Soil Water Content Prediction Analysis

David Arthur
David Lowe

March 9, 2017

Abstract

ARTHUR, B. D. LOWE, D. A. Predicting Soil Water Content Using Crop Water Stress Index **Purpose:** To a) find a way to model a nonlinear relationship between the Crop Water Stress Index (CWSI) for a crop and the Soil Water Content (SWC) for that crop and b) help farmers use this model to make accurate predictions about how much water to give a crop based off of the CWSI for that plant. **Methods:** Use a natural spline basis function expansion on CWSI to obtain a series of basis functions that allows us to model the nonlinear relationship between CWSI and SWC. Show how cross validation tests can be used to select the optimal number of knots for a natural spline basis function expansion on CWSI for predicting SWC. **Results:** Results of cross validation suggested that one knot was sufficient for using CWSI to predict SWC. A linear model using two natural spline basis functions was capable of modeling the nonlinear relationship between CWSI and SWC and could be used for predictive purposes. **Conclusion:** It is possible to use CWSI to predict SWC. As the CWSI for a plant increases, farmers need to use more water to increase the SWC for those plants. By using this model, farmers can save money and time while using water efficiently to increase their crop yields.

1 Introduction

For years, farming has played a crucial part in providing the food that keeps families alive. In the United States especially, the general population continues to depend on farmers who work year round to provide the food necessary for the nation. As a result, farmers have always been interested in learning how to increase their crop yields. In order to do this, farmers must rely on the effective use of irrigation methods.

Irrigation is the method by which a farmer controls how much water a crop receives and when that crop receives it. Knowing how much water to give and when to give it becomes especially important in times of drought or water scarcity. For this reason, farmers have learned to use two important tools when trying to determine how to use water efficiently. The first of these tools is a metric known as the Soil Water Content (SWC). This tells the farmer how much water is in the soil. The second tool is known as the Crop Water Stress Index (CWSI), which provides an estimate of the crop water status.

The CWSI is measured using widely available surface temperature thermometers. As a result, it is a relatively easy and cheap tool to use. Farmers can use the CWSI to know when their crops need water. CWSI values close to 1 mean the plant is probably already dead, and values close to 0 indicate that the plant is well hydrated. The drawback of this index is that it doesn't tell a farmer how *much* water the plant needs. In order to do this, the farmer needs to know the SWC for that area of their crop. Unfortunately, measuring SWC is expensive and takes time.

Farmers are interested in seeing if it is possible to establish a relationship between SWC and CWSI. More specifically, they want to see if they can use the CWSI of a crop to predict the SWC of that crop. If they were able to do this, they would be able to save themselves both time and money while still accomplishing their goal of knowing when to water a crop *and* how much water to use. This would benefit them greatly, as it would allow them to use water efficiently while still giving them the best crop yield possible.

1.1 Data

We are given measurements of both the CWSI and the SWC for 44 different crop locations. The CWSI is a continuous rating between 0 and 1 and can be thought of as an indicator of how hot a plant is. If the CWSI for a plant is close to 0, this implies that the plant is cool and is likely well hydrated. If the CWSI for a plant is close to 1, this means that the plant might be dead already or is in dire need of water.

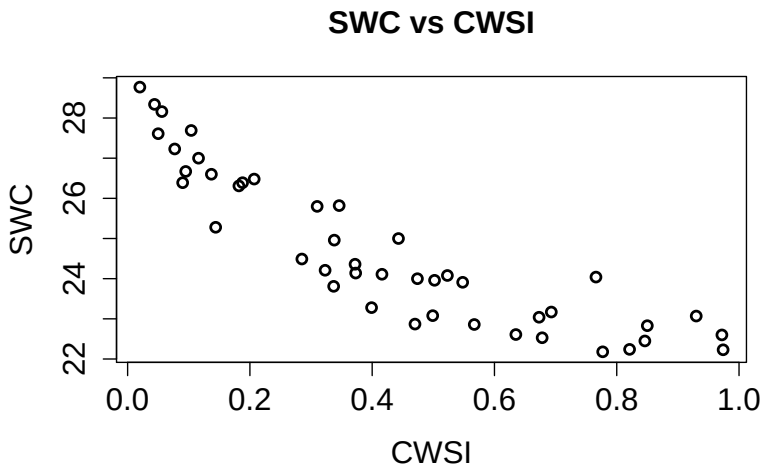

Figure 1: SWC vs CWSI

The SWC gives an idea of how much water is in the soil. For our particular data set, the smallest SWC value is 22.18 and the largest value is 28.77. Just given this data, we might infer that if the SWC for a plant is around 22, the plant might need more water. Furthermore, if the SWC for a plant is close to 28, this likely indicates the soil contains sufficient water and thus the plant is well hydrated. Thus, since it appears that the ideal SWC is around 28, this will be the target that farmers will want to reach for their crops.

In Figure 1, we can see the relationship between the CWSI and the SWC. It appears that as the CWSI increases, the SWC decreases. This goes along with our intuition that if a plant is extremely hot, it probably needs more water. However, we can also see that the relationship does not appear to be linear. Because of this, we decided to use a nonlinear regression model to help us use the information available in CWSI to predict SWC.

2 Model & Methods

2.1 Model

The model we are using is

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \epsilon_i \tag{1}$$

where

$$\epsilon_i \overset{iid}{\sim} Normal(0, \sigma^2). \tag{2}$$

Here y_i represents the SWC value for i^{th} crop location. β_0 represents the expected SWC value for crop locations where the CWSI is 0. b_1 and b_2 are used to represent two natural spline basis functions obtained via a natural spline basis function expansion on the CWSI. These natural splines are what allow us to model the nonlinear relationship between the CWSI and the SWC.

Due to the fact that we are using natural spline basis functions to model this nonlinear relationship, β_1 and β_2 are somewhat difficult to interpret in context. For example, if we wanted to give an interpretation for β_1, we would say that it is the value for how much we expect the SWC to change for every one unit increase in the values obtained after using basis function b_1 on the original CWSI values. β_2 would have a similar interpretation. As such interpretations are not very helpful, and because we are more concerned with prediction than we are inference, we will not worry about interpreting β_1 and β_2 in the future.

Finally, the ϵ_i in our model refers to the difference between the actual value and the predicted value for the SWC for the i^{th} crop location. These may be referred to as residuals or errors. We believed that these residuals would be independent and identically distributed from a normal distribution with zero mean and variance σ^2. Hence σ^2 can be thought of as a measure of how much the real SWC values vary around the predicted SWC values from our model.

2.2 Model Assumptions

By using this model, we make some assumptions about the data. First, we believe that there is a nonlinear relationship between CWSI and SWC. However, we still believe that the relationship between the SWC values and our β parameters is linear, which allows us to use a linear model. We also believe that the residuals from our model are independent and identically distributed. In other words, we believe that having knowledge about one error doesn't give us knowledge about another error, and that all of the errors come from the same distribution.

Our model shows that we believe that the errors come from a normal distribution with mean zero and constant variance. This means the errors from our model should be centered around zero and that the variance shouldn't be changing from one error to the next. Finally, by using natural splines, we were saying that we believe the relationship between CWSI and SWC is linear outside the range of our data. This is not a problem since we are primarily interested in making predictions for SWC values using only CWSI values between 0 and 1, which is where we have our data.

2.3 Nonlinear Regression and Natural Splines

As stated earlier, we want to use the information contained within CWSI to make predictions for SWC, while accounting for a nonlinear relationship between the two. In order to accomplish this, we performed a basis function expansion on our CWSI variable using natural spline basis functions.

A basis function expansion can be thought of as representing a single function, containing many characteristics, with multiple functions, each of which might be useful in modeling a certain characteristic of the function of interest. For example, we consider the following equation:

$$y = 1 + 2x + x^2 + 4x^3. \tag{3}$$

If we consider y as the function of interest, then we would say that y can be represented using four different functions. The first function would be 1 and would capture the feature about y which is constant. The second function would be $2x$, the third would be x^2, and the last would be $4x^3$, three functions that would capture the linear, quadratic, and cubic features of y respectively. In this case, the basis could be denoted as $\{1, x, x^2, x^3\}$. This is also known as a polynomial basis function expansion where each element in the basis would be an individual polynomial basis function. If these basis functions were used in regression, it would be known as a polynomial regression analysis, and each individual basis function could be used as a different "predictor variable". For our analysis, y would be represented by CWSI, and the functions that make up y would be natural spline basis functions.

Although the math is more complicated, natural spline basis functions can be used in a similar manner. The major difference is that with natural splines, you are essentially dividing the available data into sections, and then fitting a function to each section of the data. The places where you divide the data are referred to as "knots". This would be the same as fitting piecewise polynomials to the data, except that the natural spline basis functions are calculated in such a way that they are smooth and continuous at the knot points (by enforcing that the first and second derivatives be continuous at the knot points). This method also differs from using regular b-spline basis functions in that the functions are constructed assuming linearity beyond the end points of the data. This helps to reduce variability at the end points.

In order to determine the optimal number of basis functions for predictive purposes, we performed a series of cross-validation tests. Starting with one knot, we removed 15% of the data and then fit our model to the remaining 85% of the data. We then used this model to predict the 15% of the data that we removed. We did this 1000 times and calculated the mean squared error of our predictions each time. We then did this for two knots, three knots, etc. To simplify the process, we allowed the knots to be evenly spaced for each test.

By using natural splines, we were able to model the nonlinear relationship between CWSI and SWC in the same way that polynomials would without the restriction of using a single function over the entire range of the data. Natural splines allowed us the potential to improve predictions by essentially fitting piecewise polynomials. However, because natural splines enforced smoothness and continuity at the knot points, we gained the added advantage of a single formula that will be easier for farmers to use than a series of formulas for different chunks of the data. Natural splines also gave us the potential to improve predictions at the endpoints by enforcing linearity beyond the range of the data. Finally, because we used a natural spline basis function expansion on CWSI, we are still using the information about CWSI to predict SWC, which is what we wanted.

3 Model Justification & Performance Evaluation

As was mentioned before, in order for us to use natural splines for nonlinear regression, we had to select the number of knots to use. Since our goal was to be able to predict the SWC from the CWSI, we wanted to choose the number of knots that made our predictions as accurate as possible. To do this, we performed a cross-validation study to compare the mean squared error for several different knot values. To simplify where the knot points were located, we evenly spaced the knot points over the range of CWSI. Figure 2 shows the results of this cross-validation. After performing our cross-validation study, we decided to use one knot point at the 50^{th} percentile of CWSI, so as

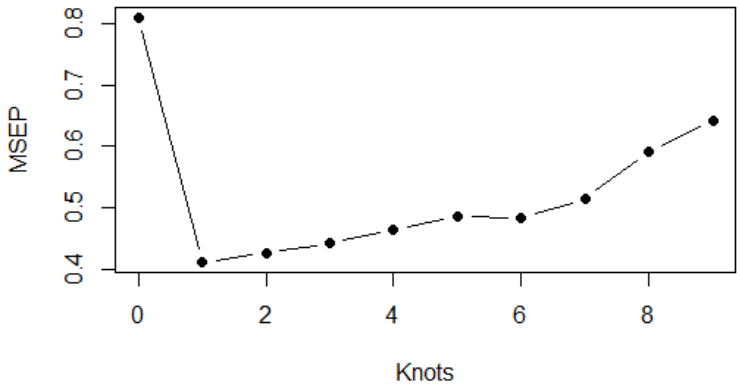

Figure 2: Predicted Mean Squared Error vs. Number of Equidistant Knots

to minimize the mean squared error of the predictions. With only one knot, the model used an intercept, along with two basis functions to represent the information contained within CWSI, to estimate SWC.

3.1 Assumptions

After fitting the model, we worked to verify the explicit assumptions we made when we decided to use our model. This model assumed linearity in the β's, which is demonstrated by Equation 1. The model also assumed that each measurement of SWC came from independent crops. The residuals plot shown in Figure 3 is ordered by observation, and with no obvious pattern displayed, it indicates no dependent relationships between observations.

The model also assumed that the errors were distributed normally. The histogram of the studentized residuals in Figure 4 shows that this assumption is met. The normal QQ plot in Figure 5 also indicates that the residuals are fairly normal, showing minimal signs of deviance from the standard normal quantiles.

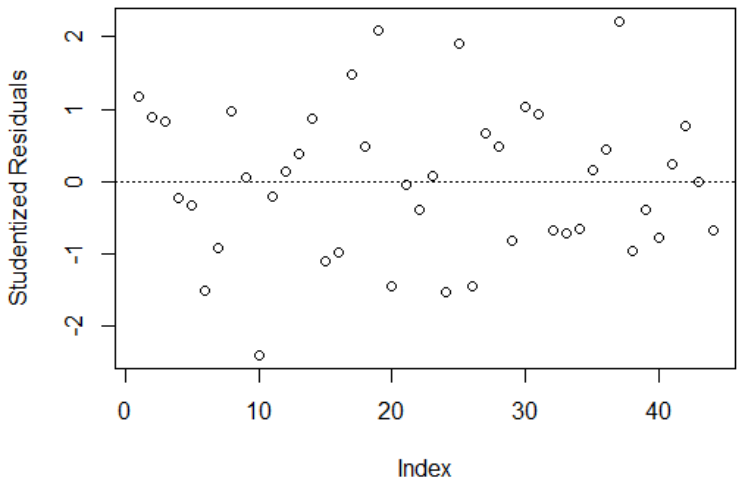

Figure 3: Fitted Values vs. Studentized Residuals

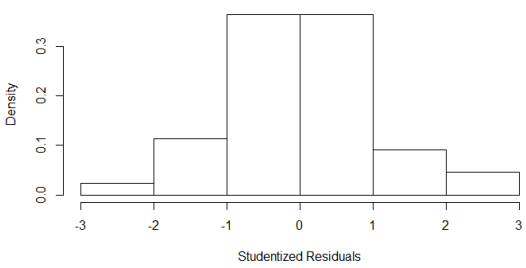

Figure 4: Histogram of Studentized Residuals

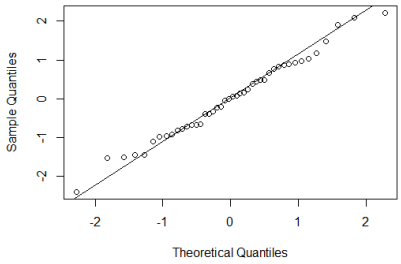

Figure 5: QQ-Plot of Studentized Residuals

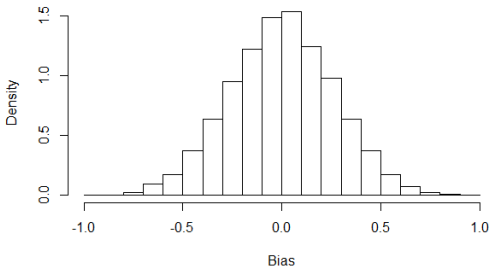

Figure 6: Cross-Validation Prediction Bias

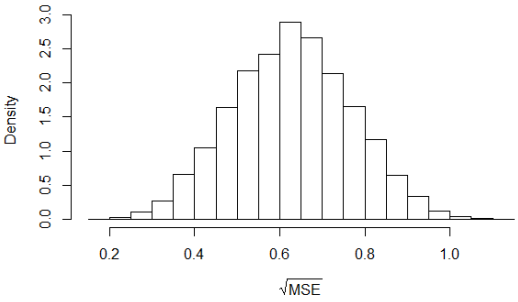

Figure 7: Histogram of Predictive Root Mean-Squared Error

Lastly, we assumed that there was a constant variance for the residuals. Figure 3 verifies this assumption, as there is no difference in the spread of the residuals. After checking these assumptions, we concluded that our model would be appropriate for this analysis. We then proceeded to use the model to help us accomplish our goal: to accurately predict SWC from CWSI measurements.

3.2 Performance

Before testing the model's ability to predict, we wanted to assess how well the model fit the data we were given. The best fit line and prediction intervals produced by this model can be seen in Figure 8. Due to the natural cubic spline, there is a nonlinear fit to the data, as we had hoped. The model gave us an R^2 value of 0.8994, indicating that 89% of the variability in soil water content could be explained by the intercept and the natural spline basis functions of CWSI. Since this indicated that the model fit the data well, we felt it would also be an efficient model when it came to making predictions.

To test the model's ability to predict, we ran cross-validation tests using 85% of the data for training, and 15% for testing. After 10,000 repetitions, the model's prediction performance could be seen more clearly. Coverage, bias, prediction interval width, and predictive mean-squared error were measured at each iteration in order to assess the predictions produced by the model.

Coverage denotes the proportion of times that the prediction intervals contained the true SWC values. The model's prediction intervals contained 93.8% of the true values, suggesting that the model came close to the true values for SWC when given the corresponding values of CWSI. The prediction interval widths were 2.63 SWC units on average, indicating that 95% of the actual SWC readings were within 1.315 SWC units of the predicted value. This interval width may seem fairly large, since the entire range of SWC readings in our data is between 22.18 and 28.77, a 6.59 SWC unit span.

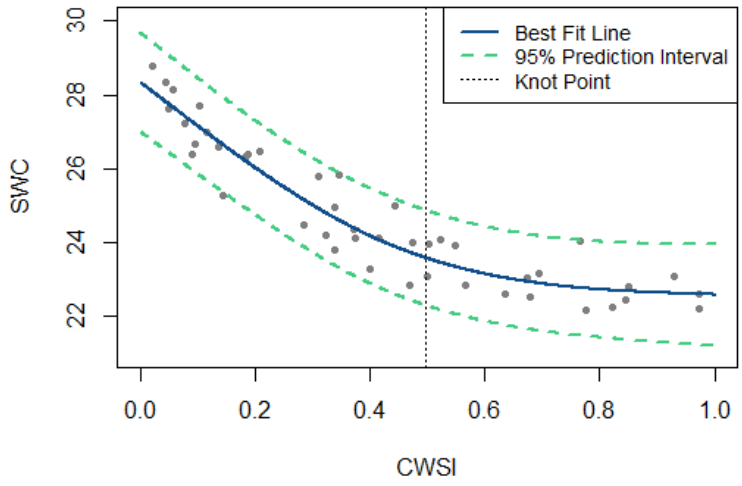

Figure 8: Natural Spline ($df = 2$)

In order to assess the bias of our predictions, we looked at the differences between the predicted values and the actual values for SWC. Figure 6 shows a histogram of the bias for our predictions. This histogram seems to suggest that our predictions were unbiased, as it is centered at zero. To further measure the precision of the model, predictive root mean-squared error was measured as well. The histogram shown in Figure 7 represents the average standard deviations produced by the differences between the predicted and actual values. The average root mean-squared error was 0.629. These predictive diagnostics helped us to conclude that the model was in fact producing accurate, albeit somewhat variable, predictions of SWC.

4 Results

At the outset of the study, we established that the goal was to predict a crop's soil water content using the crop water stress index reading for that crop. Ultimately, we wanted to produce a model that would allow farmers to know how much water to add to their crops given a specific CWSI reading. While the natural cubic spline regression model defined in equation 1 allows for a non-linear fit to the data, it does not provide ease of interpretation. This is due to the fact that the natural spline basis functions are hard to represent in some simple mathematical form. As such, it would be difficult for farmers to take a specific CWSI value and know how to transform that value into the correct prediction for SWC. Because of this, we have instead provided a more visual method of helping farmers know how to relate CWSI to SWC using our model.

Table 1 shows a chart with predicted SWC values, given incremental values of CWSI between 0 and 1. Such a chart can give farmers a quick and easy way to estimate the range of the SWC of a crop, given a general reading of CWSI. This chart also indicates the prediction interval surrounding the predicted SWC value. The last column indicates the difference between the predicted SWC at the given CWSI, and the predicted SWC at a CWSI of zero (when the crop is considered healthy). This tells the farmer how much water is needed to bring the crop to a healthy state. The farmer can easily read a chart such as this to know how many SWC units of water are needed for a crop.

For a more precise approach to SWC prediction, farmers can use a best fit line, such as the one shown in Figure 9. The figure shows an example of using the plot to predict SWC with a CWSI reading of 0.437. The predicted SWC value produced by this CWSI reading is 23.93, with a 95% prediction interval of (22.64, 25.21). Just as with Table 1, the farmer can then estimate how much water is needed given the range of soil water content produced from this model. This can be done

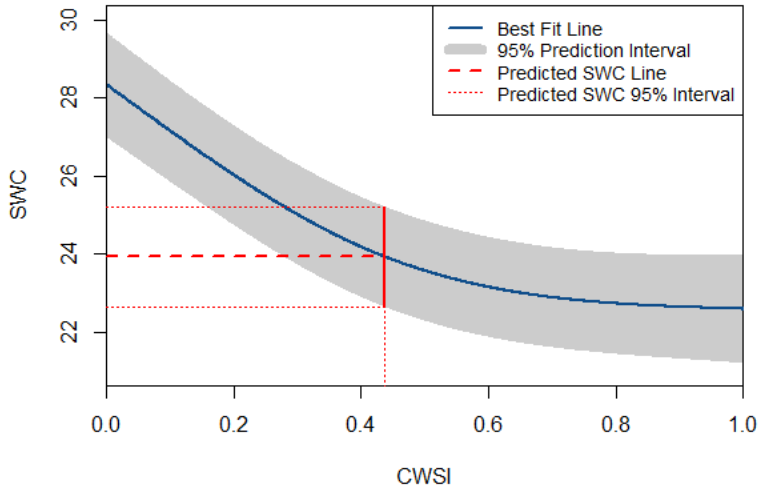

Figure 9: The natural cubic spline best fit line with 95% prediction intervals. This plot can be used to estimate SWC at specific CWSI readings, along with a 95% prediction interval of the true value of SWC.

by measuring the difference between the predicted SWC line, and the best fit line at a CWSI of zero. Farmers can use this technique for any CWSI reading, and produce predictions and intervals to estimate how much water is required to keep the crop adequately watered. A caution should be given when using this graph, however; crops with a high CWSI (about 0.8 to 1.0) may indicate that the crop is already dead, in which case watering will not help. Farmers can use their expertise to use these charts and graphs most efficiently.

This model provides a quick and easy way for farmers to predict SWC using CWSI readings. The use of the actual model may be more challenging, as some type of statistical computing tool would be needed to calculate the basis functions. The model is best utilized through visual interpretation, coupled with a solid understanding of the soil water content's relation to how much water is needed for the crop to be healthy.

5 Conclusions

For farmers, the information from this model can be invaluable to maintaining healthy crops, and being efficient with water use. The model meets our initial objective: to predict soil water content with crop water stress index. The data did not show a linear trend; however, with the natural cubic spline regression model, the predictions are able to follow a nonlinear trend. The model predicts with a fair amount of accuracy and gives farmers knowledge as to the state of the soil of a given crop, at any value of CWSI. From this information farmers can use their knowledge of soil water content, and the recommended addition of water from the chart/plot, to irrigate the field in a way that keeps crops healthy, saves money, and maintains profitability in their crop. Such techniques become necessary, especially when competing for irrigation with other farmers, and in times of water scarcity. This model can be applied and improved to refine a farmer's ability to predict soil water content.

Every model, including this one, has shortcomings and flaws. One issue mentioned earlier was the wide range that the prediction interval covered. To produce an even better fit for the model, we could assess at which location(s) to put the knot points, via cross-validation. The most commonly used selection technique for knot points is to evenly space them across the span of CWSI. However, placing knot points at different intervals may lead to a more precise fit, thus tightening the prediction intervals.

The natural cubic spline is an effective technique to account for nonlinear data, such as the relationship between SWC and CWSI. One of its greatest shortcomings, however, is its lack of interpretability. Cubic splines do not permit a simple relationship between CWSI and SWC. Interpretation, therefore, relies almost solely on the visual representation of the model. This can complicate the process of educating the farmer on applying the model to their crop, as it is not a simple plug-in relationship with CWSI.

Further study may suggest that there are other models that may fit the data better than the natural cubic spline model. Methods such as polynomial regression, or even a simple transformation of the CWSI values may provide a more interpretable model, or a better fit to the data. Possible improvements aside, this model provides farmers with a tool to accurately predict soil water content from a cheap and efficient crop water stress index measurement.

Table 1: SWC Prediction Chart

CWSI	Estimated SWC	Lower 2.5%	Upper 97.5%	Needed Water (SWC Units)
0.00	**28.34**	**26.99**	**29.68**	**0.00**
0.05	27.74	26.43	29.06	0.59
0.10	27.15	25.86	28.45	1.18
0.15	26.58	25.30	27.86	1.76
0.20	26.02	24.75	27.30	2.31
0.25	25.50	24.22	26.77	2.84
0.30	25.01	23.73	26.29	3.33
0.35	24.57	23.29	25.85	3.77
0.40	24.18	22.90	25.46	4.16
0.45	23.85	22.56	25.13	4.49
0.50	23.57	22.28	24.85	4.77
0.55	23.34	22.06	24.62	5.00
0.60	23.15	21.87	24.43	5.19
0.65	23.00	21.72	24.28	5.33
0.70	22.89	21.60	24.17	5.45
0.75	22.80	21.51	24.08	5.54
0.80	22.73	21.43	24.03	5.61
0.85	22.68	21.37	23.99	5.66
0.90	22.65	21.31	23.98	5.69
0.95	22.62	21.26	23.98	5.72
1.00	22.59	21.20	23.98	5.75