

Project 2  
Course 02445  
Project in Statistical evaluation of  
artificial intelligence

Rasmus J. P. s164564  
Nikolaj S. P. s183930

January 2020

## Summary

Data collection can be expensive and time consuming so how do we decide what to collect and how to collect it? That is the job of a data scientist to answer and in this report we examine different techniques of measuring bio-available phosphorous in soil and analyze them individually as well as determine whether the amount of bio-available phosphorous influence the yield of a barley.

We find that the yield certainly correlates with bio-available phosphorous but only up to a point whereas afterwards increasing the amount of phosphorous does not have a significant effect on the yield of barley. We conclude that the two measuring techniques were significantly different and make recommendation towards the more expensive technique.

## 1 introduction

Crops need nutrients and if the levels of certain nutrients is too low the yield will be affected. Measuring the bio-available phosphorous BAP in soil is an important task for farmers in order to provide his or hers crops with sufficient amounts of nutrients. We know that the BAP is an important nutrient for plants but how big exactly is this influence of BAP to the yield of barley if it's there at all? That is the first aim of this report - to analyze and determine whether there is a significant influence from BAP on the yield and try and determine a model to describes this proposed effect of BAP. We analyze the models proposed and evaluate them on their fit to actual data. A new and more expensive measurement technique "DGT" is proposed to be better than the older "Olsen P" technique. Anyzing these two techniques are the second aim for this report and to determine if there's a significant difference between the two in order to make a detailed and guided recommendation for farmers. We will determine ... missing information about statistic test used to choose between Olsen and DGT.

## 2 Data

An experiment was performed on nine different fields spread across Denmark and Norway and each field partitioned into 4 plots. The yield of barley was measured and soil samples were analyzed by the two measuring techniques "DGT" and "Olsen P". The data contains 4 variables, three continuous and 1 categorical. The continuous variables are *yield* ( $hkg/ha$ ), *DGT* ( $\mu/L$ ), *Olsen P* ( $mg/100g$ ) and the categorical is a unique identifikation number for each of the nine fields. Observations were collected one from each of the four plots in the nine fields resulting in 36 observation of 4 variables. We decided to impute two missing datapoints with the mean of other plots from the same field. OBS BURDE VI IKKE BRUGE ALLE PLOTS THIL AT REGNE DENNE MEAN?. Our reasoning behind imputing the data was firstly, the fields do not have a high variance of yield between the plots, which means replacing the missing values with the mean would likely not be too far off the real values. Secondly, field 11 is an important datapoint with the lowest BAP, and removing it might heavily impact our models. See appendix .... ??..

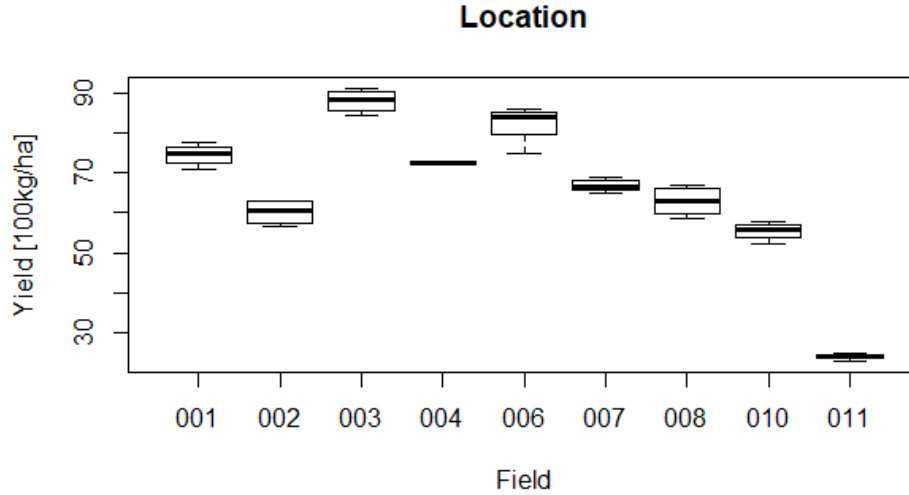


Figure 1: "Boxplot of yield for each field. We can see the variance within a field is relatively low, and location looks to have influence on yield"

Include a few good plots to highlight important features in data. You can put additional plots in the appendix.

### 3 methods and analysis

We start our analysis by assuming a linear correlation between response (yield), and our term (BAP measured with both Olsen P and DGT techniques). The resulting model is then evaluated on how well it fits our data via analysis of variance ANOVA.

#### Plots of linear fit

A non-linear model like the *Michaelis-Menten model* [see apendix for Michaelis model description](#) is often used for this kind of data so, like before, we evaluate the fit of this new model [anova on nls?](#). For now we will disregard the linear model because [reasons and plots](#) and focus on comparing the two measuring techniques using the non-linear model.

Using R's `nls()` function we fit the Michaelis-Mente models to our data and record the resulting parameters and their p-values making sure that the  $\beta$  parameter is a number larger than zero i.e. a beta of zero would result in the model *response = term*.

#### 3.1 Accuracy validation and testing

To maximize the accuracy of our estimated parameters we perform leave-one-out cross validation LOOCV and save the squared error of each fold for both models i.e

content...

Model	Linear DGT	Linear Olsen-P	Non-linear DGT	Non-linear Olsen-P
<b>R-squared</b>	0.2702	0.1542		
<b>p-values</b>	DGT = 0.000685	Olsen-P = 0.0103	Alpha = 2e-16 Beta = 0.0014	Alpha = 1e-9 Beta = 0.0432

one model where term is the "Olsen P" technique and one where it is "DGT". We then do a comparison of means with a paired two-sampled t-test.

## 4 results

Present the results. Tables and figures are good ways of illustrating results. What do your results show? Discuss your results. How reliable are they?

In table ?? we present our summed statistical analysis of variance for each separate model. The first thing we notice is that both of the standard linear models, the models fitted as: response = intercept+slope\*term turn out to perform quite well with a Residual standard error RSE of 15.37 for DGT technique and 16.55 for the Olsen P. Looking at the two non-linear models the difference becomes more apparent with RSE values of 10.58 for DGT and 16.33 for Olsen P.

Comparing each model to one another with CV and identical folds i.e. LOO-, we arrive at the tabled data in table ??.

From table ?? we learned that the non-linear model with the DGT measuring technique had the lowest value of RSE. Our subsequent test, see table ?? show that this lowest of all RSE is significantly different than all others.

## 5 discussion and conclusion

From our analysis we conclude that our investigation was well-founded, a significance level of 95% for the difference between using DGT- and Olsen P techniques in favor of the more expensive DGT. It must also be noted that had we only done a linear

Paired t-test between:	t-statistic	df	p-values
<b>Non-Linear DGT - Non-linear Olsen-P</b>	-2.694	35	0.011
<b>Non-Linear DGT - Linear Olsen-P</b>	-2.481	35	0.018
<b>Non-Linear DGT - Linear DGT</b>	-2.381	35	0.023
<b>Linear DGT - Linear Olsen-P</b>	-1.874	35	0.069
<b>Linear DGT - Non-linear Olsen-P</b>	-1.590	35	0.12
<b>Linear Olsen-P - Non-linear Olsen-P</b>	0.3065	35	0.76

regression model, we would not have come to the same conclusion.

All 4 models showed a 0.95 level of significant influence of bio-available phosphorous in the soil on the resulting yield of barley.

What are your conclusions? The conclusion should be connected to the aim of the report in the introduction. Highlight important results

If you have found interesting problems/aspects that you haven't carried out, you can specify them here as 'future work'.