# Improving the Statistical Model for Determination of Wax Appearance Temperature from Rheological Data

Reidar Barfod Schüller

Dept. of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, P.O.Box 5003, N-1432 Ås, Norway.

## ABSTRACT

The wax appearance temperature (WAT) for waxy oils and condensates can be determined from rheological data utilizing a previously published statistical data analysis method[1, 2].

The paper presents a way to improve the model making it more robust in practical use, by utilizing more than one true parallel data set. The prediction interval, PI, of the model can also become a parameter that indicates the quality of the input data.

## INTRODUCTION

A model for prediction of the wax appearance temperature from rheological data[1] has shown that it is possible to determine the wax appearance temperature of crude oils.

This method assumes a linear relationship between the logarithms of viscosity versus 1/T at temperatures above the WAT according to a normal type Arrhenius behaviour. Viscosity data are measured as the temperature is reduced, and a statistical description is made of the complete viscosity versus temperature data set from temperature $T_1$ to temperature $T_n$. An acceptance criterion, normally at 95-99% confidence level, is then made on whether the next viscosity data point at temperature $T_{n+1}$ is below the WAT.

It is, however, important to be able to say something about the robustness of this method, especially with regard to poor quality data containing noise. The creation of an artificial test data set is one way of investigating this, and a random number adjustment to ideally linear data is used in this work.

Originally the model required that the prediction interval (PI) was specified as an input parameter, and for some data sets the resulting WAT is sensitive to the selection of PI.

It is, however, possible to set a criterion that optimizes the value of PI, and pI therefore can become an output value from the model.

## BRIEF MODEL DESCRIPTION

The model has previously been published in full detail[1], but a brief description is given in the following.

A linear regression model relating a dependent variable $y$ to a single predictor variable $x$ is assumed.

$$y_i = a + bx_i + \varepsilon_i \quad (1)$$

Let $(y_1,x_1)$ $\ldots(y_n, x_n)$ be a set of experimental data assumed to represent an in-control state. The linear model is fitted to the in-control data using Least Squares Estimation, and based on the estimated regression coefficients, $a$ and $b$, a predictor for a new observation $y^*$ given observed predictor value $x^*$ will be:

$$\hat{y}^* = \hat{a} + \hat{b}x^* \quad (2)$$

In addition to the estimates of the regression coefficients an unbiased estimate of the error variance $\sigma^2$ is obtained by the Mean Sum Squares of the Error (MSE). The MSE is given by the standard formula:

$$\text{MSE} = \hat{\sigma}^2 = \frac{\sum_{i=1}^{N} e_i^2}{n-2} \quad (3)$$

Where $e_i$ is the residual of observation $i$ found as the deviation between the observed response value and the value predicted by the model:

$$e_i = y_i - \hat{y}_i \quad (4)$$

From the linear model fit a prediction interval for a new observation $y^*$ can be constructed using the standard formulas from linear regression (e.g. Montgomery *et al.*[3]). A (1-$\alpha$)100% prediction interval is an interval which with probability (1-$\alpha$) will contain a future observation $y^*$ for a given value $x^*$. The interval is defined by a lower limit *LL* and an upper limit *UL*. The upper limit is given by:

$$UL = \hat{y}^* + t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad (5)$$

where $\bar{x}$ is the mean of the in-control $x$-observations. The statistic $t_{\alpha/2,n-2}$ is the upper $\alpha/2$ percentile of the student-t distribution with $n-2$ degrees of freedom.

In order to identify a break point a linear prediction model is fitted based on the first $n$ data points where we assume that the observations are ordered and that the last observation is closest to the break point. The next observation, for which the predictor value is $x_{n+1}$, is assumed to be even closer to the break point or being at the break point.

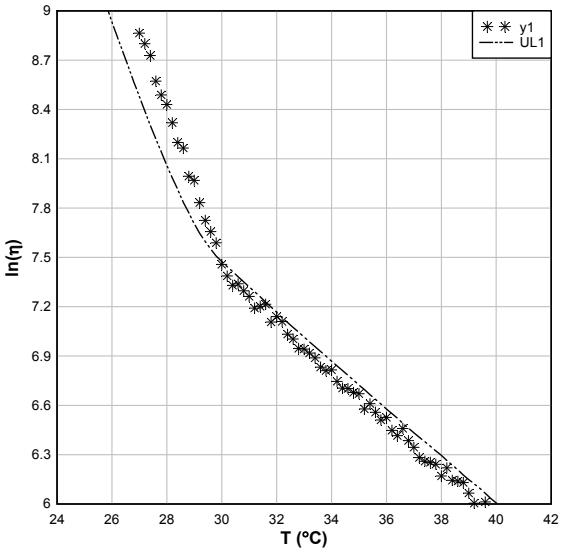


Figure 1. Example of model application showing the presence of a WAT at 30 °C. Noise level is 10%. Data points are marked with stars and the upper limit is marked with the line.

An example of applying the statistical model to artificial data is shown in Fig. 1 where the upper limit line, UL, also is shown.

## SUGGESTED MODIFICATION TO THE MODEL

By using several true parallel data sets the robustness of the model can be improved. True parallel data sets are data sets were, in this case, the temperatures are similar, forming a viscosity versus temperature table were each temperature has several measured viscosities.

It is possible to generate such true parallel data sets in modern rheometers, e.g. Paar Physica MCR301 and the software Rheoplus having a carefully controlled temperature variation during the test.

True parallel data sets may be artificially generated using a random number generating function. Each data point value,

$y_0$, was in this study assigned an adjusted value equal to:

$$y = y_0\left[1 - \frac{Noise}{2} + Noise \cdot R\#1 + R\#2 \cdot Y\right]$$
$$= y_0\left[1 + Noise\left(R\#1 - \frac{1}{2}\right) + R\#2 \cdot Y\right] \quad (6)$$
$$= y_0\left[1 + Noise \cdot \xi + R\#2 \cdot Y\right]$$

where *R#1 and R#2* are numbers between 0 and 1 generated by a random number generator, and *Noise* is the specified noise value between 0 and 1. *Y* is the magnitude of a second random noise signal.

The variable $\xi$ takes a value between -0.5 and + 0.5, so the random noise added is symmetrical about $y_0$.

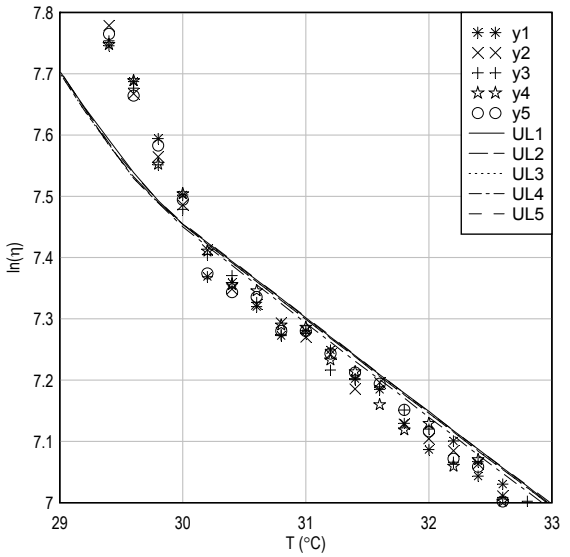


Figure 2. The five parallels with a noise level of 10%. True WAT is 30 °C. The upper boundary lines LL are also shown.

The random number generator in Excel, RAND(), was in this study used to generate the random number function. Independent random number generators were used for each parallel.

An arbitrary value of *Noise* = 10% was used, and the value of *Y* was taken as 1 in the analysis to generate the parallel data series shown in Figure 2.

Improvement algorithm

For a point to qualify to lie above the UL at a given temperature, the point must lie above the UL in more than one true parallel, typically for 2, 3, 4 or 5 parallels.

$$\pi = \pi_1 \; AND \; \pi_2 \; AND \; \pi_3 \; etc. \quad (7)$$

$$\pi = \pi_1 \cdot \pi_2 \cdot \pi_3 \cdot \ldots \cdot \pi_i \quad (8)$$

The data must consistently show that the data point is above the UL. The requirement for the point to lie above the UL can be made very strict if many occurrences are specified. The criterion is met when wax crystals start forming in the liquid below the WAT.

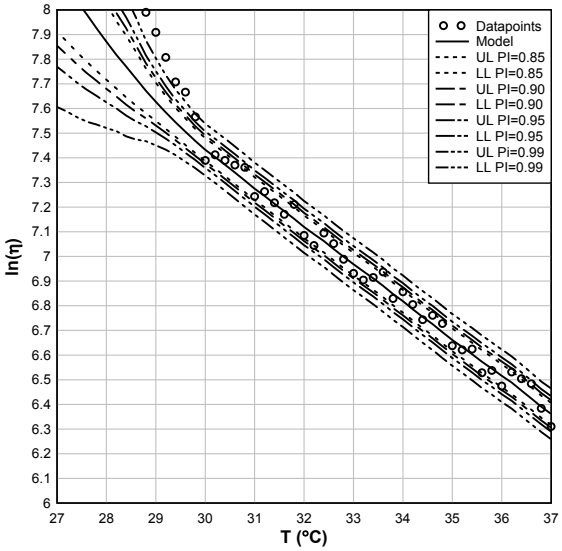


Figure 3. One data series with model line and boundaries for different values of the prediction interval PI.

A second improvement to the model can be made by trying to find the value of PI that maximizes the temperature of the continuous level TRUE data string; e.g. the sum of observations above UL = 5 of Figure 4 or the sum of observations above UL = 2 of Figure 5.

$$PI_{MIN} \Rightarrow T_{MAX,CONTINUOUS} \quad (9)$$

A low value of this limiting PI will indicate high quality data.

To minimize the under prediction of the WAT the value of PI should be minimized.

## RESULTS

The data points where y > UL can be clearly seen in Figure 2, and the effect of varying the value of PI is shown in Figure 4. As the value of PI is increased, the UL line moves further away from the model line.

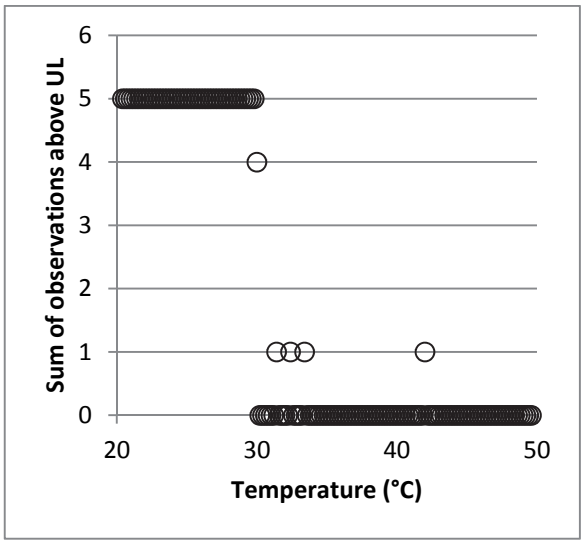


Figure 4. Illustration of logical AND algorithm showing sum of indicators at each temperature level.

Figure 4 shows an illustration of using the AND algorithm with up to 5 parallels. It is seen that there are noise signals, from Eqn. 8, above 30 °C indicating noise for combining two or more parallels, but no such noise when combining five parallels. Using 5 parallels gives a WAT just below 30 °C.

An example using two real parallel data sets, with WAT at approximately 34 °C, is shown in Figure 5 and noise occurs at sum of observations at level-1, but hardly any noise at level-2.

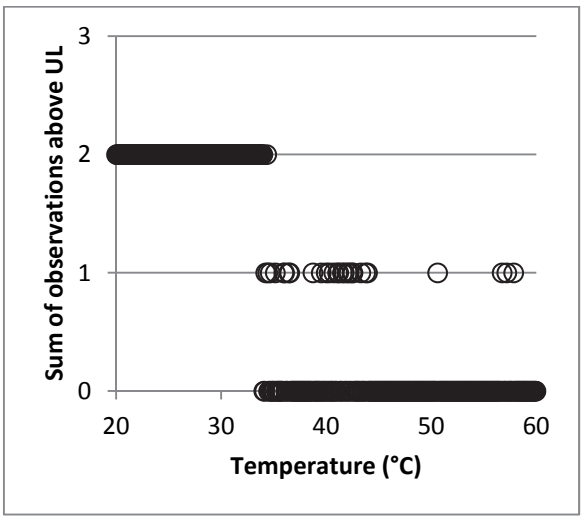


Figure 5. Example data set with two parallels showing appreciable noise at value 1, but a distinct WAT at level 2.

## DISCUSSION

A number of calculations have been made for five generated parallel date sets. The presence of noise always seems to cause an underestimation of the WAT.

The type of noise added has been made by means of a random number generator that generates grey noise without any systematic effects. The results presented here are therefore only valid for this type of noise in the data set. If the noise causes any systematic error in the data, then this may cause larger errors in the WAT than what has been shown above.

By using just two or three parallels the results show that “outliers” will be removed. The noise level itself does not seem to affect this to a large extent since an increased noise level will move the UL line further away from the average value level.

If the probability of having an outlier is 1% or 0.01, then the probability of having an outlier in the combination of n true parallels is:

$$p = (0.01)^n$$

The chance of an outlier with three parallels is consequently 0.000001. Hence

the use of three parallels should be sufficient to significantly improve the robustness of the model.

As the value of PI is increased, the upper and lower boundaries move further away from the model line as seen in Figure 3. This always results in a larger under prediction of the WAT, but as PI is decreased this causes more erroneous data at temperatures above the WAT. Therefore the PI should be kept as low as possible to minimize the under prediction.

A criterion can be made that tries to minimize the value of PI, leading to a more accurate determination of the WAT. Such a criterion would give PI as an output value from the model. A low value of PI indicates high quality data. If a high value of PI is calculated, this will indicate that the data sets are of lower quality, and there is probably an under prediction of the WAT value.

Tests using sets of real rheology data consisting of real true parallels must be performed to determine how strict the requirements need be to make the method robust enough for practical use.

## CONCLUSIONS

The results of this study can be summarized as follows:

- The statistical method for determination of wax appearance temperature from rheological data seems to be improved applying the suggested algorithm.
- The use of three parallels should improve the method significantly.
- Application of the algorithm generally gives a under prediction of the WAT.
- A prediction interval (PI) of 95% may be generally used, so there is no need to specify any parameter when using the model.
- It is also possible to make a criterion that also calculates a limiting value of PI for the data sets. PI should be minimized for a most accurate determination of the WAT.
- Tests using sets of real rheology data consisting of real true parallels must be performed to determine how strict the requirements need be to make the method robust enough for practical use.

## REFERENCES

1. Schüller, R.B., M. Tande, T. Almøy, S. Sæbø, R. Hoffmann, H. Kallevik, and L. Amundsen, (2009), "A new statistical method for determination of wax appearance temperature", *Annual Transactions of the Nordic Rheology Society*, **17**(1): p. 191-197.

2. Schüller, R.B., (2010), "Effect of random noise on the prediction accuracy of the statistical model for determination of wax appearance temperature from rheological data", *Annual Transactions - The Nordic Rheology Society*, **18**: p. 19-24.

3. Montgomery, D.C., E.A. Peck, and G.G. Vining, (2006) "Introduction to linear regression analysis", 4 ed, Wiley-Interscience. Hoboken, New Jersey: John Wiley & Sons, Inc. 612, 978-0-471-75495-4, 0-471-95495-1