Technical University of Denmark

02445: Statistical evaluation of artificial intelligence systems

Project 2: Phosphate in soil and the effect on barley production

Authors:	**Study numbers**
Karl Byberg Ulbæk	s183931
August Semrau Andersen	s183918

Summary

This report investigates two different methods of measuring amounts of bioavailable phosphorous contained in soil on the basis of a given data set. It seeks to determine which of the two methods is most precise, by evaluating its ability to predict yield on barley fields using a linear and non-linear regression model. The models are statistically compared for verifying if one method is significantly better than the other, and further the report adresses the question of whether the amount of bioavailable phosphorous in fact influences the final yield. The report finally discusses what a farmer should be observant of, when statistically evaluating his/her yield and the tools used for production.

January 21, 2020

Introduction

In this report, we will utilise the provided data set $fosfor_data.Rdata$ to investigate two different methods, DGT and Olsen-P, of measuring the amount of bioavailable phosphorous in soil on different barley fields located in Denmark and Norway.
Farmers around the world are progressively getting access to more technologically advanced tools in the pursuit of improving the yield of their fields and making their business more economically viable. It is crucial that they in this process have access to and evaluate data, that can help them make better decisions on what technology is most beneficial for, and potentially attain knowledge that enables them to improve on traditional procedures.

The report will through statistical analysis of the provided data make a statistical argument for which of the two methods of measuring bioavailable phosphorous, DGT and Olsen-P, is the best. This will include making predictive modelling based upon the given measurements and corresponding yields. Further, the report will seek to verify the causality between bioavailable phosphorous and yield, by investigating the statistical significance of these measurements. Finally, it will discuss the results and the potential consequences these might have for the farmer.

Data

The observations of the 'Phosphorous' data set consists of four attributes:

- **'location'** is nominal and determines which of the nine different fields located in either Denmark and Norway the sample is taken from. Eight of the fields all have 4 observations each, while the last field marked number '011' only has two complete observations.
- **'yield'** is ratio and measures the final yield of barley crop on each of the individual plots. The yield is measured in hundred kilograms per hectare; $\frac{100kg}{ha}$. It is this measurement that is missing for two observations from field '011'.
- **'DGT'** is ratio and is the newer and more expensive measurement of bioavailable phosphorous. DGT is measured in microgram bioavailable phosphorous per liter; $\frac{\mu g}{L}$, and was obtained before the crops of barley were sown.
- **'olsenP'** is ratio and is the older and cheaper measurement of bioavailable phosphorous. Olsen-P is measured in milligram bioavailable phosphorous per 100 gram; $\frac{mg}{100g}$. These measurements were also obtained before the crops of barley were sown.

As field '011' only has two complete observations, we have chosen to remove it completely from the analysis.

General overview

Besides the missing data of field '011', the phosphorous data set is generally easy to get an overview of.

On the boxplot, seen as figure 4 in appendix, page 8), showcasing yield for each location, it can be observed that field '004' has a much lower variance than the other fields, with a variance of $var(004) = 0.2335$ compared to the mean variance of all fields being $var(location) = 7.9056$.

Regarding the different locations, it is important to mention a disadvantage the data possesses. While there are 32 different observations of yield, meaning 4 specific yield-values for every of the 8 fields, the measures of DGT and Olsen-P are the same for the entire field. I.e., only one DGT and Olsen-P measurement has been made for every field/location. As the measurements are not inexpensive to perform, an explanation for this may be that while the yield varies a lot within the same field, before mentioned variance of around 7900 kilogram per hectare, the variance of bioavailable phosphorous could be almost constant for big areas of land, making one measurement sufficient for each field. This disadvantage is not a huge hurdle, but causes problems limits the way predictive models will be trained in the following section.

Methods and Analysis

First focus: Choosing between DGT and Olsen-P measurements

When comparing and choosing between the two measurement-methods; the newer and more expensive DGT (Diffusive Gradients in Thin film technology) and the older, cheaper Olsen-P, we first need to clarify what criteria the decision will be based upon.
Our assumption is here that a farmer having to choose between DGT and Olsen-P will seek to utilise the more precise measurement of the two, as this will assist in making better decisions on where and how much to fertilise, given there is an insufficient amount of bioavailable phosphorous ('Bio-P' from now on) available in the plants soil.
The farmer will though also be considering the costs associated with each measurement methods, and as it is informed that DGT is more expensive than Olsen-P (without knowing how much), we assume that this method will have to be significantly better in order to be preferred over the cheaper method.

Predicting yield from bioavailable phosphorous

Determining whether DGT or Olsen-P is the more precise measurement method for analysing the amount of Bio-P in soil is not completely straight forward. Even though DGT literally has more digits in its Bio-P measurement, it is not necessarily more precise.
The precision of DGT and Olsen-P will primarily be found in regards to how good they are at predicting yield. The predictions will be made based regression models. We have chosen to apply both a linear and a non-linear model. The linear model is the simplest form of regression, and will provide a good benchmark for model prediction performance. The non-linear model is potentially more fitting, as looking at the data also indicates a non-linear trends. A non-linear regression model is different in the way, that it does not try to fit the data directly, but has two parameters, alpha and beta, that are optimised to fit a regression in the best way possible. The measurements can be seen plotted in figure 5 in appendix, page 8.
The predictive regression models will be statistically evaluated using cross-validations, in which we train on the data consisting of seven locations, and test on the data consisting of the remaining last location (eight in total). We hereby circumvent the problem of having only

one distinct DGT and Olsen-P measurement for each field, though four different measurements would have been preferred. Through calculating of mean squared errors (MSE) we will be able to compare the two methods. We will further compare the predictive squared errors for DGT and Olsen-P using a paired t-test, to find if one is outperforming the other significantly.
Further, the statistic significance of Bio-P measurements in regards to yield (addressed in more detail in the following section) will be taken into account, as this statistic measure gives a quick indication as to how well DGT and Olsen-P describe yield.

Second focus: Does the amount of bioavailable phosphorous influence the harvest yield?

We are informed that there is causality between the amount of Bio-P a plant has, and its growth. Despite of this, it is still relevant to verify whether the causality is apparent in the data, or if the data contradicts this prior belief.
By performing permutation on the linear regression models of DGT and Olsen-P, the significance of Bio-P to yield will be evaluated by comparing the models actual slope to a 95% density interval of permutated linear slope coefficients.
These together with significance level p-values will indicate whether the amount of bioavailable phosphorous influences the yield.

Results

First focus: Choosing between olsenP and DGT measurements.

Linear Regression Model

The p-values of the linear regression models of DGT and Olsen-P were first inspected as they give an indication as to whether a model will be a good predictor. P-Values:

$$\text{DGT: } Pr(>|t|) = 0.0408 \qquad OlsenP : Pr(>|t|) = 0.175$$

The linear model based on DGT is significant in regards to the 0.05 significance level, while Olsen-P is not. This means that statistically, Olsen-P does not reject the null-hypothesis and is not significantly different from a linear model with a flat slope.

The linear regression models for DGT and Olsen-P were used for predicting yield.
The mean squared errors of the cross validated predictions are presented below:

$$\begin{aligned} \text{DGT: } MSE &= 197.82 & \text{Olsen-P: } MSE &= 238.66 \\ \text{DGT: } \sqrt{MSE} &= 14.065 & \text{Olsen-P: } \sqrt{MSE} &= 15.449 \end{aligned}$$

This result shows that DGT is somewhat better for predicting yield using a linear regression model, which also matches the trend seen with p-values.
Looking at the linear regression graphically in figure 1, it is clear that even though DGT is a better fit than Olsen-P, both of them only fit the data to an extend.

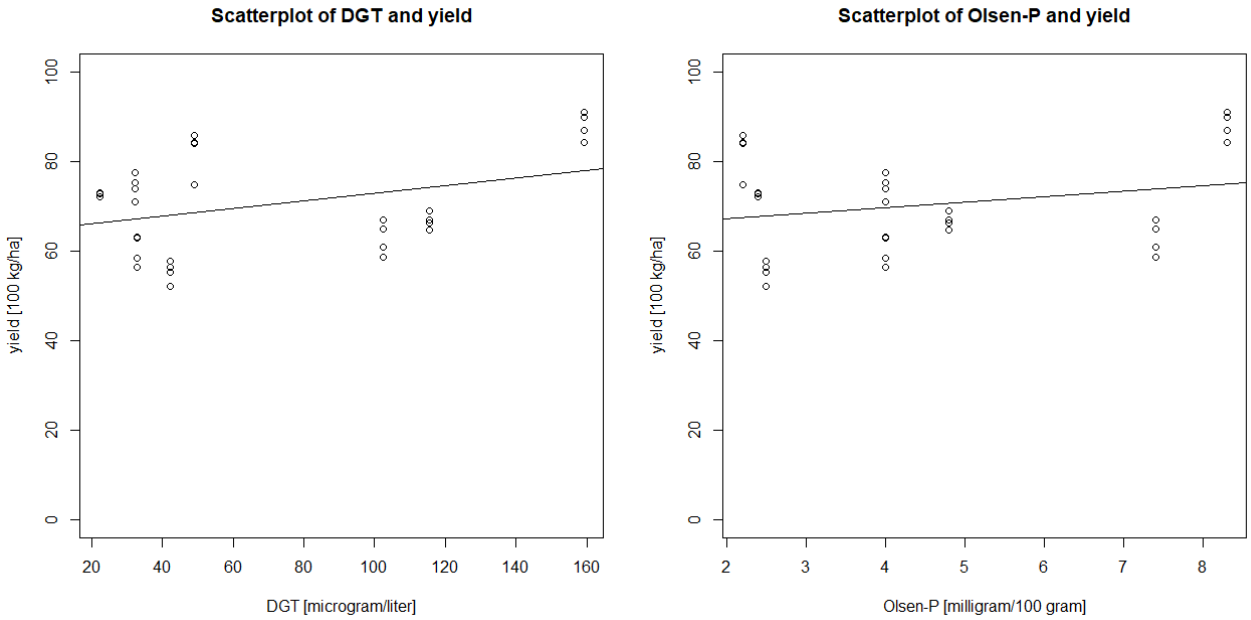


Figure 1: Scatter plots of DGT and Olsen-P respectively fitted with their linear regression.

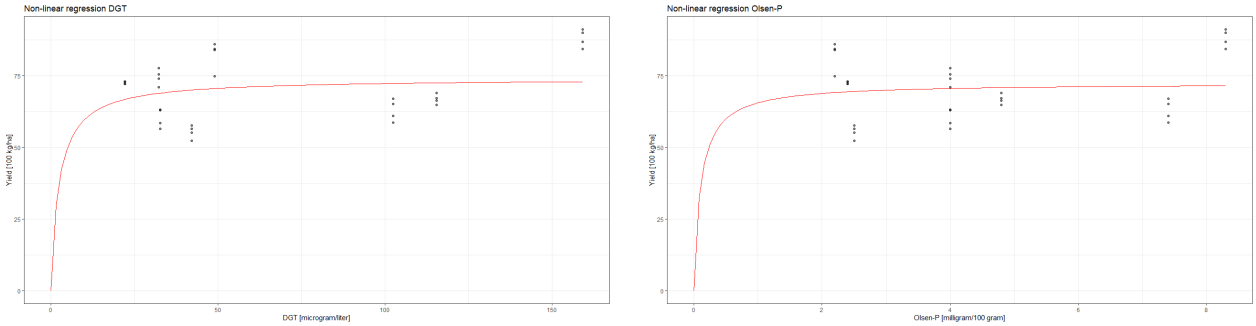
The performance of the two linear regression models is close, but when performing a paired t-test on the squared errors of 32 predictions, it was found that they are are in fact significantly different from each other, p-value of 0.01231. This means that the linear regression model based on DGT measurements is statistically better than Olsen-P.

Non-Linear Regression Model

Seeking a better fit for predictive modelling, a non-linear regression model, specifically the Michaelis-Menten model, was evaluated in the same way as the linear regression model was.

$$\text{DGT: } MSE = 194.36 \qquad \text{Olsen-P: } MSE = 236.46$$
$$\text{DGT: } \sqrt{MSE} = 13.941 \qquad \text{Olsen-P: } \sqrt{MSE} = 15.377$$

The fitted regression models can be seen plottet below in figure 2.



(a) Non-linear regression model of DGT.

(b) Non-linear regression model of Olsen-P.

Figure 2: Non-linear regression model for DGT (left) and Olsen-P (right).

They both showed a minuscule improvement in performance, in which DGT again seemingly performs better than Olsen-P. Though in this case, when performing a paired t-test we find that the p-value of 0.0522 indicates that the two non-linear regression models are not significantly different.

When comparing the linear regression model and non-linear regression model for DGT, we find that their MSE's are very close, and a paired t-test confirms this notion with a p-value of 0.3111.
There is therefore no immediate advantage in using the non-linear regression model over the linear regression model.

Second focus: Does the amount of bioavailable phosphorous influence the harvest yield?

To determine whether the amount of bioavailable phosphorous a field has, influences the yield of the field, we look into the statistical significance of the Bio-P measurements DGT and Olsen-P. From the linear regression models we found that DGT had significant effect on the yield, with p-value of 0.0408, while Olsen-P did not with a p-value 0.175.
To inspect this significance, permutation of the data was performed. For DGT, the permutation found that 95% of the linear regression coefficients were contained withing the interval -0.080218 and 0.081678. With an observed slope of 0.083971, which is outside the interval, we can once again reject the null-hypothesis. DGT is deemed statistically significant in terms of influencing yield. This can graphically be observed in figure 3 below.

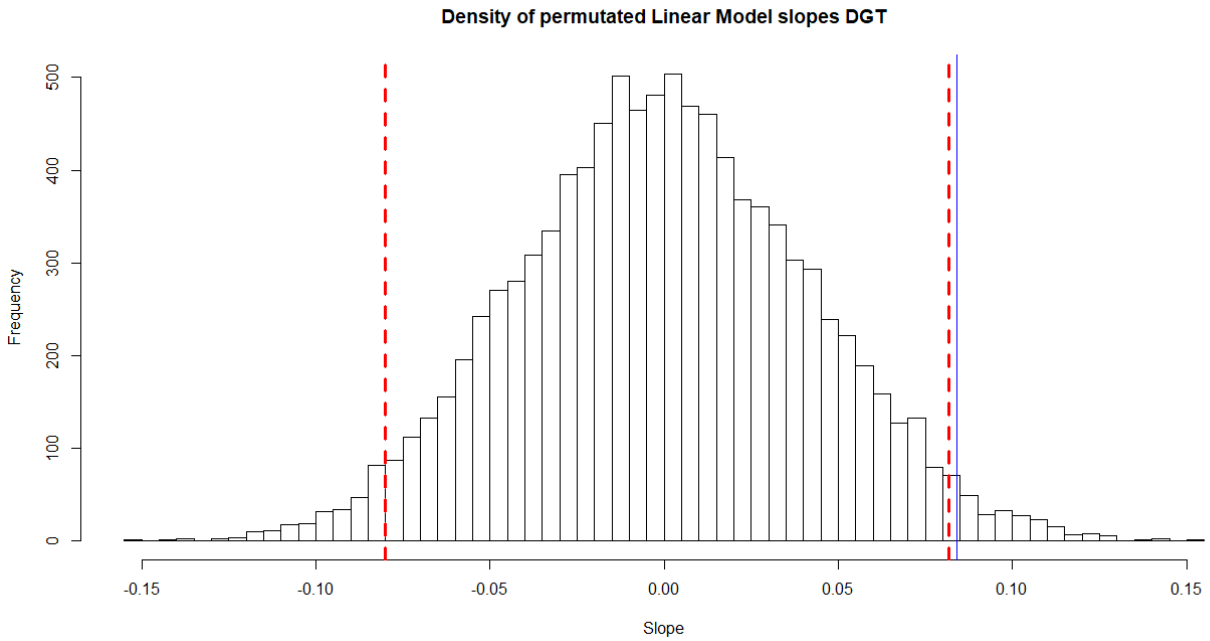


Figure 3: Red lines indicate the 95% density interval, the blue line being the actual slope coefficient.

Doing the same thing for Olsen-P, earlier deemed insignificant, i.e. accepting the null-hypothesis

of no significance in terms of influencing yield, it was found that the models slope of 1.23009 is very much contained in the permutated interval ranging from -1.742545 to 1.766044. This can graphically be seen in the density plot in figure 8 in appendix, page 10.

This result confirms that bioavailable phosphorous does in fact influence the yield of barley, when evaluated on the given data.

An important thing to note, in regards to this affirmation, is the fact that 'location' itself was also found to be statistically significant to the yield. This means that, while Bio-P influences yield, it is not negligible which field it is, in effect saying that yield is affected by a variety of factors , probably entailing where the field is located, how much sun shines at the field, rainfall, mineral composition of soils etc.

Conclusion and Discussion

The goal of this report was to statistically investigate and compare the two methods of measuring bioavailable phosphorous, DGT and Olsen-P, and further determine if the amount of Bio-P had an actual influence on the yield of barley.

It was found that, when modelled after a linear regression model, DGT was in fact significantly better at predicting yield than Olsen-P was, and it is therefore deemed the more precise measurement method.
From this we conclude that there is maybe a benefit in utilising the newer and more costly DGT method, but as the difference in predictive performance was not enormous, choosing between the two will come down to what the added cost of using DGT is, and how big the potential benefit is. For small farms, Olsen-P measurements might suffice, while large scale farms could see benefit in the added precision, despite the cost.

It was further found, while not very prominently, the amount of bioavailable phosphorous has an influence on the yield of barley fields when inferred by DGT measurements.
The prior knowledge of causality between amount of Bio-P and yield is hereby supported statistically, enforcing the notion that farmers will see benefit in controlling and supplementing the amount of bioavailable phosphorous present on their barley fields.

Appendix

Link to R-code in github:
https://github.com/augu0093/02445-StatisticalEvaluation$_{E}xamProject2$

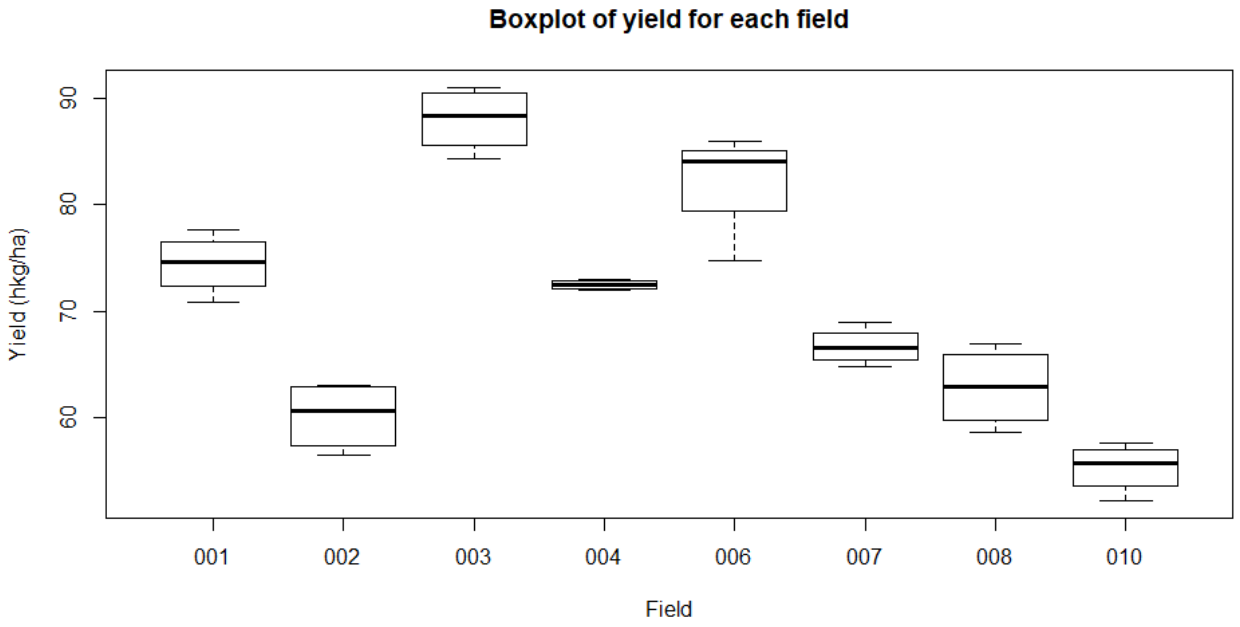


Figure 4: Boxplot showing the yield (which is spread on four subplots) for each field/location.

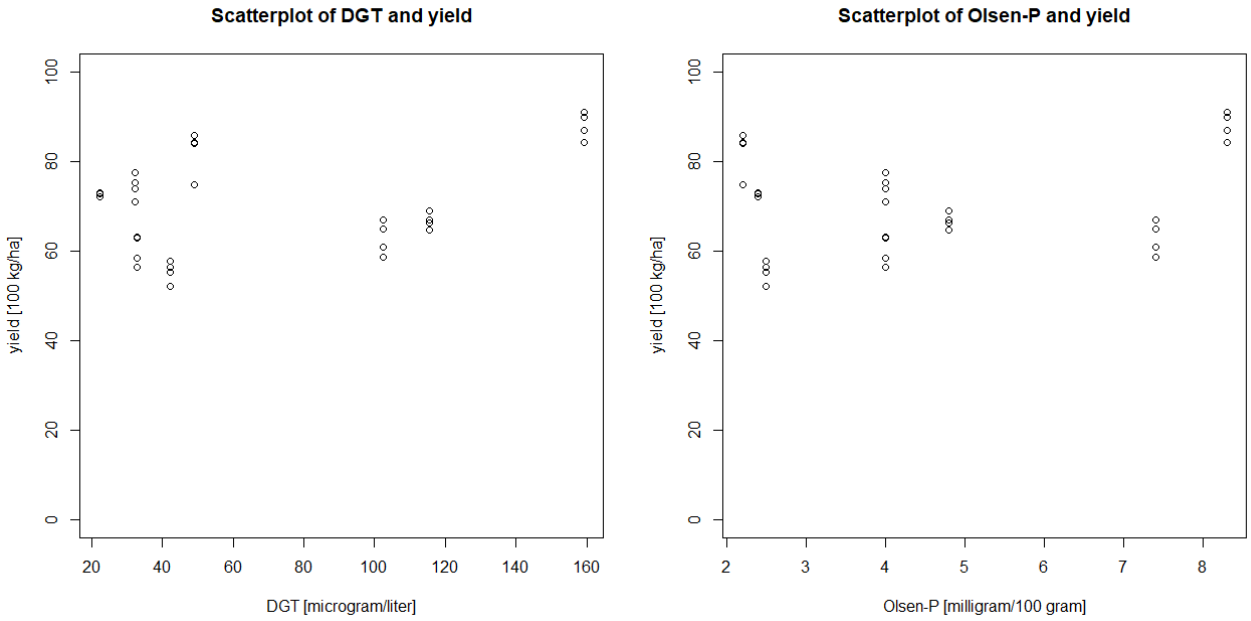


Figure 5: Scatter plots displaying data of DGT and Olsen-P respectively on the x-axis and yield on the y-axis.

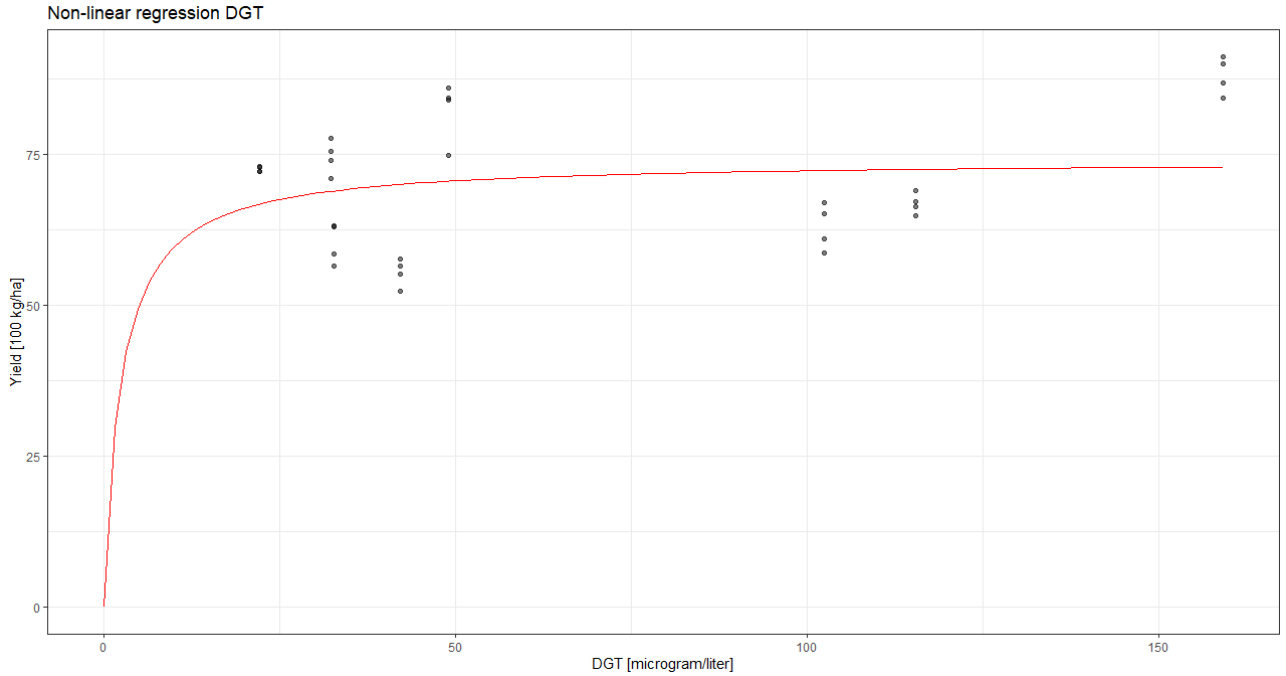


Figure 6: Non-linear regression model of DGT.

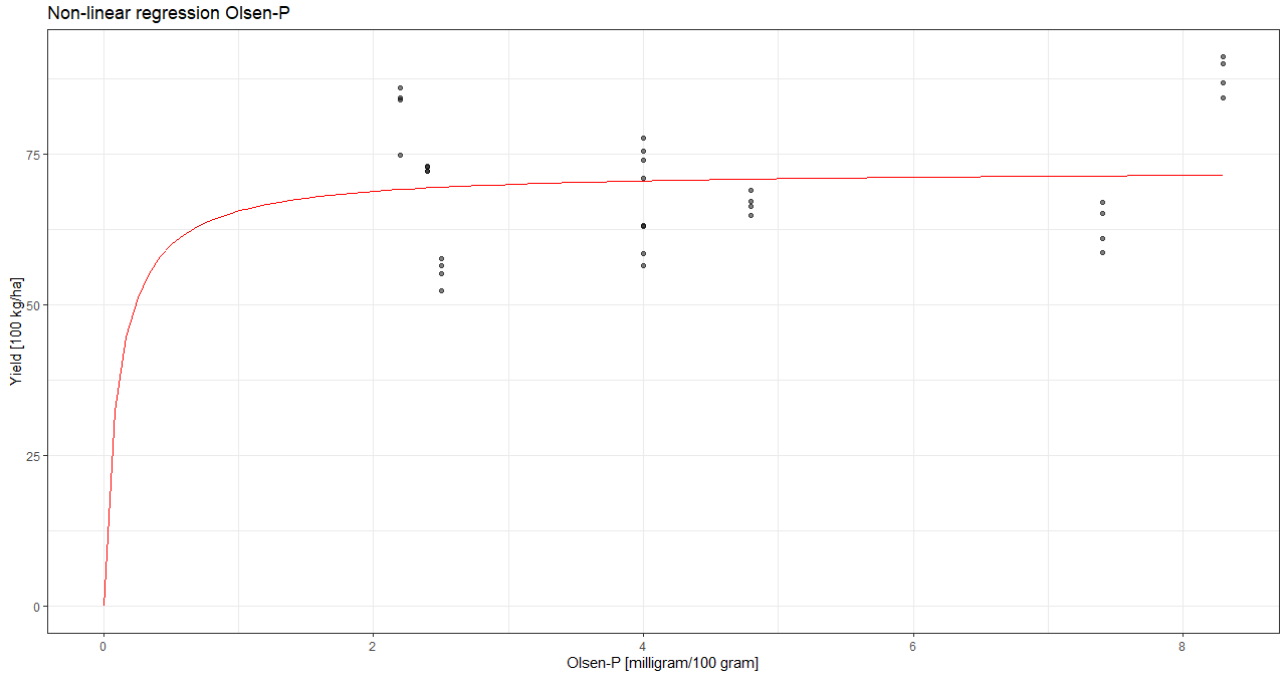


Figure 7: Non-linear regression model of Olsen-P.

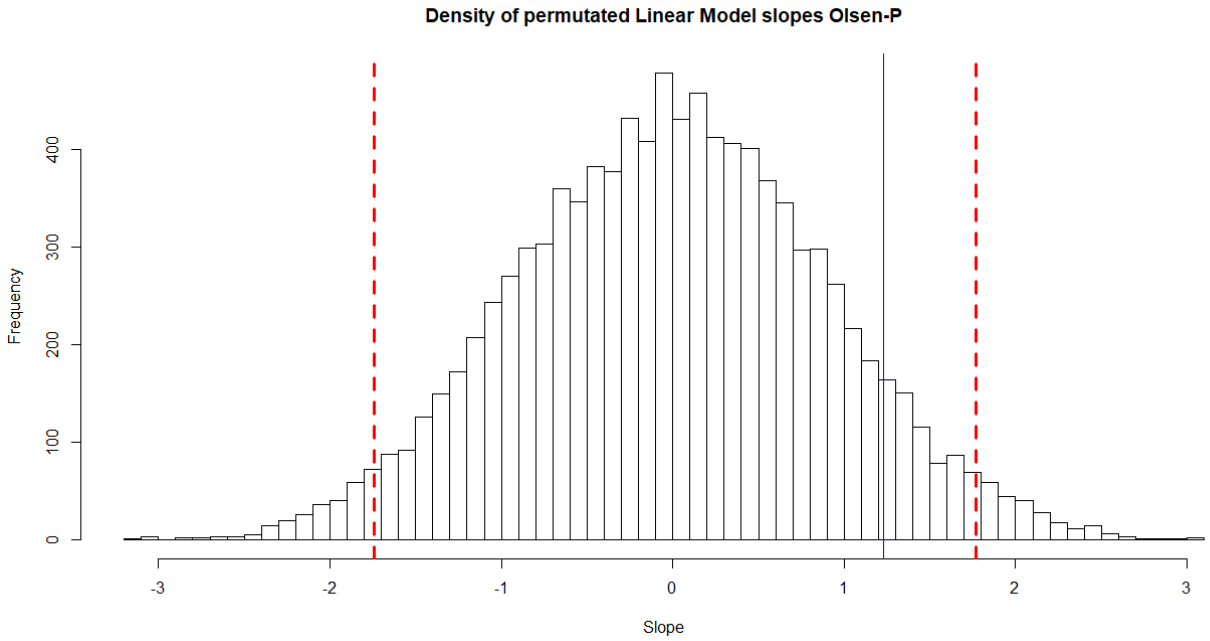


Figure 8: Red lines indicate the 95% density interval, the blue line being the actual slope.