EVOS data project - Analysis

Contents

Background

This analysis accompanies the manuscript "A funder imposed data publication requirement seldom inspired 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:

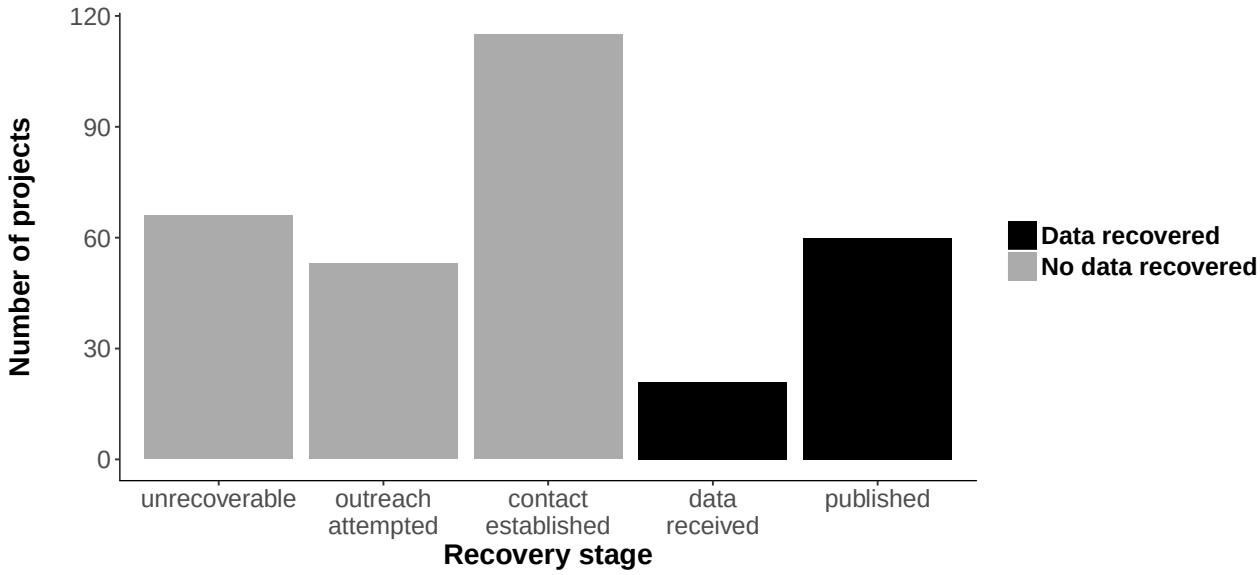
1. Twenty-five years after the EVOS, for how many projects funded by EVOSTC can we collect data?
2. Are there differences in data reporting based on characteristics of the data project?

- 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?




```
# Number of projects requested
nProj<-sum(overall)

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

Total number of projects sought = 315
Percent success = 26%

---

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

```
blrDat<-rslt2 %>%
  select(end,dataType,statSucc,agSubGrp)
blrDat$DP<-ifelse(blrDat$end<1995,0,1) # add data policy binary, formal language for data sharing was i

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

Logistic regression

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

```
summary(mod)
```

```
##
## Call:
## glm(formula = statSucc ~ ., family = binomial(link = "logit"),
##     data = blrDat)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.4941  -0.7840  -0.6218   0.9002   2.3219
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -133.13617   78.68570  -1.692   0.0906 .
## end                  0.06696    0.03951   1.695   0.0901 .
## dataTypebirds       -0.78280    0.56376  -1.389   0.1650
## dataTypefish        -1.05760    0.54034  -1.957   0.0503 .
## dataTypehabitat     -1.14491    0.68420  -1.673   0.0943 .
## dataTypemammals     -0.78966    0.62084  -1.272   0.2034
## dataTypemodeling   -16.19635  717.28730  -0.023   0.9820
## dataTypeoil         -0.13461    0.61137  -0.220   0.8257
## dataTypephysical     0.69336    0.67220   1.031   0.3023
## dataTypeplankton    -0.58023    1.36239  -0.426   0.6702
## dataTypesocial      -1.94444    0.92208  -2.109   0.0350 *
## agSubGrpakNative     1.21848    1.48195   0.822   0.4110
## agSubGrpgov_fed     -0.28315    0.41384  -0.684   0.4938
## agSubGrpgov_state   -0.28603    0.45006  -0.636   0.5251
## agSubGrpnonProf      0.28733    0.54156   0.531   0.5957
## agSubGrpprivate      0.07505    0.58653   0.128   0.8982
## DP                  -1.17639    0.51101  -2.302   0.0213 *
## ---
## 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: 319.32  on 287  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 353.32
##
## Number of Fisher Scoring iterations: 15
```

Nested logistic regression

How do our three characteristics and data policy influence each step in the recovery process?

Nest 1: Confirmed contact info ("emailed"+)

```
nest<-rslt2 %>%
  mutate(pContInf=ifelse(is.na(reason),1,ifelse(reason=="no contact info",0,1))) %>% # for NEST1: use a
  mutate(pRepl=ifelse(Status=="Emailed",0,1)) %>% # for NEST2: remove "no contact info" values when ana
  mutate(pSent=ifelse(Status=="SentData",1,ifelse(Status=="Published",1,0))) %>% # for NEST3: rm "no co
  mutate(pPub=ifelse(Status=="Published",1,0)) %>% #for NEST 4: use remaining data
  mutate(DP=ifelse(blrDat$end<1995,0,1))

nest1blmB<-glm(pContInf~end+dataType+agSubGrp+DP,family = binomial(link="logit"),data=nest)

summary(nest1blmB)
```

```
##
## Call:
## glm(formula = pContInf ~ end + dataType + agSubGrp + DP, family = binomial(link = "logit"),
##     data = nest)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.88791   0.07093   0.17943   0.37078   1.16691
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -429.9442   203.6775  -2.111   0.0348 *
## end                  0.2172     0.1023   2.123   0.0338 *
## dataTypebirds        2.3226     1.2352   1.880   0.0601 .
## dataTypefish         0.3504     0.8456   0.414   0.6786
## dataTypehabitat     -0.7184     0.9323  -0.771   0.4409
## dataTypemammals     17.3437  1705.7682   0.010   0.9919
## dataTypemodeling    16.9579  2996.9404   0.006   0.9955
## dataTypeoil         -0.1411     0.9195  -0.153   0.8780
## dataTypephysical     0.5933     1.3156   0.451   0.6520
## dataTypeplankton    15.2109  6062.8002   0.003   0.9980
## dataTypesocial      -0.2307     1.1470  -0.201   0.8406
## agSubGrpakNative    14.5623  7259.6195   0.002   0.9984
## agSubGrpgov_fed     -1.3152     1.1426  -1.151   0.2497
## agSubGrpgov_state   -1.8951     1.0962  -1.729   0.0839 .
## agSubGrpnonProf     -0.5892     1.5611  -0.377   0.7059
## agSubGrpprivate     -1.9524     1.3395  -1.458   0.1450
## DP                  -0.2331     0.8953  -0.260   0.7946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 152.76  on 303  degrees of freedom
## Residual deviance: 114.66  on 287  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 148.66
##
## Number of Fisher Scoring iterations: 18
```

Looking at just for how many contact information could be found, there is a significant positive effect of age (p=0.0337576), increasing 0.2171972 annually.

Nest 2: replied given we found contact info ("Replied"+)

```
nest2<-nest %>%
  filter(is.na(reason) | reason !="no contact info")

nest2blmB<-glm(pRepl~end+dataType+agSubGrp+DP,family = binomial(link="logit"),data=nest2)

summary(nest2blmB)
```

```
##
```

```
## Call:
## glm(formula = pRepl ~ end + dataType + agSubGrp + DP, family = binomial(link = "logit"),
##     data = nest2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2626   0.3541   0.5347   0.6583   1.3495
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -47.46804  957.32675  -0.050    0.960
## end                 0.03261    0.04654   0.701    0.484
## dataTypebirds     -15.64750  952.82542  -0.016    0.987
## dataTypefish      -16.19130  952.82537  -0.017    0.986
## dataTypehabitat   -16.50893  952.82545  -0.017    0.986
## dataTypemammals   -15.86128  952.82546  -0.017    0.987
## dataTypemodeling  -15.04480  952.82591  -0.016    0.987
## dataTypeoil       -15.56903  952.82554  -0.016    0.987
## dataTypephysical  -15.39180  952.82563  -0.016    0.987
## dataTypeplankton  -16.57684  952.82615  -0.017    0.986
## dataTypesocial    -17.94584  952.82551  -0.019    0.985
## agSubGrpakNative   -1.05876    1.47367  -0.718    0.472
## agSubGrpgov_fed     0.25915    0.44310   0.585    0.559
## agSubGrpgov_state   0.53949    0.50731   1.063    0.288
## agSubGrpnonProf     0.20735    0.60298   0.344    0.731
## agSubGrpprivate     0.31106    0.75119   0.414    0.679
## DP                 -0.60954    0.60426  -1.009    0.313
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 272.96  on 282  degrees of freedom
## Residual deviance: 246.79  on 266  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 280.79
##
## Number of Fisher Scoring iterations: 16
```

Nest 3: Sent data given we received a response ("SentData"+)

```
nest3<-nest2 %>%
  filter(!Status=="Emailed")

nest3blmB<-glm(pSent~end+dataType+agSubGrp+DP,family = binomial(link="logit"),data=nest3)

summary(nest3blmB)
```

```
##
## Call:
## glm(formula = pSent ~ end + dataType + agSubGrp + DP, family = binomial(link = "logit"),
##     data = nest3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.6854  -0.9083  -0.6925   1.0978   1.8748
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -73.98189   83.75562  -0.883   0.3771
## end                  0.03749    0.04205   0.892   0.3726
## dataTypebirds       -0.94125    0.60409  -1.558   0.1192
## dataTypefish        -0.95053    0.57669  -1.648   0.0993 .
## dataTypehabitat     -0.72173    0.75146  -0.960   0.3368
## dataTypemammals     -0.74646    0.65898  -1.133   0.2573
## dataTypemodeling   -16.48806  750.26513  -0.022   0.9825
## dataTypeoil         -0.03300    0.68701  -0.048   0.9617
## dataTypephysical     0.80760    0.74963   1.077   0.2813
## dataTypeplankton    -0.24077    1.54434  -0.156   0.8761
## dataTypesocial      -0.80480    1.03415  -0.778   0.4364
## agSubGrpakNative    17.29619 2399.54477   0.007   0.9942
## agSubGrpgov_fed     -0.34207    0.44240  -0.773   0.4394
## agSubGrpgov_state   -0.35054    0.48491  -0.723   0.4697
## agSubGrpnonProf      0.34064    0.58727   0.580   0.5619
## agSubGrpprivate      0.16107    0.63031   0.256   0.7983
## DP                  -1.07661    0.53977  -1.995   0.0461 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 297.20  on 229  degrees of freedom
## Residual deviance: 265.88  on 213  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 299.88
##
## Number of Fisher Scoring iterations: 15
```

Our variables were not significant indicators as to whether data were sent given that we received a response.

Nest 4: Data were published given we received data ("Published")

```
nest4<-nest3 %>%
  filter(Status %in% c("SentData","Published"))

nest4blmB<-glm(pPub~end+dataType+agSubGrp+DP,family = binomial(link="logit"),data=nest4)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(nest4blmB)
```

```
##
## Call:
## glm(formula = pPub ~ end + dataType + agSubGrp + DP, family = binomial(link = "logit"),
##     data = nest4)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.90043  -0.33150   0.00005   0.40330   2.43570
```

```
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.042e+02  2.241e+02   0.465   0.6420
## end                -5.151e-02  1.126e-01  -0.458   0.6473
## dataTypebirds      -1.559e+00  1.439e+00  -1.083   0.2786
## dataTypefish       -1.047e+00  1.398e+00  -0.749   0.4539
## dataTypehabitat     1.798e+01  7.395e+03   0.002   0.9981
## dataTypemammals    -1.395e-01  1.407e+00  -0.099   0.9211
## dataTypeoil        -2.906e+00  1.608e+00  -1.807   0.0708 .
## dataTypephysical    1.796e+01  4.289e+03   0.004   0.9967
## dataTypeplankton    2.387e-01  1.836e+04   0.000   1.0000
## dataTypesocial      1.712e+01  1.147e+04   0.001   0.9988
## agSubGrpakNative    1.783e+01  1.773e+04   0.001   0.9992
## agSubGrpgov_fed    -1.604e+00  1.295e+00  -1.239   0.2154
## agSubGrpgov_state  -1.621e+00  1.476e+00  -1.099   0.2719
## agSubGrpnonProf     1.660e+01  4.756e+03   0.003   0.9972
## agSubGrpprivate     1.857e+01  5.107e+03   0.004   0.9971
## DP                  3.007e+00  1.506e+00   1.997   0.0459 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 92.105  on 79  degrees of freedom
## Residual deviance: 44.678  on 64  degrees of freedom
## AIC: 76.678
##
## Number of Fisher Scoring iterations: 19
```

Our variables were not significance indicators as to whether data were complete enough to publish given data were sent.

---

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 the "randomForest" package when independent variables are different types (Strobl et al. 2009).

For the random forests the independent variable with the highest absolute value has the highest impact on the dependent variable.

Random forests

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

```
##          agSubGrp               end factor(dataType)                DP
##    -0.0020413841     0.0050292995     0.0106638069     0.0002133661
```

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

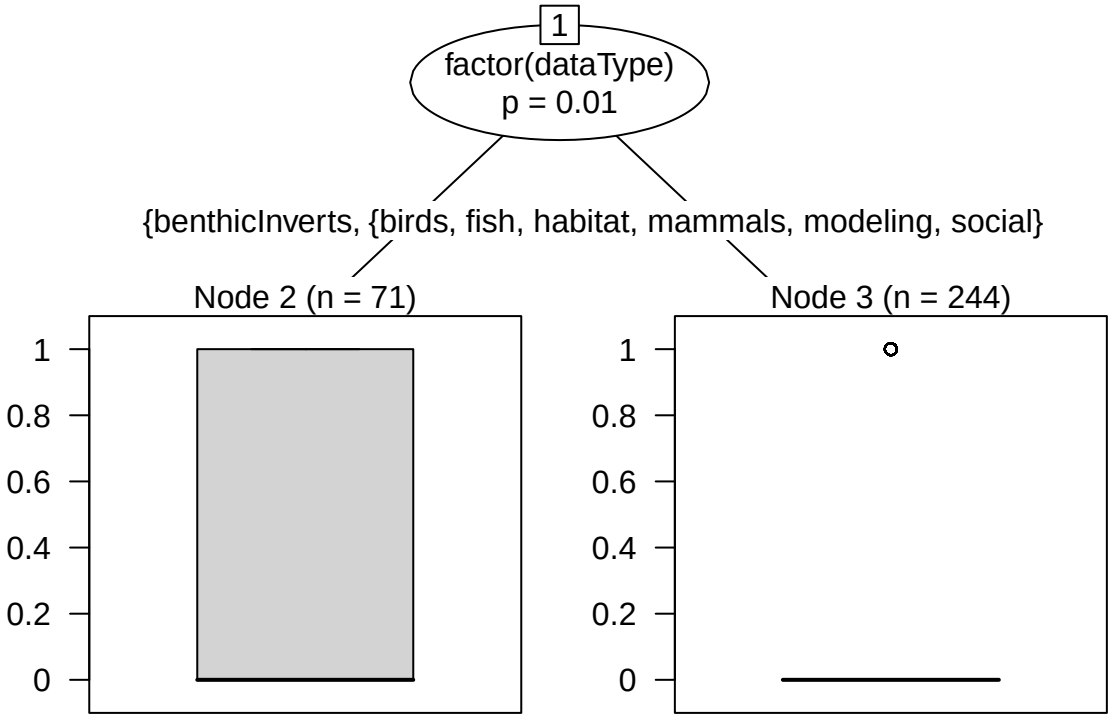
Here we run one iteration of the forest analysis above to display which variables whithin each classification determine positive or negative results.

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

partreeBio
```

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  statSucc
## Inputs:  agSubGrp, end, factor(dataType), DP
## Number of observations:  315
##
## 1) factor(dataType) == {benthicInverts, oil, physical, plankton}; criterion = 0.99, statistic = 25.49
##   2)*  weights = 71
## 1) factor(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.

Post-data policy model

We isolate the data to project ending post 1994 to assess trends based on just data under the formal data policy. This model was not used in the submitted manuscript.

```
post<-blrDat %>%
  filter(DP==1) %>%
  select(-DP)

modP<-glm(statSucc~.,family = binomial(link="logit"),data=post)
summary(modP)
```

```
##
## Call:
## glm(formula = statSucc ~ ., family = binomial(link = "logit"),
##     data = post)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.56553  -0.74643  -0.58609  -0.00017   2.08002
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -137.21640   83.01808  -1.653   0.0984 .
## end                  0.06831    0.04146   1.648   0.0995 .
## dataTypebirds       -0.71796    0.77151  -0.931   0.3521
## dataTypefish        -0.74375    0.69304  -1.073   0.2832
## dataTypehabitat     -0.77642    0.83198  -0.933   0.3507
## dataTypemammals     -0.31634    0.80365  -0.394   0.6939
## dataTypemodeling   -16.90646 1230.50642  -0.014   0.9890
## dataTypeoil          0.10992    0.80749   0.136   0.8917
## dataTypephysical     0.77199    0.79400   0.972   0.3309
## dataTypeplankton    -0.49599    1.43178  -0.346   0.7290
## dataTypesocial      -0.67638    1.07159  -0.631   0.5279
## agSubGrpakNative     1.09851    1.49276   0.736   0.4618
## agSubGrpgov_fed     -0.49860    0.47681  -1.046   0.2957
## agSubGrpgov_state   -0.33573    0.55182  -0.608   0.5429
## agSubGrpnonProf      0.56306    0.59581   0.945   0.3446
## agSubGrpprivate     -0.09940    0.66034  -0.151   0.8804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 245.77  on 220  degrees of freedom
## Residual deviance: 220.83  on 205  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 252.83
##
```

```
## Number of Fisher Scoring iterations: 16
```