

CCRP - Research Methods Support Project

**Aims**

Visual displays of research data are important for at least two reasons:

* Helping researchers understand see and understand important results.
* Communicating and explaining results to others who will use them.

Many of the datasets that CCRP projects produce are rich and complex yet most of the graphs and other visualisations drawn and reported are rather simple, a bar chart drawn in Excel being the most common. There will often be benefits from using designs display more of the complexity in the data, but even the simplest layout can be made more effective by careful attention to detail. Current software can make it easy to draw such graphs.

Hence the aim of this collection of data displays is to show some of the possibilities for displaying research results with more diverse types of graphs and data graphic. All examples use data from CCRP projects, data that is typical of that produced within the program. We aim to identify some of the principles of good graphical practice used in each example, and point out some of the results that can be seen in the displays.

We hope the collection will provide some inspiration to CCRP teams lead to the use of more insightful and imaginative graphs within the program and beyond.

**Software for data display**

It does not matter what software you use, as long as you can generate the displays you want and need, rather than only a limited range offered. That includes the ability to modify and customise any details that are important. Two other characteristics of software make it much easier to do a good job:

1. Ability to code or program all elements of the graph (rather than doing it 'by hand' by clinking and selecting options). This means you can quickly draw the graph again with new data, modify it or share the code with others.

2. Free or open source software meaning you can learn from others without buying new software.

Software we know of that meets all these criteria is R. Examples here are mostly drawn with R.

[Legume integration for soil fertility management 3](#_Toc507180773)

[Legume integration for soil fertility management – farm to farm variation 5](#_Toc507180774)

[Variation in clonal response to EXW inoculation 7](#_Toc507180775)

[FAW on maize and sorghum 9](#_Toc507180776)

[PVS of bean varieties 11](#_Toc507180777)

[Farmers' experiment on striga management 13](#_Toc507180778)

[Plant – insect interactions 15](#_Toc507180779)

[Disease resistant bean variety selection 17](#_Toc507180780)

[Finger millet yield and farmers' ranks 19](#_Toc507180781)

[Effectiveness of research stage discussions in attaining multiple workshop objectives 21](#_Toc507180782)

[Appendix 1: Code to generate examples 23](#_Toc507180783)

[Appendix 2: Alternative for plant-insect interactions 27](#_Toc507180784)

[Appendix 3: Building up to a final product 28](#_Toc507180785)

# Legume integration for soil fertility management

|  |  |  |  |
| --- | --- | --- | --- |
|  | Average maize grain yield (t ha-1) | | |
|  | Site | |  |
| Treatment | Kandeu | Manjawira | Overall |
| no input | 0.8 | 0.9 | 0.8 |
| 25% recommended fertilizer | 1.3 | 1.6 | 1.5 |
| 100% recommended fertilizer | 3.4 | 3.3 | 3.3 |
| organic fertilizer | 3.0 | 3.2 | 3.1 |
| Legume system 1 | 3.7 | 3.0 | 3.4 |
| Legume system 2 | 3.6 | 3.1 | 3.4 |
| Legume system 3 | 3.0 | 3.4 | 3.2 |
| Legume system 4 |  | 3.6 |  |
| Legume system 5 |  | 3.6 |  |
| Legume system 6 | 3.5 |  |  |
| Legume system 7 | 3.4 |  |  |
| Overall | 2.9 | 2.9 |  |
|  |  |  |  |

Figure 1

## Data

About 15 farmers in each of two sites in Malawi carried out trials comparing alternative soil fertility management options. The treatments included several 'baselines' or standard practices as well as alternative ways of incorporating legumes. Three of the legume systems were common across the two sites but there were also two site-specific options at each location. Maize grain yield was measured as an indicator of soil fertility status.

## Graphical aims

In this first example we display the means and actually use a table of numbers rather than a graph as, in this case, it is probably clearer. Note that this is a table designed to communicate the main patterns in the means. It might need modifying (for example by adding standard errors of differences, sample sizes or more descriptive names for treatments) to make a table suitable for an academic research paper.

## Principles being used

* Make the most important comparisons the easiest to see by placing them adjacent vertically.
* Reduce number of digits in each number to the minimum needed to see the patterns.
* Only use horizontal lines, not vertical, and then only light lines needed to distinguish parts of the table.
* Using horizontal and vertical spaces to make the table more readable, and justify text to keep numbers and headings aligned vertically.
* Add means to margins when they are useful.

## Messages in the picture

* Average yields without input are low but there are clear responses to fertilizer.
* All the legume options result in average yields that are similar to those with full fertilizer.

## Acknowledgements

Data from the Southern Africa Best Bets project.

Table design by Ric Coe

# Legume integration for soil fertility management – farm to farm variation

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Inputs | | | | | | |
|  |  | none | fertilizer | | | legumes | | |
| Farm | Site | 0 | F1 | F2 | OF | L1 | L2 | L3 |
| Man4 | M | 1.0 | 2.5 | 9.4 | 5.2 | 3.2 | 8.2 | 5.2 |
| Man7 | M |  | 5.6 | 6.4 | 5.3 | 6.0 | 6.5 | 6.6 |
| Kan13 | K | 1.1 | 4.8 | 6.3 | 4.2 | 6.6 | 4.4 | 6.1 |
| Man10 | M | 1.2 | 1.7 | 6.4 | 8.3 | 6.2 |  |  |
| Kan9 | K | 1.5 | 2.1 | 2.7 |  | 6.6 | 4.8 | 7.3 |
| Kan10 | K | 0.5 | 1.2 | 6.3 | 4.5 | 3.1 | 4.3 | 5.7 |
| Man17 | M | 1.2 | 1.4 | 2.3 | 2.7 | 5.4 | 5.8 | 3.5 |
| Kan12 | K | 0.9 | 1.9 | 1.7 | 3.2 | 4.6 | 5.7 | 4.5 |
| Man2 | M |  |  | 5.9 | 3.1 | 2.7 | 2.2 | 3.2 |
| Kan16 | K | 3.0 | 2.1 | 3.7 | 5.0 | 4.3 | 2.6 | 2.4 |
| Man5 | M | 1.8 | 1.8 | 2.4 | 4.1 | 4.2 | 5.3 | 2.4 |
| Kan2 | K | 0.7 | 0.8 | 3.2 | 3.1 | 3.2 | 4.4 | 5.0 |
| Kan7 | K | 0.2 | 1.3 | 5.0 | 3.0 | 4.8 | 4.5 |  |
| Man8 | M | 0.5 | 1.9 | 2.9 | 2.6 | 5.3 | 2.2 | 4.9 |
| Man1 | M | 1.1 | 1.5 | 2.9 | 2.1 | 1.6 | 2.9 | 5.2 |
| Kan3 | K | 1.6 | 1.0 | 2.0 | 3.1 | 4.4 | 2.1 | 4.3 |
| Man9 | M | 1.3 |  |  |  | 1.1 | 2.9 |  |
| Kan11 | K | 0.5 | 0.9 | 2.8 | 2.8 | 2.8 | 3.2 | 4.6 |
| Man11 | M | 1.4 | 1.7 | 1.5 | 2.2 | 2.9 | 3.0 | 2.7 |
| Kan6 | K | 0.4 | 0.5 | 5.2 | 3.5 | 1.4 | 3.0 | 1.2 |
| Kan17 | K | 0.3 | 0.5 | 2.4 | 1.7 | 1.5 | 1.5 | 3.1 |
| Kan18 | K | 0.1 | 0.2 | 3.1 | 1.6 | 3.5 | 3.0 | 2.0 |
| Man12 | M | 1.6 | 0.4 | 1.3 | 1.7 | 1.4 | 5.7 | 2.0 |
| Kan14 | K | 0.7 | 1.3 | 3.5 | 1.5 | 2.3 | 1.3 | 1.1 |
| Kan8 | K | 0.3 | 1.0 | 1.5 | 3.0 | 5.5 | 1.6 | 1.8 |
| Man14 | M | 0.1 | 2.5 | 0.9 | 1.4 | 2.6 | 1.5 | 1.8 |
| Man3 | M | 0.2 | 0.5 | 0.9 | 1.5 | 1.5 | 1.5 | 1.4 |
| Kan1 | K | 0.1 | 0.2 | 1.6 | 1.4 | 1.3 | 1.0 | 1.5 |
| Man15 | M | 0.2 | 0.2 | 1.0 | 1.8 | 0.7 | 3.5 | 1.0 |
| Man6 | M | 1.3 | 0.4 | 1.5 | 2.5 | 0.9 | 0.3 | 0.2 |
| Man16 | M | 0.2 | 0.4 |  |  |  | 0.3 |  |

Figure 2

## Data

This is the same data as in Figure 1.

## Graphical aims

The aim is again to compare the options but this time to highlight variation between farms and sites.

## Principles being used

* Only use the data that is important for the comparison – in this case, options that are common across the two sites.
* Use colours to highlight patterns that might be hidden in numbers alone.
* Make order matter. In this case the order of columns is based on the logical of the treatments. The order of the rows was chosen to be in descending order of mean yield on each farm. Thus the farms with highest average yield are at the top of the table.

## Messages in the picture

* Both sites have similar distributions of overall yields per farm – there are similar numbers of farms at the top and bottom of the table from each site.
* Yields without inputs are low everywhere except for one farm.
* Response to inputs, both fertilizer and legumes, varies across farms.
* More farms respond well to legumes than to fertilizer.
* The legume option giving the highest yield varies across farms.
* There are more farms that respond well to legumes than to fertilizer.
* About a quarter of all farms are unresponsive to either legume or fertilizer input.

## Acknowledgements

Data from the Southern Africa Best Bets project.

Graph design by Ric Coe

# Variation in clonal response to EXW inoculation

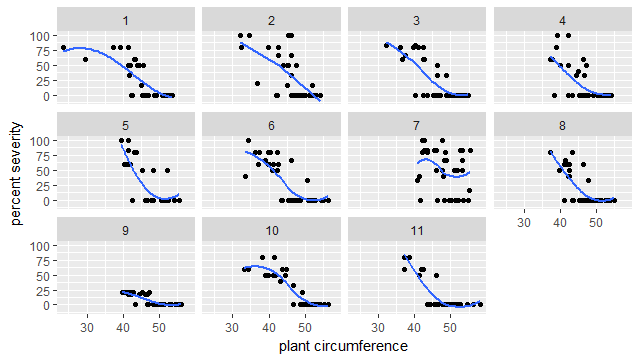


Figure 3

## Data

An experiment was done to evaluate the responses of 11 different clones of enset to inoculation with EXW. Severity of symptoms was measured and the diameter of each plant recorded at time of inoculation.

## Graphical aims

We want to show the relationship between severity and diameter, making it easy to see if there is a relationship, its nature and whether it is the same for all clones.

## Principles being used

* Small multiples (repeated graphs of identical design) reveal similarities and differences between patterns across many cases or contexts.
* Smoothing lines help highlight trends without making assumptions of straight lines.

## Messages in the picture

* There is a clear decrease in severity with increasing diameter.
* The pattern is much the same for all clones except two. Clone 7 does not show reduced severity for large diameters. Clone 9 has low severity at all diameters.

## Acknowledgements

Data from E Africa EXW project

Graph designs by Ric Coe

# FAW on maize and sorghum

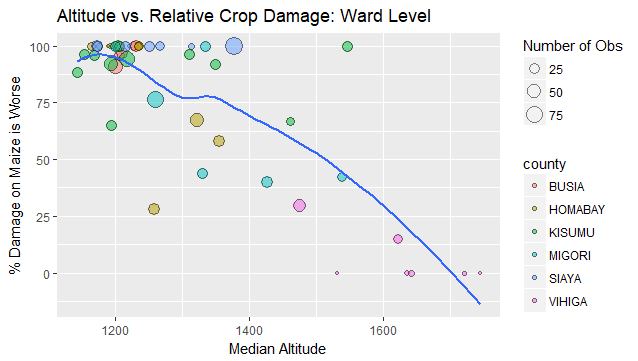


Figure 4

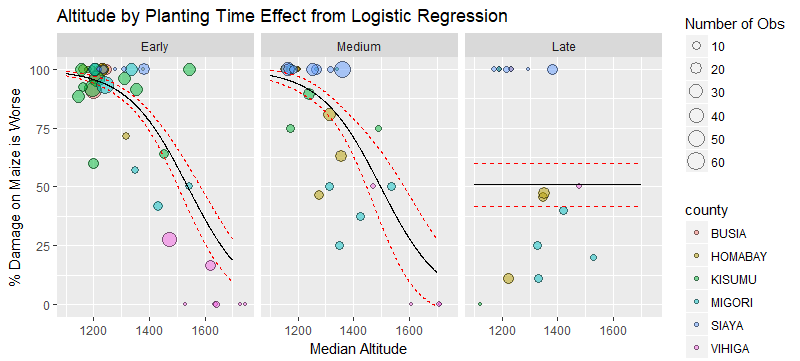


Figure 5

## Data

A rapid survey was done of nearly 1200 farmers growing both maize and sorghum during an outbreak of fall army worm (FAW). One response measured was whether the FAW damage was worse on maize than on sorghum. Context variables were measured in the hope they would explain variation.

## Graphical aims

We want to show the important patterns discovered. These an effect of altitude (a continuous variable), time of planting (categorical with three levels) and their interaction (there is no altitude effect for late planted fields). Scatter plots would be the usual approach. The response variable is a 0/1 variable at the level it is recorded hence any scatter plot with this as the y-axis has all the observations at 0 or 1, making it hard to see patterns. Hence for the purpose of displaying the data (not statistical modelling) observations are averaged to the ward level. The number of observations in each ward is not the same so needs to be shown. A statistical model has been fitted and we want to show the model along with the original data. Figure 4 is designed to show the key result of an altitude effect that applies within as well as between counties, with a smooth line to highlight trend. Figure 5 breaks it down by planting time and adds the fitted logistic regression model.

## Principles being used

* High data density (a lot of information per unit area of graph) gives rich displays.
* Make the display as unambiguous as possible by adding variables that explain and help remove ambiguity (county and n in this case).

## Messages in the picture

* At low altitude (<1200m) , maize is damaged by FAW more than sorghum in most farms. This difference decreases as altitude increases to 1400m. Higher still and sorghum is affected worse than maize.
* The altitude effect applies to early and mid planted sorghum but not late planted.

## Acknowledgements

Data from E Africa FRN-NGO project

Graph designs by Ric Coe and Sam Dumble

# PVS of bean varieties

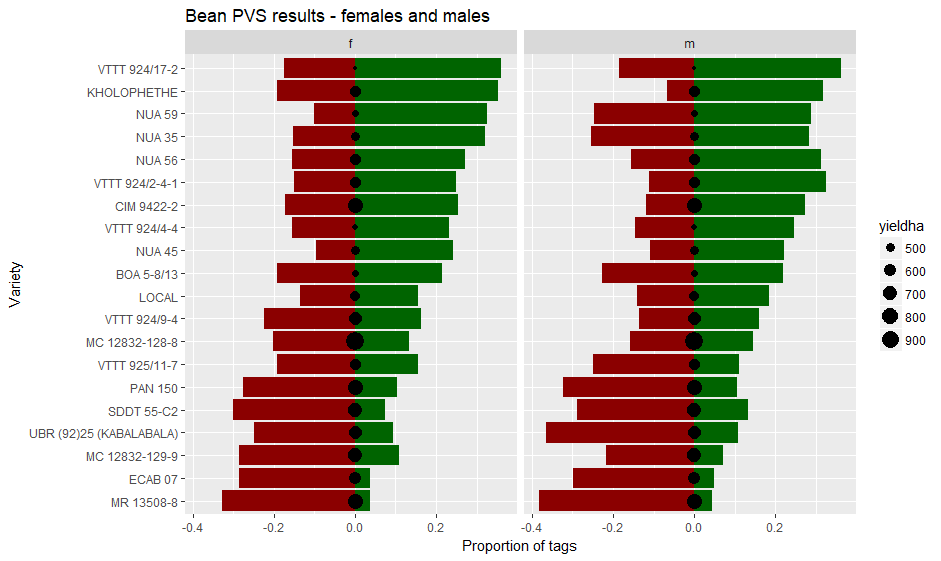


Figure 6

## Data

Participatory variety selection (PVS) exercises were carried out to collection farmers' views on bean varieties. The experiments are carried out in several locations and seasons. A variable number of female and male farmers were invited each time to assess the varieties. Each farmer was given 4 green ("good", "positive", "I like it") tags and 4 red ("poor", "negative", "I don't like it") tags to allocated to varieties but putting 0 or 1 tag on each. Mean yield of each variety was also measured.

## Graphical aims

The main aim is to compare varieties. We want to compare them in terms of farmers rating them positively or negatively. Farmers did not agree, so each variety had both positive and negative tags and that needs displaying as they do not necessarily 'cancel out' – some negative tags tell us that some farmers have reservations about a variety, even if many farmers think it good. The farmer choices do not always correspond to yield – they use other criteria as well – so we wanted to show the correspondence between yield and farm opinion. The numbers of farmers present at each assessment varied hence we express 'number of tags' as the proportion of the maximum number that could have been allocated.

## Principles being used

* Alternatives are easiest to compare when on top of each other, rather than beside each other. Hence varieties are on the y-axis. Male and female are compared side by side we want to see compare the overall patterns of variety differences for each group.
* 'Maintain the visual metaphor'. Hence we use red bars for red (negative) tags, green bars for green tags and show them counterpoised or balancing each other . Circles for yields have areas (not diameters) proportional to the data value
* Make order meaningful. The default order for varieties is alphabetical but that does not convey any information. Hence we make the order correspond to the overall mean proportion of red tags. Deviations from the expected patterns then stand out.

## Messages in the picture

* There are positive and negative assessments for every variety so no overall agreement.
* In general, varieties with many green tags have fewer red tags. But there are exceptions – NUA 45 for example. So farmers are probably considering different properties when allocating red and green tags.
* There is low correspondence between assessment indicated by the tags and yield. Yield seems not to be particularly important in determining farmer assessment.
* The overall patterns of assessments by men and women are very similar. However a few varieties show distinct differences. For example, many women are negative about KHOLOPHETHE with fewer giving negative tags to NUA 59. This pattern is reversed for men.

## Acknowledgements

Data from the Southern Africa Seed Systems project.

Graph design by Ric Coe

# Farmers' experiment on striga management

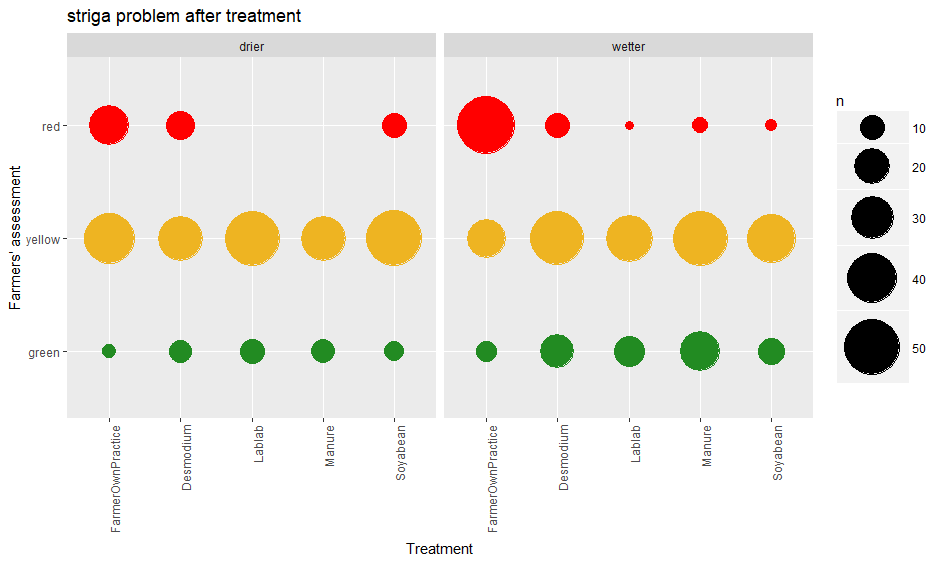


Figure 7

## Data

The data are from a trial done by farmers in a farmer research network. About 135 farmers tried 5 alternative ways of managing striga in their crop field, one of this was their usual or current practice. At the end of the first season they assessed the alternatives by giving each a 'traffic light' colour code with green for 'good' or 'go ahead', yellow for 'not sure' or 'caution' and red for 'a problem' or 'stop'.

## Graphical aims

When data are recorded on a categorical scale such as this red/yellow/green scale we want to avoid turning them into numbers and that implies we cannot use means. Hence we need to be able to compare frequencies of each category. The most important comparison for farmers is between alternative treatments so these need to be clear in the graph. Researchers also want to contexts. In the example sites classified as wetter and drier are compared. The design does not have the same number of farmers in every group so we want to show numbers not just proportions [debatable!]

## Principles being used

* 'Maintain the visual metaphor'. The image of the traffic light is used in data collection so the same concept is used in display.
* Make the most important comparisons the ones easiest to make, so that treatments are adjacent and the baseline (Farmer practice) treatment is first.

## Messages in the picture

* The alternatives tend to be rated better than farmer practice, but there are no options that are given 'green' by most farmers. There is a lot of caution about all options.
* Manure and lablab are the best rated.
* More farmers in wetter areas than in drier areas see the new options as an improvement over current practice.

## Acknowledgements

Data from E Africa FRN-NGO project

Graph design by Ric Coe

# Plant – insect interactions

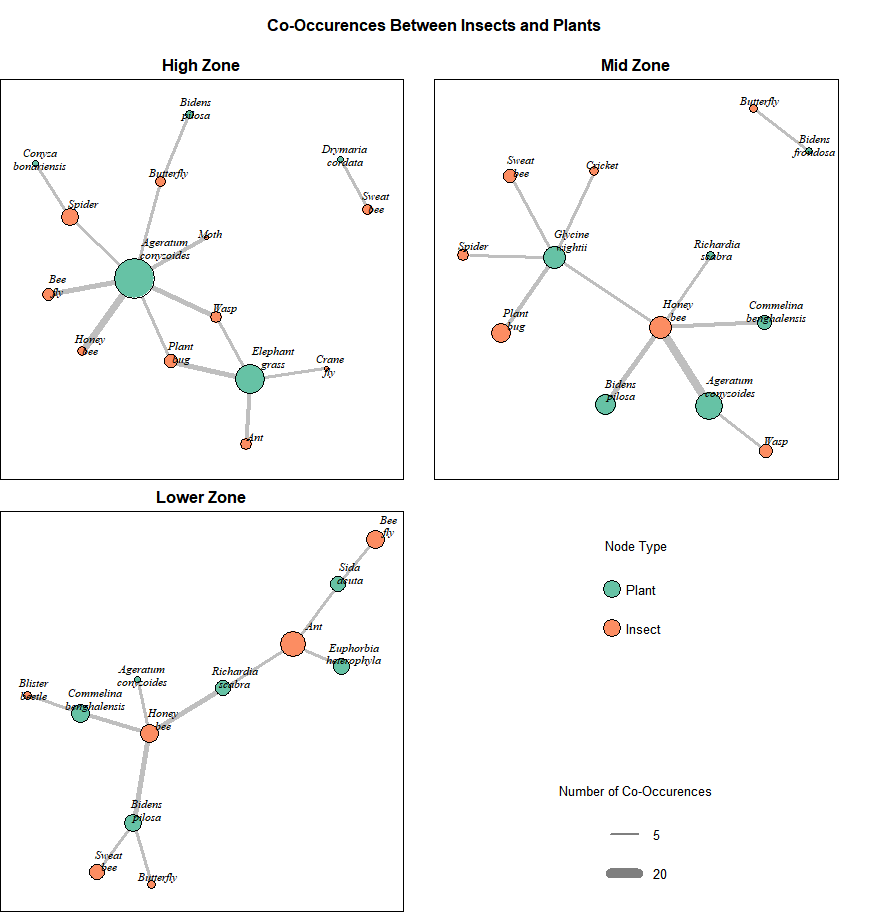


Figure 8

## Data

Observational data collected across three different agro-ecological zones (primarily distinguished by high, medium and low altitude) from 8 different sites within each zone for each relationship observed between an insect and a plant within a randomly selected transect walking from the middle of the site out to the edge.

## Graphical aims

## Visualise which insects are commonly found with which plants, and how these relationships vary in the different agro-ecological zones, to be able to understand pest pressures on certain plants. The next step could be to classify insects as beneficial/pests and plants as weeds/wanted and add this information to the display.

## Principles being used

* High data density (a lot of information per unit area of graph) gives rich displays.
* Make order/position meaningful. These graphs were produced using a ‘force-directed’ graphical algorithm, which arranges the points such that the relationships can be visualised clearly and that the structure of the network can be clearly visualised. An example of how this graph would look without a force-directed algorithm is seen in Appendix 2 as this is the more common default.
* Using size and colour as graphical dimensions to provide extra information. The sizes of the circles illustrate relative frequencies of which the insects/plants were found, and the colours illustrate whether a node refers to a plant or to an insect.
* Small multiples (repeated graphs of identical design) reveal similarities and differences between patterns across many cases or contexts.

## Messages in the picture

Patterns varied substantially across the three zones, and in all cases the complete networks between insects and plants were complex, and so easier to communicate visually than through using large amounts of text or numbers.

In the high zone the *Ageratum conyzoides*, a weed, acted as a ‘hub’ – attracting many different insects. For the middle zone, although there was still a large amount of *Ageratum conyzoides* within the site, it did not act as a ‘hub’ in the same way – instead it was the *Glycine vighetti* crop which was attractive to the largest number of different insects. There was no clear ‘hub’ plant within the lower zone.

## Acknowledgements

Data from Southern Africa Botanicals project

Graph design by Sam Dumble

# Disease resistant bean variety selection

A screenshot of a cell phone

Description generated with very high confidence

Figure 9: horizontal line shows the progressive resistance of a beanline to diseases from left to right. A beanline is resistant if the disease score is between 1 and 4, 1 being highly resistant.

## Data

The data is from an on-station trial of 97 beanlines tested for resistance against four diseases; BCMV, ALS, CBB and Rust. A beanline is selected for further crossing if it is resistant to all the four diseases.

## Graphical aims

We want to show how each of the 73 beanlines resistant to BCMV progress in their resistance to three other diseases, and their survival to final selection. The usual approach would be to use a table with conditional formatting displaying the pattern of disease scores for each beanline, see Figure 2 for an example. That however would show the overall variation of the beanlines’ resistance with no demarcations for which diseases the beanlines are eligible for selection. This makes it tricky to display the selection process in one table. With the graph in Figure 9, we see a pattern of the disease scores for each of the 73 beanlines. A beanline’s resistance is set for a score of 1 to 4 and the selection process is represented by a continuous horizontal line. A beanline having a disease score above 4 is termed as susceptible to the respective disease, hence horizontal line will not progress in the sequence. This beanline will ultimately not qualify for overall selection.

## Principles being used

* Make order matter. In this case, beanlines are ordered by level of resistant to BCMV. The diseases in order of importance by commonality, severity to beans, and researcher interests.
* Closely integrate the statistical and verbal descriptions of the dataset. In this graph, we integrate the descriptions of the selection criteria and the meaning behind the horizontal lines. Without it the graph is meaningless.

## Messages in the picture

* There is not always a one-size-fits-all approach, a trade-off is to be made. This is shown in the selection of the beanlines in the graph, where some of the most resistant beanlines to BCMV do not qualify as resistant beanlines by all diseases.

## Acknowledgements

Data from Southern Africa Bean Bruchids project

Graph design by Nuru Kipato

# Finger millet yield and farmers' ranks

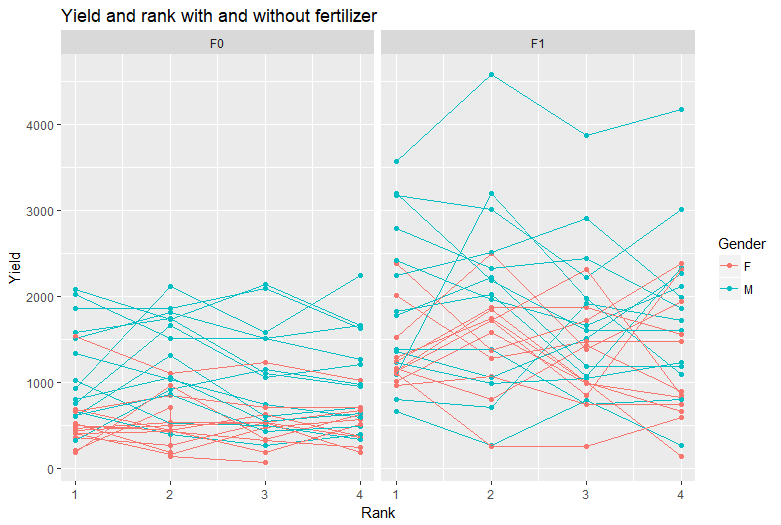


Figure 10

## Data

The data are from trials done by farmers comparing different finger millet varieties. Each farmer grew 4 varieties with and without fertilizer. The trials were done in several years and locations, with different farmers each time. At crop maturity the grain yield was measured. Farmers were also asked to rank the varieties ( 1=best).

## Graphical aims

The aim of the graph is to display the relationship (if any) between rank and yield, and the way this varies across farmers and other conditions.

Analysis and display of ranks has take their nature into account. Each farmer ranks the varieties on their farm, so we can compare those to yields on their farm. But different farmers sometimes have different sets of varieties, so comparing or averaging ranks across farms can be misleading.

If farmers are ranking based on yield then we would expect the variety with rank 1 to have highest yield, rank 2 next highest, and so on. Hence for one farmer, a graph of yield again rank would show a line that decreases from left to right.

## Principles being used

* Show the variation between farmers
* Distinguish groups that might be explain variation
* Use lines to connect points that are logically connected in an order, even if the relationship is discontinuous. A rule is sometimes given that data point should only be connected by a line if intermediate positions along the line represent something real. This is too restrictive.

## Messages in the picture

* The lines for each farmer do not generally show a decrease from rank 1 to rank 4. Farmers are not using yield (or not only yield) to determine rank.
* This is the same for plots with and without fertilizer and for men and women farmers.
* On many farms the yield of different varieties is quite similar. When that happens it is not surprising that farmers rank them on other characteristics.
* There is a lot of farm to farm variation in yield.
* Women generally have lower yields than men.

## Acknowledgements

Data from East Africa Finger Millet project

Graph design by Shiphar Mulumba

# Effectiveness of research stage discussions in attaining multiple workshop objectives

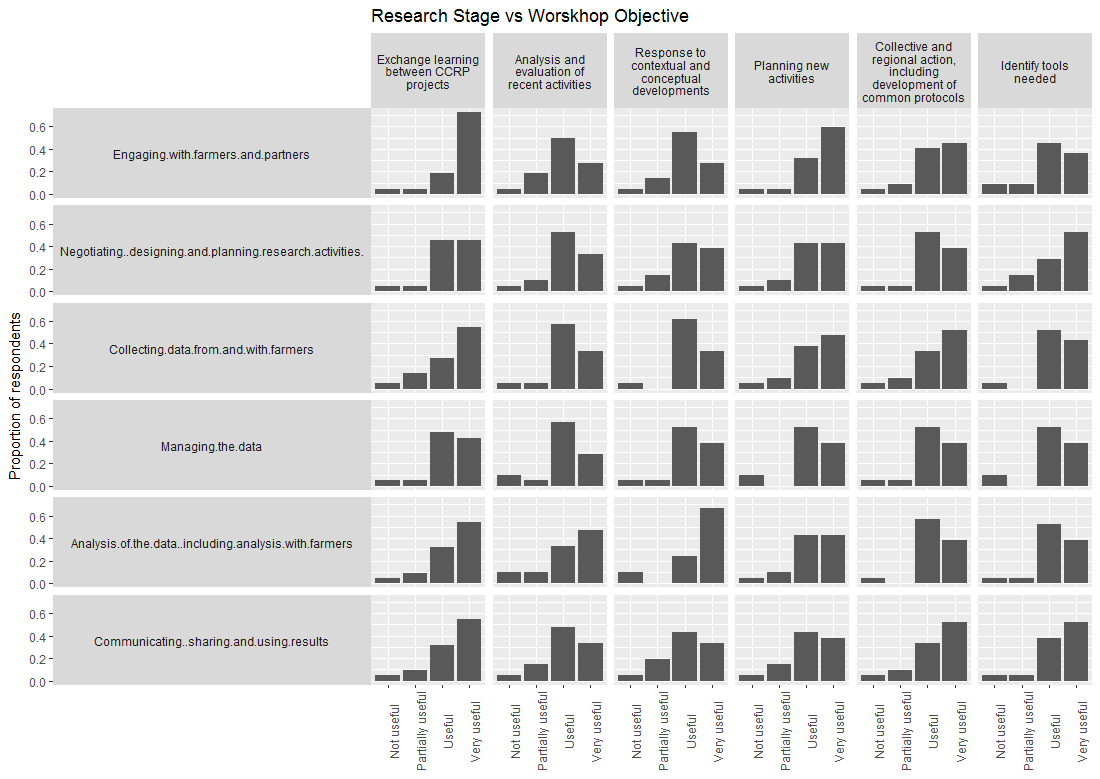


Figure 11:

## Data

The data is from a research methods workshop evaluation survey. Participants were required to rate each research stage discussions against every set objective of the workshop from a large matrix.

## Graphical aims

This graph is an example of a dense presentation of a lot of data.

## Principles being used

* Maintaining the metaphor – there was a large matrix (research stage x workshop objective) in the survey, we have maintained the same matrix in the display of results.

## Messages in the picture

## Acknowledgements

Data from Research Methods Support project

Graph design by Ric Coe and Nuru Kipato

# Appendix 1: Code to generate examples

[Not sure how much to provide. If we give the whole program we really need to provide a datafile as well and that requires permissions]

## Figure 1

Data in a spreadsheet where summarised and laid out as a table using Pivot Table in Excel. The layout was then modified in Word.

## Figure 2

The table was data were arranged as a table using Pivot Table in Excel and coloured using Conditional Format. The pivot table was copied to another sheet so it could be modified, such as inserting the Site column and colouring it.

## Figure 3

library(ggplot2)

ggplot(d, aes(x=pcirc, y=pc2))+

geom\_point()+

geom\_smooth(se=F)+

facet\_wrap(~trt)+

xlab("plant circumference")+ylab("percent severity")

## Figure 4

[how much of the data processing to show?]

library(ggplot2)

ggplot(data=meandat3,aes(x=altitude.median,y=100\*r.mean, group=plantsorghum))+

geom\_point(aes(fill=county,size=r.length),alpha=0.5,pch=21)+

geom\_smooth(se = FALSE,aes(weight=sqrt(r.length)))+facet\_grid(.~plantsorghum)+

xlab("Median Altitude")+

ylab("% Damage on Maize is Worse")+

ggtitle("Altitude vs. Relative Crop Damage: Ward Level")+

scale\_size\_continuous(name="Number of Obs")

## Figure 5

library(ggplot2)

ggplot(data=meandat2,aes(x=altitude.median,y=100\*r.mean))+

geom\_point(aes(fill=county,size=r.length),alpha=0.5,pch=21)+

geom\_line(aes(y=100\* fit,x=az),data=preddata1,inherit.aes = FALSE)+

geom\_line(aes(y=100\* (fit+1.96\*se.fit),x=az),data=preddata1,inherit.aes = FALSE,col=2,linetype=2)+

geom\_line(aes(y=100\* (fit-1.96\*se.fit),x=az),data=preddata1,inherit.aes = FALSE,col=2,linetype=2)+

ggtitle("Altitude by Planting Time Effect from Logistic Regression")+

xlab("Median Altitude")+ylab("% Damage on Maize is Worse")+

facet\_wrap(~plantsorghum)+

scale\_size\_continuous(name="Number of Obs")

## Figure 6

Drawn using R. The critical lines of code are:

library(ggplot2)

library(forcats)

ggplot(data=(subset(d, siteloc!="South")), aes(x=fct\_reorder(varname, plusprop,fun="mean", na.rm=TRUE)))+

stat\_summary(aes(y=plusprop), fun.y="mean", geom="bar", fill="darkgreen")+

stat\_summary(aes(y=-minusprop), fun.y="mean", geom="bar", fill="darkred")+

geom\_point(data=mean, aes(x=varname, y=0, size=yieldha))+

facet\_grid(.~gender)+

ggtitle("Bean PVS results - females and males")+

xlab("Variety")+

ylab("Proportion of tags")+

coord\_flip()

[Comments with explanations needed?]

## Figure 7

library(ggplot2)

ggplot(all, aes(x=relevel(Strigatreatment\_selection, ref="FarmerOwnPractice"), y=strigapest))+

geom\_count(aes(colour=strigapest))+

facet\_grid(.~climate)+

theme(axis.text.x=element\_text(angle=90,hjust=1))+

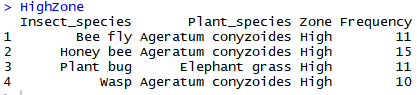
scale\_colour\_manual(values=c("green"="forestgreen","yellow"="goldenrod2","red"="red"), guide=F)+

scale\_size\_area(max\_size = 20)+

ggtitle("striga problem after treatment")+xlab("Treatment")+ylab("Farmers' assessment")

## Figure 8

Drawn using R, with the igraph library. The key step is in processing the data into the required format before trying to produce the graph.



The data should be one row for each relationship being plotted, with a column indicating the number of times that relationship was observed. The names of the columns do not matter, but the first column should be one type of node (insect in this case) and the second column should be the other type of node (plants).

A simple network graph can then be produced using the following lines, using the Fructerman Reingold algorithm, for arranging the nodes:

library(igraph)

Network\_High <- graph\_from\_data\_frame(HighZone,directed=FALSE)

plot(Network\_High, layout=layout.fruchterman.reingold)

There are a lot of customisation options for the appearances of the igraph – which work completely differently to ggplot2, and other R graph libraries.

But many of these can be added without explicitly writing R syntax, through a point-and-click editor which can be loaded using:

tkplot(Network\_High)

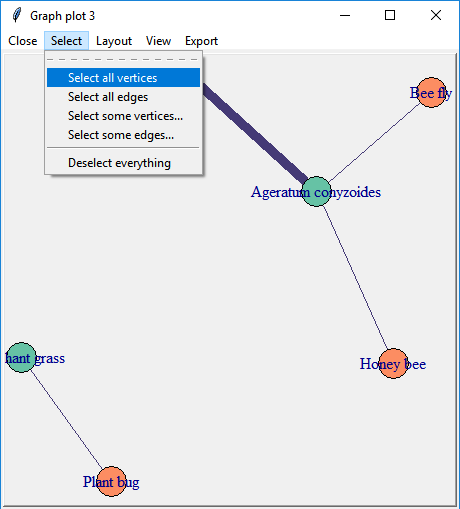


Figure 9:.

# Make sure that data is sorted by Beanline and Disease

setkey(beans.melt, Beanline, Disease)

# Is status between diseases same (`SAME == 0`)

beans.melt[, SAME := c(0, diff(grepl("\_valid", beanline\_status2))), Beanline]

#re-sort the data by disease and disease\_score

setkey(beans.melt, Disease, disease\_score)

library(ggplot2)

library(forcats)

ggplot(beans.melt, aes(x=fct\_inorder(Beanline), Disease)) +

geom\_point(aes(size = disease\_score, fill=ifelse(disease\_score<=4,"<= 4",">4")),alpha=0.5,pch=21)+

geom\_line(aes(y = ifelse(SAME == -1| SAME == 1, NA, Disease))) +

scale\_size\_continuous(name="Disease score")+

labs(x="Beanline", y="Disease", fill="Severity of disease")+

scale\_fill\_manual(values=c("black","red"))+

ggtitle("Beanline selection by sequential diseases resistance ")+

coord\_flip()

## Figure 10

ggplot(subset(fmlong,(Year==2007)|(Year==2013)), aes(x=Rank, y=Yield, group=Farmer)) +

geom\_point(aes(colour=Gender))+

geom\_line(aes(colour=Gender))+

facet\_grid(.~Fertilizer)+

ggtitle("Yield and rank with and without fertilizer")

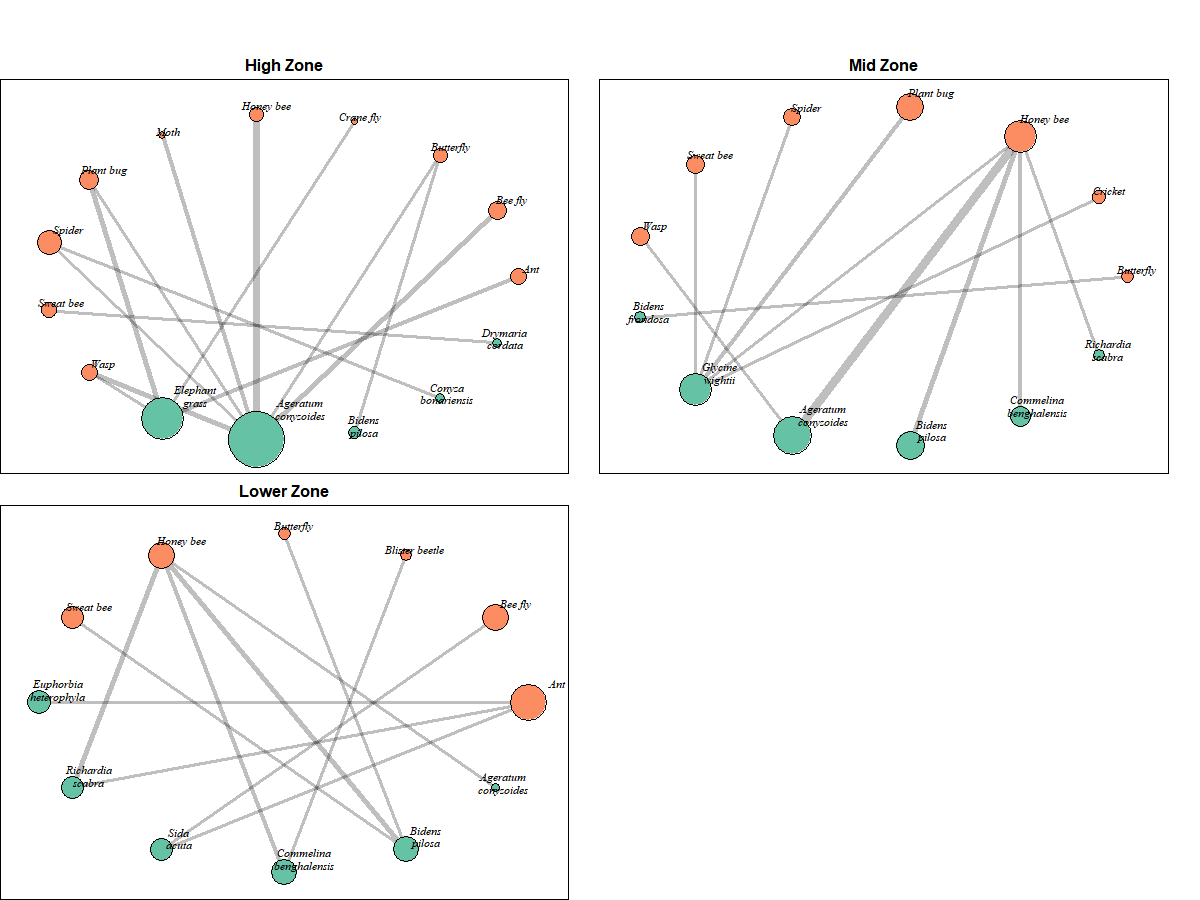
## Figure 11

ggplot(rms2, aes(x=Response))+geom\_bar(aes(group=factor(0), y=..prop..))+ facet\_grid(Research\_stage~ Objectives, switch="y", labeller=labeller(Research\_stage=label\_wrap\_gen(10) ,Objectives=label\_wrap\_gen(20)))+

theme(axis.line=element\_blank(), axis.text.x=element\_text(angle=90),axis.title.x=element\_blank(), strip.text.y = element\_text(angle=180))+

labs(y="Proportion of respondents", title="Research Stage vs Worskhop Objective")

# Appendix 2: Alternative for plant-insect interactions

Example of the same networks plotted in Figure 8, but using a standard graphical algorithm, rather than a force-directed graphical algorithm.

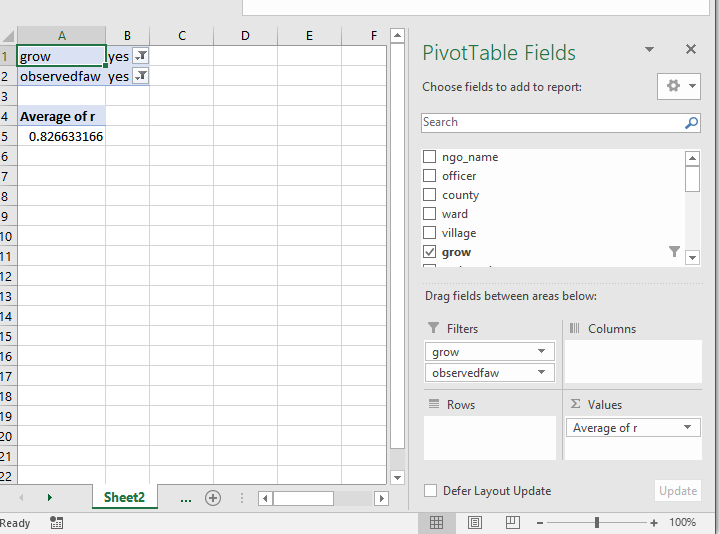
# Appendix 3: Building up to a final product

## Data

A rapid survey was done of nearly 1200 farmers growing both maize and sorghum during an outbreak of fall army worm (FAW). One response measured was whether the FAW damage was worse on maize than on sorghum. Context variables were measured in the hope they would explain variation. This is the same data presented in Figure 4.

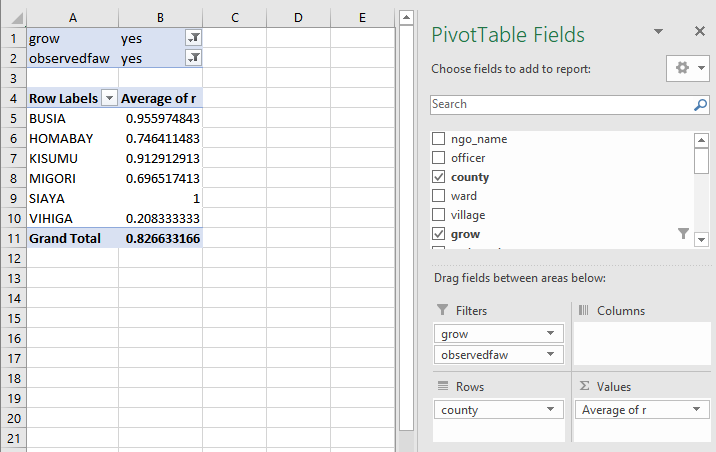
## Process

We want to show the important patterns discovered within our data. To help discovery of these trends we might possibly start with a pivot table in Excel.



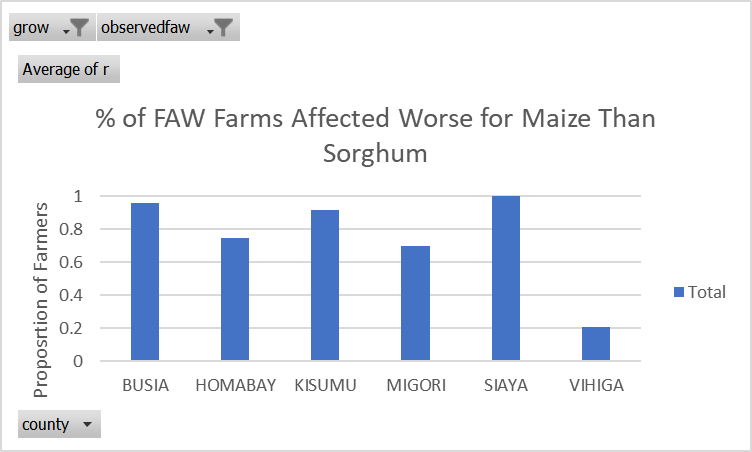
For this dataset we are only interested in farmers who grow both maize and sorghum (the column “grow” in the dataset) and farmers who observed the fall army worm (the column “observedfaw”). So we use these two fields in the filter section of the pivot table. The column “r” contains the outcome variable, which takes value 1 if maize is worse affected than sorghum and value 0 if not. Calculating an average of a 0/1 variable is the same as calculating the proportion of the data where the variable is equal to 1. So we can see that overall 83% of affected farms had been worse affected for maize than for sorghum.

Let’s start by considering if this varies by county:

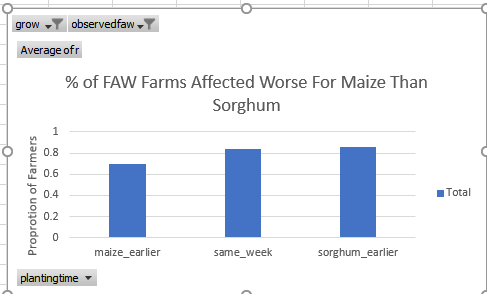


This is indeed a big factor – 100% of farmers in Siaya are worse affected on Maize, but only 21% of farmers in Vihiga. Numbers can be boring & it’s easy in excel to produce a quick and informative bar chart from these numbers using the PivotChart option.

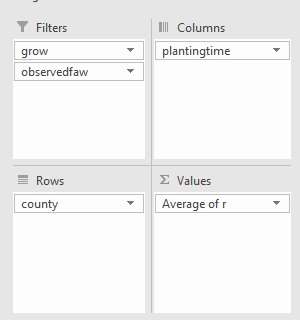




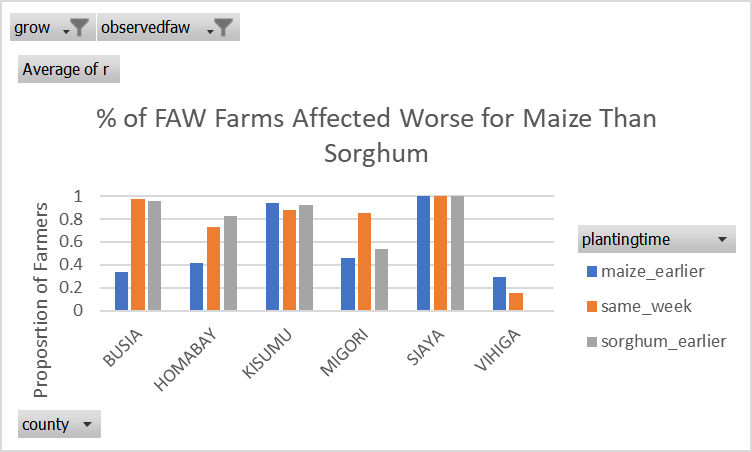
Another variable that was thought to be of interest was time of planting:



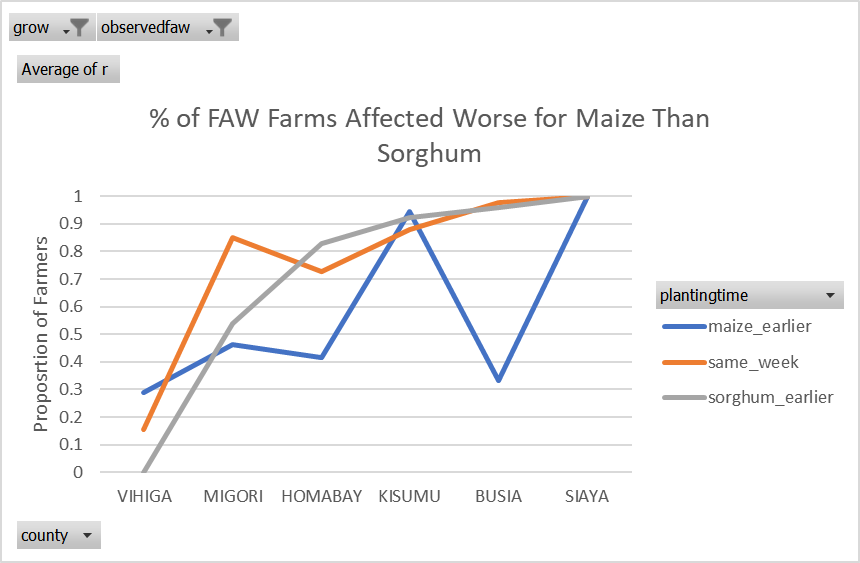
There looks like a small difference there, but not quite so obvious. Perhaps instead we want to see how county and planting time interact:



A bar chart is not the best way of visualising this as it may start to become confusing to interpret all of the different bars:



The trends will appear more clearly using lines instead of bars, and ordering the counties by severity of the response:

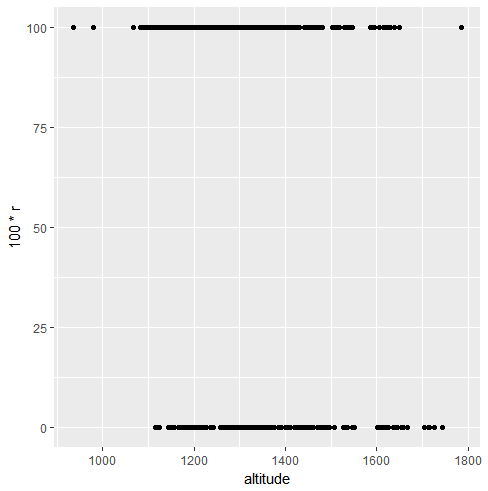


We can now see clearly that in Homabay and Busia the planting time seems to be more important, but in Kisumu and Siaya it does not appear to be important. We want to start trying to understand why that might be the case, by characterising what the ‘real’ driving factor behind the large differences across districts might be. What environmental conditions vary across the districts, and might explain why Busia and Homabay have similar results to each other, but different results to Kisumu and Siaya.

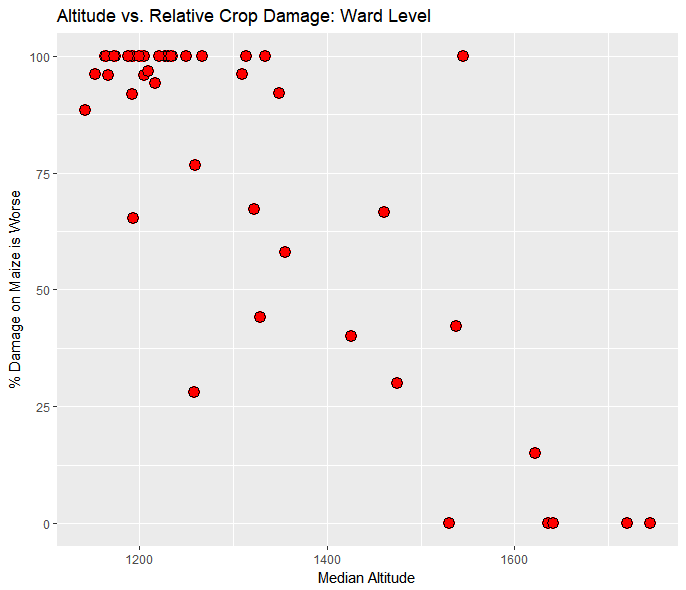
Altitude might be one such environmental factor.

At this point we are reaching the limit of what might be possible using excel – we should start thinking of using a more specialist software package instead.

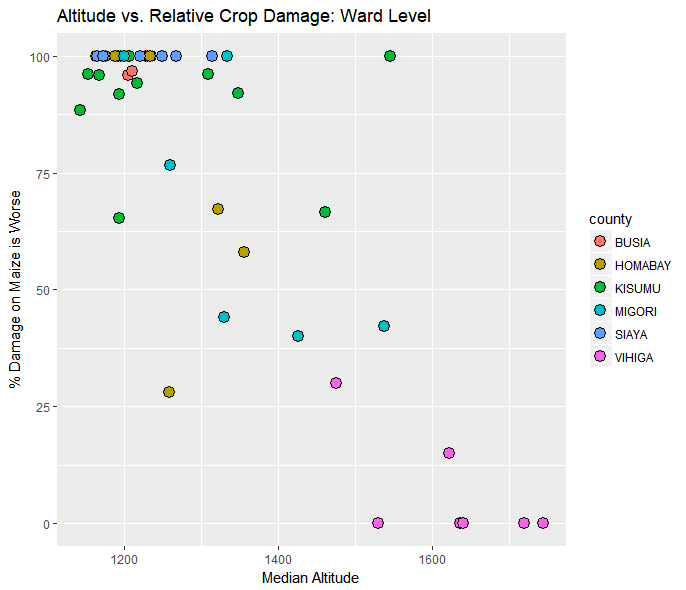
But we need to be careful with how we deal with the altitude variable.



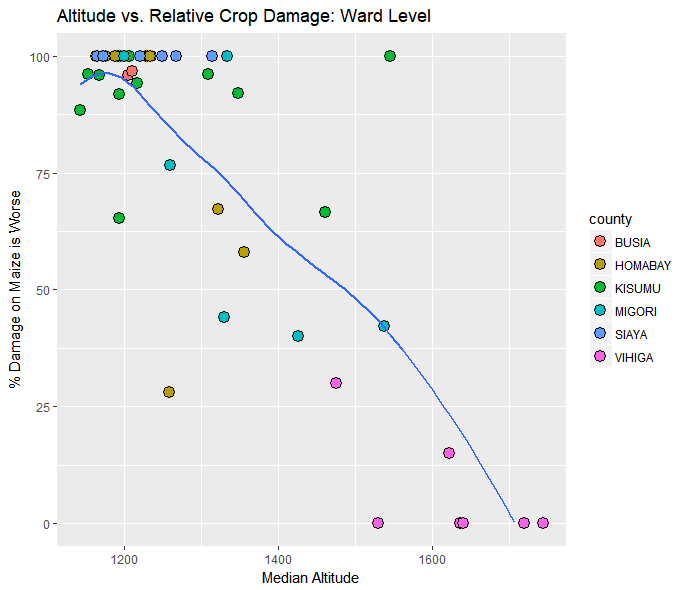
This is not a very useful graph, plotting the altitude of each farm against the outcome variable. However, we can take advantage of the fact that we have a clustered sample. The data contains farms clustered by ward. So we can calculate the average altitude of each ward and the percentage of farmers with maize worse affected in each ward and plot these two variables against each other.



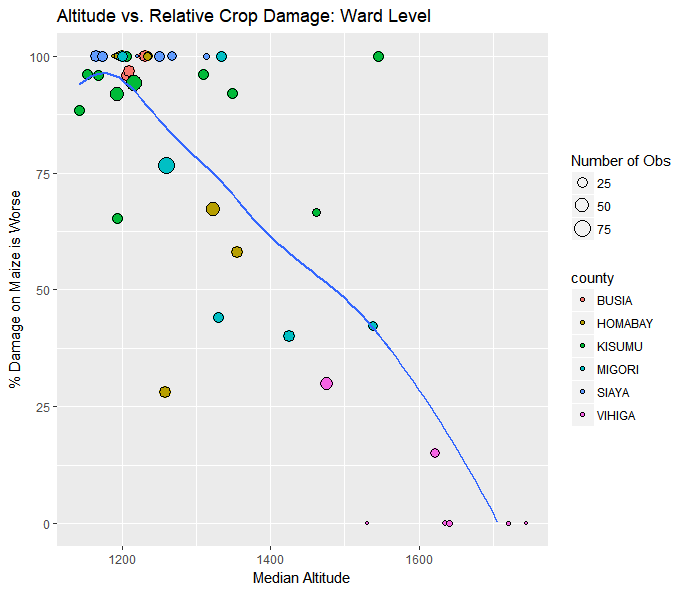
A fairly overwhelming trend! We know that each ward is entirely within one of our counties, so we can maybe set different coloured points for each county.



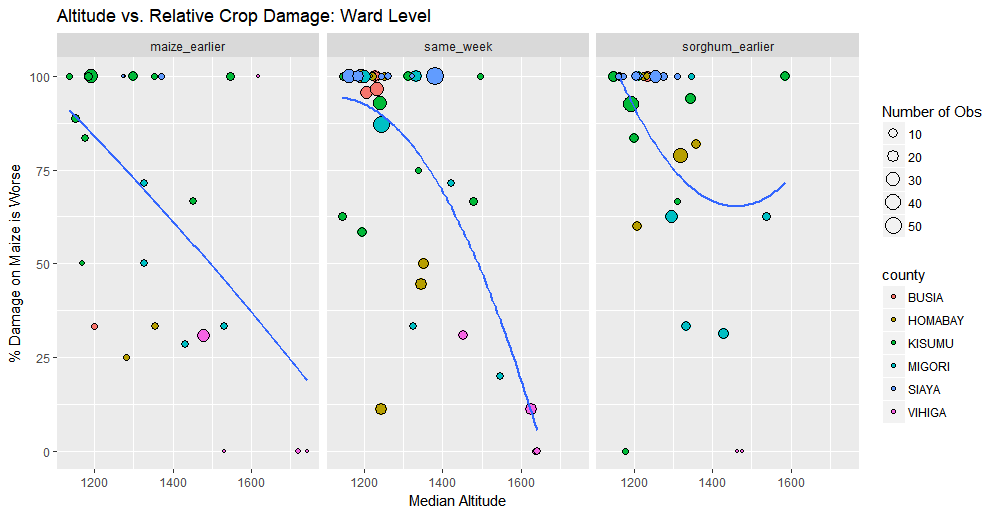
Now the county effect, particular for Vihiga is becoming clear. The counties with lower altitudes are having maize be worse affected and the high altitude counties are not. We can draw on a trend line to help guide the viewer:



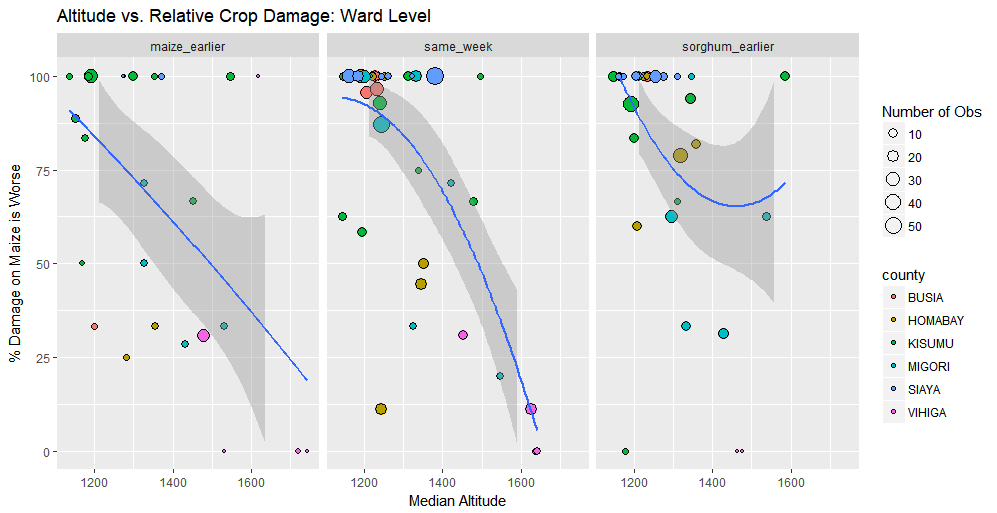
But there is still a misleading aspect of this graph; not all of our counties are equally represented. In some counties we have over 75 farmers, in other counties we only have 10. So we can set different size points proportional to the sample size to make the better represented counties more obvious.



In all this excitement we have forgotten about the planting time:



It looks like the altitude effects are strong where maize was planted earlier or at the same time as sorghum. But the trend line is not so clear for where sorghum was planted earlier. We can put on a margin of error around this trend line



And this further shows us that for where sorghum was planted earlier the trend is never consistently increasing or decreasing.