Ensemble of Cubist models for soy yield prediction using soil features and remote sensing variables

Tzvi Aviv
AvivInnovation
Toronto, ON, Canada
tzvi@avivinnovation.com

Vanessa Lundsgaard-Nielsen
University of Toronto
Toronto, ON, Canada
vanessa.nielsen@mail.utoronto.ca

ABSTRACT
The goal of this work is to develop a predictive model for selecting elite soy variants for commercial production. Current breeding practices for new soy variants require rigorous evaluation over three stages of field tests, corresponding to three successive growing seasons. We propose to leverage machine learning methods for identifying high yielding variants using remote sensing and soil features. To support this proposition, we trained an ensemble of fifteen decision tree models, one for each relative maturity band. Collectively, our models identified fifteen elite varieties from 21 predictive variables to forecast soybean yields in 2015 at 58 test locations. This method can boost commercial soy yields by about 5% and shorten the time for commercial variant development.

CCS Concepts
•CCS → Applied computing → Computers in other domains → Agriculture

Keywords
yield prediction; decision trees; cubist; remote sensing; NDVI; Land Surface Temperature; MODIS; Random Forest .

1. INTRODUCTION
Crop yields are highly variable across fields as a result of complex interactions among factors such as environmental conditions, soil properties, management practices, and disease and pest attack [3, 10, 4]. In particular, environmental conditions play a vital role in yield variability, and crop productivity is limited by water availability, light, temperature, and nutrients [15,21]. Here, we augmented the soil and weather data supplied by Syngenta with publicly available remote sensing data and selected 21 variables that were the most predictive of soybean yield. Our solution uses vegetation indices calculated from satellite images to predict crop yields on a site-by-site basis. Vegetation indices, such as the normalized difference vegetation index (NDVI) have been widely used for agricultural mapping and monitoring and are calculated using the red and near-infrared wavelengths [5]. In the last decade, remote sensing-focused yield forecasting has shifted to the National Aeronautics and Space Administration's (NASA) Moderate Resolution Imaging Spectroradiometer (MODIS), which provide spatial data at a fine resolution [8,24].

Soybean variants are classified according to relative maturity (RM), which reflects the time it takes a variety to reach physiological maturity [2]. Depending on the day length and flowering time, RM is designated with a numerical system ranging from 0 to 10 from north to south [2, 12]. Effectively, RM is an indicator of the fit of genetically coded flowering time to the environmental conditions in geographic locations. The dataset provided in this challenge includes yield data for variants mostly in RM bands ranging from 2.0 to 3.5 covering the soy regions of the US Midwest (Indiana, Nebraska, Illinois, Iowa and Minnesota). The differences in yields, weather and soil conditions across RM bands prompted us to avoid a global yield model for all RM bands, and we opted instead to generate models for each RM band, creating an ensemble of predictive models, one for each RM band. We applied decision tree methods to explore the predictive properties of these variables. Other reported work explore correlation, multiple linear regression, decision trees and neural networks to explore potential predictive models for crop yields with a varying degree of success [20, 9, 1]. Our approach differs from the existing body of published work by predicting soybean yields for specific variants of interest at the farm site level.

KDD'17, Month 8, 2017, Hallifax, N.S., Canada.

DOI: http://dx.doi.org/10.1145/12345.67890

2. CRITERIA FOR ELITE VARIETIES
An 'elite' variety is stably producing high soy yields across years and locations. Our primary goal was the development of quantitative models for predicting soy yields of each variant and the identification of elite soy variants. We used the median yield of commercial ('CHECK') varieties in each RM band as a benchmark for comparison (Table 1) and selected 15 elite varieties exhibiting yields 2.5 bu/ac higher than corresponding commercial varieties. These elite lines are expected to boost soy yields by about 4.5% relative to commercial varieties, in each RM. Time series analysis of these elite variants indicates stable productivity over the course of the study and reinforces their elite designation.

3. ESTIMATES OF TYPE I ERRORS
The current process of seed selection is exposed to potential Type I errors in which variants are designated as highly productive elite lines, yet fail to produce high yields in subsequent years. To evaluate the potential of our yield analytic procedure to reduce Type I errors we examined the potential utility of our Cubist ensemble (described in details in the next section) to reduce Type I error by applying our model to the class of 2013 to predict yields in 2014 using soil and remote sensing variables (RMSE=1.16, correlation =0.84, **Table 2**). When a similar 2.5 bu/ac gain over RM matched check lines is used as a cutoff to select elite lines, our algorithm eliminated six soy lines from the elite list due to insufficient yield consistency. We conclude that our predictive analytics method can successfully 'weed out' non-elite lines and enhance precision in soy seed selection by Syngenta.

Table 1. Predicted yields of elite soy variants (2014 class)

RM	VARIETY_ID	Predicted Yield (2015)	Check Yield	Yield Gain
2	V140364	60.9	55.1	5.8
2.1	V140393	56.9	54.2	2.8
2.2	V111237	60.5	55.9	4.6
2.5	V114553	60.7	58	2.7
2.5	V114655	61.3	58	3.3
2.6	V114569	62.9	60.2	2.6
2.7	V114530	62.3	59.8	2.6
2.8	V114564	63.8	60.9	2.9
2.8	V114585	63.6	60.9	2.7
2.8	V152312	63.5	60.9	2.6
3	V114565	62.1	59.3	2.9
3.1	V114589	62.8	59.5	3.3
3.1	V152320	64.1	59.5	4.6
3.5	V152324	62.9	59.7	3.1
3.5	V152415	62.4	59.7	2.7

Yields are reported in bu/ac

4. METHODOLOGY

We supplemented soil features provided by Syngenta with publicly available remote sensing data. Reflectance of vegetation in multiple light spectrums (vegetation indices) is a good estimator of crop yield, fruit ripening, biomass, and plant senescence [11,18, 25].Vegetation indices can be calculated from space and are conveniently supplied by NASA in Moderate Resolution Imaging Spectroradiometer (MODIS) and Normalized Difference Vegetation Index (NDVI) data products. Here we use a smoothed and gap-filled NDVI product generated for the conterminous US for the period 2012-Jan-01 through 2015-Dec-31. The data was obtained from the NASA Stennis Time Series Product Tool (TSPT) generating smoothed NDVI data from both the Terra satellite (MODIS MOD13Q1 product) and Aqua satellite (MODIS MYD13Q1 product) instruments. The data is provided at a spatial resolution of 250 m and a temporal resolution of 8 calendar days. NDVI was chosen over other vegetation indices due to its ability to cancel out a proportion of variability caused by changing sun angles, topography, clouds, and various atmospheric conditions [17]. We merged data from locations identical in their longitude, latitude, soil, and climatic variables (six sites). We then focused on sites that were present only in stage three of the class of 2014, thus NDVI data was collected for 58 locations. For each growing season per each location, we determined maximal NDVI value (NDVIMAX, in a 0 to 1 scale), the timing in the year of the NDVI peak (MAXTIME, and the sum of all NDVI values (SUMNDVI).

Surface temperature and soil moisture can be estimated successfully from space and provide valuable variables for soy yield prediction [13]. We obtained daytime and nighttime land surface temperatures (LST) from MODIS (product MOD11A1). This data was downloaded at a spatial resolution of 1 km and a temporal resolution of 8 calendar days for the 58 locations of interest. Plotting the NDVI and LST time series data reveals the expected seasonal periodicity in the remote sensing NDVI and LST data and the potential for yield prediction. The potential of remote sensing to explore correlations of surface temperatures and biomass production are illustrated in Figure 1 using remote sensing data from one location (location 4370) as an example. In this location low soy yields in 2012 are correlated with high day time temperatures and low NDVI scores (Figure 1).

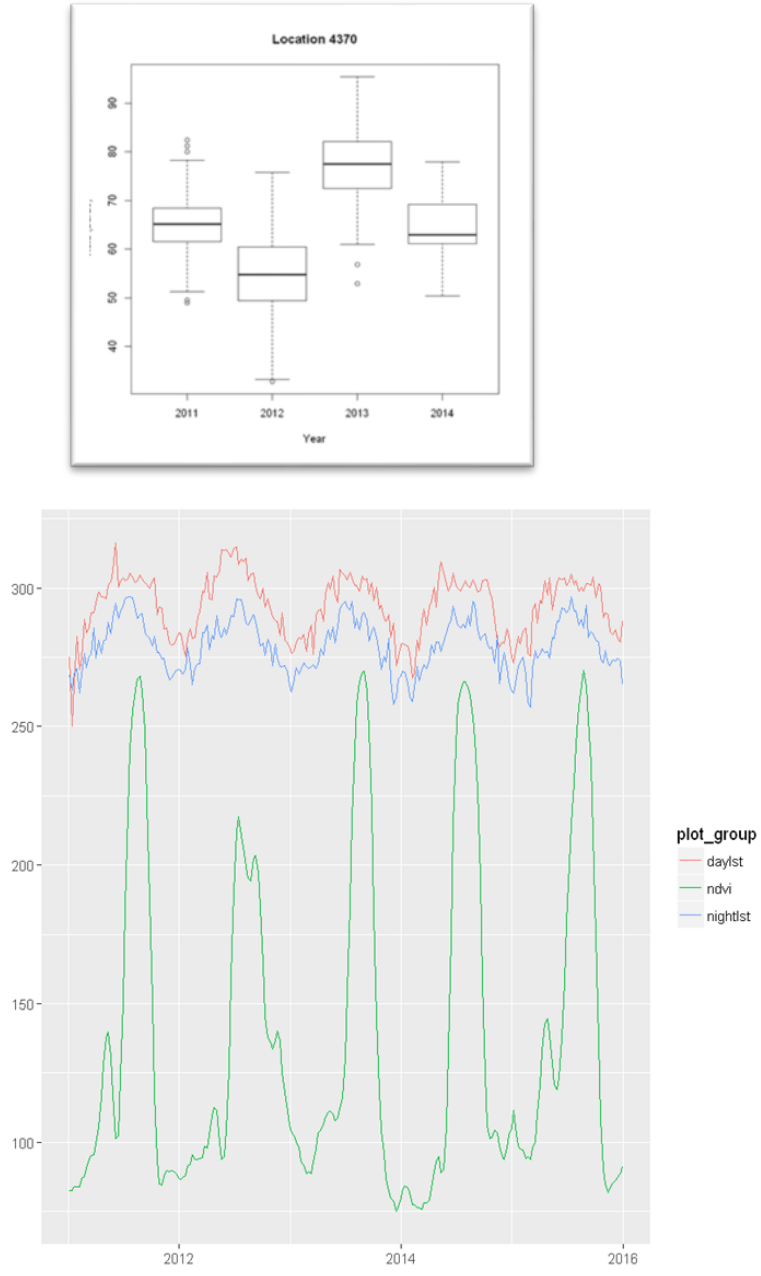

Figure 1: Yields (top), LST, and NDVI (bottom) in location 4370.

We extracted several features from the NDVI and LST time-series and performed correlation analyses to select features that correlate with soy yields (Figure 2). We found a positive correlation of NDVIMAX with soy yield and a negative correlation of JULYSUM and AUGSUM with soy yield. In addition strong

negative correlation of NDVIMAX with JULYSUM and AUGSUM are noticeable. JULYSUM and AUGSUM are also strongly correlated with each other. The sensitivity of soy to high temperature is well known, and we surveyed cumulative monthly daytime temperatures and minimum nightly temperatures as potential predictors of soy yield [6,7].We identified cumulative daytime temperature in July and August (JULYSUM, AUGSUM), and monthly minimum nighttime temperatures in May and July (MAYMIN, JULYMIN), as most inverse correlated with yields and selected these variables as predictive variables in our models (Figure 2).

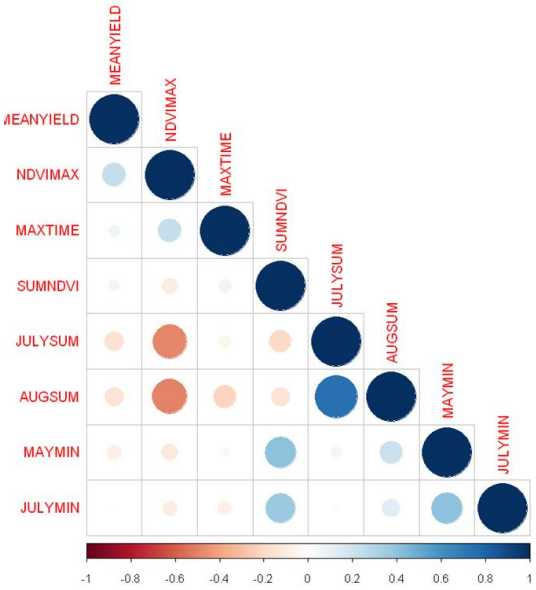

Figure 2: Correlation plots of remote sensing variables and soy yields.

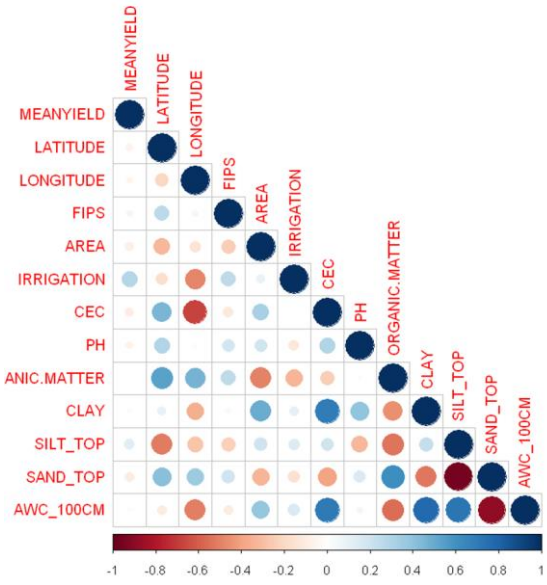

Figure 3: Correlation plots of soil variables and soy yield.

We also surveyed the correlations of soil variables with each other and with soy yields (Figure 3). The strongest correlation was observed with the irrigation variable. We decided to retain all soil variables in our models, despite strong correlations of some soil variables with each other.

Decision tree based methods are versatile and powerful machine learning techniques that are well suited for this challenge. Our preliminary analysis compared random forests [14] and Cubist [22] models for feature discovery and model tuning in the R software package [23]. We used the training data set that we developed above to compare the predictive potential of the randomForest and Cubist packages in R. With minimal tuning, randomForest models achieved lower error rates relative to Cubist models. However, when random Forest models where cross validated using data outside of the training set (or when applied to real test data on the CodaLab portal) we noticed poorer predictive power relative to Cubist models (Random Forest models achieved lower scores on the CodaLab platform). We postulate that Random Forest could be over fitting the models to the data and we therefore present here only our Cubist models.

5. QUANTITATIVE RESULTS

We used the Cubist package in R to generate regression decision trees for yields in each RM band using 21 predictor variables (VARIETY_ID, LATITUDE, LONGITUDE, FIPS, AREA, IRRIGATION, CEC, PH, ORGANIC.MATTER, CLAY, SILT_TOP, SAND_TOP, AWC_100CM, NDVIMAX, MAXTIME, SUMNDVI, JULYSUM, AUGSUM, MAYMIN, JULYMIN, FAMILY). The correlations of these models with the training data varied from 0.3 to 0.85 (Table 3). Out of fifteen RM bands with sufficient number of cases for modeling, we generated three high quality models (correlation coefficient ~0.85), eight medium quality models (correlation coefficient~0.7), and four models with poor predictive power (<0.4). Low quality models were typically generated when smaller datasets were used for training (average of 114 cases) and are thus less reliable for yield prediction. Generation of more experimental data for variants in RM bands with low data concentration is recommended to enhance RM coverage by elite soy variants (RMs: 2.3, 2.9, 3.1, 3.3 and 3.6). We validated this modeling procedure by applying this methodology to 'predict' the yields of the class of 2013 in 2014 (see section 3, Table 2). We demonstrated the utility of the 'divide-and-conquer' procedure to (a) predict yields from soil and remote data and (b) select consistently high yielding soy variants. We than used these models to predict the yields of 28 soy variants grown in 58 locations in 2015 (Table 1, using soil and remote sensing variables) and obtained high prediction accuracy on the Coda-Lab portal (0.45 FMEASURE, 0.99 ACCURACY, 0.47 MATHEWSCC) reflecting the predictive value of our approach.

One advantage of using modern decision tree tools is the interpretable ranking of variables according to their contribution to the model. Here we list top four variables in each RM model (Table 3). Notably, the high variability in ranking orders likely reflects the high heterogeneity among different RM bands and supports our decision to segment the data according to RM. Grouping RM into proximity zones (RM 2.0-2.5, RM2.6-2.9, RM 3.0-RM3.5) drove down internal RMSE values to about 4. However the Codalab score generated from RM zones models were below scores obtained with the separate RM models. Future work will focus on (a) adding additional remote sensing variables to our models to enhance precision and (b) utilizing time-series prediction methods to extrapolate future NDVI and LST values using data trends in each location to allow true predictions of future yield values.

6. ACKNOWLEDGMENTS

We thank Syngenta for providing data and funding for this project, IdeaConnection and AI For Good for organizing this competition and the R community for developing and maintaining the tools used in this analysis.

7. REFERENCES

[1] Ainong, L., Shunlin, L., Angsheng, W., Jun, Q. (2007) Estimating Crop Yield from Multi-temporal Satellite Data Using Multivariate Regression and Neural Network Techniques. Photogrammetric Engineering & Remote Sensing, 10, 1149-1157

[2] Alliprandini, L. F., C. Abatti, P. F. Bertagnolli, J. E. Cavassim, H. L. Gabe, A. Kurek, M. N. Matsumoto, M. A. R. de Oliveira, C. Pitol, L. C. Prado, and C. Steckling. 2009. Understanding Soybean Maturity Groups in Brazil: Environment, Cultivar Classification, and Stability Crop Sci. 49: 801-808.

[3] Al-Kaisi, M.M., Elmore, R.W., Guzman, J.G., Hanna, H.M, Hart, C.E., Helmers, M.J., Hodgson, E.W., Lenssen, A.W., Mallarino, A.P., Robertson, A.E., and Sawyer, J.E. 2012 Drought impact on crop production and the soil environment: 2012 experiences from Iowa. Journal of Soil and Water Conservation. 68, 19 – 24

[4] Altieri, M.A., Nicholls, C.I., Henao, A. et al. Agroecology and the design of climate change-resilient farming systems (2015) *Agron. Sustain. Dev.* **35,** 869 – 890.

[5] Das, A., Qi, J. T., Oneto, M., Shere, S. & Ma, Z. Soybean Yield Prediction: Using Satellite Imagery to Predict Commodity Yields. (2016).

[6] Djanaguiraman, M., P. V V Prasad, D. L. Boyle, and W. T. Schapaugh. 2013. "Soybean Pollen Anatomy, Viability and Pod Set under High Temperature Stress." *Journal of Agronomy and Crop Science* 199: 171–77.

[7] Djanaguiraman, M, P.V. Vara Prasad, and W. T. Schapaugh. 2013. "High Day- or Nighttime Temperature Alters Leaf Assimilation, Reproductive Success, and Phosphatidic Acid of Pollen Grain in Soybean [*Glycine Max (L.) Merr.*]." *Crop Science* 53(4): 1594–1604.

[8] Doraiswamy, P.C., Hatfield, J.L., Jackson, T.J., Akhmedov, B., Prueger, J., Stern, A., (2004) Crop condition and yield simulations using Landsat and MODIS. *Remote Sens. Environ*, **92,** 548–559

[9] Elizondo, D.A., McClendon, R. W., Hoogenboom, G (1994) Neural Network Models for Predicting Flowering and Physiological Maturity of Soybean. American Society of Agricultural and Biological Engineers 37(3): 981-988.

[10] Hartman, G.L. Chang, H.X., Leandro, L.F. (2015) Research advances and management of soybean sudden death syndrome. Crop Protection. 73, 60 – 66.

[11] Hatfield, J.L., Prueger, J.H., 2010. Value of using different vegetative indices to quantify agricultural crop characteristics at different growth stages under varying management practices. Remote Sens. 2, 562–578.

[12] Henderson Communications LLC "Syngenta Seeds Develops First Ever Soybean Relative maturity Map for Canada (2009) AgriMarketing Global Hub for Agribusiness

[13] Holzman, M. E., Rivas, R., & Piccolo, M. C. (2014). Estimating soil moisture and the relationship with crop yield using surface temperature and vegetation index. International Journal of Applied Earth Observation and Geoinformation, 28, 181-192.

[14] Liaw A. and Wiener M.(2002). Classification and Regression by randomForest. R News 2(3), 18--22.

[15] Lobell, David B, and Claudia Tebaldi. 2014. "Getting Caught with Our Plants down: The Risks of a Global Crop Yield Slowdown from Climate Trends in the next Two Decades." Environmental Research Letters 9(7): 74003.

[16] Maccherone, B. & Frazier, S. MODIS Vegetation Index Products (NDVI and EVI). Available at: https://modis.gsfc.nasa.gov/data/dataprod/mod13.php.

[17] Matsushita, B., Yang, W., Chen, J., Onda, Y. & Qiu, G. Sensitivity of the Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI) to Topographic Effects: A Case Study in High-density Cypress Forest. Sensors 7, 2636–2651 (2007).

[18] Merzlyak,M.N., Gitelson, A.A., Chivkunova, O.B., Rakitin, V.Y., 1999. Non-destructive opti- cal detection of pigment changes during leaf senescence and fruit ripening. Physiol. Plant. 106, 135–141.

[19] Myers, S. S. et al. Climate Change and Global Food Systems: Potential Impacts on Food Security and Undernutrition. Annu. Rev. Public Health 38, 259–277 (2017).

[20] Norouzi, M., Ayoubi, S., Jalalian, A., Khademi, H., Dehghani, A.A. Predicting rainfed wheat quality and quantity by artificial neural network using terrain and soil characteristics (2010) Acta Agriculturae Scandinavica 60, 341 - 352

[21] Powell, Nicola et al. 2012. "Yield Stability for Cereals in a Changing Climate." Functional Plant Biology 39(7): 539–52.

[22] Quinlan, J. R. (1992). Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343-348).

[23] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL: http://www.R-project.org/.

[24] Ren, J., Chen, Z., Zhou, Q., Tang, H. (2008) Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong China. Int. J. Appl. Earth Obs. Geoinf, 10, 403–413

[25] Yu, N., Li, L., Schmitz, N., Tian, L.F., Greenberg, J.A., Diers, B.W., Development of methods to improve soybean yield estimation and predict plant maturity with an unmanned aerial vehicle based platform (2016) Remote Sensing Environment 187, 97 - 10.

Table 2. Error estimates in the Class of 2013

RM	VARIETY_ID	FAMILY	Prediction (2014)	Check	Yield Gain	ELITE CALL	Yields (2014)
2.1	V156516	FAM14408	58.2	54.2	4.0	TRUE	60.2
2.1	V156786	FAM11169	58.2	54.2	4.0	TRUE	60.6
2.4	V156783	FAM11183	58.8	56.9	2.0	FALSE	58.8
2.5	V156565	FAM14238	63.1	58.0	5.1	TRUE	62.8
2.6	V156806	FAM11189	63.2	60.2	3.0	TRUE	64.1
2.7	V152079	FAM11179	62.3	59.8	2.5	TRUE	62.8
2.7	V156574	FAM14486	62.6	59.8	2.8	TRUE	62.1
2.7	V156807	FAM11189	62.6	59.8	2.8	TRUE	61.4
2.9	V156553	FAM14238	62.4	60.7	1.7	FALSE	61.7
3	V152053	FAM14333	60.7	59.3	1.5	FALSE	60.0
3	V156642	FAM14133	58.8	59.3	-0.4	FALSE	56.5
3.2	V156763	FAM06492	61.5	60.3	1.3	FALSE	61.4
3.2	V156797	FAM06502	61.5	60.3	1.3	FALSE	62.0
3.4	V152061	FAM14581	63.6	61.0	2.6	TRUE	63.2
3.5	V156774	FAM14774	65.1	59.7	5.4	TRUE	65.7

Yields are reported in bu/ac

Table 3: Evaluation of Soy Yield Models

RM	N	RMSE	Relative Error	Cor. Coeff.	Top Contributing Variables	Model Quality
2	115	8.9	1.01	0.48	organic.matter, augsum, area, julysum	Medium
2.1	257	4.14	0.53	0.85	sumndvi, julysum,ndvimax,maymin	High
2.2	211	7.28	0.82	0.64	ndvismax,julysum,maymin,augsum	Medium
2.3	57					
2.4	198	9.47	1.06	0.43	sand_top, silt_top, julysum,ndvismax	Low
2.5	358	5.18	0.66	0.76	julysum,clay,ndvimax,longitude	Medium
2.6	190	7.05	0.89	0.56	maymin,julysum,sumndvi,longitude	Medium
2.7	426	4.94	0.61	0.76	clay,pH,sumndvi, irrigation	Medium
2.8	402	5.11	0.65	0.76	clay,pH, irrigation, cec	Medium
2.9	143	8.79	1.09	0.29	irrigation,cec,maymin,julymin	Low
3	395	6.39	0.71	0.71	maymin,julysum,irrigation,cec	Medium
3.1	105	11.08	1.12	0.23	latitude, sumndvi,fips,julymin	Low
3.2	310	4.9	0.52	0.84	irrigation, sand_top,maxtime,augsum	High
3.3	0					
3.4	442	4.64	0.51	0.85	julysum, cec, sand_top, organic.matter	High
3.5	256	6.23	0.67	0.75	cec, julysum, ndvimax, maymin	Medium
3.6	161	9.45	1.13	0.3	julysum,awc_100cm,latitude, julymin	Low