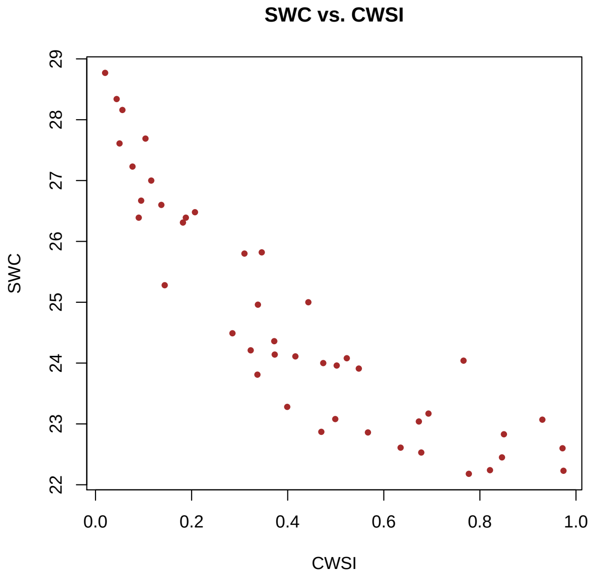

Figure 1: Scatter Plot of SWC vs. CWSI

Stat536 HW4 - Gaussian Process on Agriculture Data

Arthur Lui

26 February 2014

Introduction:

We are interested in predicting Soil Water Content (SWC) using Crop Water Stress Index (CWSI). SWC is a measure of water presence in soil, and CWSI is a measure of how severely water needs to be added (1 for severe, 0 for not severe - meaning crops are well hydrated). The motivation behind measuring CWSI and SWC is that they are good indicators of when and how much water should be supplied to crops. This is valuable information to farmers especially during a drought when water is even more scarce and the price of water increases drastically. While measuring CWSI is easy and relatively inexpensive with the aid of remote sensing devices, measuring SWC is neither easy nor cheap. Consequently, we are interested in exploring and modelling the relationship between SWC and CWSI so that we can accurately predict SWC given CWSI. Doing so will help farmers avoid the costs of measuring SWC and aid them in cutting irrigation costs. Using data consisting of 44 observations of SWC and their corresponding CWSI's from Dr. Heaton, we will model the relationship between SWC and CWSI. An important consideration in this study is that the relationship between SWC and CWSI is highly nonlinear. This suggests that the modelling methods chosen should be robust to nonlinearity.

Methods / Models Used:

A scatter plot of our data reveals the apparent nonlinearity (**Figure 1**). An alternative to the linear model should be used. A suitable model for modelling nonlinear data is the Gaussian Process.

Description of The Gaussian Process:

A Gaussian Process (GP) is a stochastic process where any finite collection observed random variables follow a multivariate normal distribution. In other words, for any set of $t_1, \ldots, t_N \in T,\ \boldsymbol{Y} = (Y(t_1), \ldots, Y(t_N))^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_Y)$.

Full Gaussian Process Model:

Let N = number of observations and,

$$\boldsymbol{Y}|\boldsymbol{W} = \begin{pmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{pmatrix} \sim \mathcal{N}(\boldsymbol{W}, \tau^2 \boldsymbol{I}_N)$$

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

which marginalizes to:

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

where,
τ^2 can be interpreted as the error variance;
σ^2 can be interpreted as the spatial variance;

$$\begin{aligned} \boldsymbol{R}_{ij} &= \tfrac{1}{2^{\nu-1}\gamma(\nu)}(2\phi\sqrt{\nu}|t_i - t_j|)^\nu K_\nu(2\phi\sqrt{\nu}|t_i - t_j|) \\ &= Matern(|x_i - x_j|, nu = \nu, alpha = \phi) \\ &= Corr(Y(x_i), Y(x_j)) \end{aligned}$$

ϕ: decay parameter. As ϕ increases, correlation (at a fixed distance) decreases.
ν: smoothness parameter. As ν increases, smoothness increases.
K: effective range. distance where correlation decays to 0.05.
We can estimate the unknown parameters $\mu, \sigma^2, \nu, \phi, \tau^2$ (using maximum likelihood or Bayes).

This is a fascinating result that reduces computation significantly. Making use of the properties of the conditional distribution of multivariate normals distributions, one can easily can predictions and prediction intervals (see Rencher & Schaalje, Theorem 4.4d, 2008).

The GP is used because it models nonlinear relationships between random variables well and gives accurate predictions. Since we are interested in predicting SWC given CWSI and we know that the two variables are linear, the GP is a good model to consider.

Results:

Parameter Estimates:

The point estimates obtained using maximum likelihood for the covariance function are:

$$\tau^2 = 0.4$$

$$\sigma^2 = 26.81$$

$$\phi = 0.65$$

ν was chosen to be 2 because the data doesn't inform us about ν and because it gave us a smooth curve. The estimate and the 95% confidence interval for the mean are:

$$\hat{\beta} = 27.19,\ 95\%\ \text{C.I.: } (18.18, 36.21)$$

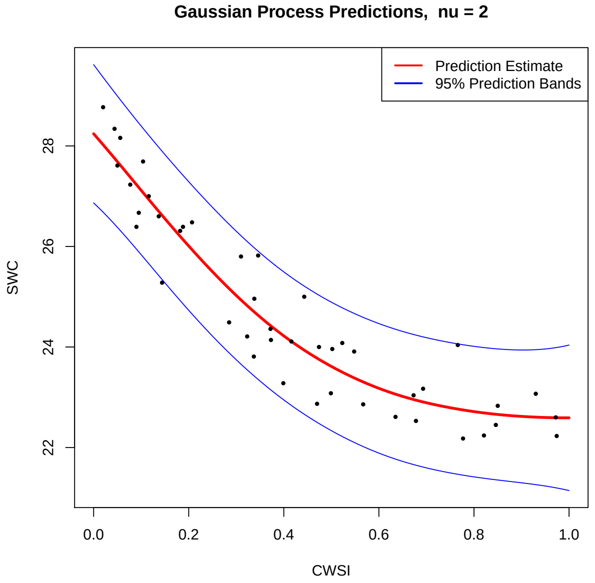

Figure 2: Plot of Data and Fitted Curve

Predictions and Predictive Accuracy:

The support for CWSI is (0,1) by definition. So, the plot of the prediction curve for CWSI $\in$ (0,1) reveals the relationship between CWSI and SWC for every possible value of CWSI (**Figure 2**). For instance, when predicted value for SWC at CWSI = 0.02 is: 28.02.

To assess the prediction accuracy of the GP model, leave-one-out cross validation was performed to obtain the coverage of the predicted values. The coverage was determined to be 0.932 (with 95% C.I.: (0.857, 1)).

Assessment of Model Assumptions:

An assumption of the GP model is that the errors are Normally distributed. The percentile-percentile plot (Q-Q plot) appears to be linear. This indicates that the residuals are normal. A histogram of the residuals also shows that the residuals are normal (**Figure 3**).

Main Points

Overall, the GP model performed well. A coverage of 93% suggests that the 95% prediction interval generated using the model contains the true response value 93% of the time. Prediction interval widths vary at different locations of the curve because the number of observations vary along the curve. The minimum prediction interval width is 2.55, which occurs at CWSI = 0.41 and $S\hat{W}C = 24.19$. The maximum prediction interval width is 2.89, which occurs at CWSI = 1 and $S\hat{W}C = 22.59$. The average prediction interval width is 2.6. Below is a table showing 20 predicted values (spread across the support of CWSI) and their prediction intervals (Table 1).

Conclusion:

Other approaches in fitting a nonlinear model include splines and local regression. A comparison of the performances of resulting models using a spline, local regression, and the GP can be done in a future analysis.

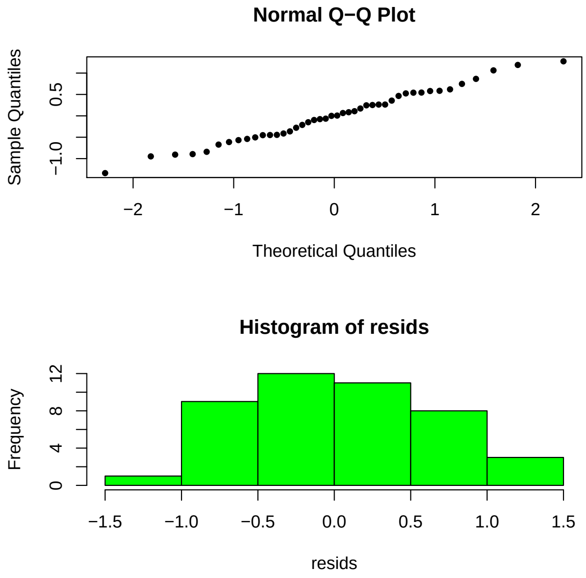

Figure 3: Residuals Diagnostics Plots

Such an analysis comparing the performance of different methods may be beneficial as different methods have different strengths. For instance, one draw back to the GP is that for large datasets, computation time becomes large as the algorithm involves the inverting of covariance matrices which increases exponentially with the number of observations.

Table 1: Prediction Values

	CWSI	Est.SWC	PI.Lo	PI.Up
1	0.01	28.24	26.87	29.62
2	0.06	27.68	26.38	28.99
3	0.11	27.10	25.83	28.38
4	0.17	26.53	25.26	27.81
5	0.22	25.99	24.71	27.27
6	0.27	25.37	24.09	26.65
7	0.32	24.91	23.63	26.19
8	0.37	24.49	23.22	25.76
9	0.43	24.12	22.85	25.40
10	0.48	23.81	22.53	25.08
11	0.53	23.49	22.21	24.77
12	0.58	23.27	21.98	24.56
13	0.64	23.09	21.80	24.38
14	0.69	22.95	21.65	24.24
15	0.74	22.83	21.54	24.13
16	0.79	22.73	21.43	24.03
17	0.84	22.67	21.37	23.97
18	0.90	22.63	21.31	23.94
19	0.95	22.60	21.25	23.95
20	1.00	22.59	21.14	24.04