EVOS data project - Analysis

Jessica Couture

3/15/2017

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:

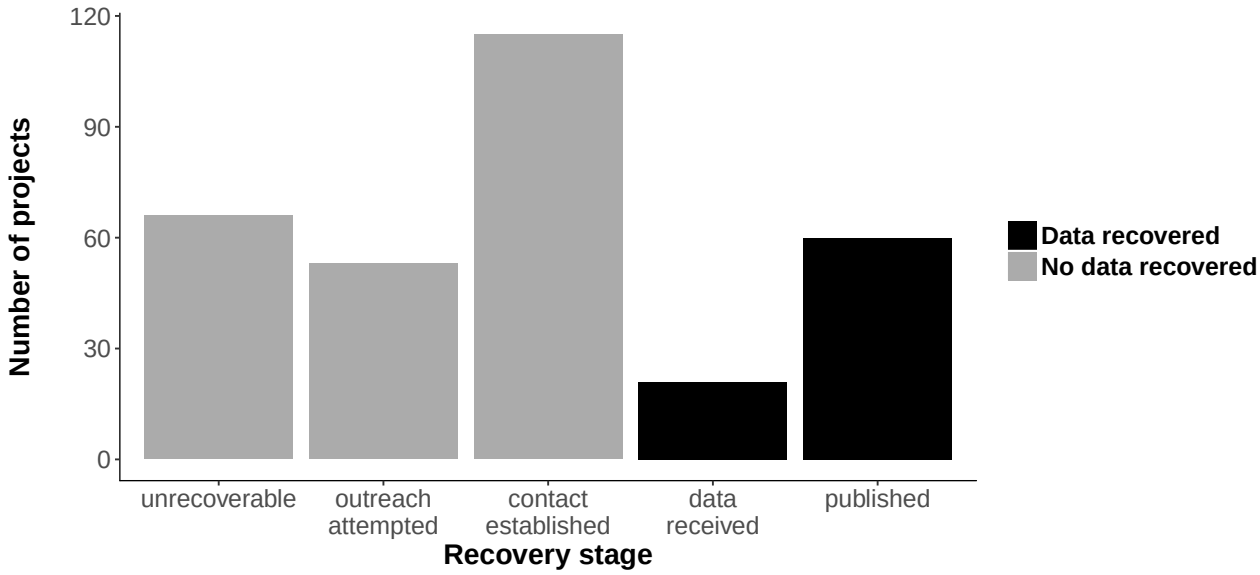
1. Twenty-five years after the EVOS, for how many projects funded by EVOSTC can we collect data?
2. Are there differences in data reporting based on characteristics of the data project?

- 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)

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

#rm factors that are not significant

dtMod<-glm(statSucc~dataType,family = binomial(link="logit"),data=blrDat)
#summary(dtMod)
```

Logistic 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.

```
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
```

Nested logistic regression

How do our three characteristics 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))) %>% #use all date
  mutate(pRepl=ifelse(Status=="Emailed",0,1)) %>% # remove "no contact info" values when analyzing -->
  mutate(pSent=ifelse(Status=="SentData",1,ifelse(Status=="Published",1,0))) %>% # rm "no contact info"
  mutate(pPub=ifelse(Status=="Published",1,0))

nest1blm<-glm(pContInf~end+dataType+agSubGrp,family = binomial(link="logit"),data=nest)
```

```r
summary(nest1blm)
```

```
##
## Call:
## glm(formula = pContInf ~ end + dataType + agSubGrp, family = binomial(link = "logit"),
##     data = nest)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.85983   0.07142   0.18621   0.36853   1.15510
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -388.47901  123.13885  -3.155  0.00161 **
## end                  0.19636    0.06181   3.177  0.00149 **
## dataTypebirds        2.32861    1.23456   1.886  0.05927 .
## dataTypefish         0.33886    0.84288   0.402  0.68767
## dataTypehabitat     -0.72367    0.93124  -0.777  0.43710
## dataTypemammals     17.35407 1702.26609   0.010  0.99187
## dataTypemodeling    16.95447 2986.76416   0.006  0.99547
## dataTypeoil         -0.13590    0.91810  -0.148  0.88232
## dataTypephysical     0.52688    1.28743   0.409  0.68236
## dataTypeplankton    15.19007 6082.80101   0.002  0.99801
## dataTypesocial      -0.24821    1.14481  -0.217  0.82836
## agSubGrpakNative    14.56793 7307.22851   0.002  0.99841
## agSubGrpgov_fed     -1.31854    1.14454  -1.152  0.24931
## agSubGrpgov_state   -1.89192    1.09699  -1.725  0.08459 .
## agSubGrpnonProf     -0.54850    1.55260  -0.353  0.72388
## agSubGrpprivate     -1.97359    1.33763  -1.475  0.14009
## ---
## 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: 114.72  on 288  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 146.72
##
## Number of Fisher Scoring iterations: 18
```

```r
summary(nest1b<-glm(pContInf~end,family = binomial(link = "logit"),data=nest))
```

```
##
## Call:
## glm(formula = pContInf ~ end, family = binomial(link = "logit"),
##     data = nest)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9880   0.2209   0.3195   0.5015   0.8321
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -372.83221   89.28161  -4.176 2.97e-05 ***
## end            0.18789    0.04476   4.197 2.70e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 198.13  on 314  degrees of freedom
## Residual deviance: 175.78  on 313  degrees of freedom
## AIC: 179.78
##
## Number of Fisher Scoring iterations: 6
```

Looking at just how many contact information could be found, there is a significant positive effect of age (p=0.0014886), increasing 0.1963598 annually.

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

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

nest2blm<-glm(pRepl~end+dataType+agSubGrp,family = binomial(link="logit"),data=nest2)

summary(nest2blm)
```

```
##
## Call:
## glm(formula = pRepl ~ end + dataType + agSubGrp, family = binomial(link = "logit"),
##     data = nest2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2003   0.3901   0.5337   0.6952   1.4633
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        21.060981 957.952334   0.022    0.982
## end                -0.001897   0.030884  -0.061    0.951
## dataTypebirds     -15.678202 955.959452  -0.016    0.987
## dataTypefish      -16.269361 955.959396  -0.017    0.986
## dataTypehabitat   -16.540762 955.959477  -0.017    0.986
## dataTypemammals   -15.825293 955.959490  -0.017    0.987
## dataTypemodeling  -15.153335 955.959932  -0.016    0.987
## dataTypeoil       -15.579178 955.959567  -0.016    0.987
## dataTypephysical  -15.478911 955.959653  -0.016    0.987
## dataTypeplankton  -16.651691 955.960175  -0.017    0.986
## dataTypesocial    -17.928804 955.959535  -0.019    0.985
## agSubGrpakNative   -0.990523   1.467694  -0.675    0.500
## agSubGrpgov_fed     0.273330   0.442469   0.618    0.537
## agSubGrpgov_state   0.551394   0.504859   1.092    0.275
## agSubGrpnonProf     0.254547   0.604737   0.421    0.674
## agSubGrpprivate     0.291050   0.746252   0.390    0.697
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 272.96  on 282  degrees of freedom
## Residual deviance: 247.82  on 267  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 279.82
##
## Number of Fisher Scoring iterations: 16
```

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

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

nest3blm<-glm(pSent~end+dataType+agSubGrp,family = binomial(link="logit"),data=nest3)

summary(nest3blm)
```

```
##
## Call:
## glm(formula = pSent ~ end + dataType + agSubGrp, family = binomial(link = "logit"),
##     data = nest3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6692  -0.8693  -0.7501   1.1413   1.8204
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         52.94372   55.62708   0.952   0.3412
## end                 -0.02641    0.02782  -0.950   0.3423
## dataTypebirds       -0.96473    0.59901  -1.611   0.1073
## dataTypefish        -1.05393    0.57058  -1.847   0.0647 .
## dataTypehabitat     -0.77314    0.74530  -1.037   0.2996
## dataTypemammals     -0.69820    0.65316  -1.069   0.2851
## dataTypemodeling   -16.70451  753.74117  -0.022   0.9823
## dataTypeoil         -0.05485    0.68086  -0.081   0.9358
## dataTypephysical     0.63270    0.73738   0.858   0.3909
## dataTypeplankton    -0.21743    1.52265  -0.143   0.8865
## dataTypesocial      -0.75982    1.01131  -0.751   0.4525
## agSubGrpakNative    17.71301 2399.54476   0.007   0.9941
## agSubGrpgov_fed     -0.24549    0.43598  -0.563   0.5734
## agSubGrpgov_state   -0.22639    0.47696  -0.475   0.6350
## agSubGrpnonProf      0.40959    0.58000   0.706   0.4801
## agSubGrpprivate      0.16341    0.62721   0.261   0.7944
## ---
## 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: 269.95  on 214  degrees of freedom
##   (2 observations deleted due to missingness)
```

```
## AIC: 301.95
##
## 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"))

nest4blm<-glm(pPub~end+dataType+agSubGrp,family = binomial(link="logit"),data=nest4)
```

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

```
summary(nest4blm)
```

```
##
## Call:
## glm(formula = pPub ~ end + dataType + agSubGrp, family = binomial(link = "logit"),
##     data = nest4)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.62533  -0.44066   0.00006   0.38304   2.19043
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.755e+02  1.423e+02  -1.937   0.0528 .
## end                1.394e-01  7.139e-02   1.952   0.0509 .
## dataTypebirds     -1.159e+00  1.343e+00  -0.863   0.3879
## dataTypefish      -2.981e-01  1.281e+00  -0.233   0.8160
## dataTypehabitat    1.828e+01  7.308e+03   0.003   0.9980
## dataTypemammals    1.230e-01  1.372e+00   0.090   0.9285
## dataTypeoil       -2.036e+00  1.479e+00  -1.376   0.1687
## dataTypephysical   1.868e+01  4.369e+03   0.004   0.9966
## dataTypeplankton  -6.657e-01  1.837e+04   0.000   1.0000
## dataTypesocial     1.828e+01  1.062e+04   0.002   0.9986
## agSubGrpakNative   1.656e+01  1.773e+04   0.001   0.9993
## agSubGrpgov_fed   -2.012e+00  1.237e+00  -1.626   0.1040
## agSubGrpgov_state -1.921e+00  1.400e+00  -1.372   0.1702
## agSubGrpnonProf    1.679e+01  4.799e+03   0.003   0.9972
## agSubGrpprivate    1.799e+01  5.440e+03   0.003   0.9974
## ---
## 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: 49.214  on 65  degrees of freedom
## AIC: 79.214
##
## 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+dataType,data=rslt2,controls = cforest_unbiased(mtry = 2, nt
varimp(partyForBio)
```

```
##     agSubGrp          end     dataType
## -0.002476064  0.005728681  0.014975199
```

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

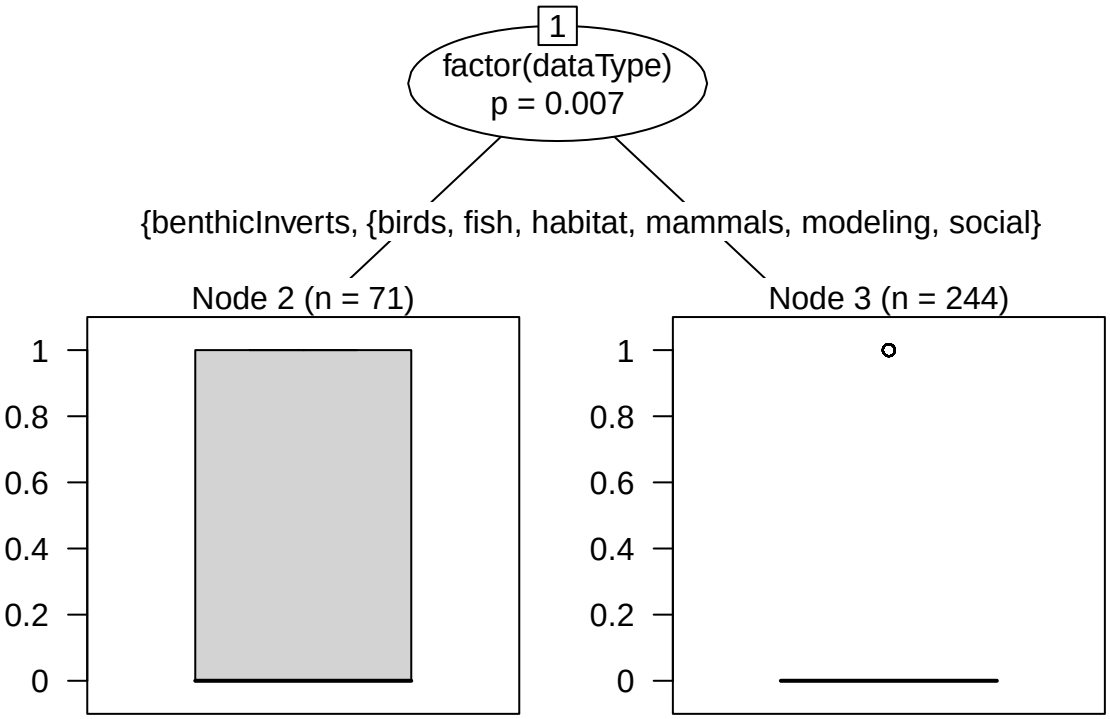
Here we run one iteration of the forest analysis above to display which variables whithin each classification determine positive or negative results.

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

partreeBio
```

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  statSucc
## Inputs:  agSubGrp, end, factor(dataType)
## Number of observations:  315
##
## 1) factor(dataType) == {benthicInverts, oil, physical, plankton}; criterion = 0.993, statistic = 25.
##   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.