Predicting Soil Water Content from Crop Water Stress Index

Stephen Merrill & Maddie Phan

Introduction

Irrigation assists farmers in distributing water in an efficient manner. As water is limited and therefore expensive, the efficient use of water is vital to the success of farmers. In addition to using water for a productive crop yield, producers can also sell surplus water at an increased price. The goal of this analysis is to predict soil hydration from the appearance of the plant. A low cost system to measure how much water a plant needs will aid the farmers in maximizing profits.

The hydration of the soil is not easily measured; however, the soil hydration can be estimated by visually measuring the approximate hydration of the plant. Soil hydration is referred to as soil water content, or SWC. CWSI is the crop water stress index which measures how dehydrated the plant appears to be. The purpose of this analysis is to predict SWC from CWSI.

Data Exploration

In order to accurately predict SWC, we need to understand the relationship between CWSI and SWC. Figure 1 reveals that the relationship appears to be non-linear. Non-linearity restricts the type of analysis that can be used for prediction. In addition, we are not assuming the plant samples are independent. It is logical that if the soil is dehydrated in a location, the soil near the original sample will have a similar level of hydration. Dependence will also limit the type of analysis that can be conducted. Analysis that can compensate for both dependence and non-linearity will be discussed in future sections.

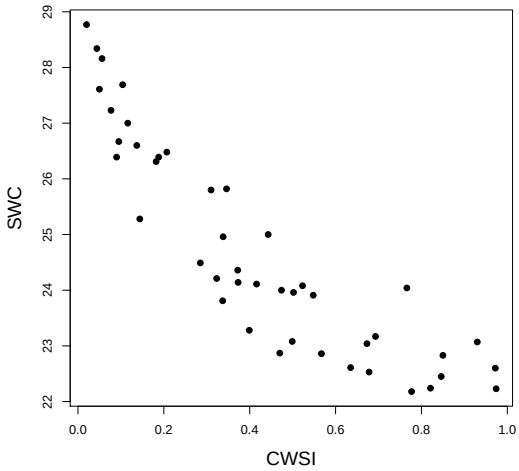


Figure 1: Note the non-linear relationship between CWSI and SWC. The goal of the analysis is to predict the soil hydration from that appearance of the plant's hydration.

The data comes from land in Southern Colorado and contains information on 44 samples. CWSI can only take on values from 0 to 1, where 0 indicates virtually completely dehydrated and 1 indicates full hydration. The mean CWSI value for the 44 observations is 0.42. The mean SWC value is 24.7. Table contains other summary statistics.

	X	CWSI	SWC
Min.	1.00	0.02	22.18
1st Qu.	11.75	0.17	23.06
Median	22.50	0.39	24.12
Mean	22.50	0.42	24.70
3rd Qu.	33.25	0.64	26.39
Max.	44.00	0.97	28.77

Table 1: Summary Statistics for SWC & CWSI

Methods

Gaussian process regression is a viable method for analyzing data that is non-linear with correlated observations. A Gaussian process is a type of Stochastic process. A Stochastic process is a collection of random variables over a specified interval, where only a finite collection of random variables are observed. In this case, we observe 44 observation of SWC which correspond to CWSI. A Gaussian process is a stochastic process where any finite collection of random variables follow a multivariate normal (MVN) distribution. When we observe SWC at various CWSI, we need to find a function such that the residuals are small and the function is smooth. If we let the function at each CWSI be random, then we can model the function with a distribution. This can be notated as follows:

$$Y(x_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_1 00 x_{100i} + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

x_{1i} : Closest predicted CMAQ value for location i

x_{2i} : Second closest predicted CMAQ value for location i

$\vdots$

x_{100i} : 100^{th}closest predicted CMAQ value for location$i$

$$Y(x_i) = \beta_0 + z_{1i}\theta_1 + z_{2i}\theta_2 + \cdots + z_{36i}\theta_{35} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

z_{1i} : The first principal component for ozone i

z_{2i} : The second principal component for ozone i

$\vdots$

z_{36i} : The 36^{th} principal component for ozone$i$

$$\boldsymbol{Y} \mid \boldsymbol{W} = \begin{bmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{bmatrix} \sim N(\boldsymbol{W}, \tau^2 \boldsymbol{I}_N) \tag{1}$$

Where $y(x_i)$ denotes the SWC for the i^{th} plant or CWSI.
x_i denotes the i^{th} CWSI measurement.
$\boldsymbol{Y} \mid \boldsymbol{W}$ represents the distribution of the of the SWC given the function.
$\boldsymbol{W}$ is the collection of random variables that describes the non-linear relationship of CWSI with SWC.

τ^2 is the variance in the distribution of SWC given W.
σ^2 is the variance in the distribution of W.
μ is the constant mean of the Gaussian Process.
$\boldsymbol{R}$ is the variance of $\boldsymbol{W}$.

$$\boldsymbol{W} = \begin{bmatrix} w(x_1) \\ \vdots \\ w(x_N) \end{bmatrix} \sim N(\mu \mathbf{1}_N, \sigma^2 \boldsymbol{R}) \tag{2}$$

Where $w(x_i)$ denotes the function for the i^{th} CWSI. x_i denotes the i^{th} CWSI measurement. $w(x)$, the Gaussian Process, also needs to identify a covariance structure. We use a Mattern function, which satisfies the requirements of being positive definite and being strongly correlated over small distances. To this end we use a Mattern function with parameters α and ν which govern decay, or change in correlation, and smoothness, respectably.

By definition, we can obtain the joint distribution of the SWC and the function. After we have the joint distribution, we can integrate over $\mathbf{W}$ to obtain the marginal distribution of $\mathbf{Y}$. The distribution of $\mathbf{Y}$ can be expressed as:

$$\boldsymbol{Y} \sim N(\mu \mathbf{1}_N, \sigma^2 \boldsymbol{R} + \tau^2 \boldsymbol{I}_N) \tag{3}$$

Since the goal of the analysis is to predict SWC based on CWSI, we need to use our model to predict. First, we decide the CWSI values where we are intending to predict. Then, by the properties of the MVN, we can obtain the conditional distribution of $\boldsymbol{Y^*} \mid \boldsymbol{Y}$. By the process described above, we know the mean and variance of the conditional distribution. In other words, we can find the mean and variance of the predicted values, given the observed SWC values. Additionally, we can obtain prediction intervals by taking the corresponding quantiles of the conditional distribution using the empirical rule.

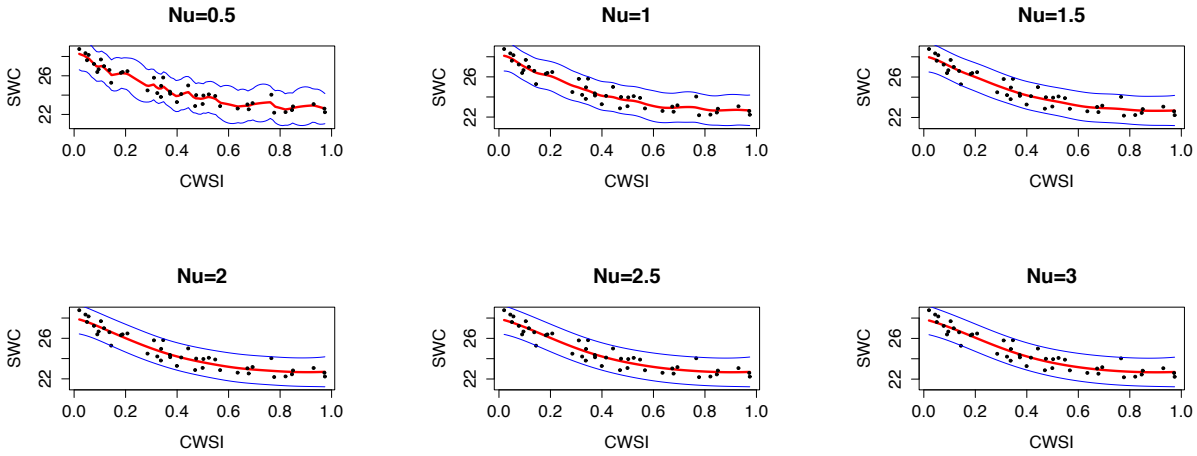
Gaussian process regression (GPR) does not carry the same assumptions as linear regression. GPR is not restricted to model only linear relationships, which indicates that linearity is not an assumption of the method. The method also predicts relationships when observation are not independents. This indicates that dependence will be present and the variance will be relative to the amount of data. Therefore, the only assumption that is made with GPR is that the residuals are normally distributed.

Model Justification

Gaussian Process Regression allows us to model a non-linear relationship between SWC and CWSI. Additionally, this method allows for us to model data where various CWSI are dependent on each other. The model was chosen through the GPR described above.

As discussed previously, GPR assumes that the residual values are normally distributed. In other words, the observed SWC values minus the predicted SWC values is normally distributed. However, this assumption is not easily verified as we only have one draw from the multivariate normal. Since we modeled all observed SWC values as one draw from the multivariate normal, one draw will yield a vector of length 44. Therefore, it is not possible to verify that a single draw comes from the MVN. We will move forward assuming that our data comes from a MVN distribution with the knowledge that the distribution is robust to misclassification. We will be aware that a violation on the normality assumption will increase the width of the prediction intervals. Additionally, our use of the t distribution for constructing the prediction intervals would be incorrect.

GPR also uses a decay and smoothness parameter. The decay parameter, α, is chosen using cross-validation. However, the smoothing parameter, ν, is not easily chosen. A small ν value is very sensitive to the data and the fitted line will be very jagged. Whereas a large ν value will may be too robust to the information in the data and ignore important trends. The figures below show predicted SWC for various ν values. We chose a ν of 1.5 as it seemed to be robust to smaller trends, while still capturing the global pattern of the data.



Performance Evaluation

This study is solely concerned with prediction of SWC, rather than inference, so we evaluate our model based on prediction performance. Since there are only n=44 observations we used a Leave-one-out Cross Validation (LOOCV) method to simulate predictions and obtain results (see Table 1 and Figure 2).

Table 2: Prediction Diagnostics

	Estimate	Lower	Upper
Bias	-0.001713	-1.385079	1.381653
RMSE	0.697747	-0.789005	1.263421
Coverage	0.931818	0.437794	1.425842
Interval Width	2.830984	2.624461	3.037506

Of particular note in Table 1 are bias and coverage. Although the bias interval ranges fairly widely, the overall prediction nature of our model is remarkably unbiased. This can be seen by comparing average distances between actual points and the predicted fit line in Figure 3 below, and is due to a property of the Multivariate Normal Distribution, which we assumed fit our data. The coverage statistic also supports the validity of our model in that it is so close to .95, the theoretical value with our assumed error rate. This means that the correct number of actual points were contained within the intervals of error for our predicted SWC values. The coverage can be seen graphically in Figure 2. All of the intervals in the table are quite wide and sometimes uninterpretable. That is a function of the low number of trials our sample size restricts us to and is a weakness of our results.

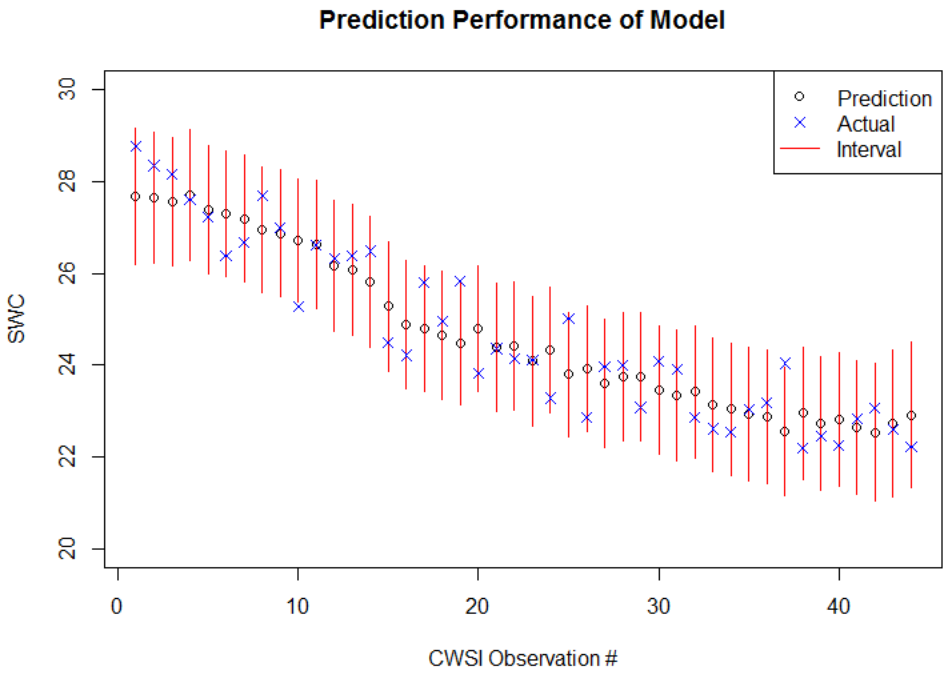


Figure 2: 93.18% of the prediction intervals generated through the cross-validation method contained the true SWC response value.

Results

Our results include estimates of the covariance parameters and a confidence interval for μ, presented below in Table 2.

Table 3: Parameter Estimates

σ^2	τ^2	α	ν	μ	μ CI
2.392	0.440	4.972	1.5	24.982	(24.970, 24.994)

These parameters are defined as: σ^2 and τ^2 are variances in Multivariate Normal Distributions that make up the model. α is the decay parameter, which governs the correlation of the regression line and ν is the smoothness parameter, which governs the smoothness of the regression line. ν cannot be estimated and is assigned based on visual inspection of different values. Both are parameters in the Mattern covariance function. μ is the mean of the regression line, or the mean of the predicted values of SWC.

The model addresses the motivation for the study by making accurate prediction of SWC and CWSI. The predictive accuracy was validated through a simulation. The final results, the effect of varying levels of CWSI on SWC, can be seen in Figure 3. There is a curved negative trend, fit by the red Gaussian Process regression line. In the future, farmers can use Figure 3 to predict values of SWC from observed CWSI levels.

Conclusions

In order for farmers to most efficiently use their irrigation water, it is of interest to be able to predict the SWC level of the soil from the CWSI level. Our proposed Gaussian Process model meets this goal by making

unbiased predictions on the non-linear relationship. Analysis of the prediction results shows a very effective approach to modeling the desired relationship.

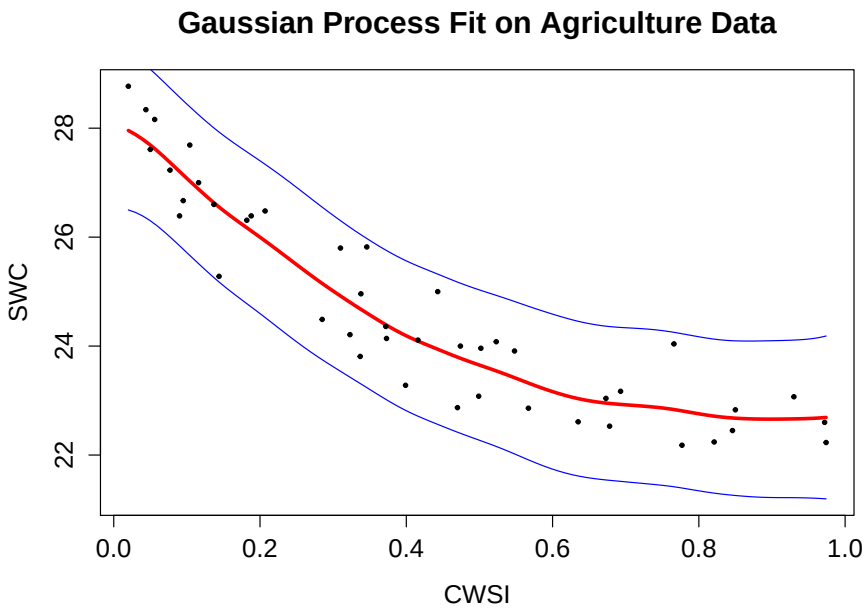


Figure 3: The red fitted line is nearly unbiased on average and the blue confidence intervals appropriately contain the data.

One shortcoming of the study is the low amount of sample data that creates wide uncertainty intervals in our results. In the future, if more data were gathered our results would be able to support the model more strongly. We could also consider a model with more covariates, such as location of the land plots, type of crops being grown, etc. However, in order for Gaussian Process Regression to continue to be accurate we will need to ensure there are no problems with high dimensional data.

Team Work

Stephen and I shared the work equally. Mostly, I told jokes and he laughed out of obigation.