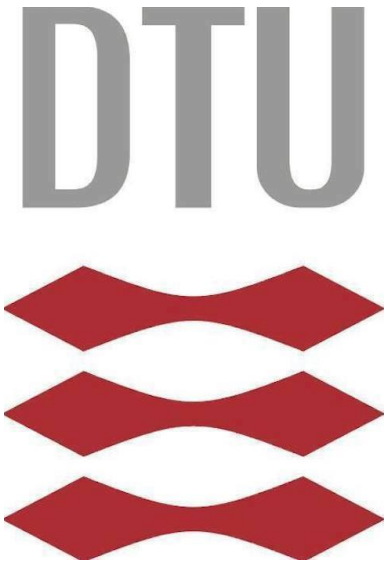
Danmarks Tekniske Universitet



Peter Grønning, s183922
Rasmus Bryld, s183898
Lukas Leindals, s183920

Statistical evaluation of artificial intelligence systems - 02445

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

January 2020
Danmarks Tekniske Universitet

Abstract

Phosphorus is a vital nutrient for plants and it would therefore be valuable for a farmer to know the amount of bio available phosphorus in the soil before planting his seeds. This report uses the data set "fosfor_data" to analyse whether the traditional "olsenP" or the newer, more expensive "DGT" method is the best measuring method and whether or not phosphorus actually has an influence on the yield. As the farmer is interested in the outcoming yield, the methods are evaluated by their ability to predict the yield with the Michaelis - Menten regression model. We found that the models for DGT and olsenP had a MSEs of 179.5 and 474.3 respectively. When comparing the difference in performance via Rank Sum Test and non-parametric bootstrapping we saw that the DGT method was significantly superior at a 5 % significance level. To investigate if phosphorus actually has an effect on the yield we used a permutation test, which showed no significant effect of phosphorus at a 5 % significance level.

1 Introduction

Phosphorus is an essential nutrient for plants and many farmers tend to use NPK fertiliser to get a higher yield, where the P represents phosphorus. Whether the farmer simply wishes to cut costs by using just the right amount of phosphorus, or to grow his/her plants organically on the best possible location, it is vital to know how much bio-available phosphorus is in the ground, how this possibly influences yield, and lastly how to get the most accurate measurements with the least cost. In this report we will look at two different measures of phosphorus, which are the traditional "olsenP" method and the newer, more expensive method DGT. We will investigate which of the two methods are better at predicting the yield and whether or not the amount of bio available phosphorus have an influence on the yield. This will be done by carrying out different statistical methods on the data set "fosfor_data".

2 Data

The "fosfor_data.Rdata" file contains soil samples from 9 different fields in Denmark and Norway. The categorical location variable contains the ID of the location where barley was sown and each field was split into 4 plots. For each plot the yield, DGT and olsenP was measured, where the last two are two different measures of phosphate, making these 3 variables continuous. Yield was measured in hkg/ha, DGT in $\mu g/L$ and olsenP in $mg/100g$. In table 1 a summary of quantiles and the mean for each of the 3 numeric variables can be seen. Furthermore two NA's were found in the yield column.

Summary statistic	yield	DGT	olsenP
Min.	23.00	3.072	2.000
1st Qu.	58.25	32.300	2.400
Median	67.01	42.132	4.000
Mean	65.17	62.062	4.178
3rd Qu.	74.95	102.440	4.800
Max.	91.05	159.186	8.300

Table 1: A summary of the 3 numeric features

In figure 1 different plots of the 3 numeric variables are shown. The diagonal contains density plots of the variables, below the diagonal a scatter plot of each of the combinations is shown and above the diagonal the correlations are shown.

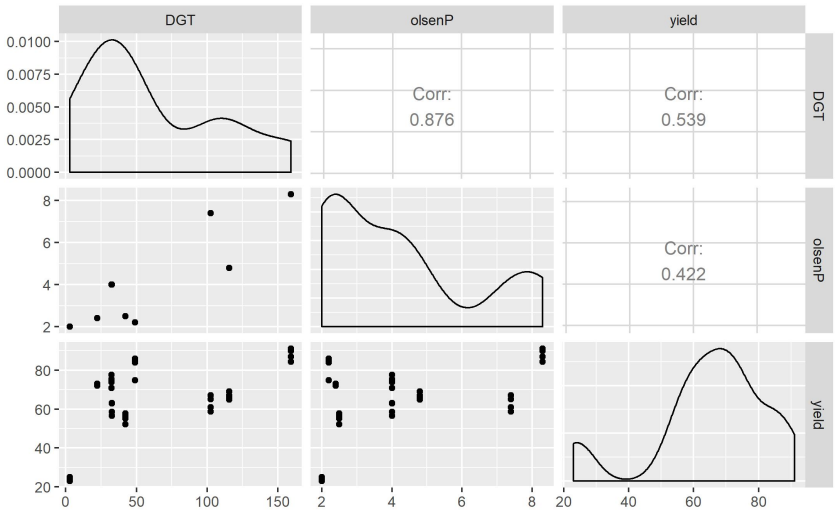


Figure 1: Relationship between the numeric variabels in the phosphorus data

3 Methods and analysis

3.1 Preprocessing the data

First the data was loaded into R and the location variable was turned into a factor. To handle the NA's in the data set we chose to replace these with the mean of the values in the same location. This was done because DGT and olsenP measurements did not change within the location.

3.2 Comparing regression models

The Michaelis-Menten regression model (1) will be used to fit the data as this model has been recommended in the report description. There are good reasons to disregard linear regression, as both the correlation between yield and DGT and olsenP (0.539, 0.422, see figure 1) and R-squared values (0.29, 0.18) (see appendix A.2) are low. The Michaelis-Menten model also has a limitation on the yield as function of phosphorus, described by an asymptote of the curve equal to the parameter α, which translates well to the real world, where plants do not have the possibility of returning infinite yield. To get a visual understanding of the difference between the two models, both have been plotted in figure 2 where it appears that the Michaelis Menten model seems to capture more of the pattern in the data.
For all of the above reasons the Michaelis-Mennten model is fitted to the data in two different seutps; both with Yield as the response variable but one with DGT as the explanatory variable and the other with olsenP as the explanatory variable. Hence two models are trained - $\mathcal{M}_A$ and $\mathcal{M}_B$.

$$y \approx \frac{\alpha \cdot x}{\beta + x}, \quad x > 0 \tag{1}$$

The parameters of each model are obtained by fitting the models to the data where each location is reduced to the mean observation of yield and the 95% confidence intervals of these are estimated using parametric bootstrapping. By reducing each location to the mean, the groups will be fixed and parametric bootstrapping is possible under the assumption that errors are normally distributed.
The models are tested using 9-fold cross validation with the Squared Error (SE) as the loss function. In each fold an entire location is left out, as we are interested in how well the model will predict the yield at a new location based on the level of bioavailable phosphorous. The

model is then used to predict the yield of each of the four left out observations for each location resulting in 36 predictions. The loss of each prediction is concatenated into a vector for each model.

$$z_A = \begin{bmatrix} (y_1 - \hat{y_1}_A)^2 \\ \vdots \\ (y_n - \hat{y_n}_A)^2 \end{bmatrix}, \quad z_B = \begin{bmatrix} (y_1 - \hat{y_1}_B)^2 \\ \vdots \\ (y_n - \hat{y_n}_B)^2 \end{bmatrix}$$

From these vectors the Generalisation Error (GE) of the models can be estimated as the Mean Squared Error (MSE) simply by calculating the mean of z_A and z_B. 95 % confidence intervals of the MSE of the two are calculated with non-parametric bootstrapping.

One could use a paired t-test with the test statistic t_{obs} calculated as in equation (2) to test if there is a significant difference between the GE of the model. However, since many observations $z_i = z_{i_A} - z_{i_B}$ are highly correlated due to the grouped locations and z_i do no appear to be normally distributed (see appendix A.3) different non parametric approaches are chosen in this report.

$$\hat{z} = \frac{1}{N}\left(\sum_{i=1}^{N} z_{i_A} - z_{i_B}\right) \tag{2}$$

A p-value stating if there is a significant difference between z_A and z_B on a 5 % significance level is obtained from the Wilcoxon Rank Sum Test (RST) with the null hypothesis $z_A = z_B$. A confidence interval for the true difference in MSE, z, is estimated using non-parametric bootstrapping.

3.3 Testing the effect of phosphorus

To test whether or not the amount of bio available phosphorus had an effect on the harvest yield, a two tailed permutation test was used. To do this, a number, 10000, of Michaelis-Menten regression models were trained where the harvest yield was assigned to random level of phosphorus, making sure the yield observations belonged within the same groups (locations) as orignially after the permutations and that phosphorus measures were assigned the same harvest yield. The model with the lowest MSE will be deemed as the one describing the effect of phosphorus on the yield the best and thus the one chosen for the permutation test.
For each permutation the parameters of the fitted Michaelis-Menten models $\tilde{\alpha}$ and $\tilde{\beta}$ were stored in a data-frame in order to create a distribution for each of the two parameters of $\mathcal{M}_A$ and $\mathcal{M}_B$ when fitted to randomly permuted data.
The likelihood of obtaining parameter values from these distribution as or more extreme as the ones observed when fitting the Michaelis-Menten model to the original data will serve as p-value in the two-tailed permutation test with the null hypothesis: there is no effect of phosphorus on the harvest yield, ie. the relation between phosphorus and yield described by the fitted Michaelis-Menten model is due to change

$$H_0: \quad \mu_{\tilde{\alpha}} = \alpha_{obs}, \quad \mu_{\tilde{\beta}} = \beta_{obs}$$

4 Results

4.1 Comparing regression models

Fitting the non - linear Michaelis Menten regression model to the full dataset using respectively DGT and olsenP as the explanatory variable resulted in two models. The parameters and their 95% confidence intervals are listed in table (2) below. Further visualisation of the fit of the models is found in figure (2) - to emphasise on the non-linearity of the data linear regression models fitted to the data are shown in figure (2) as well.

Parameters	α	β		α	β
$\mathcal{M}_A$	78.114 [68.0 ; 94.4[	5.494 [1.66; 16.5]	$\mathcal{M}_B$	87.875 [56.1 ; 159.8]	1.2073 [-0.34, 5.40]

Table 2: Parameter of both Michaelis Menten models

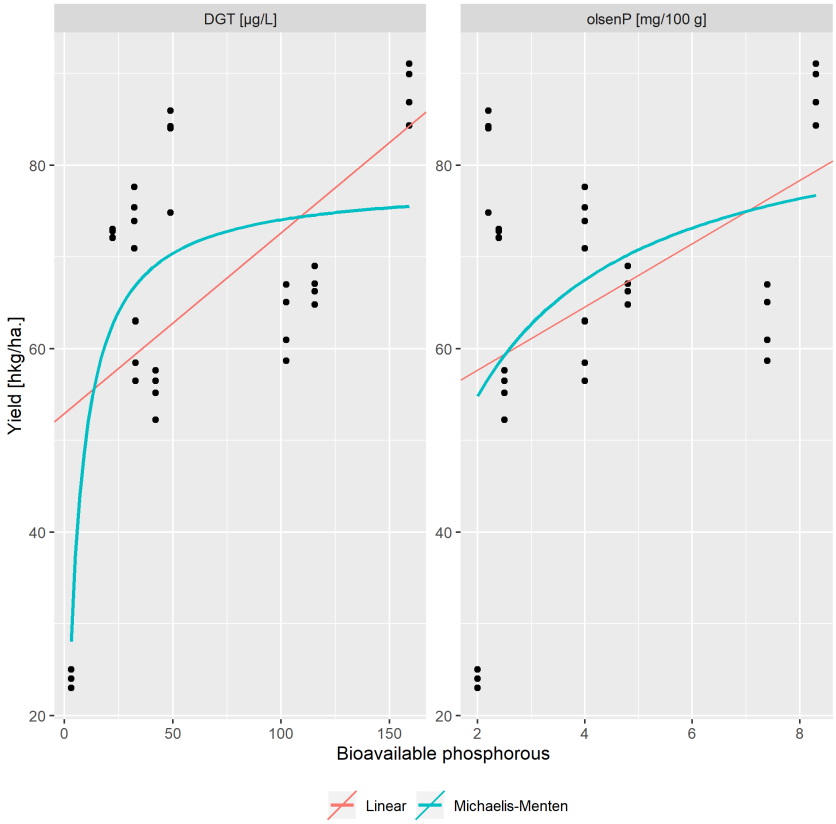


Figure 2: The two regression models for the DGT and olsenP measurement

The GE and their 95% confidence intervals (computed by non parametric bootstrap) are listed in table (3) below.

Michaelis - Menten	DGT	olsenP
MSE	179.50	474.30
95 % confidence	]147;212[	]282; 693[

Table 3: Generalisation errors of the models

The difference in GE ($\hat{z}$) of the models is calculated to be

$$\hat{z} = -294.93$$

Under the null-hypothesis: there is no difference in GE between the models, ie. $\hat{z} = 0$, p-values and confidence intervals of $\hat{z}$ are calculated using RST and non-parametric bootstrapping, both listed in table (4) below.

Difference in performance	RST	Bootstrapping
p-value	0.014	-
95 % confidence	-	]-495 ; -120]

Table 4: Statistical test results of difference in performance

4.2 The effect of phosphorus

In figure (3) below the distributions of α and β of $\mathcal{M}_A$ fitted to the permuted data are shown. With the distributions vertical lines are plotted indicating the 95% confidence intervals and the observed values of the parameters.

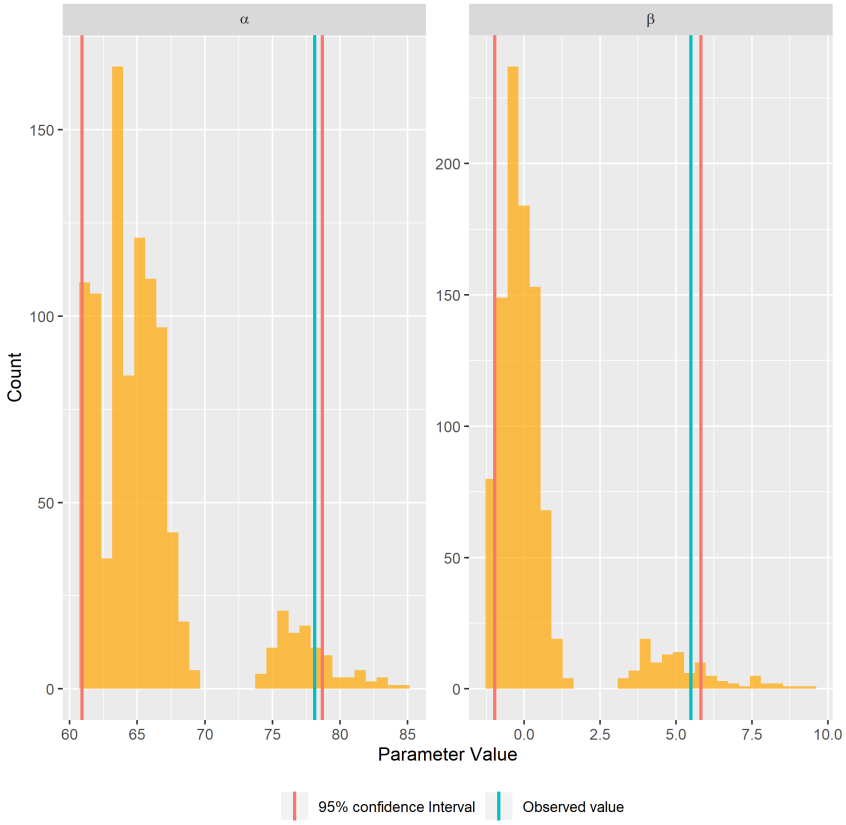


Figure 3: Distributions of the parameters of $\mathcal{M}_A$ by permutation test

The results of the permutation test are listed below in table (5).

Model	95% α-span	95 % β-span	α_{obs}	β_{obs}	$\alpha_{p-value}$	$\beta_{p-value}$
$\mathcal{M}_A$	[60.9 ; 78.7[	]-0.96 ; 5.82[	78.1	5.49	0.066	0.066

Table 5: Results from permutation test

5 Discussion and conclusion

According to the statistical evaluation of the two measuring methods, there was a significant difference in performance between DGT and olsenP at predicting yield based on phosphorous using the Michaelis-Menten model, as the p-value belonging to the null hypothesis (no difference in MSE) was 0.014. DGT not only had a much lower MSE (179.5 against 475.3) but also a more narrow confidence interval (]147;212[against]282;693[).
Using the Michaelis Menten model fitted to DGT one should expect predictions of yield based on bio available phosphorus to be $\approx \pm\sqrt{179.5} \approx \pm 13$ hkg/ha and using the model fitted to olsenP $\approx \pm\sqrt{475.3} \approx \pm 22$ hkg/ha - with a greater uncertainty than with DGT. The α parameter of DGT model was around 78, which, according to the model, means that the amount of phosphorus can maximally increase yield to 78 hkg/ha.
When testing whether phosphorus actually influences yield, however, we could not reject the null hypothesis, that effect of phosphorus on the yield described by the Michaelis-Menten model was due to chance, at a 5% significance level ($p = 0.066$). This is most likely due to many factors impacting the yield other than phosphorus. The few number of data points, which are grouped in 4 yield measurements per phosphorus measurement, and the fact that yield seems to be highly sensitive to location makes the task a difficult regression problem, and could explain why we could not reject the null hypothesis, even though phosphorus is a popular fertiliser. In many cases a significance level of 5% might not be necessary and the question of whether the farmer should upgrade to the DGT measurement on a 7% significance level is also influenced by the price difference between the two methods.

When looking at figure 1 we see that there is a group which might be considered an outlier since the yield is much lower for this group than the others. If this group was removed the linear model may be a better choice, but DGT and olsenP both seem to agree on the low values, so this has not been done.

A Appendix

A.1 R-markdown used for project 2

https://github.com/s183920/02445_Statistical_evaluation_of_AI/blob/master/Proj_2/proj2.Rmd

A.2 Linear models

```
Call:
lm(formula = yield ~ DGT, data = data_fos)

Residuals:
     Min       1Q   Median       3Q      Max
-30.5574  -8.7287  -0.4578  14.6196  23.3168

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.95788    4.15290  12.752 1.64e-14 ***
DGT          0.19676    0.05267   3.736 0.000685 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.37 on 34 degrees of freedom
Multiple R-squared:  0.291,     Adjusted R-squared:  0.2702
F-statistic: 13.96 on 1 and 34 DF,  p-value: 0.0006852


Call:
lm(formula = yield ~ olsenP, data = data_fos)

Residuals:
    Min      1Q  Median      3Q     Max
-34.651  -7.389  -0.677  12.005  27.565

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.756      5.979   8.489 6.49e-10 ***
olsenP         3.450      1.270   2.717   0.0103 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.55 on 34 degrees of freedom
Multiple R-squared:  0.1784,    Adjusted R-squared:  0.1542
F-statistic: 7.382 on 1 and 34 DF,  p-value: 0.01029
```

Figure 4: Summary of fitted linear models

A.3 Q-Q plot of losses

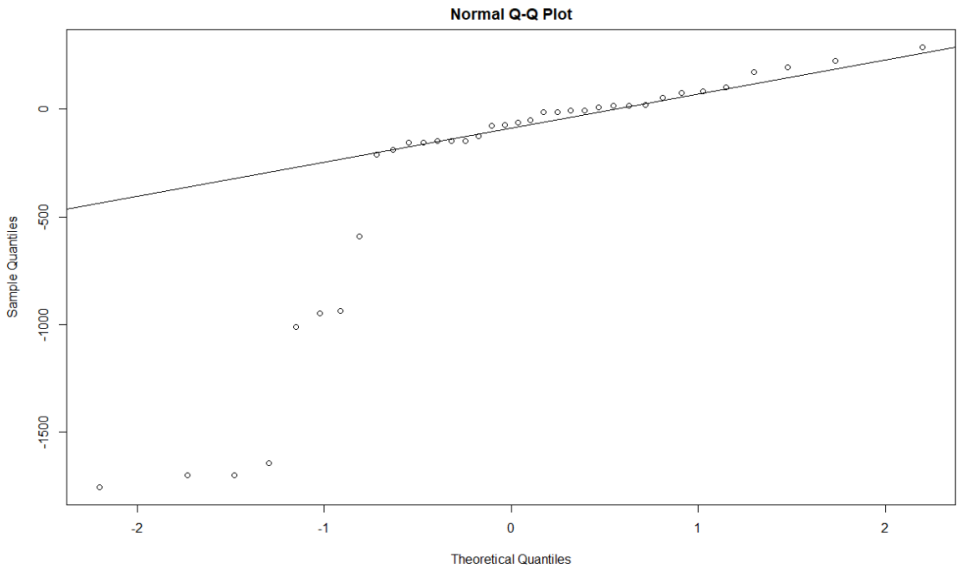


Figure 5: A Q-Q plot of the difference in losses for the regression model