Agriculture Case Study

Spencer Ebert & Jeremy Meyer

Brigham Young University

1 Introduction

Irrigation water efficiency is a common interest for farm managers in water-scarce environments. The Crop Water Stress Index (CWSI) can be used to measure when crops need water. CWSI is measured on a scale of 0 to 1, with a 1 being crops that need the most water. This is easily measured from surface temperature readings. While this is useful for knowing when to add water, quantifying how much water to add requires information about the Soil Water Content (SWC). The SWC metric is much more expensive and arduous to obtain. However, if a relationship can be found between CWSI and SWC, farmers can gain information about SWC without many expenditures. Thus, the goal of this analysis is to find a unique relationship between SWC and CWSI so farmers can gain information how much water to add from measurements that are less expensive.

1.1 The data

The data we will use for the analysis is shown in Figure 1. It contains 78 different measurements of both SWC and CWSI. Although the relationship is mostly negative, it is non-linear because it appears to follow a curve. However, there is still a strong association between both variables, so information from CWSI can reasonably be used to determine SWC.

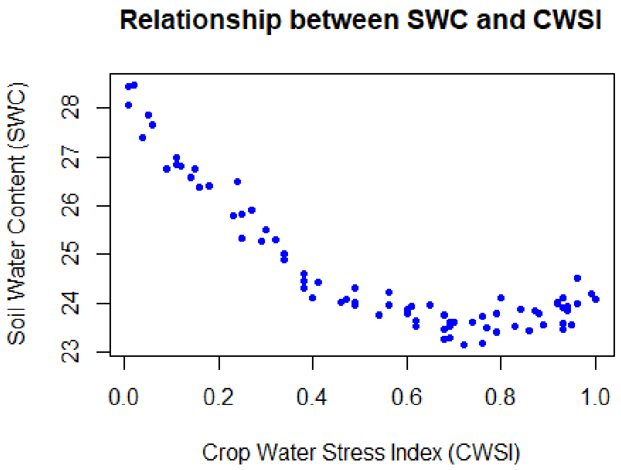

Fig. 1.

2 The model

A classic approach to capture non-linear trends is to use natural cubic splines. Cubic splines work by splitting the data into a number sections and fitting cubic polynomials in each section. The boundary points of these sections are called knots and the entire curve is continuous and differentiable twice at those points. The result is a smooth curve that is far more flexible than a simple linear fit. We chose natural splines because of their ability to capture the curved nature of the CWSI data and thus help find a relationship between CWSI and SWC.

Natural cubic splines, as opposed to just cubic splines, have linear constraints at the ends of the data so the behavior outside the range of the data is greatly controlled. Even though CWSI is only between 0 and 1 and we won't have to predict outside the range of the data, using natural splines does not harm the analysis and is generally a good practice for non-linear data sets. Since natural cubic splines are an extension of the linear model, we can generate predictions for any value of CWSI to predict Soil Water Content. This will allow farmers to get estimates of the amount of water to add from CWSI information.

Cubic splines use basis function expansions to generate the smooth, fitted curve. Basis function expansions are functions that can extract features or attributes from the explanatory variables. The basic form of a basis function expansion is as follows:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + ... + \beta_p b_p(x_i) + \epsilon_i$$

Where y is the response variable (SWC), the β terms are the linear model coefficients, ϵ is the error term for the model, x_i is the predictor for the i^{th} observation and $b_p(x_i)$ is the nth basis function expansion for x_i. The model of a cubic spline we will consider to model the data is structured follows:

$$y_i = \beta_0 + \sum_{p=1}^{K+3} b_p(x_i)\beta_p + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \tag{1}$$

where the p^{th} basis function expansion for the predictor x_i, $b_p(x_i)$, is equal to:

$$b_p(x_i) = \begin{cases} x_i^p & p \leq 3 \\ (x_i - \xi_{p-3})_+^3 & p > 3 \end{cases} \tag{2}$$

The k^{th} break point value or knot is represented by ξ_k and K represents the total number of knots. The plus sign subscript on the basis function represents a truncated power basis function. This behaves similarly to an indicator function in that:

$$(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & x_i > \xi_k \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Putting everything together, the model for the cubic spline turns into:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \sum_{k=1}^{K} (x_i - \xi_k)_+^3 \beta_{k+3} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \tag{4}$$

Thus the cubic spline model acts as a multiple regression on K+3 predictors with an intercept. The values for each of the predictors is determined by a basis function expansion. Natural cubic splines impose additional constraints in that the curve beyond the range of the data becomes linear.

3 Justification & Performance

Before we fit the final model with natural cubic splines, we must choose the optimal number of knots. We accomplished this by cross validation. We sampled 70% of the data to be used as a training set for 1000 repetitions. We built the natural cubic spline model with a fixed number of knots from the training set and the mean squared error (MSE) was computed across the test set. These values were averaged across all 1000 iterations and done for a varying number of knots. We tested this for 1-14 knots and compared by finding the lowest mean squared error (MSE). The results are shown in the graph in Figure 2.

Although the natural spline with 7 knots had the lowest MSE, the resulting spline plotted with the data is very jittery. We decided to choose 2 knots as optimal because it has a smoother fit and due to the relatively small size of our data, it runs a lower risk of overfitting to the data we observed. The difference in MSE is very minimal (.0013). Comparisons between 2 knots and 7 knots can be seen in Figure 3.

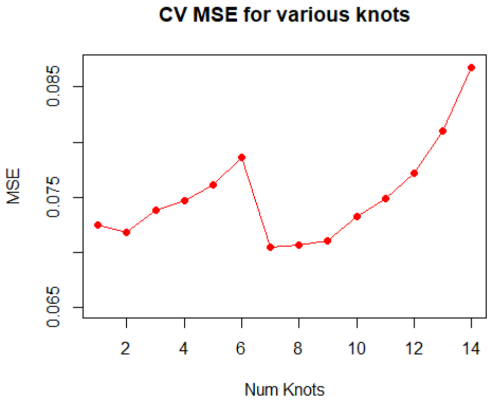

Fig. 2.

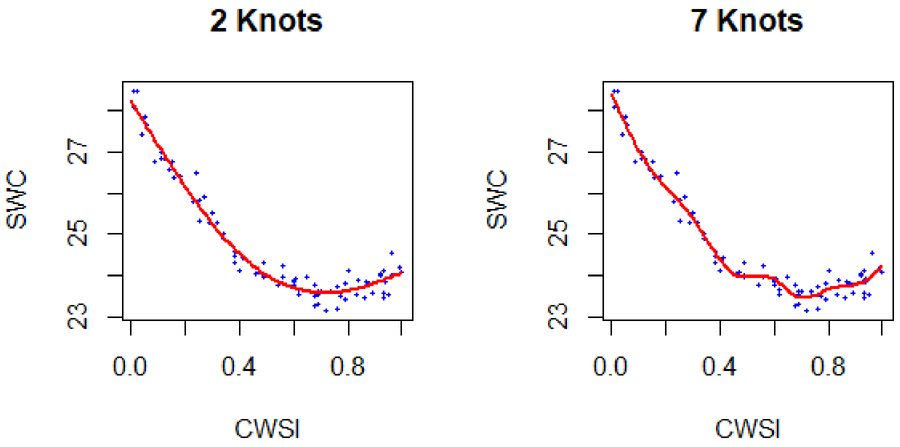

Fig. 3. Note that the 7 knot spline does not look as smooth

3.1 Assumptions

There are three assumptions that must hold to use natural cubic splines. They are independence between residuals, and normality and homoscedasticity of the residuals. For the independence and homoscedasticity (equal variance) assumptions we look at a fitted residuals plot in Figure 4.

This plot doesn't seem to show any correlation between the residuals and we have no reason to assume that there is correlation from the way the data was gathered so we are confident in assuming that there is independence between the residuals. The red dotted lines are used to look at equal variance. They are 2 standard deviations of the residuals away from 0. From this graph, it looks as though the residuals have equal variance for all of the fitted values so we are confident in our homoscedasticity assumption.

To check our normality assumption we look at a histogram of the residuals in figure 5. The residuals look mostly normal and there don't seem to be any outliers so we are confident in our normality assumption.

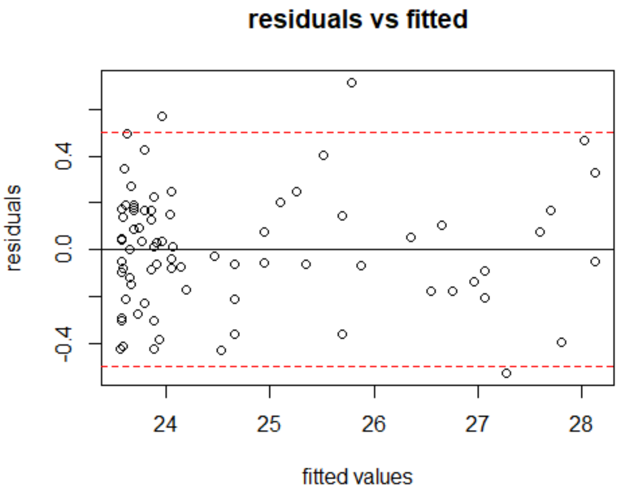

Fig. 4. For checking assumptions of independence and equal variance of residuals

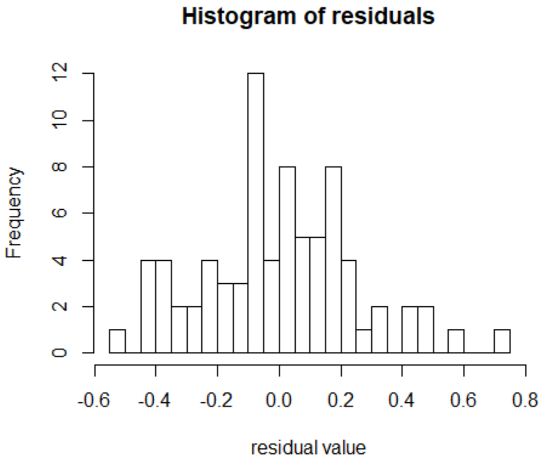

Fig. 5. For checking assumption of normality with residuals

3.2 Prediction Performance / Model Fit

Two important questions with the model is how well it predicts and how well it fits the data. To answer the former, we tested for bias, rMSE, prediction interval width and coverage. This was done through cross-validation by randomly selecting 70% of the data to train a model and holding out the rest for testing. We found the average bias of the testing set estimates, the rMSE of the trained model on the testing set, the average prediction interval width at each testing set point, and the average coverage of the prediction intervals. This process was repeated 1000 times and the results were aggregated by averaging across all iterations. The results are in table 1.

Table 1. Results for prediction performance

Bias	rMSE	PI Width	Coverage
.0812	0.2647	1.056	94.42%

From the table, we can see that the predictions are unbiased and that the prediction intervals are only 1.05 units wide. That is only 19% of the total range in the data, which shows that our model does a great job at narrowing the possible range of predicted values. The coverage is as expected for a 95% confidence interval and the rMSE is 0.2647, which is the standard deviation estimate of the error around the line. All of these values suggest that our model predicts well. To evaluate model fit, we computed the testing set R^2 value for each of the 1000 iterations. The average R^2 was 0.961, which means that our model accounts for about 96.1% of the variation of the SWC around its mean. The high R^2 value can be seen from Figure 3 because of how close the data points are to the line.

3.3 Comparison to Other Methods

The other method we looked at in our analysis was kernel smoothing using a normal kernel. The basic idea behind kernel smoothing is to fit a function locally to predict specific points. We used a normal kernel to find our predictions, because it gives a smooth line for predictions and uses all the points for each prediction (the weights are adjusted for each point). The formula for normal kernel smoothing is as follows:

$$\hat{y}(x_0) = \frac{\sum_{i=1}^{N} K_\delta(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\delta(x_0, x_i)} \tag{5}$$

$$K_\delta \propto exp(\frac{1}{2\delta^2}(x_i - x_0)^2) \tag{6}$$

In this equation, $\hat{y}(x_0)$ is the predicted SWC value for the particular CWSI and x_i are all of the other CWSI values.

To assess the optimal value of δ, we used leave one out cross validation and compared MSE's. We got $\delta = 0.034$, but we wanted our curve to be smoother so we adjusted δ to 0.07. To calculate uncertainty for our predictions, we used bootstrapping to find 95% confidence intervals and prediction intervals.

One of the issues when using kernel smoothing is bias at the limits. At the end points there aren't any values on the other side of the limits to influence the prediction. In the case for CWSI and SWC, when CWSI is around 0 there aren't any points less than 0 to influence the prediction, thus all of the prediction is based on points greater than that CWSI making the prediction a little lower than it should be. Figure 8 illustrates this point and shows the values tapering off when CWSI is between 0 and 0.2.

On the other hand, when using natural cubic splines predictions values near the limits aren't biased so we get better predictions on the CWSI interval 0 to 0.2. Farmers are more interested in finding predictions for higher SWC values which correspond to CWSI between 0 and 0.2, thus we chose to use a natural cubic spline to get better predictions in that interval.

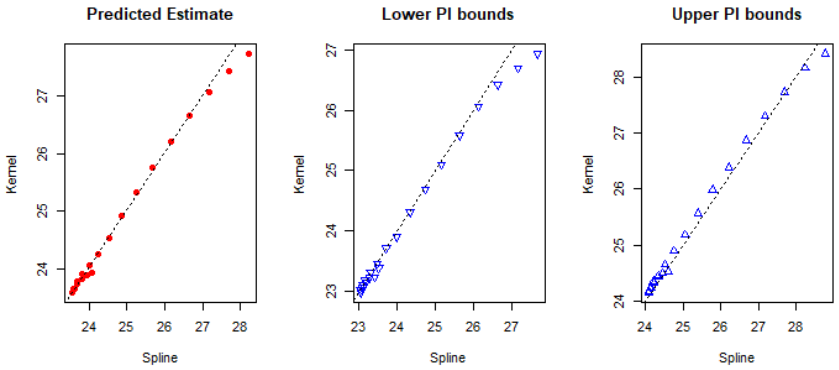

Fig. 6. These graphs compare the predictions and 95% prediction intervals using the two different methods. As can be seen in the graphs, for higher values of SWC (27-28) kernel smoothing tends to predict lower than natural cubic splines. That represents the bias of kernel smoothing near the limits. Overall though, the predictions tend to be pretty similar.

We also compared the cubic spline model and normal kernel smoothing method by looking at R^2, MSE, and average Prediction interval width. Table 2 gives the results. The natural cubic spline model did better than the normal kernel smoothing for both the MSE and prediction interval width. R^2 was about the same for both methods. These diagnostics give us more confidence in using a natural cubic spline because predictions tends to be a little more accurate.

Diagnostic	Natural Cubic Spline	Normal Kernel Smoothing
R^2	0.961	0.962
MSE	0.070	0.091
PI Width	1.062	1.244

Table 2. This table gives diagnostics for kernel smoothing and natural cubic splines

4 Results

By using natural cubic splines we were able to come up with SWC estimates for given CWSI values and confidence and prediction intervals associated with that estimate. Tables 3 and 4 give predictions using a cubic spline and kernel smoothing. We suggest that farmers use the table for cubic splines because those predictions tend to be more accurate. If a farmer wants a particular SWC value, they don't have to spend extra money to get that measurement. Instead, a farmer can look at their target SWC value found in the table (Pred SWC) and look at the CWSI value corresponding to that prediction. Then they can adjust how much they water to raise or lower CWSI to reach that value which would result in their desired SWC.

Parameter estimates in the case for cubic splines aren't very interpretable so it is more helpful for farmers to look at the predicted values rather than relying on the interpretation of the β coefficients. A farmer can look at SWC prediction uncertainty by looking at the intervals. While there are confidence intervals (CIs) for the estimates, the prediction intervals (PIs) may prove more useful to farmers. This is because prediction intervals are interpreted as a range of values that the next new observation could be. In repeated sampling at a CWSI level of 0, we would expect that the next observation will lie in the interval 27.69 and 28.78 about 95% of the time. Whereas, confidence intervals only asses uncertainty in the expected SWC value.

Table 3. This table gives the predicted values when using a cubic spline

CWSI	Pred SWC	2.5% CI	97.5% CI	2.5% PI	97.5% PI
0.00	28.237	28.042	28.432	27.691	28.783
0.05	27.702	27.547	27.857	27.169	28.235
0.10	27.173	27.051	27.295	26.648	27.698
0.15	26.657	26.555	26.759	26.137	27.177
0.20	26.162	26.063	26.260	25.642	26.681
0.25	25.694	25.588	25.799	25.173	26.215
0.30	25.261	25.147	25.375	24.738	25.784
0.35	24.871	24.752	24.990	24.347	25.395
0.40	24.530	24.415	24.646	24.007	25.054
0.45	24.243	24.137	24.349	23.722	24.764
0.50	24.008	23.912	24.104	23.489	24.527
0.55	23.824	23.733	23.916	23.306	24.343
0.60	23.690	23.595	23.785	23.171	24.209
0.65	23.605	23.502	23.707	23.084	24.125
0.70	23.566	23.459	23.674	23.045	24.088
0.75	23.574	23.470	23.678	23.053	24.095
0.80	23.623	23.528	23.718	23.104	24.142
0.85	23.705	23.616	23.794	23.187	24.223
0.90	23.812	23.712	23.913	23.292	24.332
0.95	23.936	23.804	24.068	23.409	24.463
1.00	24.068	23.891	24.245	23.528	24.608

Figures 7 and 8 can also be used by farmers as a visual for their predictions. These figures show predicted estimates, the data, and prediction and confidence intervals for both methods. Based on their CWSI a farmer can adjust watering to move to SWC points they desire as seen on the graph.

Table 4. This table gives the values when using normal kernel smoothing

CWSI	Pred SWC	2.5% CI	97.5% CI	2.5% PI	97.5% PI
0.00	27.739	27.219	28.069	26.934	28.404
0.05	27.444	26.979	27.819	26.703	28.155
0.10	27.077	26.773	27.424	26.425	27.725
0.15	26.667	26.435	26.892	26.062	27.297
0.20	26.213	25.941	26.453	25.580	26.871
0.25	25.756	25.494	26.007	25.101	26.386
0.30	25.326	25.050	25.600	24.690	25.984
0.35	24.909	24.664	25.185	24.309	25.561
0.40	24.537	24.354	24.766	23.897	25.183
0.45	24.251	24.128	24.398	23.717	24.899
0.50	24.053	23.963	24.154	23.452	24.648
0.55	23.904	23.811	24.000	23.307	24.456
0.60	23.771	23.667	23.872	23.173	24.357
0.65	23.655	23.554	23.762	23.104	24.234
0.70	23.582	23.490	23.696	23.020	24.151
0.75	23.577	23.475	23.699	22.971	24.175
0.80	23.642	23.532	23.761	23.056	24.232
0.85	23.737	23.633	23.834	23.149	24.326
0.90	23.820	23.709	23.927	23.214	24.433
0.95	23.883	23.745	24.011	23.229	24.493
1.00	23.936	23.770	24.078	23.381	24.525

5 Conclusions

This analysis adequately addressed the research goal of this study. We were able to model the relationship between CWSI and SWC and calculate useful predictions for SWC based off of CWSI. Based off of these predictions, a farmer would not have to just rely on spending money on SWC measurements, they can use the predicted values to accomplish their desired SWC. They could simply look at Table 2 for an estimate and uncertainty interval for SWC.

Although the cubic spline model fit relatively well, one big shortcoming is simply that 78 data points are not enough for us to be really confident in making large-scale decisions in the amount of water to be used. Careful study of additional data can help us be even more precise in defining the relationship between the two variables. Another major shortcoming from cubic splines is that we have lost interpretability of the model: it becomes a black box that spits out predictions. It would be nice to quantify general patterns while studying the relationship between SWC and CWSI.

It would useful in the analysis going forward to receive more data on the relationship between CWSI and SWC and then see how well the model predicts. It would also be interesting to include more covariates in our model. Is the relationship between CWSI and SWC different depending on the type of soil or the crops being planted? Is there a spacial relationship between where CWSI is measured and SWC values? Another thing we would be interested in investigating is the relationship between SWC and crop yield to find the optimal SWC to maximize yield.

6 Teamwork Statement

Jeremy performed the cubic spline analysis and wrote the introduction, model definition, prediction performance/model fit section, and the shortcomings part of the conclusion.

Spencer performed the normal kernel smoothing analysis wrote the assumptions, comparison to other methods, results, and conclusion.

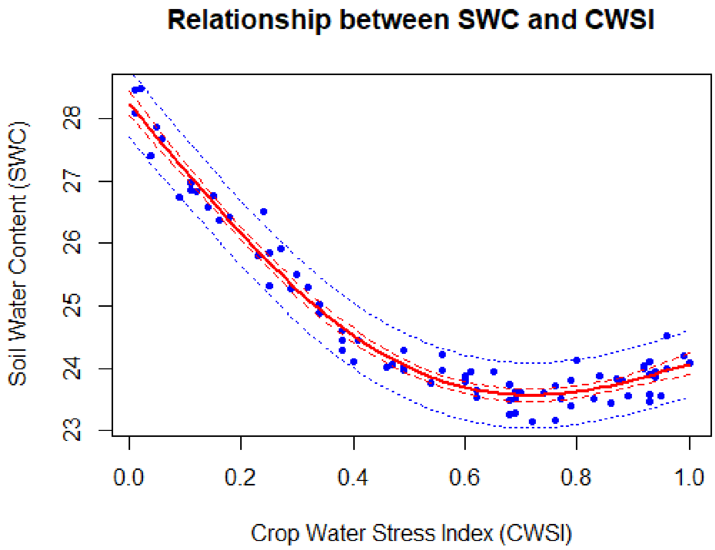

Fig. 7. Natural Cubic Spline

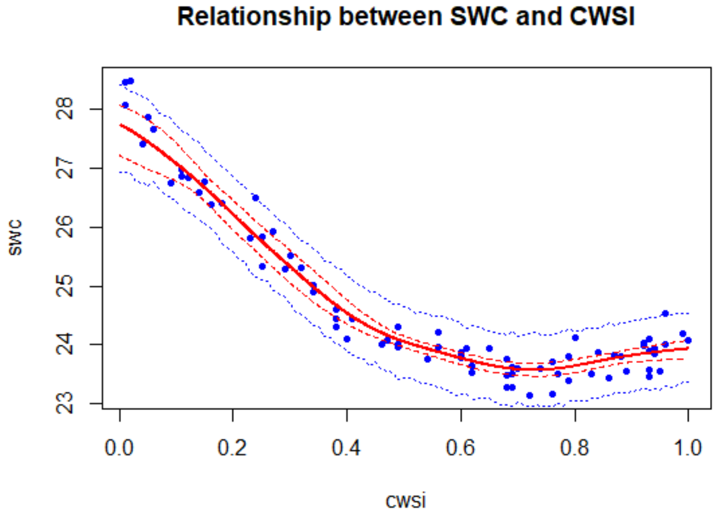

Fig. 8. Kernel Smoothing