EVOS data project - Analysis

Jessica Couture

3/8/2017

Contents

Background

This analysis accompanies the manuscript "Funder imposed data publication requirements seldom inspire data sharing", assessing results of a two year data archiving effort by a group of researchers and students at the National Center for Ecological Analysis and Synthesis at UC Santa Barbara. The Exxon Valdez Oil Spill Trustee Council (EVOSTC) was formed following the Exxon Valdez oil spill in Alaska in 1989. Since then, the EVOSTC has funded hundreds of projects and in 2012 an effort was initiated to recover and archive the data collected through these EVOSTC funded projects. The recovery effort spanned two years.

For this paper we use the results of that effort to ask 3 main questions about the data collected from the Exxon Valdez Oil Spill Trustee Council funded projects:

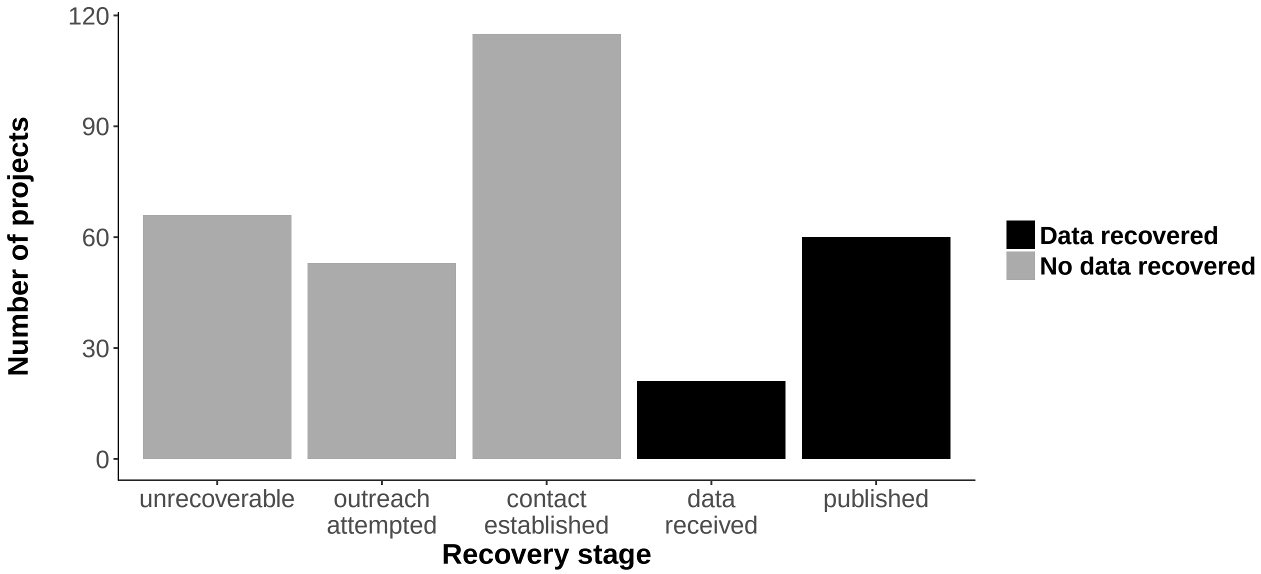
1. Twenty-five years after the EVOS, for how many projects funded by EVOSTC can we collect data?
2. Are there differences in data reporting based on characteristics of the data?

- Research field
- Sector of researching body
- Year data projects ended

3. Which of these characteristics are most *important* in determining if a dataset will be successfully recovered and how do the important characteristics influence the output (success)?

1. Project Status Reporting

Twenty-five years after the EVOS, for how many projects funded by EVOSTC can we collect data?



```r
# Number of projects requested
sum(overall)
```

```
## [1] 315
```

```r
# percent successful
sum(overall[c("Published","SentData")])/sum(overall)
```

```
## [1] 0.2571429
```

---

2. Are there differences in data reporting based on data characteristics?

Chi-square tests for each variable (characteristic)

Research field

Are there certain research fields that are more likely to make data available than others?

Chi-squared test for equal proportions between research fields

```
## 
##     benthicInverts birds fish habitat mammals modeling oil physical
##   0              11    44   75      22      24       11  18        8
##   1               9    14   19       5       9        0  11       11
## 
##     plankton social
##   0        2     19
##   1        1      2
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  bioProps
## X-squared = 25.58, df = 9, p-value = 0.002392
```

SIGNIFICANT: Reject the H0 that there are no differences in recovery in different research fields

---

Research sector

Test for equal propporTions between PI's home institution type/sector:

Government and private split in to local, federal, and non-profit, for-profit sectors: govFed, govState, nonProf, forProf

```
##
##     academia akNative gov_fed gov_state nonProf private
##   0       42        1      87        56      20      17
##   1       14        1      29        18      11       8
```

```
##
##  Pearson's Chi-squared test
##
## data:  subSecProps
## X-squared = 2.6061, df = 5, p-value = 0.7604
```

NOT SIGNIFICANT: there are no differences in recovery in different sectors

---

Year project ended

Test for equal proportions between years:

Chi-squared test for equal proportions between age of data. Since many projects span multiple years we base "age of data" on the last year that the project received funding as a conservative, assuming they had 1 year to publish/make available.

```
##
##     1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002
##   0   10    5   12   17   10    9   14    8    6   17   24   22   13   17
##   1    5    2    3   14    1    2    4    2    2    6    6    4    6    1
##
##     2003 2004 2005 2006 2007 2008 2009 2010
##   0   12   14    4    3    7    1    4    5
##   1    1    5    2    0    3    3    6    3
```

```
##
##  Pearson's Chi-squared test
##
## data:  tempProps
## X-squared = 30.577, df = 21, p-value = 0.08099
```

NOT SIGNIFICANT: there are no differences in recovery based on when data were collected

---

Binomial regression

In order to assess how the percent recovery is influenced by time, data type and agency we are running an logistic regression on all 3 factors.

```
blrDat<-rslt2 %>%
  select(end,dataType,statSucc,agSubGrp)

mod<-glm(statSucc~.,family = binomial(link="logit"),data=blrDat)

summary(mod)
```

```
##
## Call:
## glm(formula = statSucc ~ ., family = binomial(link = "logit"),
##     data = blrDat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4839  -0.7544  -0.6347   1.0350   2.1540
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        7.688787  50.452266   0.152   0.8789
## end               -0.003926   0.025240  -0.156   0.8764
## dataTypebirds     -0.815963   0.559783  -1.458   0.1449
## dataTypefish      -1.170844   0.535555  -2.186   0.0288 *
## dataTypehabitat   -1.190769   0.678245  -1.756   0.0791 .
## dataTypemammals   -0.717978   0.615339  -1.167   0.2433
## dataTypemodeling -16.382176 720.906131  -0.023   0.9819
## dataTypeoil       -0.140457   0.606569  -0.232   0.8169
## dataTypephysical   0.477153   0.657545   0.726   0.4680
## dataTypeplankton  -0.654774   1.336036  -0.490   0.6241
## dataTypesocial    -1.881887   0.907154  -2.074   0.0380 *
## agSubGrpakNative   1.348768   1.465440   0.920   0.3574
## agSubGrpgov_fed   -0.195064   0.408662  -0.477   0.6331
## agSubGrpgov_state -0.170233   0.443209  -0.384   0.7009
## agSubGrpnonProf    0.391655   0.534362   0.733   0.4636
## agSubGrpprivate    0.024819   0.583796   0.043   0.9661
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 352.45  on 303  degrees of freedom
## Residual deviance: 324.77  on 288  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 356.77
##
## Number of Fisher Scoring iterations: 15
```

```
ageMod<-glm(statSucc~end-1,family = binomial(link="logit"),data=blrDat)
#summary(ageMod)
```

Chi-sq tests from Regression:

```
library(aod)

timeChi<-wald.test(b = coef(mod), Sigma = vcov(mod), Terms = 2)

typeChi<-wald.test(b = coef(mod), Sigma = vcov(mod), Terms = 3:11)

agChi<-wald.test(b = coef(mod), Sigma = vcov(mod), Terms = 12:16)
```

Chi-sq for age of data = 0.8763739 Chi-sq for data type = 0.0465513 Chi-sq for agency = 0.7601266

INTERESING NOTE: When these tests are run on the "start" dates rather than the "end" we get significance in AGE and TYPE

3. Which characteristics are most important in determining if a dataset will be successfully recovered?

We use the "party" package in R to run a random forests analysis to determine which variables are most important. I use the same model as the glm, then create a classification tree below to show how the important variables influence the outcome. This package is be better than randomForest when independent variables are different types (Strobl et al. 2009)

Random forests

```
rslt2$dataType<-as.factor(rslt2$dataType)
partyForBio<-cforest(statSucc~agSubGrp+end+dataType,data=rslt2,controls = cforest_unbiased(mtry = 2, ntr
varimp(partyForBio)
```

```
##     agSubGrp          end     dataType
## -0.002583888  0.005530835  0.014187998
```

Based on these results the most important variable in determining the outcome is research field

---

How do the important characteristics influence the output (success)?

Classification tree

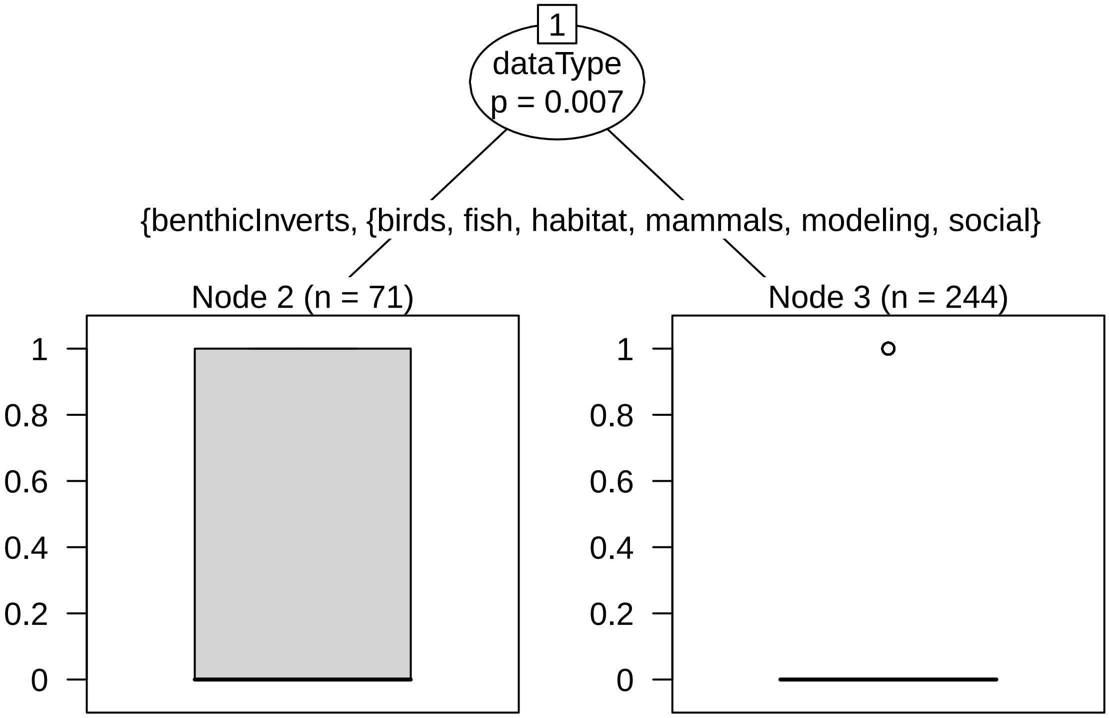
Here we run one iteration of the forest analysis above to display which variables whithin each classification determine positive or negative results.

```
partreeBio<-ctree(statSucc~agSubGrp+end+dataType,data=rslt2)

partreeBio
```

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  statSucc
## Inputs:  agSubGrp, end, dataType
## Number of observations:  315
##
## 1) dataType == {benthicInverts, oil, physical, plankton}; criterion = 0.993, statistic = 25.499
##   2)*  weights = 71
## 1) dataType == {birds, fish, habitat, mammals, modeling, social}
##   3)*  weights = 244
```

```
plot(partreeBio)
```



Birds, fish, habitat, mammal and *modeling data* result in negative results. *Benthic invertebrates, plankton, oil,* and *physical data* result in positive results.