Yield prediction modeling: When? How? Why?

Maciej Zwieniecki

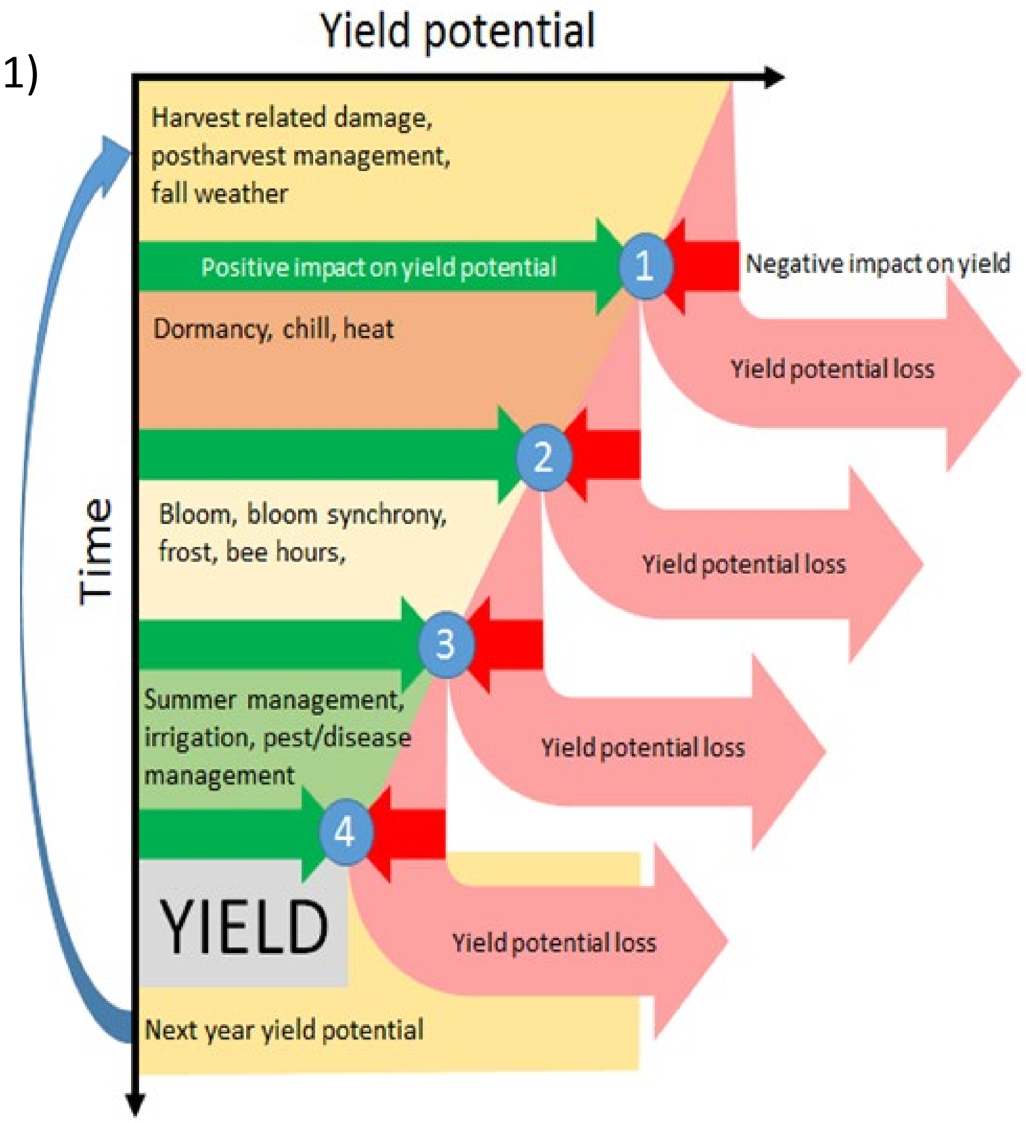
Paula Guzmán-Delgado

Jessica Orozco

This is a part of research conducted by Carbohydrate Observatory (Zwieniecki Lab at UC Davis)

When can we estimate yield?

Summer (n-1)



Continuous loss of yield potential is a result of many fixed factors:

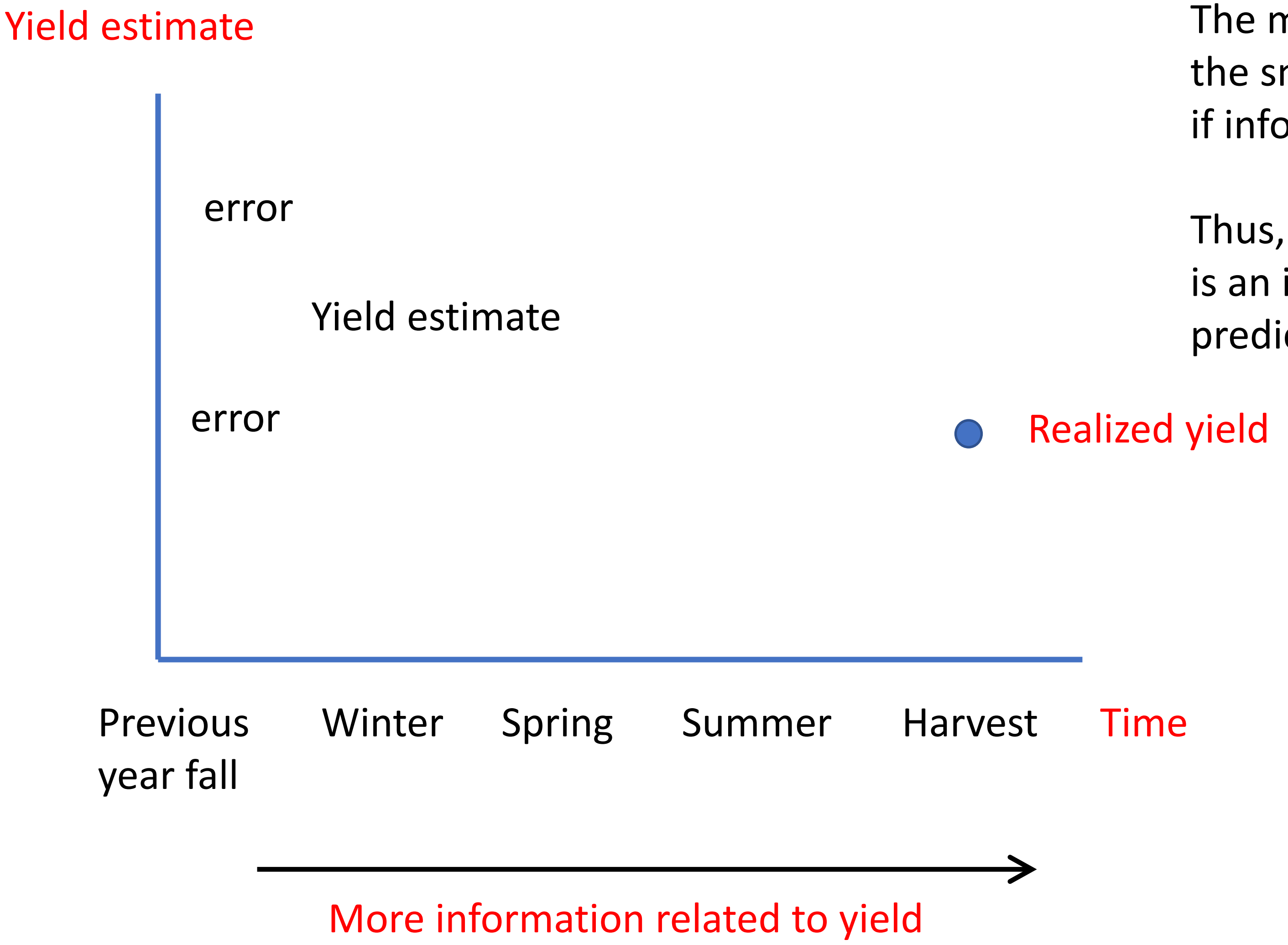
-soil, orchard age, variety, geographical location etc..

semi-stochastic factors:
-temperature, rainfall, pathogens, occurrence of fires etc…

and applied practices:
-irrigation, fertilizations, pest management plan, etc…

When can we estimate yield?

We can estimate yield anytime, however our estimate will have a significant error related to semi-stochastic factors



The more information about the orchard we have the smaller is the error of yield estimate especially if information is relevant to the yield.

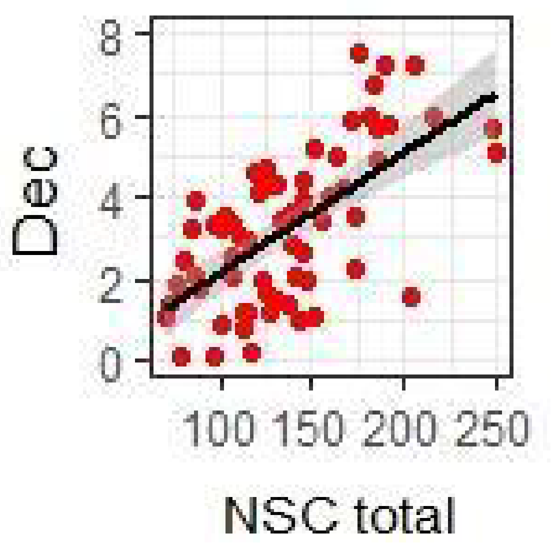
Thus, finding factors that are correlated with yield is an important step in constructing yield prediction models.

How can we predict yield?

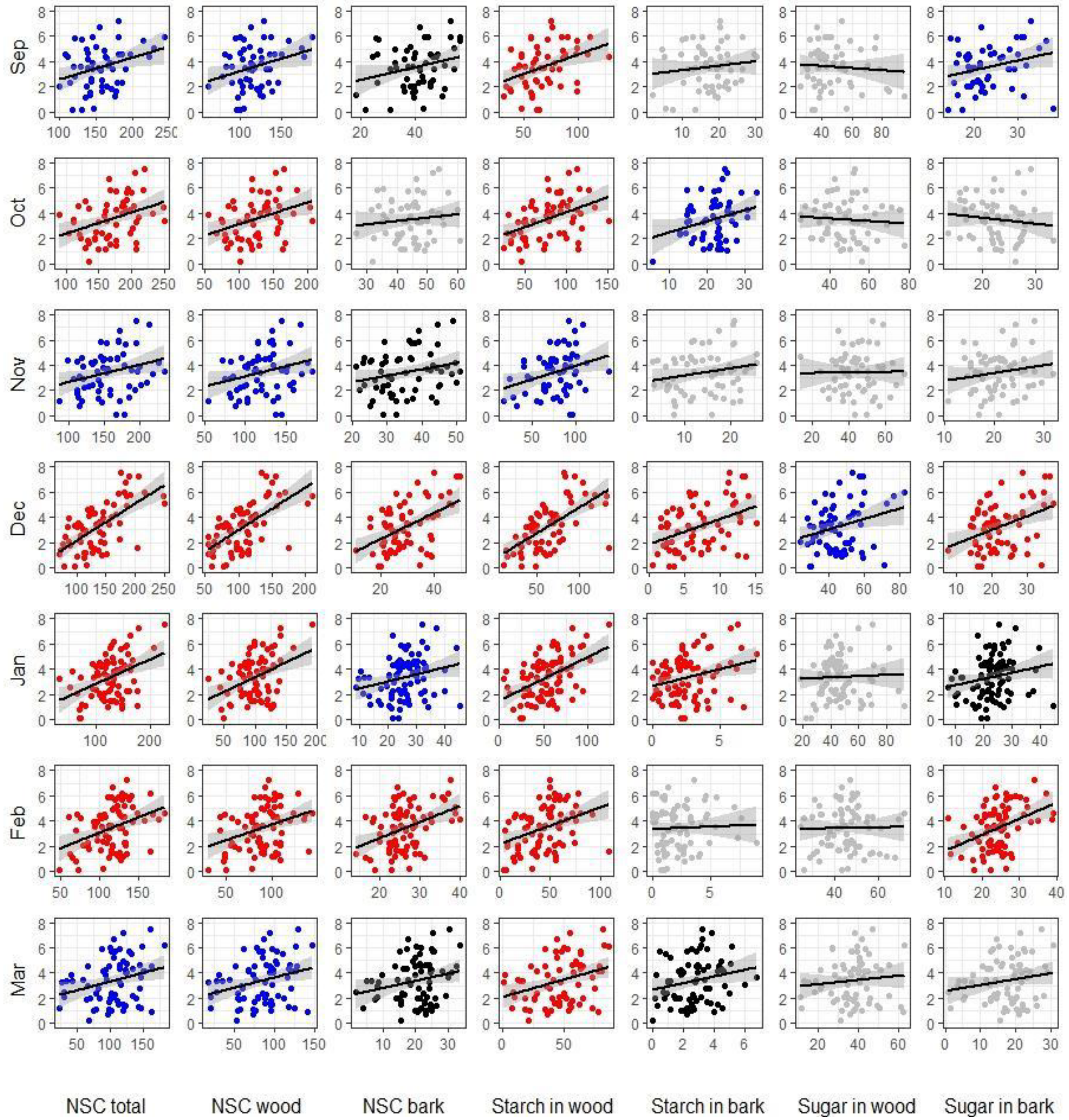
First, we can look for factors that correlate with yield.

We can also check for correlations between other factors, like soluble sugars or starch in September, October, November, ... and yield.

For example, we test for correlation between amount of nonstructural carbohydrates in December (a factor) with yield:

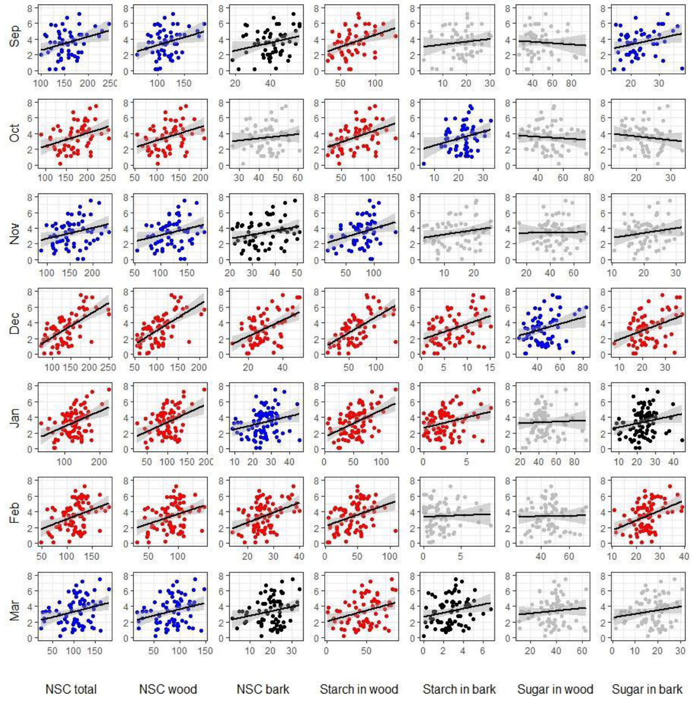


BUT there is a large error



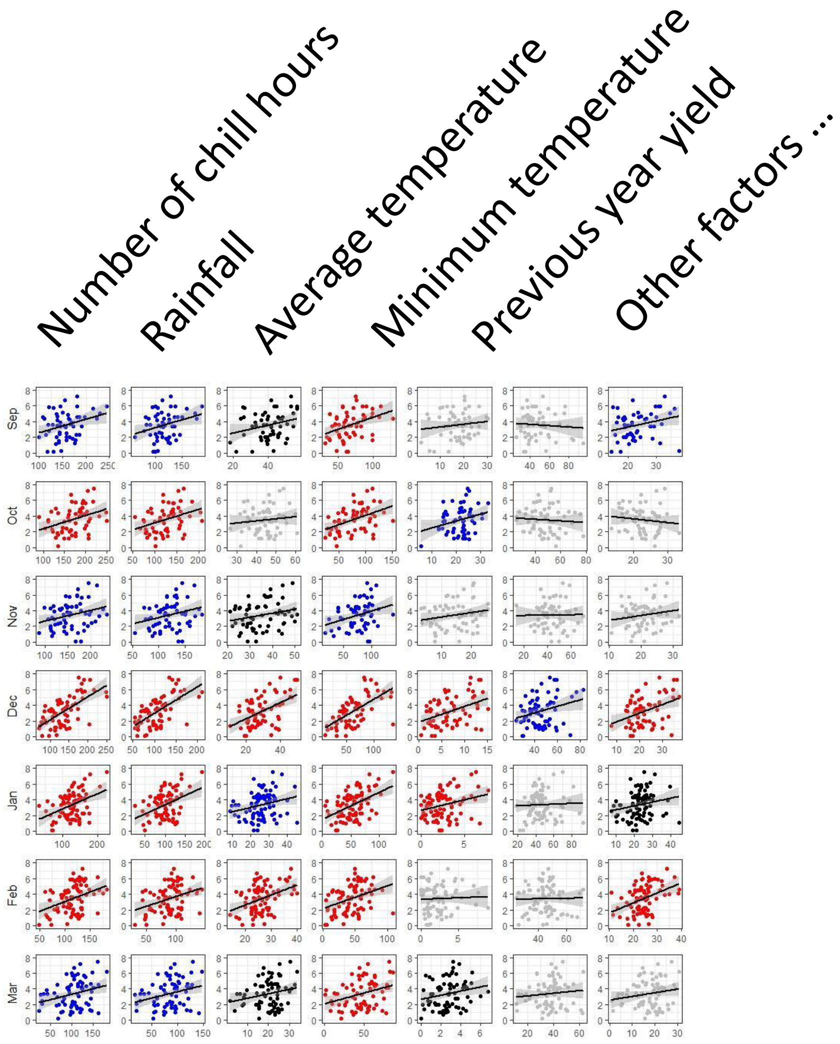
How can we predict yield?

Each correlation has large error. However, if we combine the correlations into a single model and if factors are relatively independent, there is a high probability that we can reduce error of yield estimate



+

We can add other factors:

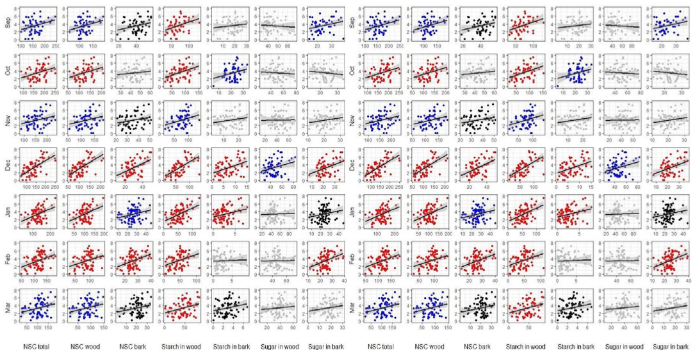


Now we can complicate the picture and realize that selection of factors is almost infinite, and it is very difficult to decide on what is important and what to select for the analysis

How can we predict yield?

We need help of ‘deep learning’ that is multidimensional combinations of linear fits and estimation of parameters of a complex linear model. It is linear correlation on steroids.

Yield(model) = a(1)*factor(1) + a(2)*factor(2) + …. + a(n)*factor(n)



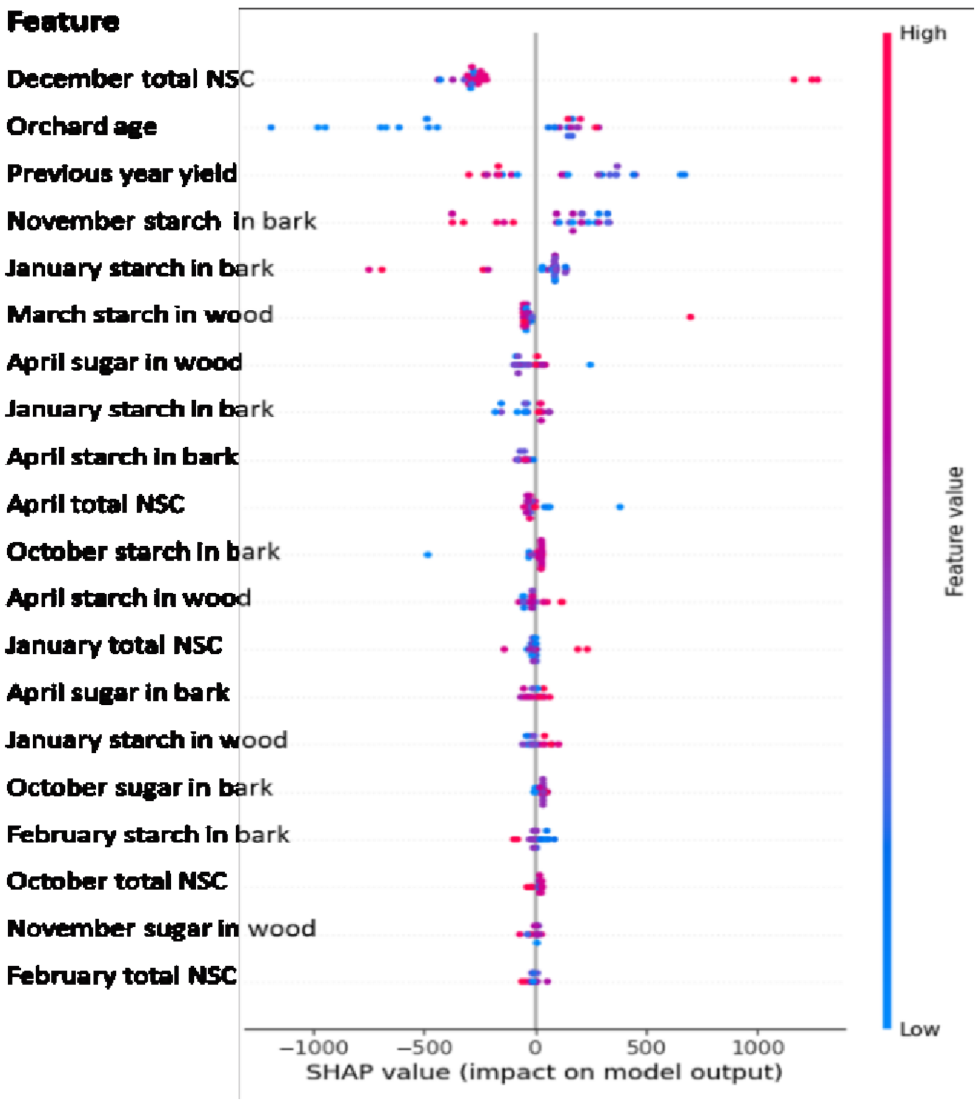
Or if this does not help we can use artificial intelligence algorithms that have ability to test for combination of seemingly unrelated factors and events into a coherent model

Yield(model) ~ Yield(real) ------ estimate error
Change a(i)s and see if program notices improvement, it keeps new parameters and iterates till no improvement occurs.
This part is called model training.

There are ~200 methods and algorithms to perform this analysis.

Regression Model	MAE	MSE	RMSE
Extreme Gradient Boosting	8.383432e+02	1.144238e+06	9.929191e+02
Gradient Boosting Regressor	9.041745e+02	1.261718e+06	1.084073e+03
AdaBoost Regressor	9.887913e+02	1.547970e+06	1.182868e+03
CatBoost Regressor	9.168102e+02	1.344820e+06	1.125314e+03
Random Forest	9.794012e+02	1.516015e+06	1.193861e+03
Extra Trees Regressor	9.729235e+02	1.475905e+06	1.162145e+03
Light Gradient Boosting Machine	1.202051e+03	2.124878e+06	1.434624e+03
Huber Regressor	1.215772e+03	2.020300e+06	1.404907e+03
Elastic Net	1.248803e+03	2.561780e+06	1.494331e+03
Support Vector Machine	1.262265e+03	2.383184e+06	1.510696e+03
Bayesian Ridge	1.277563e+03	2.510589e+06	1.548797e+03
Orthogonal Matching Pursuit	1.178133e+03	2.254673e+06	1.444456e+03
Lasso Least Angle Regression	1.324538e+03	2.886303e+06	1.582410e+03
K Neighbors Regressor	1.396300e+03	3.104862e+06	1.713018e+03
Lasso Regression	1.527999e+03	3.925164e+06	1.835618e+03
Ridge Regression	1.534934e+03	3.796636e+06	1.821250e+03
Decision Tree	1.349621e+03	2.934957e+06	1.610878e+03
Passive Aggressive Regressor	2.274575e+03	9.477599e+06	2.606202e+03
Random Sample Consensus	3.206556e+03	5.316786e+07	4.849691e+03
Linear Regression	5.386159e+03	4.365378e+08	1.02E+04
TheilSen Regressor	6.329980e+03	6.639100e+08	1.228249e+04
Least Angle Regression	9.759026e+07	4.186598e+17	2.31E+08

How can we predict yield?



Testing for best factors for early prediction of yield?

Why - Potential targets for modification through management

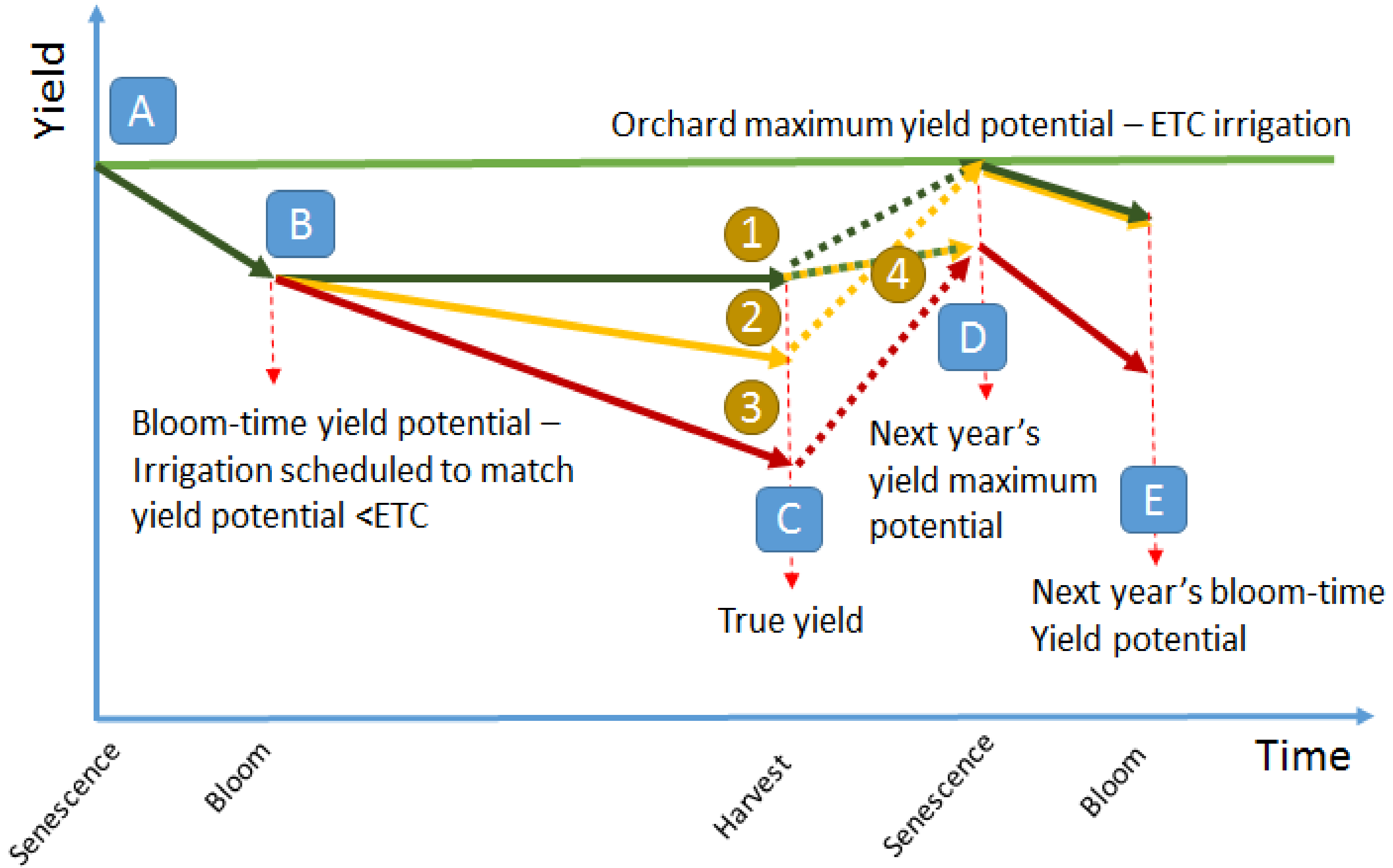
Here we used September - April data to determine few most important variables that explain most variance in yield prediction model.

Figure 4. SHAP (Shapley additi explanation) values of features fr Extreme Boosting Model. Twenty t features with the most explanatc power are listed with relati importance from highest to lowe December total NSC content in twi orchard age and previous year yie are the most important variab followed by the presence of starch bark (interestingly lower starch in bc predicts higher yield). Feature value red denotes high yield and in bl denotes low yield.

How can we predict yield? Limitations

Available yield data	Number of fields with known yields >> number of factors n the model
Coverage	To improve estimates, we need to improve geographical coverage, increase coverage of varieties, rootstock, age, and years to account for yearly variation of climate
Missing information	Clouds – reduced information from satellites, problems with sampling, missing information on irrigation, fertilization, and other management practices reduces our capacity to provide useful information on what works and what is not working
Impact of stochastic events	Yearly variation - including the alternate bearing, weather, fires, rain distribution, occurrence of pathogens ….

Why to predict yield? Provides info on impact of managements practices on yield



At the point (A), in fall, maximum orchard yield potential is established based on the previous yield and modeling approach. (B) This maximum yield is decreased due to winter weather patterns and a new current year's yield potential is determined (NSC-based yield predictive model - from this proposal). The irrigation schedule is adapted to reflect reduced yield expectations. At harvest (C), four basic potential outcomes can be expected:

(1) Best outcome - reduced irrigation allowed to achieve the current year's yield potential, and next year's yield potential was not affected (D) - allows for water saving without short- or long-term impacts on orchard performance.

(2) Moderate outcome - reduced irrigation negatively impacts current year's yield potential but does not reduce long-term orchard maximum yield potential (D).

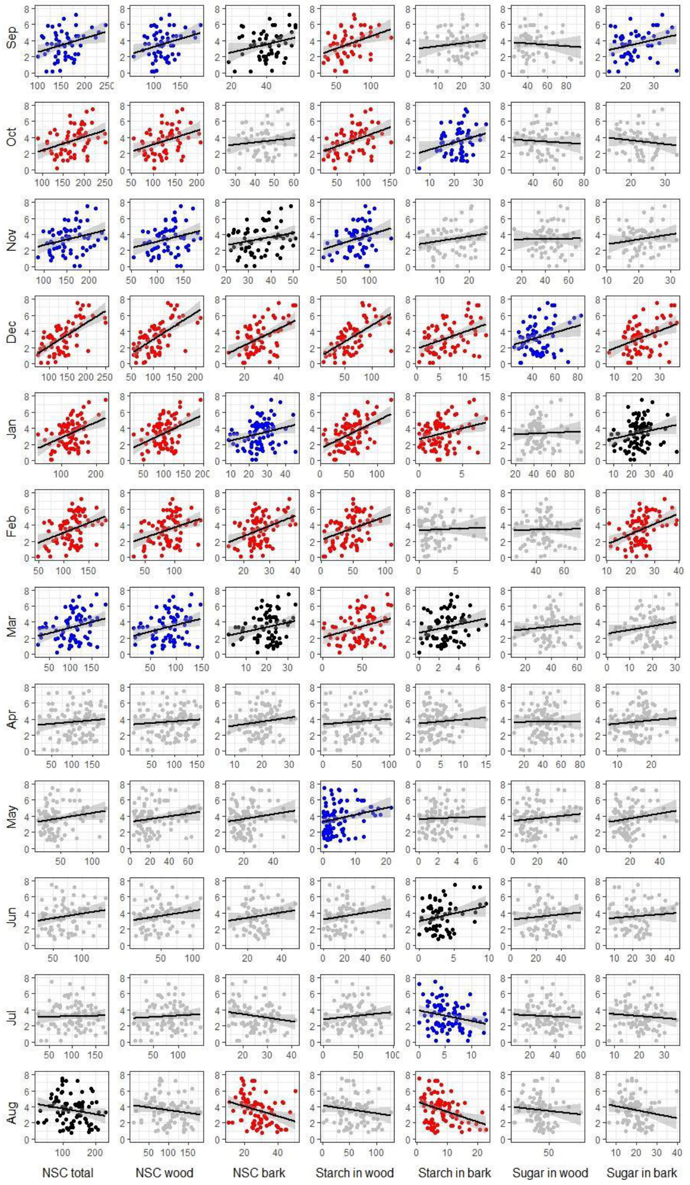
(3) Worst outcome - reduced irrigation negatively impacts both current year's yield potential and the long-term orchard maximum yield potential (D)

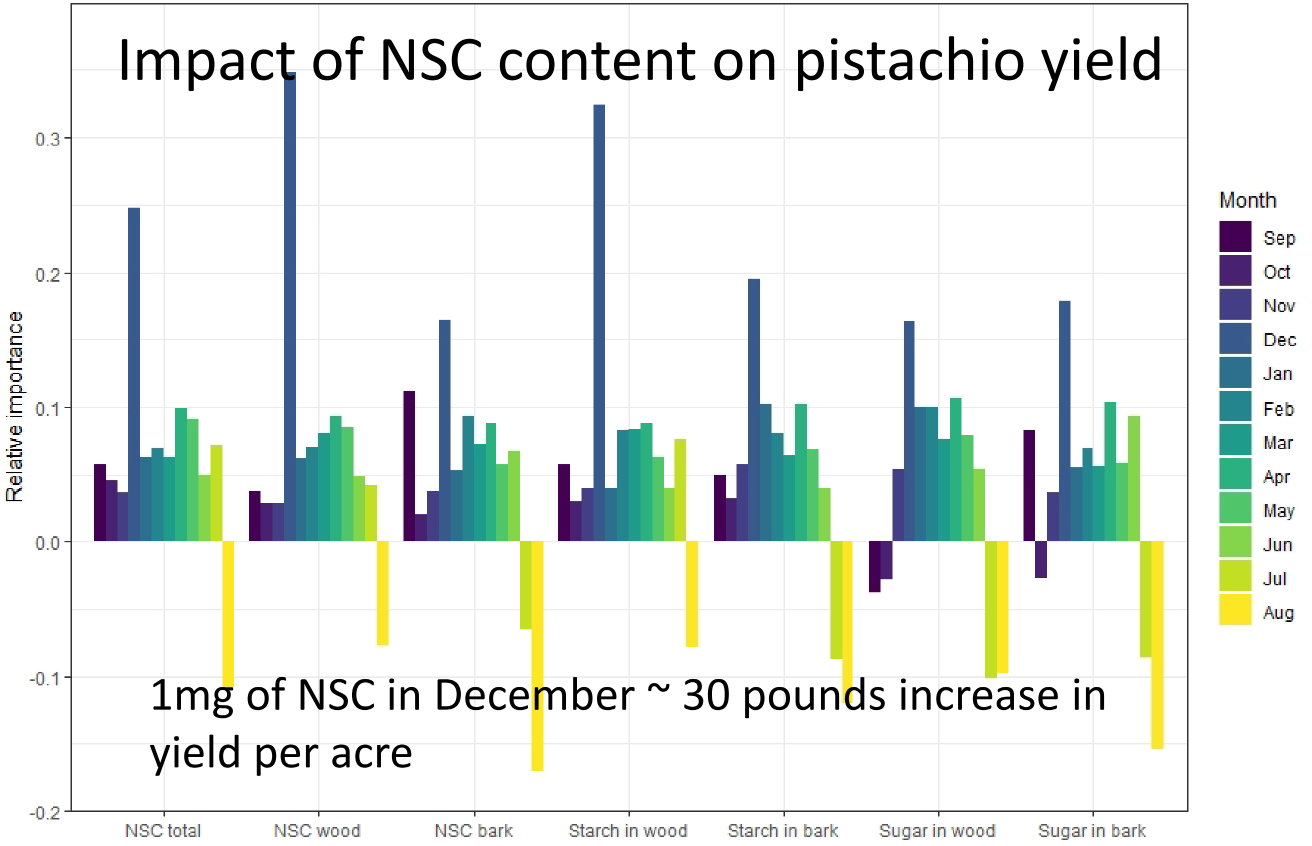
(4) Mixed outcome - reduced irrigation achieves the current year's yield potential but reduces long-term orchard maximum yield potential (D).

At (E) a new year's yield potential is established.

Why to predict yield?

Provides info on when to collect data





Other data:
Orchards' features: geography, age, previous yields, variety, management, salinity etc.

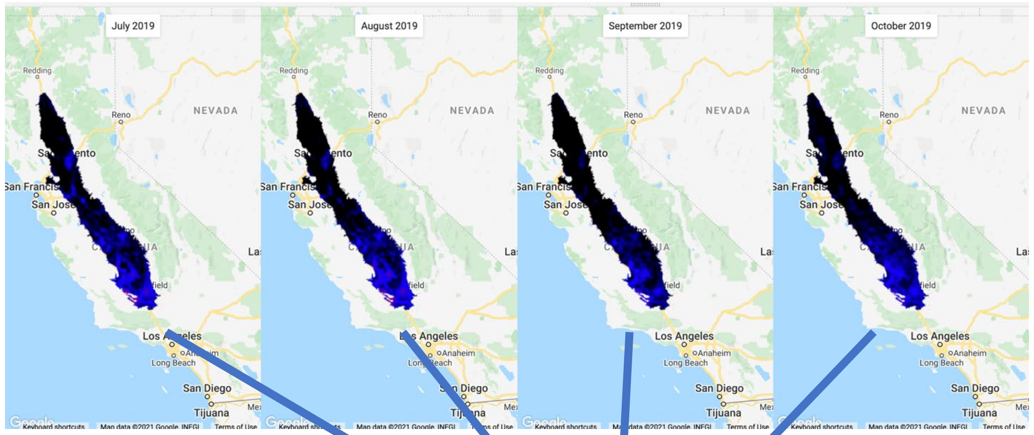
Dynamic variable: weather, bloom time, etc.

http://zlab-yield-model.herokuapp.com/

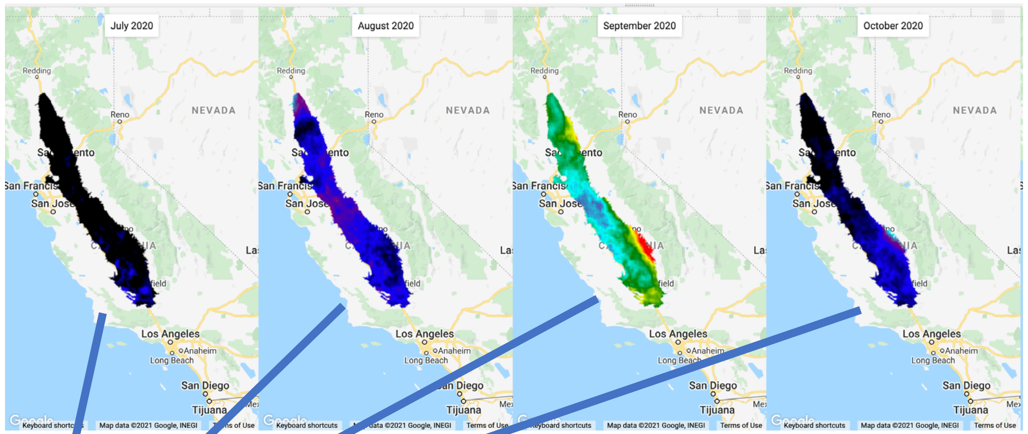
Why to predict yield?

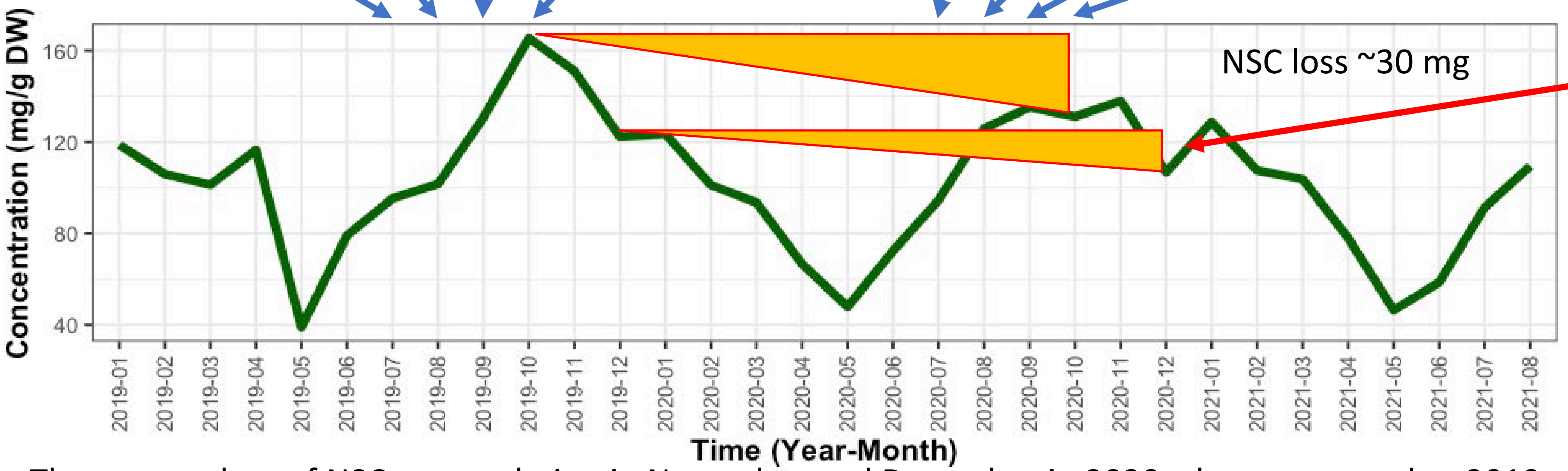
Potential impact of smoke on accumulation of NSC in pistachio (smoke data analyzed by Jessica Orozco)

Smoke levels Central Valley in 2019 (July-October)



Smoke levels Central Valley in 2020 (July-October)





Impact of smoke ~20 mg of NSC in December ~~~

Loss of 600 pounds per acre (so who knows what it could be 2021

There was a loss of NSC accumulation in November and December in 2020 when compared to 2019

Why to predict yield?

Many open questions:

If predicted crop is low - can we save on irrigation, fertilization, protection?
Are short term savings further reducing expected crop?
Do savings translate to lower crop in following years or orchards have short term memory?
How can we reduce loss of potential crop?
When and what can we do to maximize crop?
Can we predict how climate will impact production in 2, 5, 10 years?
How to choose future orchard sites?
…

It seems that having a good prediction model we can benefit pistachio industry

What we do and can you help?

What we do – we try to develop a yield prediction model using available data with beta version available on:

http://zlab-yield-model.herokuapp.com

Carbohydrate Observatory

Yield prediction model - Do not use for making management decisions - model is for research purpose only

IF YOU WANT TO LEARN MORE ABOUT THE USE OF THE MODEL Please contact Zwieniecki lab for details

Model was developed on data from Central Valley California only, use of geographical locations outside the Central Valley California is not recomended

Model was trained on limited data made available to Zwieniecki lab by California growers. Model quallity would increase over time as more dat can be used in model training

We used weathr information from PRISM Climate group, OSU. Oct-Apr data were used to train model. If model is used before aend of April availble info from current winter is used and missing data is take from last winter.

Not all information is needed but error of prediction will increase. Privided error assume all information is entered.

Choose a species and please fill as many fields in tables as you can. Then click the (submit) button.

Initial values of carbohydrates are state averages. If you know your specific NSC contents plese edit the entries. If you do not enter location it is assumed to be lat=36, lon=-119.

Almond	Pistachio	Walnut

Submit

What we do and can you help?

Yield prediction model - Do not use for making management decisions - model is for research purpose only

IF YOU WANT TO LEARN MORE ABOUT THE USE OF THE MODEL Please contact Zwieniecki lab for details

Model was developed on data from Central Valley California only, use of geographical locations outside the Central Valley California is not recomended

Model was trained on limited data made available to Zwieniecki lab by California growers. Model quallity would increase over time as more dat can be used in model training

We used weathr information from PRISM Climate group, OSU. Oct-Apr data were used to train model. If model is used before aend of April availble info from current winter is used and missing data is take from last winter.

Not all information is needed but error of prediction will increase. Privided error assume all information is entered.

Choose a species and please fill as many fields in tables as you can. Then click the (submit) button.

Initial values of carbohydrates are state averages. If you know your specific NSC contents plese edit the entries. If you do not enter location it is assumed to be lat=36, lon=-119.

Almond	Pistachio	Walnut

Information	Orchard data
Latitude West Coast USA latitude(32.3 49.3)	33
Longitude West Coast USA (-124.7,-119.7)	-120
Last year yield in pounds per acre	3000
Orchard age in years	10

Information	Oct	Nov	Dec	Jan	Feb	Mar	Apr
NSC total in mg/DW	167	154	134	120	108	102	93
NSC in wood in mg/DW	184	175	155	139	122	121	111
Starch in wood in mg/DW	80	73	57	48	41	45	40

Submit

Predicted yield is [3705] pounds per acre

This is predicted yield based on data you have entered and available weather data at the time of entry

In April Accuracy= 70.75 and RMSE = 820 pounds/acre

The model can be improved and you can help:

-**provide information** – we keep it confidential

-**send samples** – as long as the pistachio commodity supports us, analysis is free and easy

-**talk to us** – express your needs, suggest directions, share your opinion

Talk to us (email us):
Maciej Zwieniecki
mzwienie@ucdavis.edu

Paula Guzmán-Delgado
pguzmandelgado@ucdavis.edu

Jessica Orozco
jsorozco@ucdavis.edu

Model can be improved and your help is a key component

THANK YOU

Talk to us (email us):
Maciej Zwieniecki
mzwienie@ucdavis.edu

Paula Guzmán-Delgado
pguzmandelgado@ucdavis.edu

Jessica Orozco
jsorozco@ucdavis.edu

Citizen Science – Accelerating research with help from growers

What we provide:

- Envelopes for twigs collection and shipping
- Analysis of samples
- Online database of starch level in samples
- Real-time interactive map of starch content in trees across Central Valley
- Graphical depiction of starch level for each participating orchard