Analyzing indonesian rice farms

Jonas Kernebeck, Alexander Flick, Felix Lehner

07/16/2021

Contents

1 Introduction

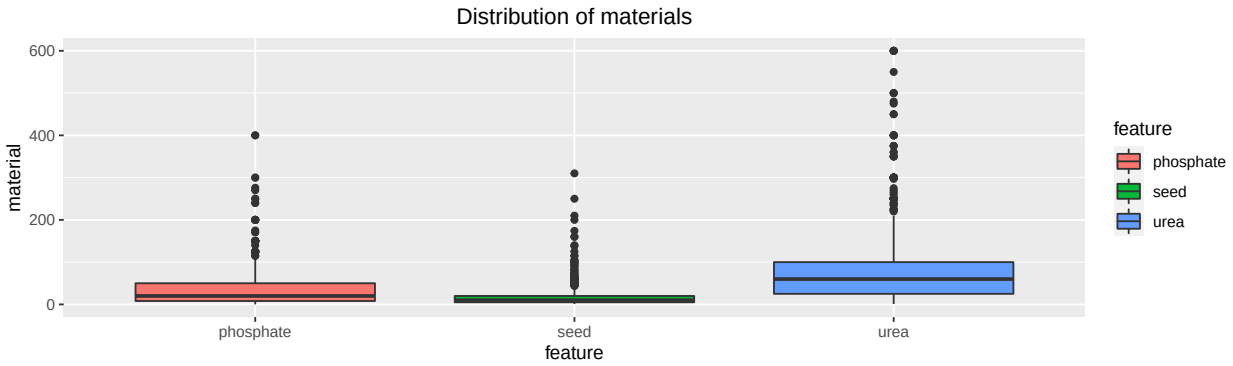
The present data set includes production data for 171 indonesian rice farms. The dataframe contains the following variables:

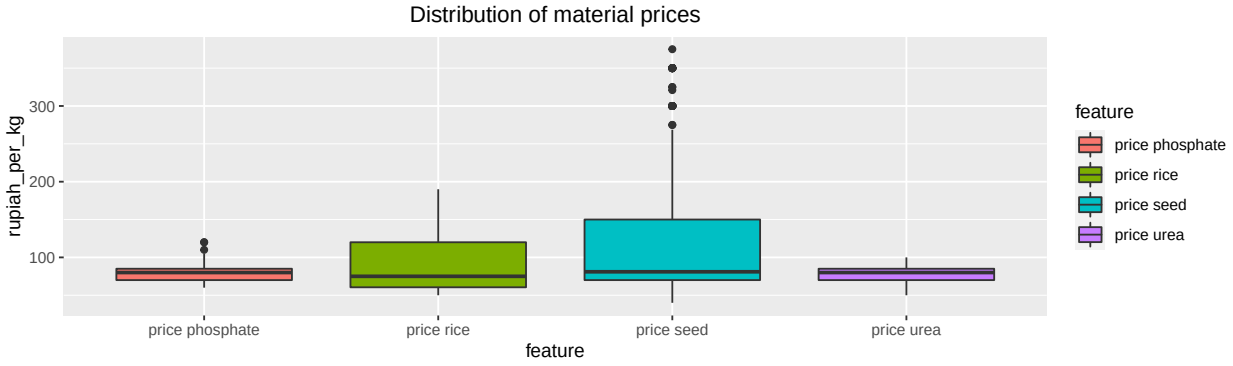
variable	description	expressions
id	unique identifier for a farm	unique id
time	unique identifier for a specific growing season	1 - 6
size	total production area in hectares	0.01 - 5.322
status	status of property rights	“owner”, “share”, “mixed”
varieties	rice seed varietes	“trad”, “high”, “mixed”
bimas	bimas-status of the farmers	“no”, “yes”, “mixed”
seed	seed in kilogram	1 - 1250 kg
urea	urea in kilogram	1 - 1250 kg
phosphate	phosphate in kilogram	0 - 700 kg
pesticide	pesticide cost in Rupiah	0 - 62600 r
pseed	price of seed in Rupiah per kg	40 - 375 r/kg
purea	price of urea in Rupiah per kg	50 - 100 r/kg
pphosph	price of phosphate in Rupiah per kg	60 - 120 r/kg
hiredlabor	hired labor in hours	1 - 4536 h
famlabor	family labor in hours	1 - 1526 h
totlabor	total labor (excluding harvest labor)	1 - 4774 h
wage	labor wage in Rupiah per hour	30 - 175.35 r/h
goutput	gross output of rice in kg	42 - 20960 kg
noutput	gross output minus harvesting cost	42 - 17610 kg
price	price of rough rice in Rupiah per kg	50 - 190 r/kg
region	region of the farm	unique region

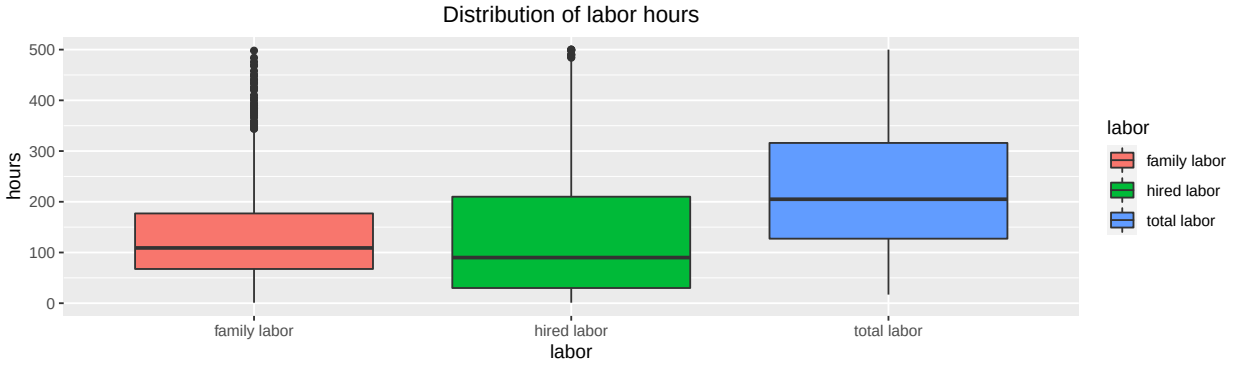
As present in the table, the data set consists of 16 numeric variables and 4 categorical variables. The target variable for the regression modeling will be *goutput*, what represents the gross output of rice in *kg* for the respective rice farm. In the following some explorative data analysis will be made to get to get a first impression of the distribution of the individual variables.

1.1 Numerical Variables

The following figure shows boxplots for the used materials and the prices paid for the materials of the respective rice farms. The boxplots for the materials show, that the distribution of all materials is right-skewed. The spread width of seed is the lowest, followed by phosphate and urea. Therefore *urea* also has the highest variance with 16166 followed by *phosphate* with 2264 and *seed* with 2048. The distribution of *urea* indicates that rice farms in Indonesia may use urea very different, caused by e.g. the bimas-status. The bimas program is a rice intensification program by the government to support local rice production by providing high-yield rice seeds as well as technical assistance. If we look at the prices for phosphate *pphosph* and urea *purea*, we can see a slight left-skewed distribution with low variance (75 for *purea* and 86 for *pphosph*). In contrast to that, the prices for seeds scatter much. The distribution of *pseed* is strongly right-skewed as well as the distribution for the rice price *price*. The price for the rice also scatters, but less than *pseed*. The two prices have a correlation of 0.67. Of course, the price of seeds affects the selling price of rice. The prices may fluctuate due to seasonal or regional factors and have an impact on each other. The distribution of labor hours is also slightly skewed to the right. Overall, the dispersion is lowest for the *famlabor*. For *hiredlabor* and *totlabor* we have a similar spread, but *totlabor* has a higher level overall. This is caused by the *hiredlabor* which is a subset of *totlabor*.

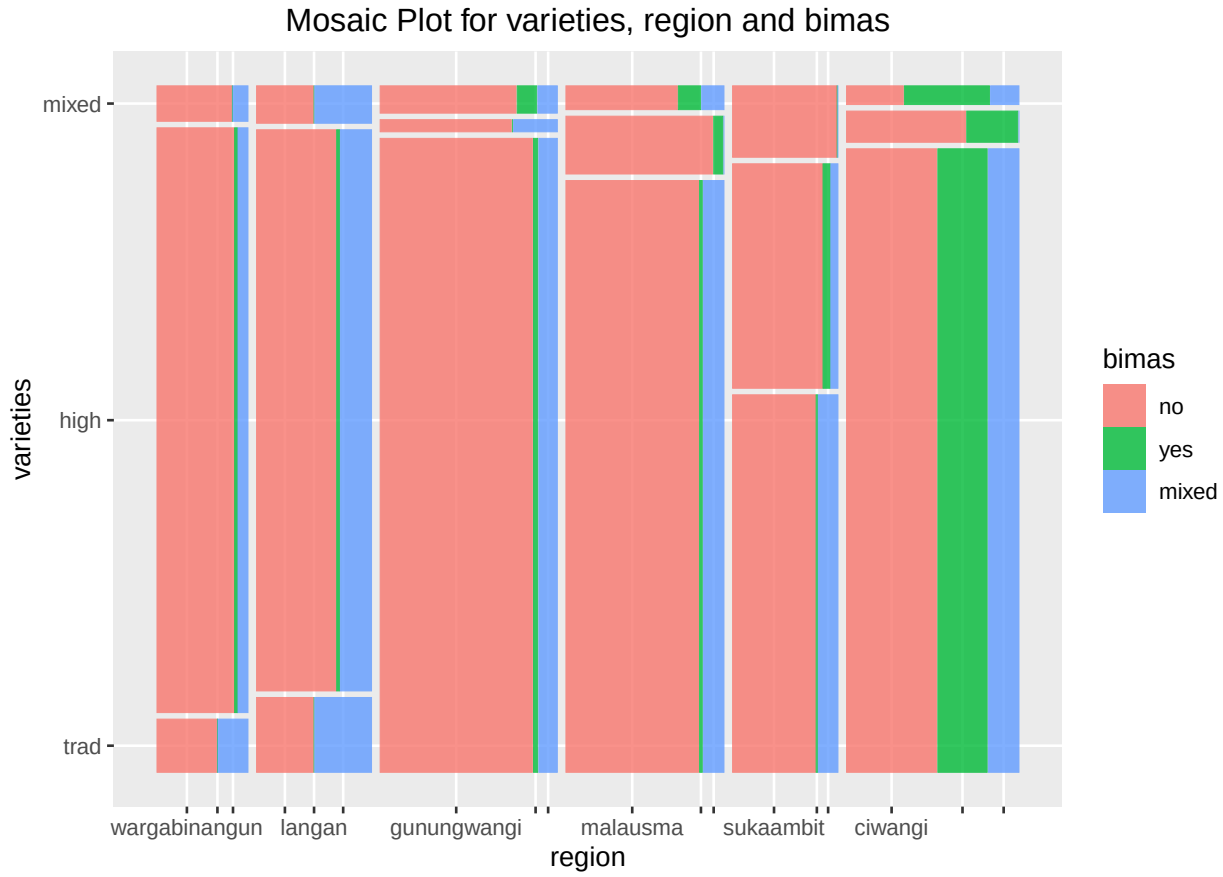

Distribution of materials
material
600
400
200
0
phosphate
seed
urea
feature
feature
phosphate
seed
urea


Distribution of material prices
rupiah_per_kg
300
200
100
price phosphate
price rice
price seed
price urea
feature
feature
price phosphate
price rice
price seed
price urea


Distribution of labor hours
hours
500
400
300
200
100
0
family labor
hired labor
total labor
labor
labor
family labor
hired labor
total labor

1.1 Categorical Variables

The following mosaic plot shows the distribution of of the categorical variables *varietes*, *region* and *bimas*. Overall, all regions are roughly equally represented in the data set. We can detect, that most of the farmers with the bimas status *yes* and *mixed* are located in the region *ciwangi*. The distribution of the different varieties is strongly dependent on the region. While the *high* varieties have the biggest share in the regions *wargabinangun* and *langan*, the *traditional* varieties are dominating the regions *gunungwangi*, *malausma* and *ciwangi*. The *mixed* varities are only used slightly in all regions.



To test wheter the categorical variables have impact on our target variable *goutput*, one- and two-sided anovas are performed. The results of these are summarized in the following table:

formula	F-value	p-value	significant
region	22.981	$<$ 2e-16	yes
varieties	11.764	8.94e-06	yes
bimas	14.817	4.57e-07	yes

formula	F-value	p-value	significant
region+varietes	3.847	3.96e-05	yes
region+bimas	5.651	2.94e-08	yes
varieties+bimas	0.791	0.531	no
region+varieties+bimas	0.860	0.580	no

The anova outputs show, that all of the categorical variables have a significant effect on *goutput*. The null hypothesis, that the mean of *goutput* is the same across the groups is rejected. The results of the two-sided anovas also show a significant interaction effect on *goutput*. While the interaction effect from the *region* with *varieties* and *bimas* is significant, the interaction effect of *varieties* and *bimas* and the interaction effect of all three variables is not.

1.3 Variable selection and transformation

The performance of the regression modeling is highly dependent of the variable selection and transformation. Therefore a suitable choice is very important. The variable *noutput* is a linear transformation of *goutput* as it represents *goutput* decreased by the harvesting costs. Therefore it is not used for the modeling because it would violate the multicollinearity assumption.
The variable *size* also correlates *strongly* with the target variable. This can be intuitively explained by the fact that a larger rice field naturally always produces a higher yield. Since the variables *seed*, *urea*, *phosphate* and *pesticide* are dependent on size, they are transformed into per-hectare sizes by dividing them with the respective hectare size of the farm. The *size* variable is not used for further modeling.
The variables *famlabor* and *hiredlabor* are subsets of the variable *totlabor* and are therefore transformed into the share of *totlabor* by dividing them with the amount of *totlabor*. The variable *totlabor* is after that transformed to a per-hectare size by dividing it with the *size*. The variable *wage* follows a bimodal distribution. Therefore it is transformed into a binary variable, which indicates if the respective value is over or under 100.

1.4 Model evaluation

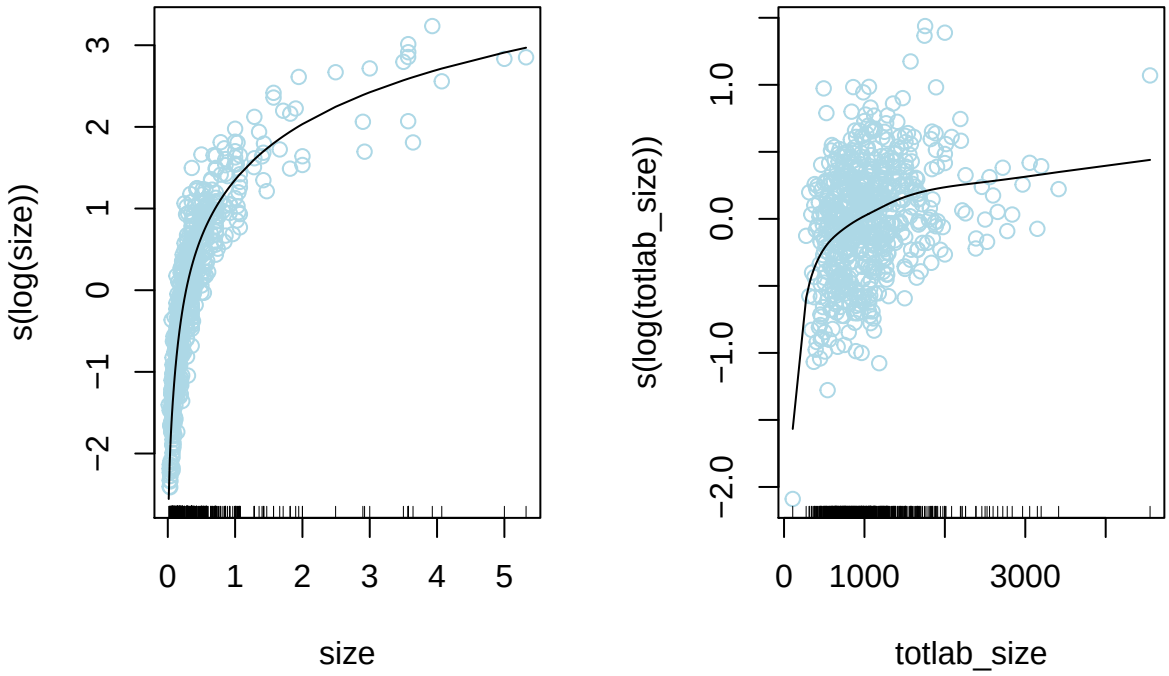
The data set will be splitted in 60-20-20 parts, where 60% of the data is used for training the model, and 20% for testing and validating respectively. In the modeling part, also cross-validation is used. To evaluate the models and compare them, different metrics will be used. The numeric metrics used are the *MSE*, which stands for the mean squared error and the *AIC*, which stands for the Akaike information criterion. Beside these metrics, also graphical analysis plots like a residual plots are used for evaluation.

3 Outliers most likely due to typos for observation 110, 947 and 1004

2 First Model

3 Second Model

GAM (Generalized Additive Model)



Explanation: The left-hand panel indicates that holding totlab_size fixed, goutput increases with size. The right-hand panel indicates that holding size fixed, goutput increases drasticall with the increased proportion of labour per size up to a proportion of 500 h / hectar and then flattens out.

GAM Comparison

table for model selection: The following table summarizes the variable selection process for the final GAM model. Starting from the top of the table we add for each grouped row in the table another variable with different representations of the variable for example checking whether log transformation yields any better results than using the variable as is. The results are mainly compared by using the p-value from the model summary, which is telling if the variable is significant important. We will use a p-value of 5% for the evaluation of variable significance. To compare models we will use the deviance. In special cases additional anova tests are performed to compare larger models with smaller ones. In addition we list other metrics like MSE on the training and validation data as well as AIC. For each grouped row, the last model is the chosen model for further analysis.

Forward selection (from James p. 79) : We begin with a model, that contains the variable with highest correlation to our dependent variable goutput, which is size. We then add different representations of one variable and add those to the model, which result in the lowest deviance for the new two-variable model. This approach is continued until all variables have been tried out. Forward selection is a greedy approach, and might include variables early that later become redundant.

Table 3: GAM comparison for variable selection

var	df	MSE.train	MSE.val	dev	aic	p_val_p	p_val_np	df_np
size								
s(size)	609	924954	535892	126.3	783.5	0.0000	0.0000	3
s(log(size))	609	928878	546896	107.5	684.3	0.0000	0.0016	3
size+labour								
s(totlab_size)	605	794773	540756	95.1	617.3	0.0000	0.0001	3
s(log(totlab_size))	605	796454	532193	93.7	608.2	0.0000	0.0140	3
size+labour+urea								
s(urea_size)	601	785027	554113	77.6	500.3	0.0000	0.0031	3
s(log(urea_size))	601	785928	543432	77.7	501.3	0.0000	0.0000	3
size+labour+urea+phosphor								
s(phosph_size)	597	640702	494023	71.7	460.0	0.0000	0.2846	3
s(log(phosph_size + 1))	597	637849	493065	71.7	459.9	0.0000	0.0000	3
size+labour+urea+phosphor+seed								
s(log(seed_size))	593	643332	475996	70.7	459.2	0.0077	0.1453	3
s(seed_size)	593	646291	475982	70.6	458.1	0.0245	0.0851	3
size+labour+urea+phosphor+seed+pesticide								
s(pest_size)	589	609355	448837	68.9	451.8	0.0006	0.2926	3
s(pest_size, df = 1)	592	619311	450226	69.4	449.4	0.0010	0.0024	0
size+labour+urea+phosphor+seed+pesticide+price								
s(price)	585	490980	374465	64.6	419.7	0.0007	0.0000	3
size+labour+urea+phosphor+seed+pesticide+price+family_labour								
s(fam_ratio)	581	486748	367829	63.9	421.6	0.1542	0.2932	3
size+labour+urea+phosphor+seed+pesticide+price+price_info								
s(pseed)	581	478646	372191	63.1	413.3	0.9013	0.0013	3
s(pphosph)	581	456162	356139	59.4	376.0	0.0000	0.0000	3
s(pphosph)+s(purea)	577	442872	360534	58.9	378.7	0.0711	0.0108	3
size+labour+urea+phosphor+seed+pesticide+price+price_info+wage								
s(wage)	577	412261	352955	57.5	364.1	0.0000	0.2517	3
wage_cat>100	580	423262	358715	58.0	363.5	0.0000	NA	NA
size+labour+urea+phosphor+seed+pesticide+price+price_info+wage+categorical								
bimas	578	427986	373423	57.0	357.0	0.0041	NA	NA
bimas+varieties	576	412347	351919	56.6	356.5	0.1174	NA	NA
bimas+status	576	412929	372755	56.8	359.1	0.3627	NA	NA
bimas+region	573	420793	367757	56.4	360.4	0.1572	NA	NA
size+labour+urea+phosphor+seed+pesticide+price+price_info+wage+categorical+excluded								
s(purea)	574	411251	384322	56.2	356.9	0.0367	0.0019	3
s(purea)+varieties	572	395042	361190	55.8	356.2	0.1137	NA	NA

Table columns:

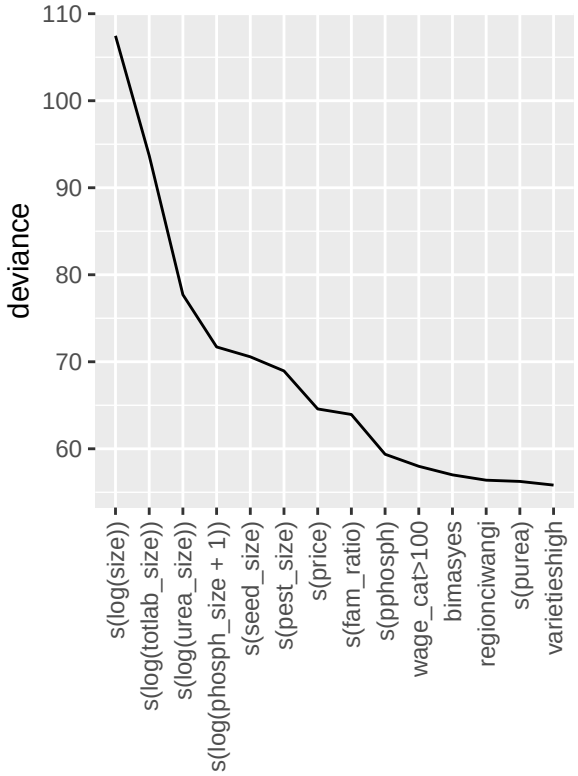
- var: the variable that is added
- df: degrees of freedom
- MSE.train: MSE on training data
- MSE.val: MSE on validation data
- dev: residual deviance (goodness of fit)
- aic: Akaike information criterion
- p_val_p: parametric p-value for last added variable
- p_val_np: non-parametric p-value for last added variable
- df_np: non-parametric degrees of freedom for last added variable

Variable selection explanation:

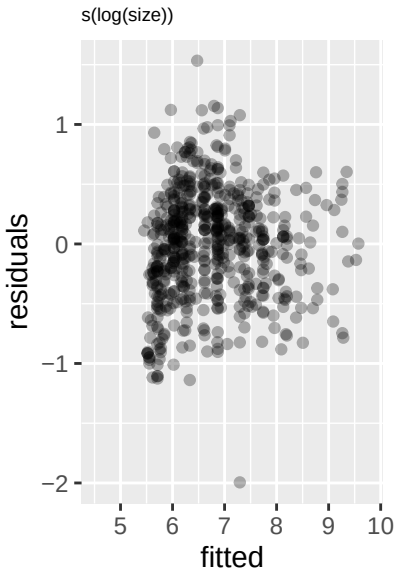
- size, labour, and urea: we will use logarithm of these variables, because of a smaller deviance
- phosphor: this variable has lots of zeros as values. So we can not use directly log transformation because log of 0 is -inf. log(x+1) transformation is the best way to avoid errors created by log transformation and is widely used among data scientists. So we will use this approach. We will use logarithm of phosph_size, because of a smaller deviance.
- seed: we will use seed as is, because of a smaller deviance.
- pesticide: non-parametric p-value is > 0.05 so we decrease df to achieve more suitable fit.
- price, family labour: Both variables are significant important for the model.
- material prices:
  - pseed: p-value > 0.05, thus we do not use this variable
  - pphosph: p-value < 0.05
  - pphosph + purea: p-value < 0.05, but the anova test comparing smaller including only pphosph with larger model pphosp + purea yields a p-value of
- wage:
- categorical variables:
  - bimas and bimas+region: p-value > 0.05,
  - bimas+varieties and bimas+status: p-value < 0.05
- final model: we check again for excluded variables after some more variables have been added if the results change. purea has a p-value of 0.05 and performing an anova test for varieties we can see that there is a significant difference between the two models.

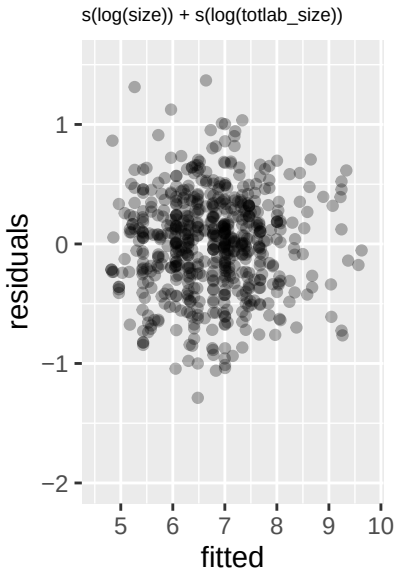
```
## Analysis of Deviance Table
##
## Model 1: log(goutput) ~ s(log(size)) + s(log(totlab_size)) + s(log(urea_size)) +
##     s(log(phosph_size + 1)) + s(seed_size) + s(pest_size) + s(price) +
##     s(pphosph) + wage_cat + bimas + s(purea)
## Model 2: log(goutput) ~ s(log(size)) + s(log(totlab_size)) + s(log(urea_size)) +
##     s(log(phosph_size + 1)) + s(seed_size) + s(pest_size) + s(price) +
##     s(pphosph) + wage_cat + bimas + s(purea) + varieties
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       574     56.247
## 2       572     55.826  2  0.42118   0.1156
```

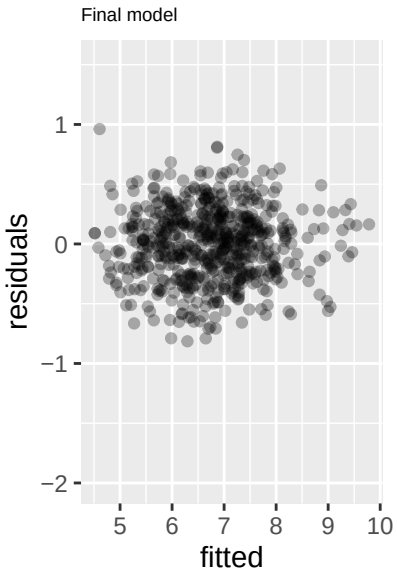
Model performance



Residual comparison







Final GAM

```
## log(goutput) ~ s(log(size)) + s(log(totlab_size)) + s(log(urea_size)) +
##     s(log(phosph_size + 1)) + s(seed_size) + s(pest_size) + s(price) +
##     s(pphosph) + wage_cat + bimas + s(purea) + varieties
```

R-squared (R^2): is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Essentially, an R-Squared value of 0.9 would indicate that 90% of the variance of the dependent variable being studied is explained by the variance of the independent variable.

$R^2 = 1 - \frac{Unexplained\,Variation}{Total\,Variation}$

$R^2 = 0.9102197$

Adjusted R-Squared: The adjusted R-squared compares the descriptive power of regression models that include diverse numbers of predictors. Every predictor added to a model increases R-squared and never decreases it. Thus, a model with more terms may seem to have a better fit just for the fact that it has more terms, while the adjusted R-squared compensates for the addition of variables and only increases if the new term enhances the model above what would be obtained by probability and decreases when a predictor enhances the model less than what is predicted by chance.

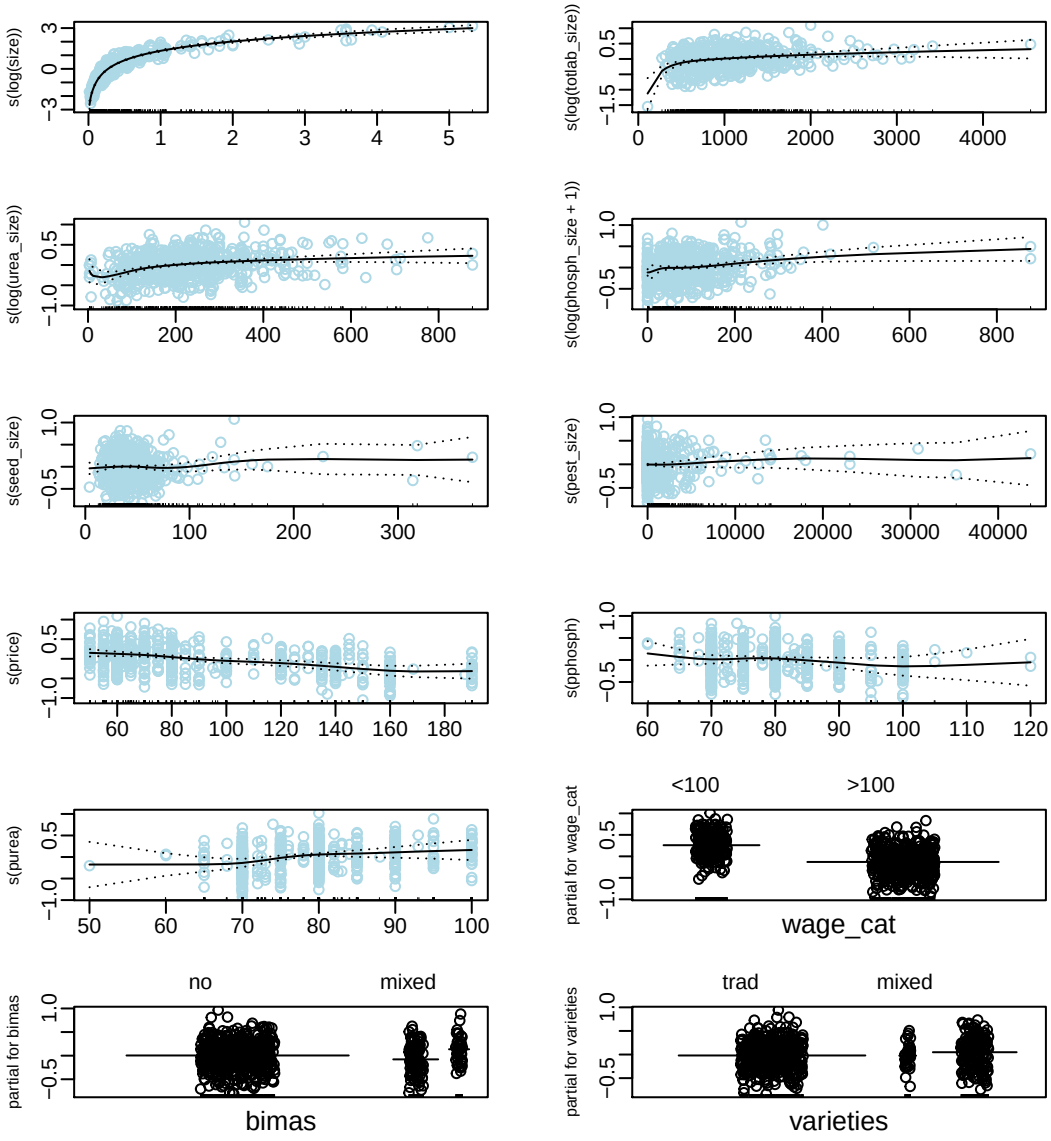
Source: https://www.investopedia.com/terms/r/r-squared.asp https://www.researchgate.net/post/How-can-I-get-the-adjusted-r-squared-value-of-GAM-model

$adjR^2 = 1 - \frac{(1-R^2)*(n-1)}{n-p-1}$

n = total sample size p = number of predictors

$adjR^2 = 0.9079677$

Final GAM visualization



4 Comparision

5 Conclusion