Visualization of RiceFarms dataset

Henrik Leifsson

2020-04-06

Contents

1 Introduction

1.1 Introduction to the dataset

The dataset chosen for the given report is called `RiceFarms` and is from a package called `plm`, accessible here. The dataset gives an overview of rice production in Indonesia and consists of **1026** observations and **21** variables:

See a first `glimpse` of the dataset:

```
## Rows: 1,026
## Columns: 21
## $ X          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,~
## $ id         <int> 101001, 101001, 101001, 101001, 101001, 10~
## $ size       <dbl> 3.000, 2.000, 1.000, 2.000, 3.572, 3.572, ~
## $ status     <chr> "owner", "owner", "owner", "owner", "share~
## $ varieties  <chr> "mixed", "trad", "high", "high", "high", "~
## $ bimas      <chr> "mixed", "mixed", "mixed", "mixed", "no", ~
## $ seed       <int> 90, 40, 100, 60, 105, 105, 50, 20, 15, 7, ~
## $ urea       <int> 900, 600, 700, 600, 400, 400, 120, 100, 15~
## $ phosphate  <int> 80, 0, 150, 100, 400, 400, 0, 0, 50, 0, 50~
## $ pesticide  <int> 6000, 3000, 5000, 5000, 10200, 10200, 0, 0~
## $ pseed      <dbl> 80, 70, 140, 90, 350, 250, 60, 50, 130, 15~
## $ purea      <dbl> 75, 75, 70, 70, 80, 80, 75, 75, 70, 70, 82~
## $ pphosph    <dbl> 75, 75, 70, 70, 80, 80, 75, 75, 70, 70, 82~
## $ hiredlabor <int> 2875, 2110, 980, 2081, 3889, 3519, 670, 80~
## $ famlabor   <int> 40, 45, 95, 10, 1, 1, 140, 50, 80, 69, 20,~
## $ totlabor   <int> 2915, 2155, 1075, 2091, 3889, 3519, 810, 8~
## $ wage       <dbl> 68.5, 60.1, 52.0, 57.0, 152.0, 154.5, 54.8~
## $ goutput    <int> 7980, 4083, 2650, 4500, 16300, 17424, 3840~
## $ noutput    <int> 6800, 3500, 2242, 3750, 13584, 14520, 3200~
## $ price      <dbl> 60, 60, 65, 70, 120, 140, 60, 50, 62, 60, ~
## $ region     <chr> "wargabinangun", "wargabinangun", "wargabi~
```

The most relevant variables are further elaborated below, but for an explanation of all of the variables, reader may read the documentation of the dataset from here.

- `ID`: the farm identifier
- `size`: the total area cultivated with rice, measured in hectares
- `status`: land status, on of 'owner' (non sharecroppers, owner operators or leaseholders or both), 'share' (sharecroppers), 'mixed' (mixed of the two previous status)
- `varieties`: one of 'trad' (traditional varieties), 'high' (high yielding varieties) and 'mixed' (mixed varieties)
- `pseed`: price of seed in Rupiah per kg
- `purea`: urea in kilogram
- `phosphate`: phosphate in kilogram
- `pesticide`: pesticide cost in Rupiah
- `goutput`: gross output of rice in kg
- `region`: one of 'wargabinangun', 'langan', 'gunungwangi', 'malausma', 'sukaambit', 'ciwangi'

1.2 Research Question

As we are dealing with rice production, it makes sense to look at the outcome of the said production, which is why the chosen research question for the given data is:

- *What are the most significant factors to influence the gross output of rice in Indonesia*?

Gross output of rice (in kg) was chosen because the purpose is to look into what variables affect the actual amount of rice and not the profits. In addition, net output is also depicted in terms of rice, harvesting costs were subtracted with no information about what constituted the said costs.

1.3 Preparing the Data

In order to make more accurate conclusions, new variables were created. The variable `size` often held more than one hectare of land whereas several other variables were showcasing the total for the land, lumping results of big farms together with small ones. This is why, to make adequate conclusions about the data, it was decided to look into each variable in terms of per hectare. It's also interesting to note that the regions are located on the same island of Java. Following variables were created:

- goutput_perhec = goutput/ size,
- totcost_pesticide_perhec = pesticide/size,
- totcost_seed_perhec = seed * pseed / size,
- totcost_urea_perhec = urea * purea / size,
- totcost_phosp_perhec = phosphate / size,
- totallabor_perhec = totlabor / size

Gross output per hectar or `goutput_perhec` was also cut into three categories to make it easier to compare to other variables:

1. **High Output**: 5000 - 27500 kg per hectare
2. **Medium Output**: 2500 - 5000 kg per hectare
3. **Low Output**: 0 - 2500 kg per hectare

2 Analysis

2.1 Categorical Variables

Due to the amount of categorical variables in the data, it was decided to first compare each of them to the gross output per hectare to see whether there was any correlations. Several plots were experimented with but mosaic plots were chosen in the end due to their ability to reflect the relative proportions. Here will be depicted mosaic plots of gross output per hectare with varieties of rice and region as they seemed to have most significance towards the gross output. Status did not seem to have an effect on the output and bimas intensification program was implemented on too few farms and didn't yield high enough results to be considered significant.

2.1.1 Gross Output per Hectare and Varieties of Rice

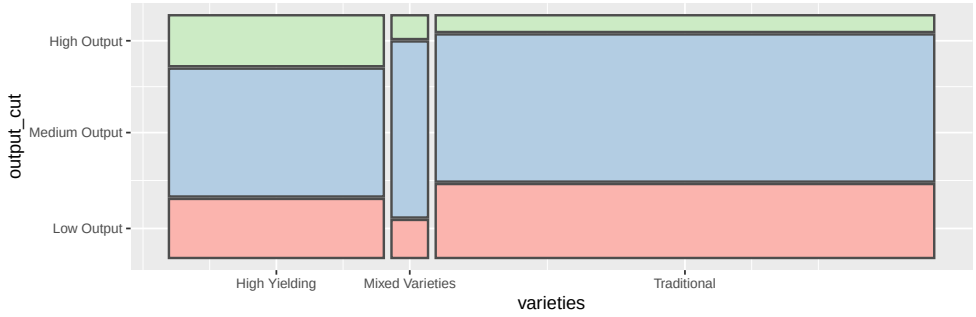

Figure 1: Mosaic plot of associations between gross output per hectare and varieties of rice.

2.1.2 Gross Output per Hectare and Region

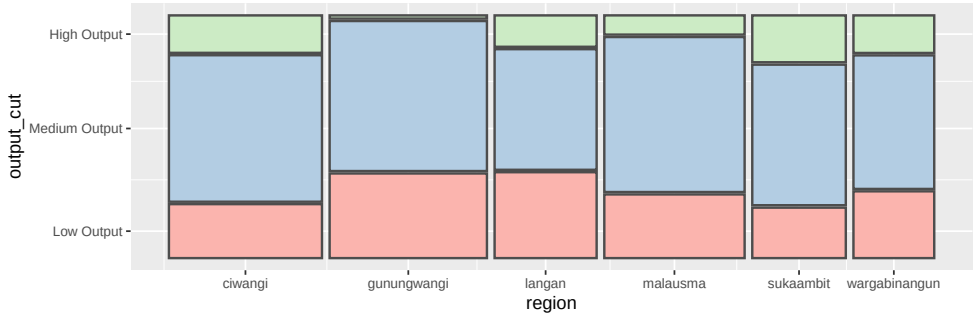

Figure 2: Mosaic plot of associations between gross output per hectare and region.

2.1.3 Varieties of Rice and Region

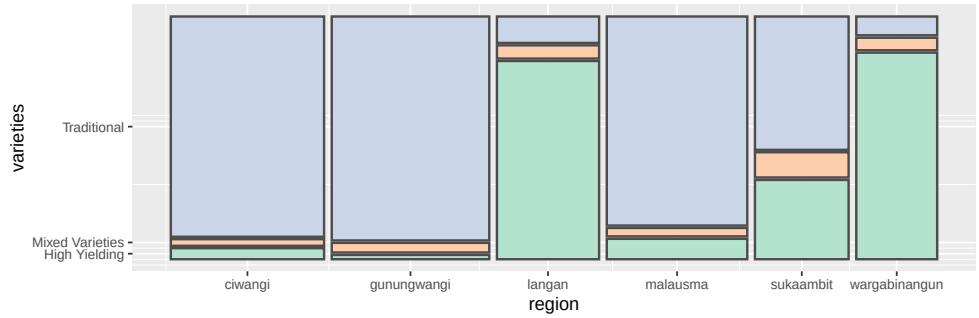

Figure 3: Mosaic plot of associations between region and varieties.

A mosaic plot of Rice and Region was also produced to see whether both of them were significant factors to influence gross output per hectare. As can be seen from the graph, some regions are high on high yielding rice type, some on the traditional rice type. Both variables varieties and region appear to be significant.

2.2 Numerical

Next, the numerical variables were looked at. Gross output per hectare was changed back to a numerical variable to be more compact. The visualizations were done in grid format, using the two previously determined categorical variables `varieties` and `region` to facet. Total cost for pesticide per hectare remained color coded in all of the graphs. Due to there being a large number of 0 values in pesticide, qualitative color palette was used to be able to better notice the color differences, even so, no significant patterns could be observed. Both y and x scales are also put into logarithmic scale to see the patterns of the data clearly.

The graph depicting gross output per hectare, total cost for phosphate per hectare, pesticide, varieties and region shows a positive relationship in all cases except for one region in traditional type of rice, which might indicate that the less total cost of phosphate per hectare would mean higher gross output per hectare. Due to stronger associations in other variables, this graph was added to the appendix.

In the cases of total cost of urea and total cost of seed per hectare, both variables showcased positive relationship with the gross output per hectare, except for one or two regions. This means that generally, assuming that more expensive total cost for urea means more urea and higher total cost of price for seed per hectare means higher quality - more urea and better seeds generally also produce a higher gross output per hectare.

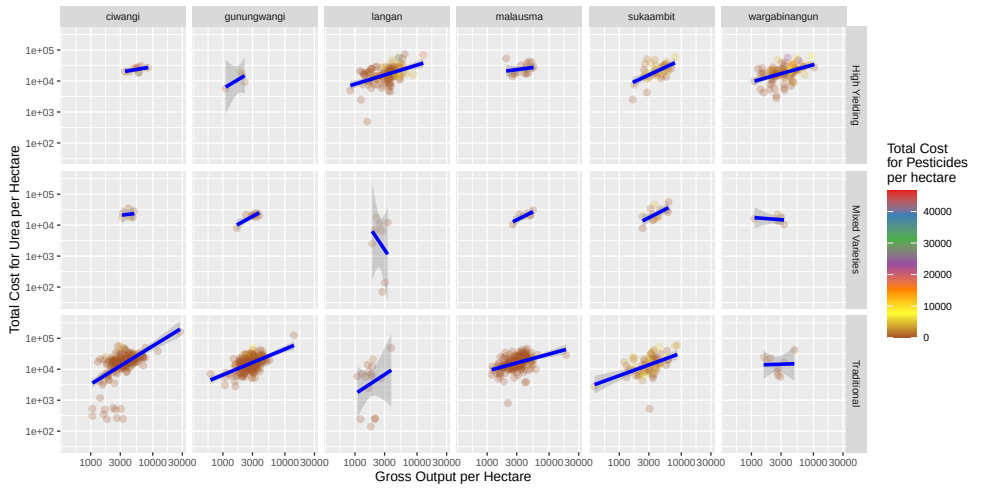

Figure 4: Scatter plots of the relationships between gross output per hectare, total cost of urea per hectare, total cost of pesticide per hectare, varieties of rice and region.

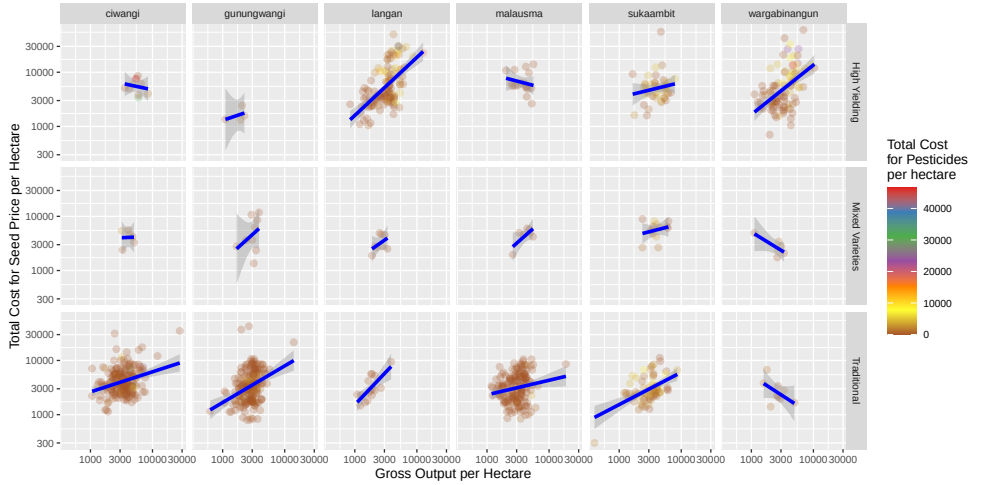

Figure 5: Scatter plots of associations between gross output per hectare, total seed cost per hectare, varieties and region

3 Conclusion

In conclusion, the categorical variable `varieties` meaning different varieties of rice impacted gross output the most. Even so, `region` also played a large role. Total cost for Phosphate per hectare, total cost for urea per hectare and total cost for price for seeds per hectare generally improved the gross output per hectare, but depended on the region and type as to what extent. To make any strong conclusions is hard as there is little information about the dataset, for example, over what period it was taken and how many periods it includes. One should look over a longer period of time, and also consider if there are geographical differences between the regions which can affect rice yields.

4 Appendix

Gross Output, Total Cost for Phosphate and Total Cost for Pesticide - all per Hectare - with Varieties and Region

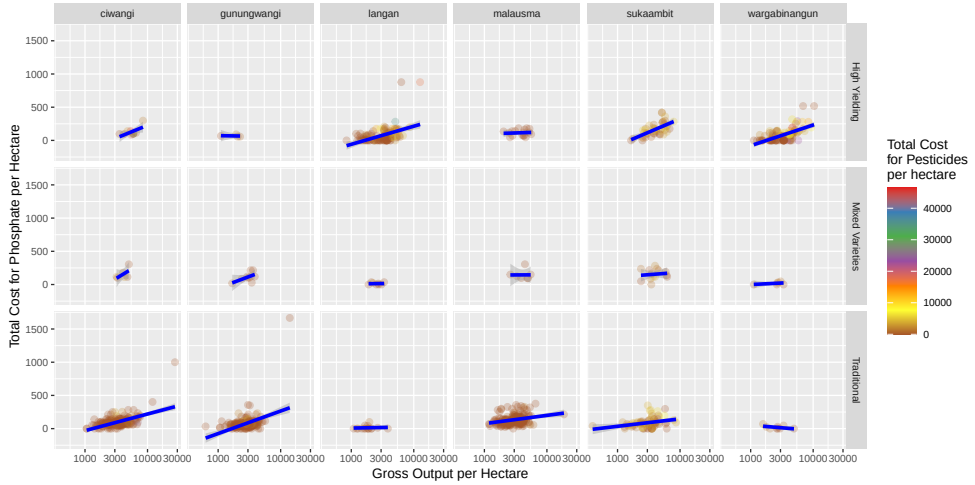

Figure 6: Scatter plots depicting the relationships between varieties of rice, region, gross output per hectar, total cost for phosphate per hectare and total cost for pesticide per hectare.