Lucas De Vrieze (r0665032)

Predicting how many output an agricultural field would yield, is a yearly recurring question. Farmers would like to estimate the size of their harvest to be able to feed themselves and their communities, and to sustain their families with the revenue. Many factors are into play, ranging from the meteorological circumstances and the soil composition over the type of seed to the costs of labour and the market price. This study shows that is possible to get a relatively good general prediction of the output at a certain location with a limited number of macro-economic variables such as price and amount of inputs. These variables, however, are usually correlated, which poses a challenge.
This analysis presents a group of PLSR models for predicting the output of rice farms in Indonesia. A first exploratory section searches for latent variables in the numerical data using PCA. Next, via biplots and random forest classifications, it assesses whether the impact of each nominal variable is strong enough to underpin a grouping structure. Subsequently, the dataset is split so that each part comprises the numeric data for one combination of strong nominal variables. In the second section, a PLSR model is fit for each one of these subsets. Finally, the predictions by all models are compared.

1 Proceedings

1.1 Data wrangling

The dataset is retrieved from the `plm` R package and consists of 1026 observations for 20 variables. The field id and ownership status were considered irrelevant and therefore omitted. Regarding the output variable, the net output was selected as this also captures the harvesting costs, in contrast with the gross output.
Next to the total labour, the distribution between family labour and hired labour is provided as well as the wage of hired labour. Family labour does not inhere a cost, so a corrected wage was defined as the proportion of hired labour in the total labour times the wage. As such, this corrected wage reflects the average wage that was paid for the total labour. Eventually, only the corrected wage and the total labour variables were retained.
Variables `varieties` and `bimas` represent the sown seed variety and the application of an intensification program. The level `mixed` was dropped for both variables as the composition of the mix is not known and is consequently expected to be too heterogeneous.
At this point, three nominal variables are remaining: the region, the seed variety and the presence of an intensification program. Count tables show that there are few intensified fields in comparison with the non-intensified ones (742 vs. 80). Especially when viewing their count distribution over the regions to assess the impact of region-specific factors later on, the data become scarce and unbalanced. Therefore, only non-intensified fields were retained. These count tables also show that, except one, all regions have a clearly preferred seed type. Consequently, only regions preferring the same seed type should be directly compared to each other.
The final dataset contains 742 observations for 14 variables, of which two are nominal.

1.2 Exploratory analysis

This exploratory analysis aims to find hidden patterns in the data and to determine whether the impact of the nominal variables is strong enough to underpin a grouping structure. If there are, separate PLSR models will have to be made in the next phase.

1.2.1 Correlation analysis

The correlation analysis consists of calculating the pair-wise Pearson correlation coefficient for all numeric variables. By plotting these in a heatmap in Figure 1, it is visible that the variables are highly correlated and that a PCA will certainly help in finding latent structures.
For example, the net output is highly correlated with the inputs (size, seed, urea, phosphate and total labour), while the market price is correlated with the costs of these inputs (seed price, urea price, phosphate price and the corrected wage). The total labour itself is correlated with the inputs as well.

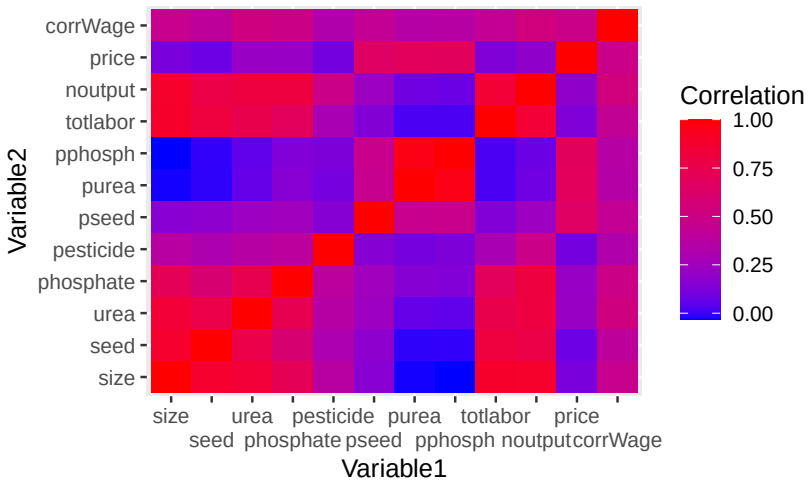


Figure 1: Pearson correlation heatmap of all numeric variables.

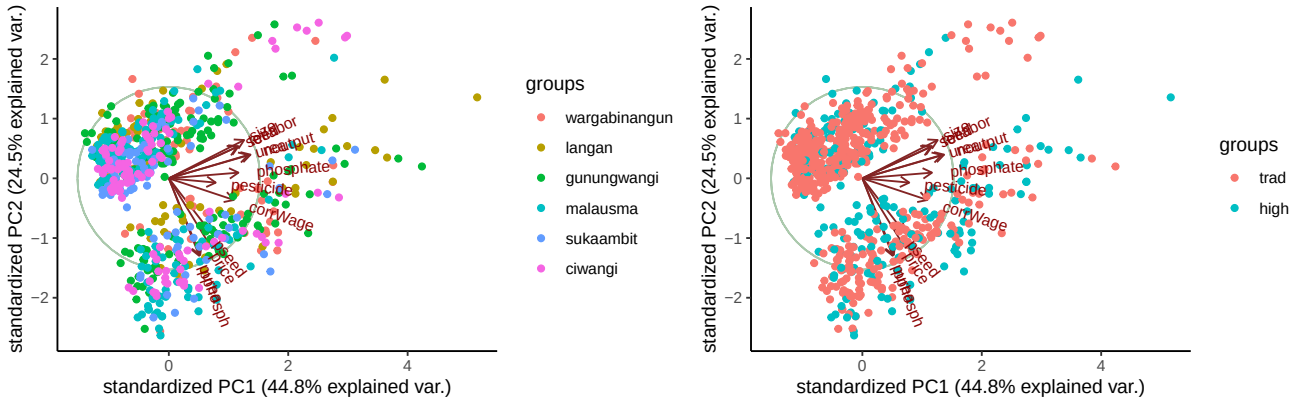
1.2.2 PCA

Due to the existence of correlated variables, a dimension reduction via PCA using the correlation matrix might be beneficial. The screeplot shows a kink at the inclusion of three PCs, which cover about 80 % of the variance. Two additional PCs increase the cumulative coverage to about 90 %. The loadings of these PCs have some interesting patterns. The first PC, which covers about half of the variance, takes all variables into account, but the inputs are weighted more heavily than the costs. The second PC, which covers 25 % of the variance, has negative loadings for the costs, while the third PC, covering about 7 % of the variance, has negative loadings for pesticides.
A first biplot shows two clear clusters and that there are some outliers which can be removed via PCA outlier detection. As the screeplot indicated that three PCs cover 80 % of the variance, a data point is removed if it is classified as an outlier by the PCA outlier detection method using any pair of PCs taken from the first three PCs. Therefore, the function `pca.outlier` of the `mt` package was modified so that it allows to set any pair of PCs for outlier detection instead of hard-coding for the first two PCs. As such, using a significance threshold of 99 %, 22 observations were removed, after which the PCA was redone and a new biplot was obtained. The same conclusions still hold, although the loadings slightly shifted. The first PC covers about 45 % of the variance, the second one about 25 % and the third one about 9 %. The screeplot still levels off at 3 PCs, but 6 components are now required to cover 90 % of the variance.

1.2.3 Biplots

Figure 2 shows these new biplots with a different group labelling according to the regions and the seed variety. The two clusters are again clear, but the regions nor the seed varieties are able to capture this group structure. Also for other PCs, there is no clear colour pattern.
Regarding the biplot itself, it shows that the net output aligns with all input variables (size, seed, labour, urea, phosphate), so the more inputs, the more outputs. The input prices are orthogonal to the inputs, but they align with the market price. The cost of the inputs is thus reflected in the selling price, but expensive inputs do not necessarily yield more output. The corrected wages appear to be correlated with both. A wage obviously is a cost, which explains its correlation with the other prices. The correlation with the output is less clear. Has hired labour a higher quality than family labour, and hence the higher net output? Another possibility is that it reflects that more labour results in more output, forcing the family to hire labour when the total labour requirements exceed their own capabilities.

Lucas De Vrieze (r0665032)



(a) Region group colouring **(b)** Seed variety group colouring

Figure 2: Biplot of the first two PCs with a group colouring based on one of the nominal variables.

From the PCA loadings, it was observed that the market price is one of the main distinguishing variables of the first two PCs, next to its correlates, the input costs (cfr. positive loadings in PC1, negative in PC2). Drawing a histogram also shows that its distribution has some bimodal tendencies. Hence, the market price might be the cause of this clustering. To verify this, a new nominal variable `premium` was introduced. It was defined as 'Cheap' if the market price was below the average price and 'Premium' otherwise. The average price was taken as threshold as it happens to be located near the valley between the two modes. The biplot using a premium group colouring in Figure 3 shows that this indeed is a good discriminator. However, this distinction is mostly orthogonal to the net output, so it is questionable whether this has an impact on the net output.

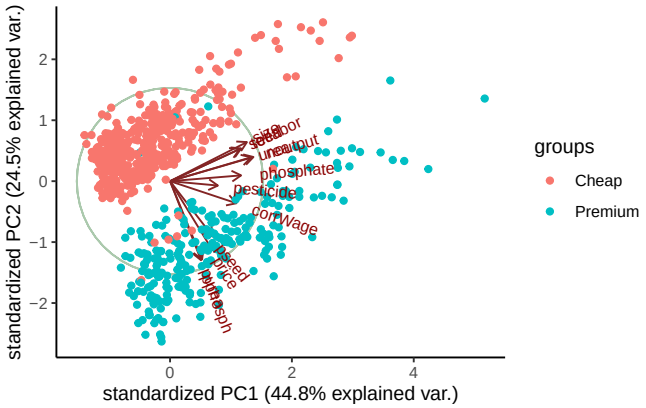


Figure 3: Biplot of the first two PCs with premium group colouring

1.3 Verifying group structures using random forests

There now are three nominal variables. For only one of these, there was a clear segregation at the biplot (`premium`). In order to decide for which nominal variables separate PLSR models will have to be constructed, a default random forests classifier is trained and tested using a 3:1 training/test set ratio. If a certain nominal variable indeed underpins a grouping structure, it should reflect in a good classifier performance. So, in case the classification performance is good, the importance of the variables is inspected for effects on the net output that are specific for this nominal variable. For example, a high importance of the net output implies that it is necessary to discriminate the groups and, hence, that it adheres some region-specific features that the other variables in the dataset do not account for. In that case, the nominal variable is retained.

1.3.1 Regions

From the count tables, it was observed that most regions have a preferred seed variety. Therefore, it is necessary to verify region grouping structure twice, once for regions preferring the traditional variety and once for regions preferring the high-yield one. This preference, however, is not absolute, so fields with the high-yield seed variety in traditional variety regions should be excluded and vice versa.
A random forest for the high variety regions has a training accuracy of 77.8 % and a test accuracy of 71.4 %, which is not that impressive. The variable importance plot shows that the net output is not one of the most important variables. Hence, the net output is not impacted by a region-specific factor that is not covered by one of the numeric variables in this dataset. The most important variables here are the market price, the price of urea and phosphate and the amount of seed sown. These might, for example, reflect a region-specific market dynamic or a local seeding habit, which the respective numeric variables should be able to capture.
For the traditional seed variety, the training accuracy was 76.9 % and the test accuracy 78.9 %. Again, the net output is not one of the most important variables. On the contrary, these are the market price and the pesticides usage. So, the same conclusion holds for the traditional seed variety regions. As a result, the region variable is not retained.

1.3.2 Seed varieties

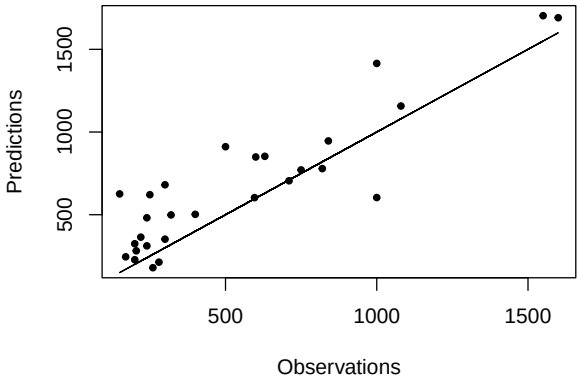
The same procedure was applied for verifying the seed variety group structure. The training accuracy amounts to 84.6 % and the test accuracy 82.8 %. The variable importance plot has an interesting feature: next to the price of the seed, also the phosphate consumption has a very high importance. A possible interpretation is that the high-yield seed variety has different nutrient needs, with an impact on the net output per applied input, especially for phosphate. So, the seed variety variable is retained.

1.3.3 Premium

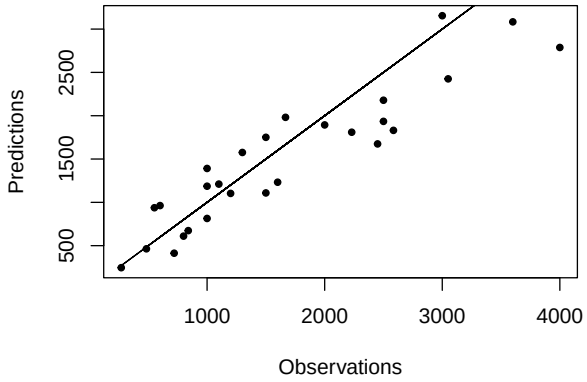
Although the biplot showed that the premium category is an excellent discriminant at the background of the two dominant PCs, the same procedure is repeated for the sake of completeness. The market price variable was left out as the premium category was constructed from it. The training accuracy is an excellent 97.4 % and the test accuracy 97.7 %. The variable importance plot shows that only the input costs (the price of urea, phosphate and the seeds, and the wages) are important, which are mostly orthogonal to the net output in the biplot. Although the impact on the net output thus has not been proven, the premium variable was retained because it is a good discriminator of the two clusters. Its effect, if any, should arise when comparing the PLSR models.

1.4 PLSR model construction

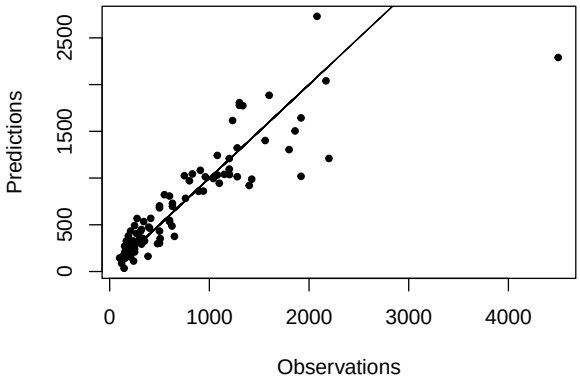
All records in the dataset were assigned a group number based on their seed variety and price category. A count table shows that the data are fairly well balanced over the four resulting groups.
For each group, the following procedure was executed. First, a PLSR model of the net output was fit to all other numeric variables for all records in that group and validated using 10-fold cross-validation. A validation plot determined the number of components to retain. Then, the group was split in a training and test set in a 3:1 ratio. A PLSR model was fit to the training set and its prediction performance was assessed using the test set. Finally, the test set predictions were plotted against the original observations in Figure 4. To facilitate comparing the model performances quantitatively, a R^2-like metric was determined from the total and the residual sum of squares of the identity line. It reflects the goodness of fit of the identity line in the prediction-observation point cloud.
For each group, at least a relatively good predictive performance was obtained according to the R^2 values. All point clouds generally align with the identity line. These PLSR models appear to be able to grasp the general trends in net output using only a limited number of correlated variables. The remaining variance might be due to more localised agronomic factors not taken into account in this study, e.g., the local weather, the exact soil composition, cultivation habits, fertiliser composition etc.



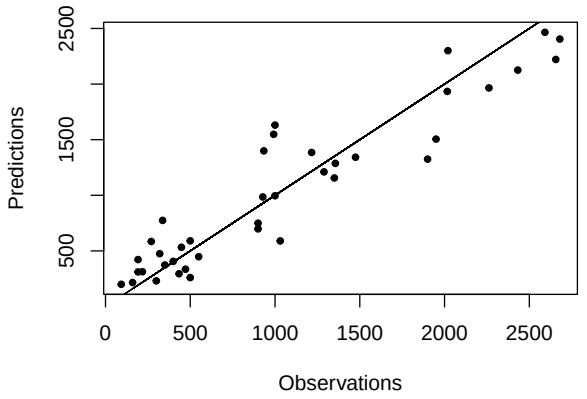
(a) Group CH (Cheap & high-yield): $R^2 = 0.70$



(b) Group PH (Premium & high-yield): $R^2 = 0.80$



(c) Group CT (Cheap & traditional): $R^2 = 0.75$



(d) Group PT (Premium & traditional): $R^2 = 0.88$

Figure 4: Visual comparison of the predicted and observed net output of the test datasets for all four groups. The more accurate the prediction, the closer to the identity line. In this regard, a R^2 coefficient was calculated, which reflects how well the identity line fits the point cloud. From a minimal R^2 of 0.7, it is concluded that the PLSR models perform quite well.

1.5 Comparing model predictions

Each one of the four constructed PLSR models predicts a net output from a set of input variables for a certain group of this dataset, but the weighting of these input variables in the latent model components is dependent on the combination of seed variety and price category of that group. To assess the impact of these two variables, each model is applied to predict a net output for all fields in the dataset. As such, four predictions are obtained for each field, which allows to compare the behaviour of the models directly.

Figure 5 shows the distribution of all predictions by all four models. The high-yield seed boxplots are slightly shifted upwards in comparison with the traditional seed boxplots. High-yield seeds are thus expected to give rise to a higher net output. However, remembering the importance of the phosphate variable for the random forest classification, this higher net output might come at the cost of higher nutrient requirements, especially for phosphate. At the other hand, such a shift is absent when comparing cheap and premium rice models. This implies that higher input prices result in more expensive rice without any substantial impact on the net output. Hence, the premium categorisation might rather reflect the local market dynamics instead of more expensive high-quality inputs giving rise to more output.

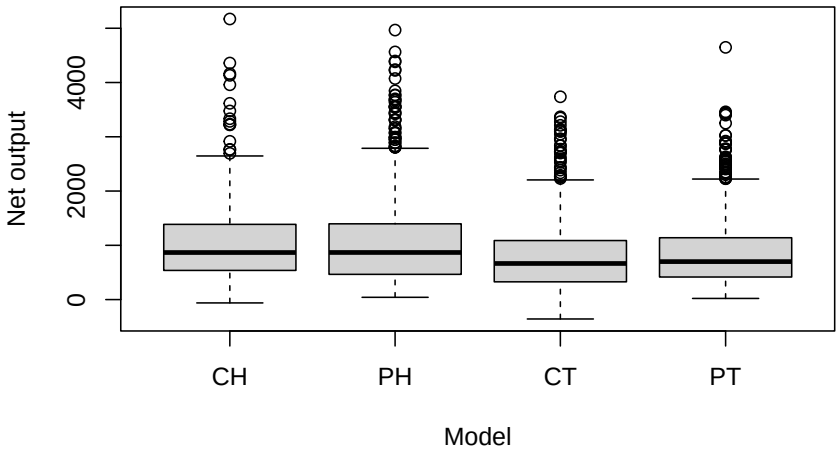


Figure 5: Comparative boxplot of the net output predictions by the four PLSR models. The high-yield seed models (CH & PH) generally predict a higher net output than the traditional seed models (CT & PT). Interestingly, there is no clear difference when contrasting the predictions by cheap and premium rice models (CH vs. PH & CT vs. PT).

2 Conclusion

This analysis showed that a PLSR is able to capture the general trends in predicting the net output of rice farms in Indonesia using only a limited number of correlated variables. After a data wrangling step, an exploratory analysis showed that the numeric variables in this dataset are highly correlated. A PCA and biplots demonstrated that these mainly are organised in two groups: input quantities such as nutrients, seeds and labour at one hand, and input costs such as wages, buying prices and the selling price of rice at the local market at the other hand.
The binary nominal variables supported no clear grouping structure, but the biplots shed light on a new binary market premium category as being a good discriminant of the two perceived clusters. Via random forests classifications and the associated variable importances, the seed variety and this premium category were retained as a grouping structure. Consequently, these two binary variables underpinned a split of the dataset into four subsets. For each one of these subsets, a PLSR model was constructed to account for the highly correlated variables. A comparison of the predictions by these models showed that high-yield seeds are expected to yield a higher net output, but potentially at the cost of a higher phosphate consumption.