3.5 Reasons to Switch from Excel to R

Patrick Burns
http://www.burns-stat.com

March 2009

The title speaks of “Excel”. I really mean “spreadsheet” but Excel is the one that is almost universally used.

The subtitle is: Preaching to the choir with an ulterior motive.

This talk was given 2009 March 31 at the initial meeting of the London R User’s group.

Outline

- Similarity
- Reasons to Switch
- Why Not R?
- My Ulterior Motive

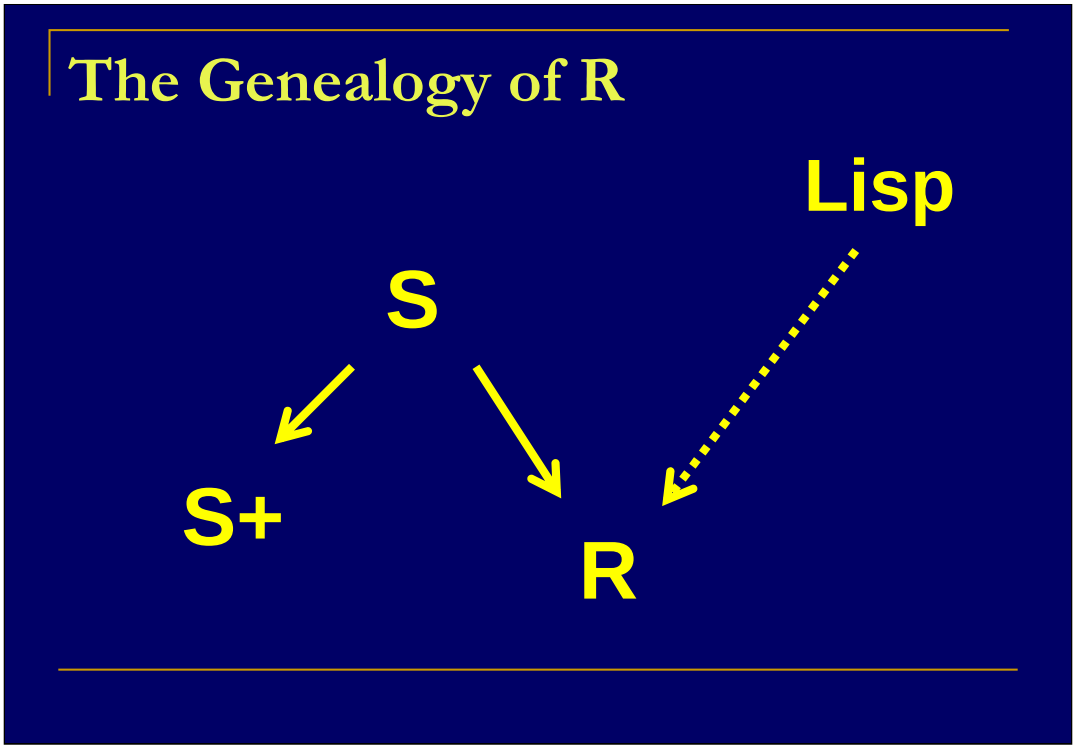
Applications of Spreadsheets

- Data storage
- Data manipulation
- Graphics
- Simple analyses
- Not-so-simple analyses

This list of what is done in spreadsheets looks very much like what I do in R. Hence it makes sense to think of R as a substitute for spreadsheets.

The color of the last item is an indication that this is where I think there are serious problems. This talk only touches on that issue. "Spreadsheet Addiction" has more of my thoughts on the topic.

http://www.burns-stat.com/pages/Tutor/spreadsheet_addiction.html



The father of R is Lisp and the mother of R is S.

Some people think that “S” stands for statistics. That is wrong. John Chambers is very clear that S did not stand for statistics. John says that there were a number of suggestions for names, none of the names resonated, but the intersection of all the suggestions was “S”.

The R Remit

- **"S" does not stand for statistics**
- **Data analysis and graphics**
- **R is doing okay**

While R comes out of the statistics community, it should not be thought of as a statistics package. It's remit is – and should be – much broader than that.

R is about data analysis in a very wide sense, and graphics.

A lot of people will think that R’s glass is half full:

• There are *x* hundreds of packages on CRAN, where *x* is a fairly rapidly increasing number.

• SAS and SPSS are adjusting their strategies because of R

• …

I’m an optimist.

I think R’s glass is about 99.9% empty.

The **VAST** majority of people who could benefit from R are not using R.

Outline

- Similarity
- Reasons to Switch
- Why Not R?
- My Ulterior Motive

There are many reasons to make the switch from spreadsheets to R. But for me there is only one elephant in the room. The elephant is safety.

African savannah elephant. Picture copyright Jean Ryder.

Reason 1: Safety

- **This is the best reason**

An Image of Computing

- **Data = Water**
- **Functionality = Earth**

The homework assignment was to add fire and air into the analogy. No reports of anyone having completed the assignment.

In R we bake functionality into functions.

In R there is a clear separation between data and functionality.

In spreadsheets data and functionality are mixed together.

This mixing is the great strength of spreadsheets.

The mixing gives spreadsheets immediacy.

Immediacy is the source of the great popularity of spreadsheets.

But this mixing is also a great weakness.

The results can be explosive.

Picture from USGS.

Development Cycle: R

- **Write function**
- **Debug function**
- **Use function**
- **Fix bugs as they periodically appear**

Ideally writing and debugging a function should be done in parallel rather than sequentially. I wasn't clever enough to show that on the slide.

Development Cycle: Spreadsheet

- **Write spreadsheet**
- **Debug spreadsheet**
- **Populate spreadsheet**
- **Debug this instance of spreadsheet**

Spreadsheets need to be debugged every time they are used.

They are usually **NOT** debugged every time they are used.

Time Series of Bugs: R

- Number of bugs hardly ever increases
- Bounded below by zero
- Implies convergence
- We have hope for convergence at zero

Time Series of Bugs: Spreadsheet

- **Sort of like Brownian Motion**

Some uses of the spreadsheet will increase the number of bugs, some uses will decrease the number of bugs.

Reason 2: Speed

- **Spreadsheet took most of the night, often fell over**
- **Transferred to S (for safety)**
- **Took a few minutes**

We usually think of running away from R to get speed. But spreadsheets can be very slow.

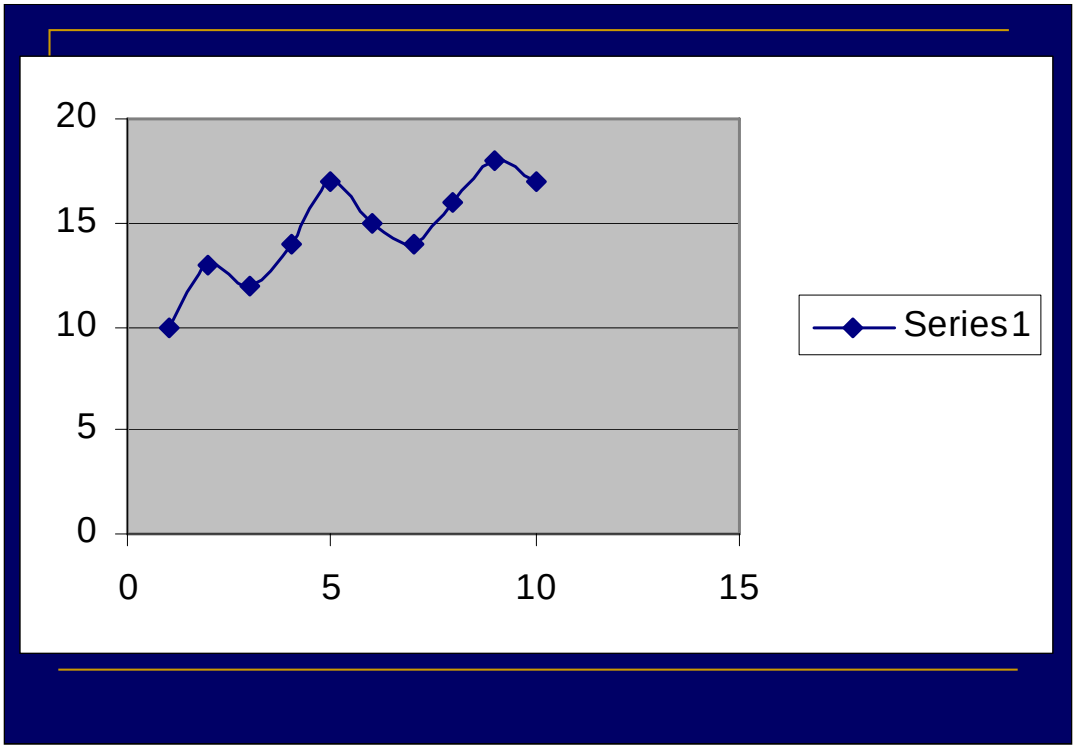
I was looking for expertise from the audience on this point as my personal sample size is 1. But my hopes were dashed.

I'm not sure of the inherent relative speeds, but certainly one problem with spreadsheets is that the same calculation can be carried out multiple times.

Trying for efficient as well as safe computation in spreadsheets seems like an overwhelming demand.

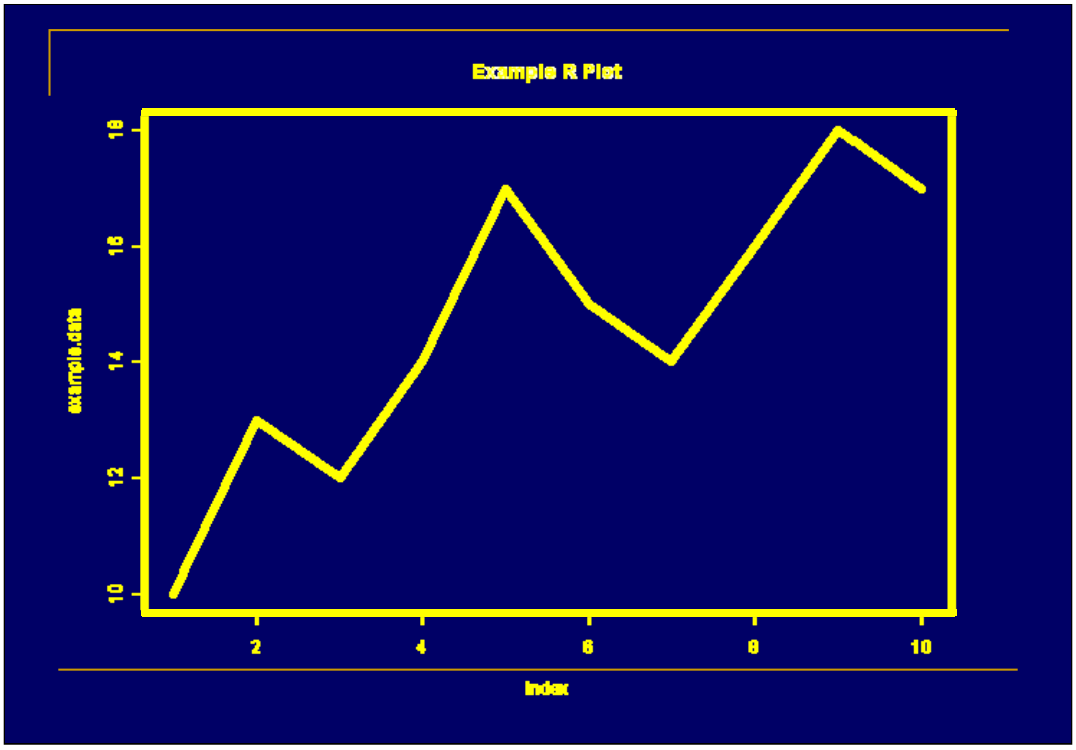
Reason 3: Graphics

- No reason that spreadsheets should have poor graphics
- But good graphics haven't happened



It continuously amazes me to witness an otherwise polished presentation and up pops a slide like this.

For a bit of an investment there could have been a graph like the next slide that looks like it belongs.



The R function that produced this graphic is:

```
function (file = "graphexamp.png")
{
   if (length(file)) {
      png(file = file, bg = "transparent", width = 650)
   }
   par(col = "yellow", col.lab="yellow", col.axis="yellow", lwd = 5)
   plot(example.data, type = "l", axes=FALSE)
   axis(1, col="yellow", cex=3)
   axis(2, col="yellow", cex=3)
   box()
   title(main="Example R Plot", col.main="yellow")
   if (length(file)) {
      dev.off()
   }
}
```

Reason 3 and a half

- …

The function for the previous slide is very complicated relative to what you need for graphing while you are doing data analysis. However, it gives an indication of the fine control you have over presentation graphics if you need it.

Reason 3.5 is whatever you want it to be. Two particular reasons given by the audience were:

1) Spreadsheets are inherently two-dimensional and thus can be very limiting.
2) Somewhat the opposite: the layout of a spreadsheet need have no discipline at all (and hence there be dragons).

Outline

- Similarity
- Reasons to Switch
- Why Not R?
- My Ulterior Motive

People should immediately accept all this wisdom.

But perhaps that’s a bit quixotic.

Why Not? – A Bad Reason

- **Command line**
  - **Can write functions to overcome memory lapses**
  - **Can build custom menus**

A bad reason not to switch is that R is command line driven, not menu driven.

Menus are essentially just a memory storage device. In R it is possible to write your own functions that serve the purpose of memory storage.

It is also possible to build custom menus in R. This is a vastly superior technology for creating applications for the occasional user than a fixed menu system.

Even though I think this is a bad reason not to switch, that does not mean it is the majority view. A lot of Excel users will think this is an excellent reason not to switch. This is an issue that needs to be addressed, but I'm not sure of the best approach.

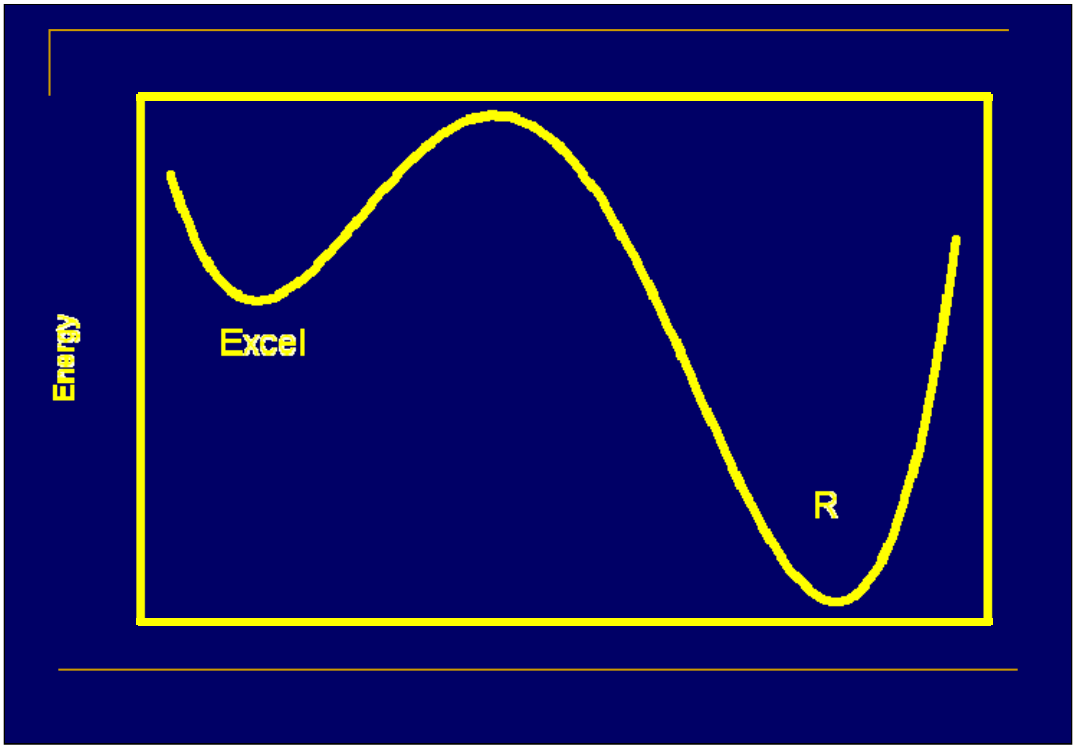
Why Not? – A Good Reason

- **Conversion cost**

The cost of conversion is a serious issue. The next slide is a sketch of the issue: A task takes less effort in R than in Excel, but effort needs to increase in order to make the switch.

The sketch shows a hill, but most Excel users will experience it as a big mountain.

Our task is to bulldoze that mountain.



The function to create this slide is:

```
function (file = "energy.png")
{
   if (length(file)) {
      png(file = file, bg = "transparent", width = 650)
   }
   par(col = "yellow", col.lab="yellow", lwd = 5, cex.lab=2)
   plot(function(x) x^4 - 3 * x^3 - 76 * x^2, -7, 10,
        axes = FALSE, xlab = "", ylab = "Energy")
   box()
   text(-5, -1100, "Excel", cex=2)
   text(7.2, -1900, "R", cex=2)
   if (length(file)) {
      dev.off()
   }
}
```

Reducing Transition Energy

- **Improve R usability**
  - **Probably little scope**
  - **Easier custom menus would help**
- **Improve R documentation**

There are two ideas for reducing the cost of conversion: Make R easier to use, and improve documentation.

I don't think there is much scope for making R easier to use (though trying is always a good thing to do). That is not because R is perfectly easy to use, it isn't. It is because we are locked into most of the rough spots because of backward compatibility issues.

There is a lot of room to improve custom menus, and that could be quite helpful.

Existing documentation can always be improved. We should also be experimenting with new forms that are aimed at satisfying particular needs.

Outline

- Similarity
- Reasons to Switch
- Why Not R?
- My Ulterior Motive

A question from the floor was: What about using packages that combine Excel and R? My answer was that that is sometimes a good solution, but often it is similar to the uncomfortable position of having one foot on shore and one foot on the boat.

On reflection I'd like to make a much stronger statement: Almost always it is better to go cold turkey. When you combine the two, you have the complications of Excel, the complications of R and the complications squared of the combination. Thinking that the combination will provide a security blanket is almost surely counterproductive.

The Ulterior Motive

- We need a book
- Possible title: “Switch from Excel to R”
- It ain’t me, babe

I think we need a book that gives specific directions of switching from Excel to R. Here I do mean Excel and not any spreadsheet – the instructions should be very specific.

I’m not going to write such a book because I don’t have the knowledge, or the time, to do it.

So my ulterior motive is to get others to write the book. It could change the history of computing, and it could result in a bit of money.

There was a suggestion on the night that I serve as an initial coordinator, and I’m happy to do that.