

# Using R and STATA for Data Analysis

Sept. 22, 2016

Glenn Wright  
University of Alaska Southeast

# Two Questions:

1. How does *decentralization* affect *deforestation*?
2. What causes *piracy*?

# Decentralization and Forests

Some arguments:

Decentralization → Deforestation

Decentralization → Forest Stabilization

A third variable: Community Engagement

# The Methods

- Propensity score/Mahalanobis matching techniques
- Regression techniques
- Simple tests (t-tests, etc.)

# Our tool: STATA

The screenshot displays the STATA 13.1 software interface. The main window is titled "Stata/SE 13.1 - Matching\_ForAnalysis\_13-07-18.dta". The interface is divided into several panes:

- Review Pane:** Shows the command window with the following text:

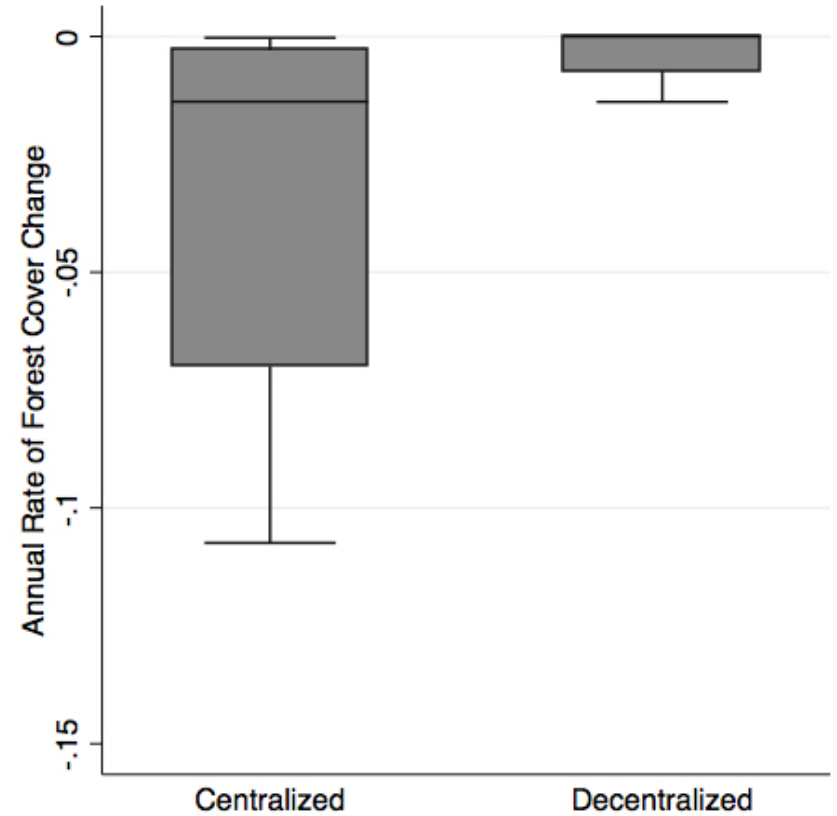
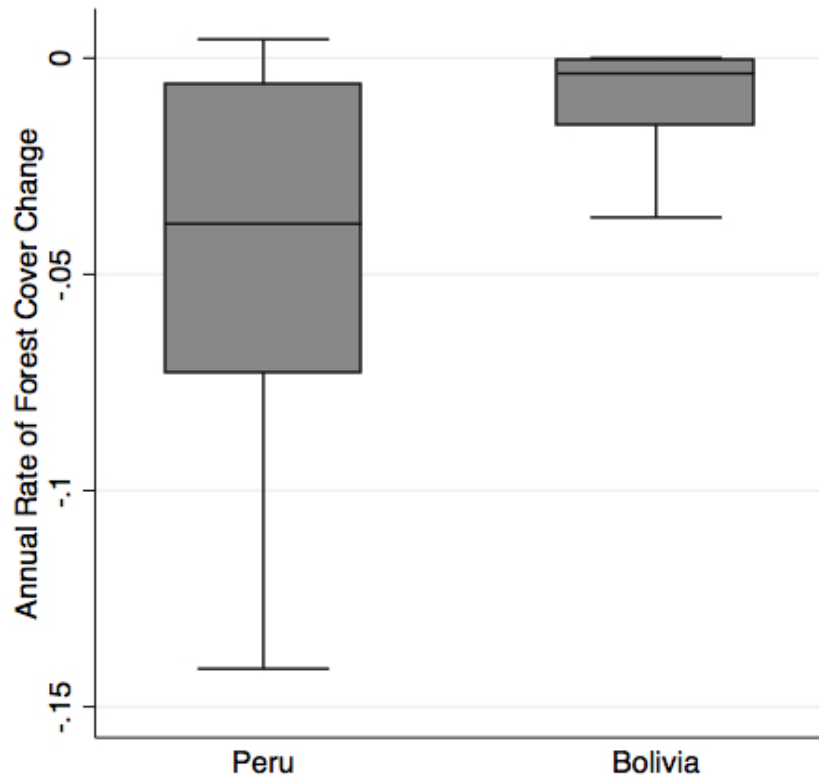
```
1 use "/U...  
...  
Statistics/Data Analysis  
Special Edition  
Single-user Stata perpetual license:  
Serial number: 401306004434  
Licensed to: Glenn Wright  
University of Alaska  
Notes:  
1. (-set maxvar-) 5000 maximum varia  
. use "/Users/glennwright/Google Drive/Matc  
.  
Command
```
- Results Pane:** Shows the command window output, including the Stata logo and the following text:

```
(R)  
...  
13.1 Copyr  
Stata  
4905  
Colle  
800-S  
979-6  
979-6  
Single-user Stata perpetual license:  
Serial number: 401306004434  
Licensed to: Glenn Wright  
University of Alaska  
Notes:  
1. (-set maxvar-) 5000 maximum varia  
. use "/Users/glennwright/Google Drive/Matc  
.  
Command
```
- Command Window:** Shows the following commands:

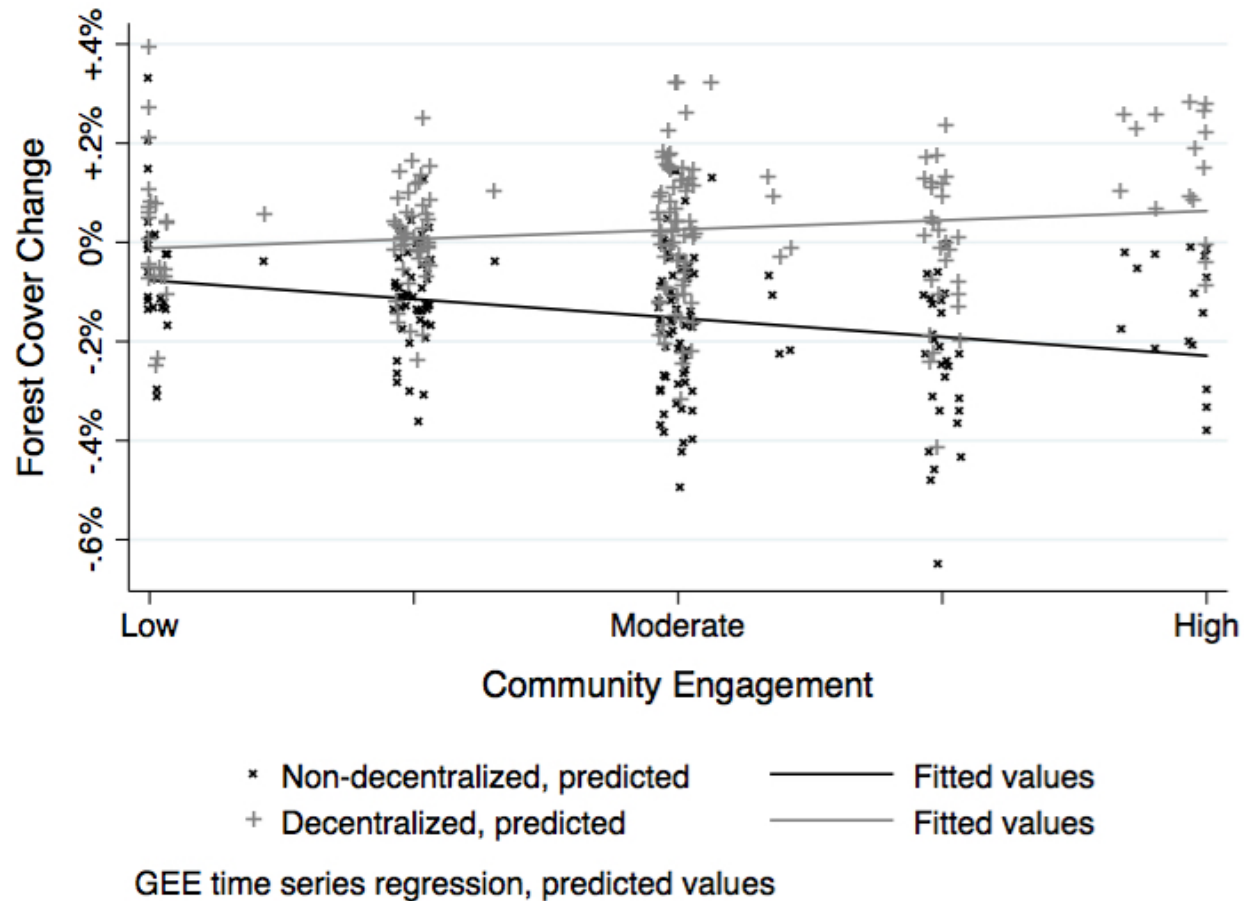
```
clear  
import excel "/Users/glennwright/Google Drive/Matching  
Deforestation/defor_data_with_new_Hansen_dv_feb_2016.xls", sheet(  
"defor_data_with_new_Hansen_dv_f") firstrow  
cd "/Users/glennwright/Google Drive/Matching Deforestation"  
destring decentralization for_cover_lag1 pa_sl12 road_density_logged ///  
ln_budget_mill pop_ln munic_ha_ln ann_pct_chg pct_chg corg_opin_for ///  
pct_chg_lag for_cover_kmsq_ln, ignore("NA") replace  
save "MatchingDataWHansen_2016_02-29.dta", replace  
logit decentralization pct_chg_lag pa_sl12 road_density_logged ///  
ln_budget_mill pop_ln  
predict p_score_pr, pr  
/*  
psmatch2 decentralization, ///  
outcome(ann_pct_chg) ///  
mahal(for_cover_lag1 pa_sl12 road_density_logged munic_ha_ln pop_ln  
ln_budget_mill) ///  
p_score(p_score_pr) ateq common trim(52)  
*/  
psmatch2 decentralization, ///  
outcome(pct_chg) ///  
mahal(pct_chg_lag pa_sl12 road_density_logged munic_ha_ln pop_ln  
ln_budget_mill) ///  
p_score(p_score_pr) ateq common trim(35)  
gen decent_X_corg = decentralization*corg_opin_for  
gen interact_base = corg_opin_for  
gen interact = decent_X_corg  
xtset country_geoid wave  
xtgee pct_chg ///  
decentralization interact_base interact ///  
pa_sl12 road_density_logged pop_ln munic_ha_ln ///  
ln_budget_mill for_cover_lag1 for_cover_kmsq_ln if _support == 1,  
family(gaussian) corr(uns)  
save "Matching_FOR_GRAPHICS_ONLY.dta", replace  
***Generating graphics
```
- Variable List Window:** Shows a list of variables and their labels:

| Variable               | Label                |
|------------------------|----------------------|
| country_geoid          | geographic iden...   |
| pa_id                  | geographic iden...   |
| ave                    | survey wave: 1 o...  |
| apartment              | department           |
| ovince                 | province             |
| strict                 | district or munic... |
| r_cover                | Forest cover (pct.)  |
| port_own_income        | De facto decent...   |
| rg_opin_for            | Community pres...    |
| jo_op_for_mean         | NGO pressure         |
| centralization         | De jure decenta...   |
| imp_for                | Importance of fo...  |
| elite_mean             | Elite education      |
| idget                  | Muni. budget (T...   |
| r_cover_lag1           | Forest cover (la...  |
| ad_density_logged      | Road density (lo...  |
| sl12                   | Slope above 12...    |
| ad_density_gua_revised | km of roads/km^...   |

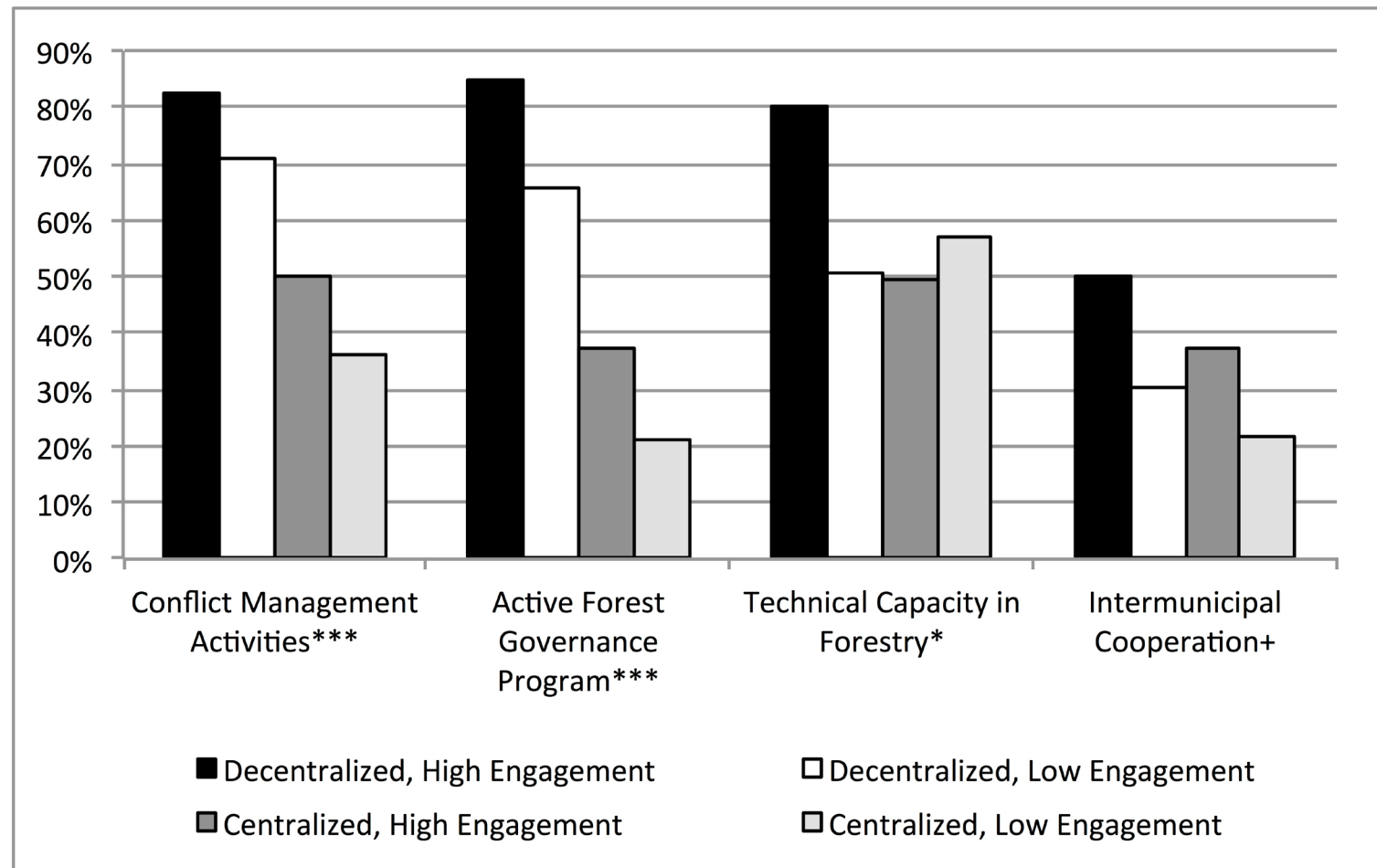
# Matching Techniques



# A Time Series Regression Model



# Simple Cross-Tabulations



# Using STATA

## Pros

- Complex modeling possible
- Excellent time-series features
- Excellent for regression
- Relatively inexpensive

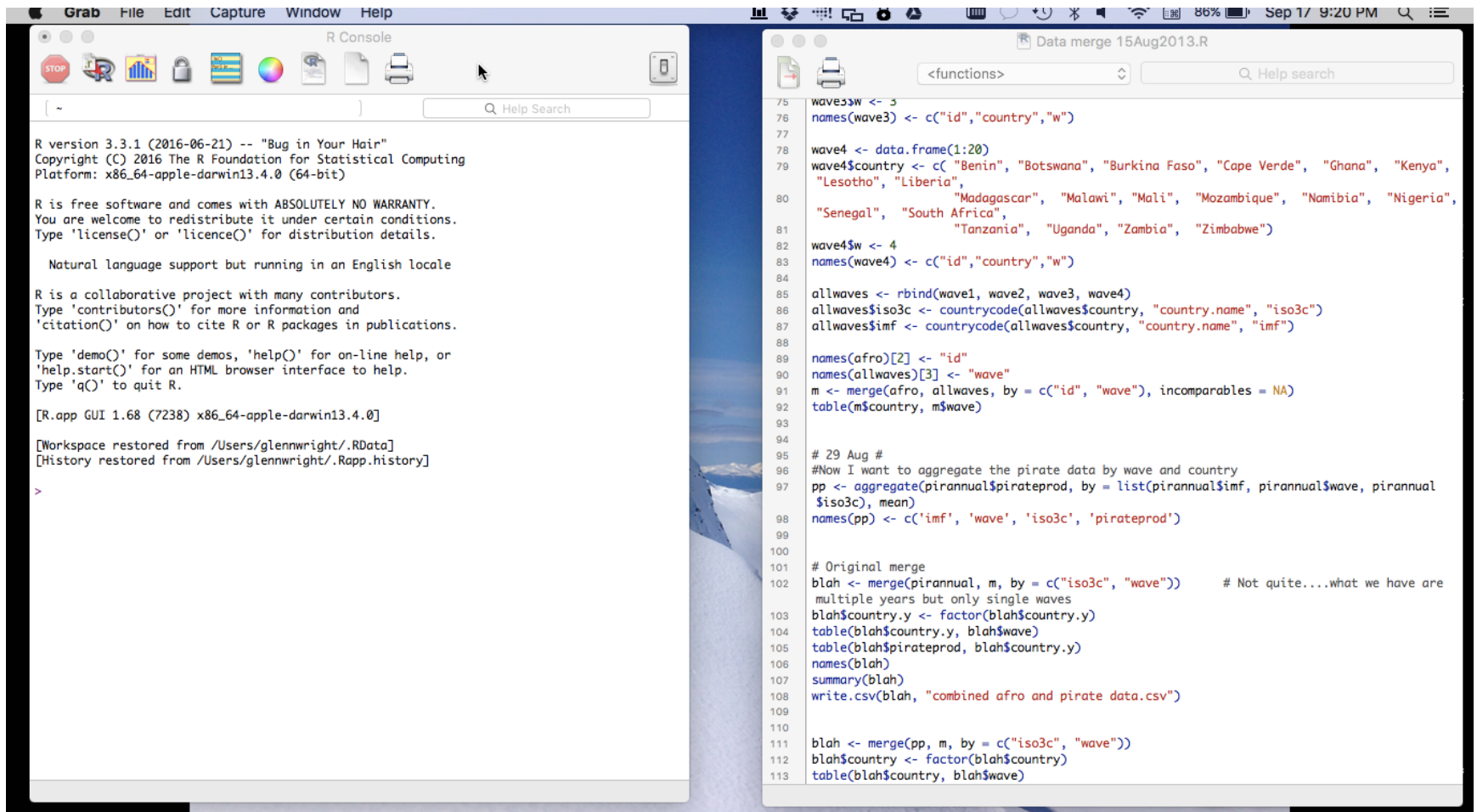
## Cons

- Steep learning curve
- Command line, clunky GUI
- Some modeling strategies impossible or under-developed
  - Structural equation modeling
  - Spatial statistics
- Costs money

# A Second Project: Maritime Piracy

Pirates form democracies, but does democracy  
create pirates?

# Using R



The screenshot displays the R Studio environment. The left pane shows the R Console with the following text:

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.68 (7238) x86_64-apple-darwin13.4.0]

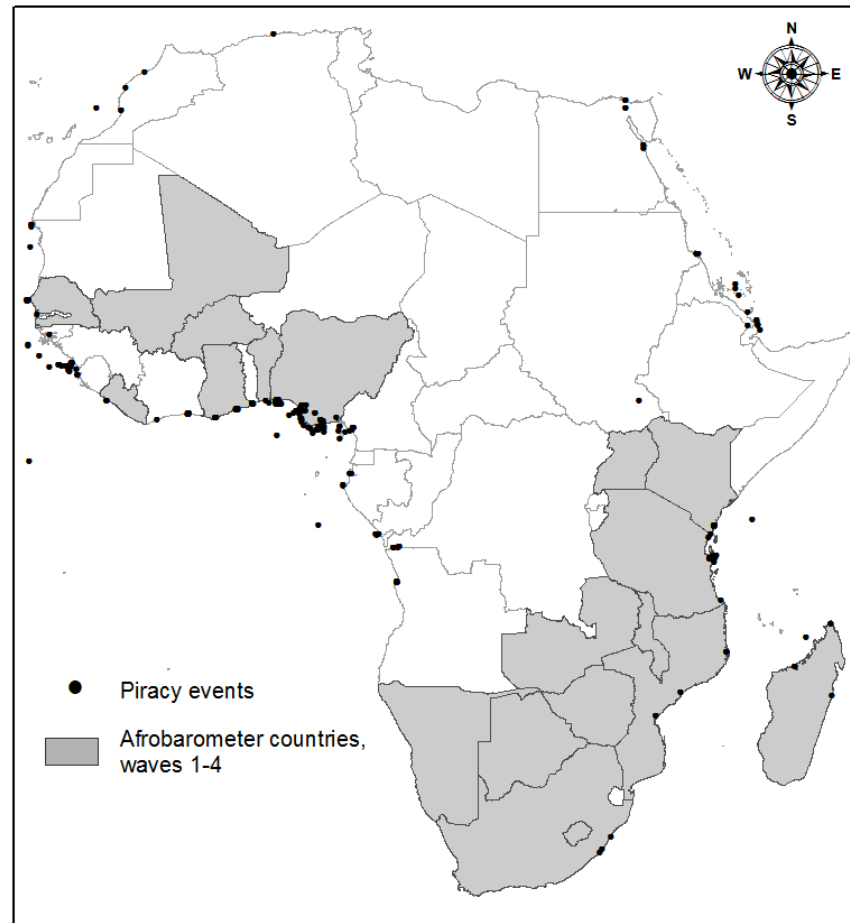
[Workspace restored from /Users/glennwright/.RData]
[History restored from /Users/glennwright/.Rapp.history]

>
```

The right pane shows the Source Editor with the following R code:

```
75 wave3$w <- 3
76 names(wave3) <- c("id", "country", "w")
77
78 wave4 <- data.frame(1:20)
79 wave4$country <- c("Benin", "Botswana", "Burkina Faso", "Cape Verde", "Ghana", "Kenya",
80 "Lesotho", "Liberia",
81 "Madagascar", "Malawi", "Mali", "Mozambique", "Namibia", "Nigeria",
82 "Senegal", "South Africa",
83 "Tanzania", "Uganda", "Zambia", "Zimbabwe")
84
85 wave4$w <- 4
86 names(wave4) <- c("id", "country", "w")
87
88 allwaves <- rbind(wave1, wave2, wave3, wave4)
89 allwaves$iso3c <- countrycode(allwaves$country, "country.name", "iso3c")
90 allwaves$imf <- countrycode(allwaves$country, "country.name", "imf")
91
92 names(afro)[2] <- "id"
93 names(allwaves)[3] <- "wave"
94 m <- merge(afro, allwaves, by = c("id", "wave"), incomparables = NA)
95 table(m$country, m$wave)
96
97 # 29 Aug #
98 #Now I want to aggregate the pirate data by wave and country
99 pp <- aggregate(pirannual$pirateprod, by = list(pirannual$imf, pirannual$wave, pirannual
100 $iso3c), mean)
101 names(pp) <- c('imf', 'wave', 'iso3c', 'pirateprod')
102
103 # Original merge
104 blah <- merge(pirannual, m, by = c("iso3c", "wave")) # Not quite...what we have are
105 multiple years but only single waves
106 blah$country.y <- factor(blah$country.y)
107 table(blah$country.y, blah$wave)
108 table(blah$pirateprod, blah$country.y)
109 names(blah)
110 summary(blah)
111 write.csv(blah, "combined afro and pirate data.csv")
112
113 blah <- merge(pp, m, by = c("iso3c", "wave"))
114 blah$country <- factor(blah$country)
115 table(blah$country, blah$wave)
```

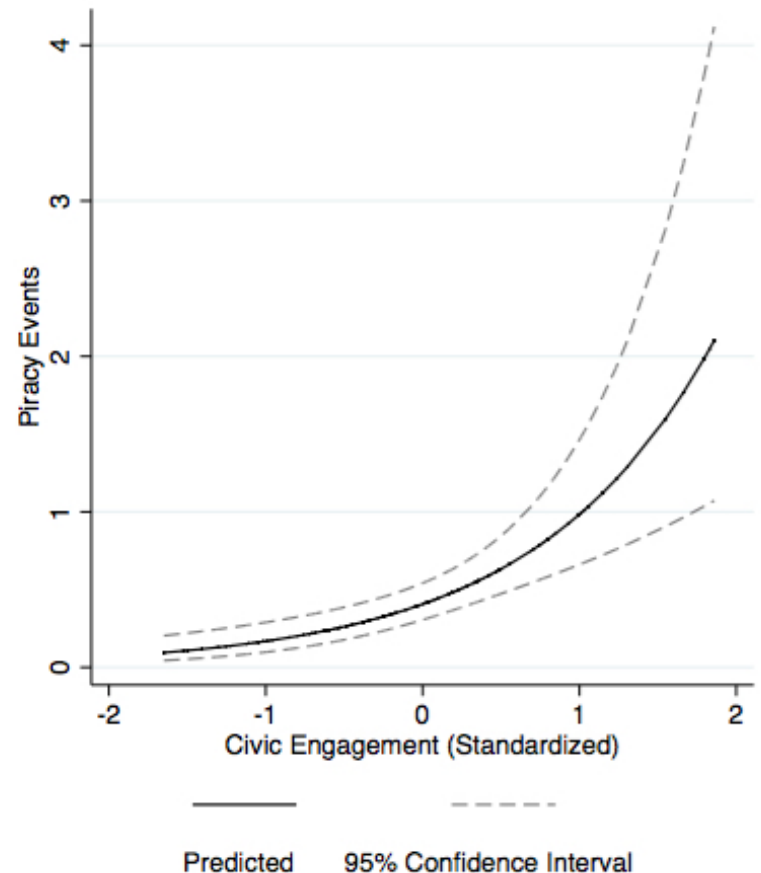
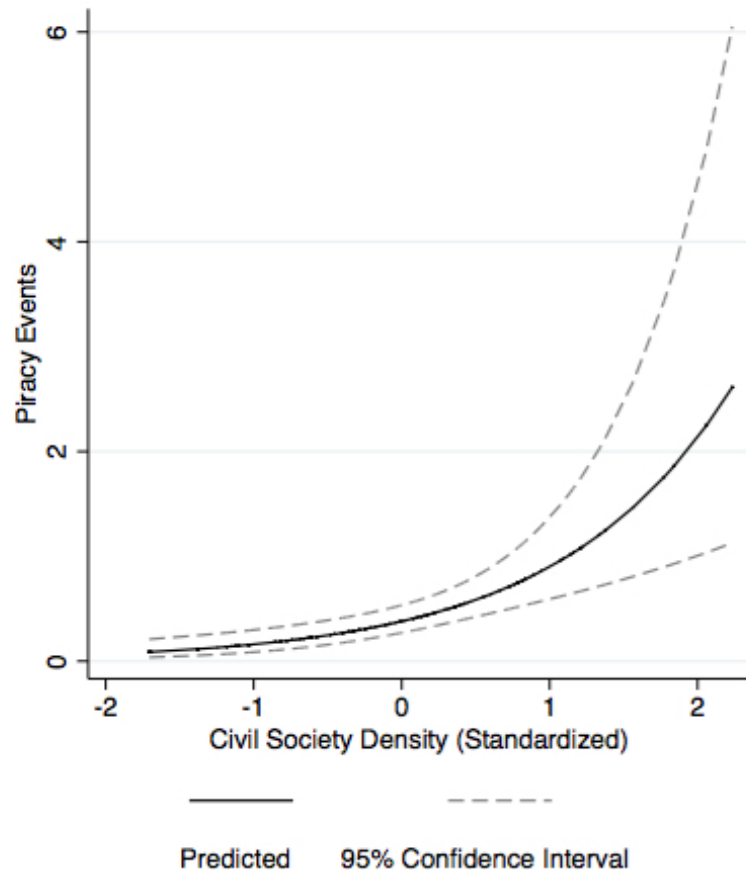
# Piracy Incidents off Africa



# More Regression Models

|                           | 1                   | 2                   | 3                   | 4                   | 5                   |
|---------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Civic Engagement          | 0.52<br>(0.18)**    |                     |                     |                     | -0.02<br>(0.27)     |
| Civil Society<br>Density  |                     | 0.83<br>(0.20)***   |                     |                     | 0.83<br>(0.24)***   |
| Anti-<br>Authoritarianism |                     |                     | 0.14<br>(0.20)      |                     | -0.14<br>(0.25)     |
| Trust in<br>Government    |                     |                     |                     | 0.43<br>(0.17)*     | 0.18<br>(0.26)      |
| Poverty                   | -0.21<br>(0.18)     | -0.24<br>(0.17)     | 0.04<br>(0.20)      | -0.19<br>(0.18)     | -0.26<br>(0.23)     |
| Fatalities                | 0.10<br>(0.06)      | 0.08<br>(0.06)      | 0.09<br>(0.06)      | 0.10<br>(0.07)      | 0.09<br>(0.06)      |
| Coastline                 | -0.09<br>(0.09)     | -0.07<br>(0.10)     | -0.11<br>(0.10)     | -0.10<br>(0.09)     | -0.05<br>(0.11)     |
| Population (ln)           | 1.00<br>(0.18)***   | 1.07<br>(0.18)***   | 1.01<br>(0.18)***   | 1.06<br>(0.18)***   | 1.07<br>(0.18)***   |
| Pop. Growth (%)           | -0.02<br>(0.19)     | 0.08<br>(0.19)      | 0.03<br>(0.20)      | 0.07<br>(0.20)      | 0.09<br>(0.19)      |
| State Capacity            | -0.53<br>(0.16)***  | -0.43<br>(0.15)**   | -0.43<br>(0.15)**   | -0.58<br>(0.16)***  | -0.48<br>(0.17)**   |
| Merchandise<br>Trade      | 1.15<br>(0.31)***   | 0.64<br>(0.31)*     | 0.99<br>(0.31)**    | 1.14<br>(0.32)***   | 0.69<br>(0.34)*     |
| Constant                  | -21.77<br>(3.56)*** | -21.32<br>(3.54)*** | -21.24<br>(3.59)*** | -23.04<br>(3.62)*** | -21.73<br>(3.62)*** |
| N                         | 524                 | 524                 | 524                 | 524                 | 524                 |

# Some Fancy Pictures



# Our Tool: R

## Pros

- Probably the most powerful widely available tool for statistical data analysis Free
- Open source
- Excellent for spatial statistics
- Just about anything is possible
- R help network is broad and deep

## Cons

- Steep learning curve
- Some compatibility issues on some platforms
- User-programmed packages of uneven quality
- R help network is kind of snobby and arrogant

Thoughts? Questions?