Reproducibility and R: Neotoma

Simon J. Goring
Williams Lab Meeting
University of Wisconsin - Madison
19 November 2013
Why reproducible code?

• You start projects and let them sit
  • You need to be able to come back & know what you’re doing
• Code is the most accurate representation of the analytic work you’ve done
• Publishing code acts as an incentive to improve your coding habits
• Reproducibility & open data improve your citation rate
Why reproducible code?

You start projects and let them sit

Why did I do this?
Why reproducible code?

**Code is the most accurate representation of the analytic work you’ve done**

- Compare:
  
  “We used a generalized additive model to predict rainfall from species composition (gam: package mgcv in R, Wood 2013).”

```r
gam(formula, family=gaussian(), data=list(), weights=NULL, subset=NULL, na.action,
offset=NULL, method="GCV.Cp", optimizer=c("outer","newton"), control=list(), scale=0,
select=FALSE, knots=NULL, sp=NULL, min.sp=NULL, H=NULL, gamma=1, fit=TRUE,
paraPen=NULL, G=NULL, in.out,...)

plant.models[[i]] <- gam(formula=paste(colnames(plants)[set[i]],
    '~ s(x, y, I(z*100))', sep=''),
knots = list(x = seq(551561, 1861700, by = 50000),
y = seq(364400, 1736700, by = 50000),
z = seq(0, 2700, length.out = 50)),
data=plants, family = poisson)
```
Why reproducible code?

Publishing code acts as an incentive to improve your coding habits

- Code has purpose & is clean & commented
- Removes superfluous code (trying things out)
- Makes sure that code outputs figures & tables you need
Why reproducible code?

Reproducibility & open data improve your citation rate

- Piwowar & Vision (2013) show a 9% citation improvement with open data, but may be up to 30% (time varying)
Code is data
So treat it like data
How to make reproducible code?

The Hierarchy of Reproducibility

**Good** Use an integrated development environment (IDE)

**Better** Use version control

**Best** Use embedded code
How to make reproducible code?

**Use an integrated development environment (IDE)**

*Keep your code in one place, let it do what it’s supposed to.*

- Rstudio (my preferred tool: [link](#))
- Tinn-R ([link](#))
- Eclipse with StatET ([link](#))
- Emacs ESS ([link](#))
- Jedit ([link](#))

Stop coding in the R console and saving the history!
How to make reproducible code?

Use version control

Help yourself keep track of changes, fix bugs and improve project management.

• git (distributed file content management system: [link](https://git-scm.com))
• subversion (centralized version control system: [link](https://subversion.apache.org))

My GitHub [account](https://github.com).
How to make reproducible code?

Use embedded code

Explicitly link code and text, save yourself time, save reviewers time, improve your code.

- Sweave/LaTeX (link)
- R Markdown (link)
Example Time

“Example time, it’s example time
it’s a sweet example makin’ you feel fine.”
Example time

![GitHub repository](http://github.com/SimonGoring/Neotoma-Workshop_Oct2013)

- IDE? RStudio
- Version control? git with GitHub
- Embedded code? RMarkdown with *knitr*
Example time

A live example happens here and is not reproduced in this presentation.

I will do it for you if you invite me somewhere nice (or somewhere not so nice, just invite me).

In the meantime, you can get started with RStudio, Git and GitHub using this worked example on downwithtime.