in Soil–crop Systems with Hermes.” In *Modelling Water and Nutrient Dynamics in Soil–crop Systems*, 147–60. Springer.

Li, C. 2009. “User’s Guide for the Dndc Model (Version 9.5).” *Institute for the Study of Earth, Oceans, and Space. University of New Hampshire, Durham, NH*.

McIsaac, Gregory F, Mark B David, George Z Gertner, and Donald A Goolsby. 2001. "Eutrophication: Nitrate Flux in the Mississippi River." *Nature* 414 (6860). Nature Publishing Group: 166.

Meinshausen, Nicolai. 2006. "Quantile Regression Forests." *Journal of Machine Learning Research* 7 (Jun): 983–99.

Molnar, Christoph. 2018. "Interpretable Machine Learning." *A Guide for Making Black Box Models Explainable*.

Robertson, G Philip, and Peter M Vitousek. 2009. "Nitrogen in Agriculture: Balancing the Cost of an Essential Resource." *Annual Review of Environment and Resources* 34. Annual Reviews: 97–125.

Schimel, Joshua P, and Jennifer Bennett. 2004. "Nitrogen Mineralization: Challenges of a Changing Paradigm." *Ecology* 85 (3). Wiley Online Library: 591–602.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. "Conditional Variable Importance for Random Forests." *BMC Bioinformatics* 9 (1). BioMed Central: 307.

Štrumbelj, Erik, and Igor Kononenko. 2014. "Explaining Prediction Models and Individual Predictions with Feature Contributions." *Knowledge and Information Systems* 41 (3). Springer: 647–65.

Team, R. 2013. "R Development Core Team." *RA Lang Environ Stat Comput* 55: 275–86.

References – Chapter 2

Asci, Serhat, John J VanSickle, Curtiss J Fry, and John Thomas. n.d. “Risk Management and Fumigation Choice in Tomato Production.” *2015 FLORIDA TOMATO INSTITUTE PROGRAM*, 19.

Butler, David M, Erin N Rosskopf, Nancy Kokalis-Burelle, Joseph P Albano, Joji Muramoto, and Carol Shennan. 2012. “Exploring Warm-Season Cover Crops as Carbon Sources for Anaerobic Soil Disinfestation (Asd).” *Plant and Soil* 355 (1-2). Springer: 149–65.

Carter, Colin A, James A Chalfant, Rachael E Goodhue, Frank M Han, and Massimiliano DeSantis. 2005. “The Methyl Bromide Ban: Economic Impacts on the California Strawberry Industry.” *Review of Agricultural Economics* 27 (2). Oxford University Press: 181–97.

Ebihara, Yoshiyuki, and Seiji Uematsu. 2014. “Survival of Strawberry-Pathogenic Fungi Fusarium Oxysporum F. Sp. Fragariae, Phytophthora Cactorum and Verticillium Dahliae Under Anaerobic Conditions.” *Journal of General Plant Pathology* 80 (1). Springer: 50–58.

Franzin, Alberto, Francesco Sambo, and Barbara Di Camillo. 2016. “Bnstruct: An R Package for Bayesian Network Structure Learning in the Presence of Missing Data.” *Bioinformatics* 33 (8). Oxford University Press: 1250–2.

Friedman, Nir, Moises Goldszmidt, and others. 1996. “Discretizing Continuous Attributes While Learning Bayesian Networks.” In *ICML*, 157–65.

Friedman, Nir, Moises Goldszmidt, and Abraham Wyner. 1999. “Data Analysis with Bayesian Networks: A Bootstrap Approach.” In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 196–205. Morgan Kaufmann Publishers Inc.

Hewavitharana, Shashika Shivanthi, David Ruddell, and Mark Mazzola. 2014. “Carbon Source-Dependent Antifungal and Nematicidal Volatiles Derived During Anaerobic Soil Disinfestation.” *European Journal of Plant Pathology* 140 (1). Springer: 39–52.

Koike, Steven T, and Thomas R Gordon. 2015. “Management of Fusarium Wilt of Strawberry.” *Crop Protection* 73. Elsevier: 67–72.

Mazzola, Mark, Joji Muramoto, and Carol Shennan. 2018. “Anaerobic Disinfestation Induced Changes to the Soil Microbiome, Disease Incidence and Strawberry Fruit Yields in California Field Trials.” *Applied Soil Ecology* 127. Elsevier: 74–86.

Momma, Noriaki, Yuso Kobara, Seiji Uematsu, Nobuhiro Kita, and Akinori Shinmura. 2013. “Development of Biological Soil Disinfestations in Japan.” *Applied Microbiology and Biotechnology* 97 (9). Springer: 3801–9.

Muramoto, Joji, Carol Shennan, Margherita Zavatta, Graeme Baird, Lucinda Toyama, and Mark Mazzola. 2016. “Effect of Anaerobic Soil Disinfestation and Mustard Seed Meal for Control of Charcoal Rot in California Strawberries.” *International Journal of Fruit Science* 16 (sup1). Taylor & Francis: 59–70.

Rosskopf, Erin N, Paula Serrano-Pérez, Jason Hong, Utsala Shrestha, María del Carmen Rodríguez-Molina, Kendall Martin, Nancy Kokalis-Burelle, Carol Shennan, Joji Muramoto, and David Butler. 2015. "Anaerobic Soil Disinfestation and Soilborne Pest Management." In *Organic Amendments and Soil Suppressiveness in Plant Disease Management*, 277–305. Springer.

Scutari, Marco. 2009. "Learning Bayesian Networks with the Bnlearn R Package." *arXiv Preprint arXiv:0908.3817*.

Shennan, C, J Muramoto, S Koike, G Baird, S Fennimore, J Samtani, M Bolda, et al. 2018. "Anaerobic Soil Disinfestation Is an Alternative to Soil Fumigation for Control of Some Soilborne Pathogens in Strawberry Production." *Plant Pathology* 67 (1). Wiley Online Library: 51–66.

Shennan, Carol, Joji Muramoto, Graeme Baird, Steven Koike, Mark Bolda, and Mark Mazzola. 2013. "Optimizing Anaerobic Soil Disinfestation for Soilborne Disease Control." In *2013 Proceedings of Annual International Research Conference on Methyl Bromide Alternatives and Emissions Reductions. San Diego (ca)*, 13–11.

Team, R Core, and others. 2013. "R: A Language and Environment for Statistical Computing." Vienna, Austria.

Tsamardinos, Ioannis, Laura E Brown, and Constantin F Aliferis. 2006. "The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm." *Machine Learning* 65 (1). Springer: 31–78.

Wang, ZP, RD Delaune, WH Patrick, and PH Masscheleyn. 1993. "Soil Redox and pH Effects on Methane Production in a Flooded Rice Soil." *Soil Science Society of America Journal* 57 (2). Soil Science Society of America: 382–85.

Yonemoto, K, K Hirota, S Mizuguchi, and K Sakaguchi. 2006. "Utilization of the Sterilization by Soil Reduction in an Open Air Field and Its Efficacy Against Fusarium Wilt of Strawberry." *Proc Assoc Pl Protec Shikoku* 41: 15–24.

References – Chapter 3

Ashby, Jacqueline A, and Louise Sperling. 1995. "Institutionalizing Participatory, Client-Driven Research and Technology Development in Agriculture." *Development and Change* 26 (4). Wiley Online Library: 753–70.

Birner, Regina, and Jock R Anderson. 2007. *How to Make Agricultural Extension Demand Driven? The Case of India's Agricultural Extension Policy*. Vol. 729. Intl Food Policy Res Inst.

Browne, Greg, Leigh Schmidt, Terry Prichard, and Wes Hackett. 2006. "Biology and Management of Phytophthora Crown and Root Rot of Walnut." *Walnut Research Reports* 2005: 335–44.

Chatterjee, Diti, Ariel Dinar, and Gloria González-Rivera. 2016. "The Contribution of the University of California Cooperative Extension to California's Agricultural Production." UCR SPP Working Paper Series, September 2016, WP.

Cook, NB, JP Hess, MR Foy, TB Bennett, and RL Brotzman. 2016. "Management Characteristics, Lameness, and Body Injuries of Dairy Cattle Housed in High-Performance Dairy Herds in Wisconsin." *Journal of Dairy Science* 99 (7). Elsevier: 5879–91.

Daberkow, Stan G, and William D McBride. 2003. "Farm and Operator Characteristics Affecting the Awareness and Adoption of Precision Agriculture Technologies in the Us." *Precision Agriculture* 4 (2). Springer: 163–77.

Garforth, Chris, Brian Angell, John Archer, and Kate Green. 2003. "Improving Farmers' Access to Advice on Land Management: Lessons from Case Studies in Developed Countries." *Agricultural Research and Extension Network Paper* 125.

George, Melvin R, and W James Clawson. 2014. "History of University of California Rangeland Extension, Research, and Teaching." *Rangelands* 36 (5). Elsevier: 18–24.

Guajardo, J, S Saa, R Camps, and X Besoain. 2017. "Outbreak of Crown and Root Rot of Walnut Caused by Phytophthora Cinnamomi in Chile." *Plant Disease* 101 (4). Am Phytopath Society: 636–36.

Guajardo, J, S Saa, Natalia Riquelme, Gregory Browne, Cristian Youlton, Mónica Castro, and Ximena Besoain. 2019. "Characterization of Oomycete Species Associated with Root and Crown Rot of English Walnut in Chile." *Plant Disease*. Am Phytopath Society, PDIS–07.

Hall, Andrew, Geoffrey Bockett, Sarah Taylor, MVK Sivamohan, and Norman Clark. 2001. "Why Research Partnerships Really Matter: Innovation Theory, Institutional

Arrangements and Implications for Developing New Technology for the Poor." *World Development* 29 (5). Elsevier: 783–97.

Hauser, Michael, Mara Lindtner, Sarah Prehsler, and Lorenz Probst. 2016. "Farmer Participatory Research: Why Extension Workers Should Understand and Facilitate Farmers' Role Transitions." *Journal of Rural Studies* 47. Elsevier: 52–61.

Hillger, David E, Stephen C Weller, Elizabeth Maynard, and Kevin D Gibson. 2006. "Weed Management Systems in Indiana Tomato Production." *Weed Science* 54 (3). Cambridge University Press: 516–20.

Huang, Anna. 2008. "Similarity Measures for Text Document Clustering." In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (Nzcsrsc2008), Christchurch, New Zealand*, 4:9–56.

Kaufman, Leonard, and Peter J Rousseeuw. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol. 344. John Wiley & Sons.

Kidd, AD, JPA Lamers, PP Ficarelli, and Volker Hoffmann. 2000. "Privatising Agricultural Extension: Caveat Emptor." *Journal of Rural Studies* 16 (1). Elsevier: 95–102.

Leeuwis, Cees. 2013. *Communication for Rural Innovation: Rethinking Agricultural Extension*. John Wiley & Sons.

Levidow, Les, Michel Pimbert, and Gaetan Vanloqueren. 2014. "Agroecological Research: Conforming—or Transforming the Dominant Agro-Food Regime?" *Agroecology and Sustainable Food Systems* 38 (10). Taylor & Francis: 1127–55.

Liu, Tingting, Randall Bruins, and Matthew Heberling. 2018. “Factors Influencing Farmers’ Adoption of Best Management Practices: A Review and Synthesis.” *Sustainability* 10 (2). Multidisciplinary Digital Publishing Institute: 432.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. “Visualizing Data Using T-Sne.” *Journal of Machine Learning Research* 9 (Nov): 2579–2605.

MacMillan, Tom, and Tim G Benton. 2014. “Agriculture: Engage Farmers in Research.” *Nature News* 509 (7498): 25.

Mircetich, SM, ME Matheron, and others. 1983. “Phytophthora Root and Crown Rot of Walnut Trees.” *Phytopathology* 73 (11): 1481–8.

Reynolds, Alan P, Graeme Richards, Beatriz de la Iglesia, and Victor J Rayward-Smith. 2006. “Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms.” *Journal of Mathematical Modelling and Algorithms* 5 (4). Springer: 475–504.

Rousseeuw, Peter J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20. Elsevier: 53–65.

Savary, S, FA Elazegui, K Moody, JA Litsinger, and PS Teng. 1994. “Characterization of Rice Cropping Practices and Multiple Pest Systems in the Philippines.” *Agricultural Systems* 46 (4). Elsevier: 385–408.

Team, R. 2013. “R Development Core Team.” *RA Lang Environ Stat Comput* 55: 275–86.