Interdiscplinary ECOLOG

Simon Goring and Sarah Supp

22 November, 2015

# Training and Hiring Patterns for Interdisciplinary Ecologists

Simon J. Goring1

Sarah Supp2

1 Department of Geography, University of WIsconsin - Madison

2 School of Biology and Ecology, University of Maine

Corresponding author: Simon Goring - [goring@wisc.edu](mailto:goring@wisc.edu)

## Abstract

*Increasing emphasis has been placed on training interdisciplinary scientists in the ecological sciences. While ecology has historically been viewed as an interdisciplinary science, there is an increasing push to borrow deeply from related disciplines in an effort to tackle the "wicked problems" that require interdisciplinary solutions.*

*Using postings from ECOLOG, a mailing list strongly associated with the Ecological Society of America, we examine patterns of hiring for graduate and tenure track positions. Using text mining and a machine learning algorithm (random forest), trained on manually classified data, we estimate rates of posting for interdisciplinary positions at the tenure track and graduate/postdoctoral level.*

*Clear patterns of hiring seasonality are apparent in the data that may be of use to other researchers. Our results indicate that since 2000 the number of pre-tenure and tenure track positions advertising for interdisciplinary scholars has increased. Rates of advertising for interdisciplinary tenure track positions are lower than rates of advertising for interdisciplinary graduate and postdoctoral positions. While the rate of advertising tells us little about the actual hiring practice, it is a vetted process that often results from a committee of individuals making explicit decisions about how the position is to be filled. Our results point to an evolving system of interdisciplinarity within the ecological science, interdisciplinarity by proxy, whereby disciplinary researchers hire graduate or postdoctoral students to conduct research into interdisciplinary hypotheses, but, possibly, seek out disciplinary researchers as colleagues at the tenured level.*

## Introduction

Interdisciplinarity has been receiving increased attention in ecology, both as a method of studying key processes, but also as a subject of interest in and of itself. Part of this increased attention is the result of the increasingly complex set of problems facing modern ecologists, intertwined as they are with land use change, climate change, changes in biogeophysical processes, and human interactions (whole raft of references). To help support the increased need for interdisciplinary approaches in research a number of training programs and funding sources have been structured to address training at the graduate and post-doctoral level. These include SESYNC, IGERT, CIEE in Canada and other stuff. . . But given the challenges of interdisciplinary research (references)

## Methods

We used a web scraper written in R (**???**) to read messages from ECOLOG from the years 2000 - 2015. This program pulled the ECOLOG message, message date and message subject for each posting. Messages from ECOLOG were used as the basis for the document corpus, which was then transformed into a term matrix using text analysis tools in the tm package for R (Meyer et al. 2008, Feinerer and Hornik 2012). Data pre-processing included the removal of excess whitespace, punctuation, and HTML tags and entities (all processing is preserved in the file load\_terms.R). Because of the volume of messages, we removed all sparse terms, retaining only the top 1% by count.

We classified a number of records by hand using a Shiny application (available publically [here](http://ec2-52-0-223-101.compute-1.amazonaws.com:3838/ESA_Shiny/)). Messages were identified as non-academic job ads (generally), classified into graduate, postdoctoral, and tenure track positions, and then also classified as interdisciplinary or not. We also provided the opportunity to classify salary type, but do not present this data here.

Using the surveyed messages we built a model using a random forest to classify messages as tenure track jobs, postdoctoral positions and graduate positions, and to test whether messages indicate interdisciplinarity. Based on the model constructed we then predict the class of the remaining messages to build a time series of job ads. Since hand coding messages was an itterative process we used the confusion matrix from the predictive models to weight the messages selected for display and classification. Messages that were predicted to be interdisciplinary from the partial models were weighted most highly, as were messages with high rates of confusion for tenure track, postdoctoral or graduate positions.

From the final model we then tracked the rate of change in the classified messages. Modeling uses a binomial generalized additive model predicting the probability that a message is tenure track, graduate or postdoctoral, and whether that message is interdisciplinary.

## Results

Once we pull the messages from ECOLOG we find that we obtain 50144 messages. We clean the text in these messages to remove punctuation, extra whitespace, HTML tags, stop words and numbers using the tm\_map function from the tm package (Meyer et al. 2008, Feinerer and Hornik 2012). This leaves us with a set of 1733 terms.

The frequency of terms in the ECOLOG corpus follows a log distribution. The most frequent terms ('will', n=8.235510^{4}; 'research', n=6.195110^{4}; university, n=3.899610^{4}; 'field', n=3.500410^{4}; and 'ecology', n=3.376810^{4}) have much higher frequencies than the mean ( = 2615.9088286). Because the term matrix was constrained to the top 1% of all terms, the least common words in the term matrix ('enjoy', 'helpful' and 'speaking') still occur several hundreds of times (all n=589).

**Table 1**. *Cross-over between terms in the top 95%ile of importance for the random forest model between classification models for Tenure Track, Postdoctoral, and Graduate positions, as well as for Interdisciplinary positions.*

##   
## -----------------------------------------------------  
## &nbsp; Tenure PostDoc Graduate   
## ----------------------- -------- --------- ----------  
## \*\*PostDoc\*\* 34.5 NA NA   
##   
## \*\*Graduate\*\* 23 39.1 NA   
##   
## \*\*Interdisciplinary\*\* 37.9 46 47.1   
## -----------------------------------------------------

Four independent random forest models were also constructed to enable us to assess term importance for each individual employment class and for interdisciplinary postings. While the overall model (7 classes) has an out of bag error estimate of 18.4%, the higest error rates within any one class occur between classes and their "interdisciplinary" alternative. This indicates higher rates of uncertainty within classes, and particularly, potential, differences in the terms used for interdisciplinarity between job classes.

**Table 2.**

##   
## -----------------------------------------------------  
## &nbsp; GR GR\_Int Other PD PD\_Int TT   
## ------------ ---- -------- ------- ---- -------- ----  
## \*\*GR\*\* 132 18 12 5 0 2   
##   
## \*\*GR\_Int\*\* 15 34 0 3 1 0   
##   
## \*\*Other\*\* 125 41 1712 96 34 39   
##   
## \*\*PD\*\* 0 0 7 77 11 1   
##   
## \*\*PD\_Int\*\* 0 1 5 20 27 0   
##   
## \*\*TT\*\* 2 0 9 0 0 84   
##   
## \*\*TT\_Int\*\* 1 0 2 1 0 8   
## -----------------------------------------------------  
##   
## Table: Table continues below  
##   
##   
## -----------------------------------  
## &nbsp; TT\_Int class.error   
## ------------ -------- -------------  
## \*\*GR\*\* 1 0.22   
##   
## \*\*GR\_Int\*\* 0 0.36   
##   
## \*\*Other\*\* 6 0.17   
##   
## \*\*PD\*\* 1 0.21   
##   
## \*\*PD\_Int\*\* 1 0.5   
##   
## \*\*TT\*\* 3 0.14   
##   
## \*\*TT\_Int\*\* 23 0.34   
## -----------------------------------

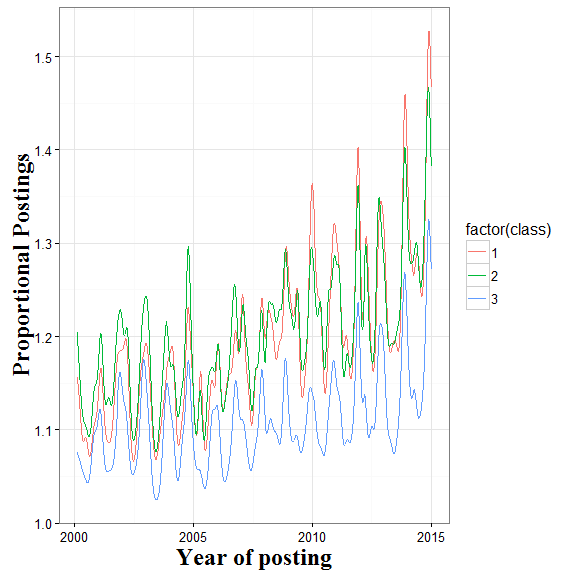
The random forest model provides an importance table, indicating the utility of a term in classifying individual messages into the various classes. The importance value represents the mean decrease in the Gini index across trees when the particular variable is used as a classifier. The tables themselves are very long, but rank of individual terms can provide some insight into the processes underlying the hiring trends in ecology. Tenure Track and Postdoctoral positions show higher cross-over of high-importance terms than do Tenure Track and Graduate positions (or Postdoctoral and Graduate positions). Both Graduate and Postdoctoral positions show much higher cross-over with Interdisciplinary terms than do terms of importance for Tenure Track postings.

**Table 2**. *Importance terms for the classification of Tenure Track, Postdoctoral, and Graduate student positions. Bolded terms represent terms that are unique within the top 99%ile of terms. Many of the top key terms for positions relate specifically to those positions (tenure track, graduate). Interestingly, among "position" terms, postdoctoral is non-unique due to its high importance within the "interdisciplinary" classification model.*

##   
## ---------------------------------------------------  
## Tenure PostDoc Graduate   
## ----------------- -------------- ------------------  
## \*\*tenuretrack\*\* postdoctoral graduate   
##   
## \*\*assistant\*\* position \*\*student\*\*   
##   
## \*\*professor\*\* research \*\*assistantship\*\*   
##   
## \*\*undergraduate\*\* university research   
##   
## \*\*teaching\*\* available available   
##   
## applications modeling \*\*assistantships\*\*  
##   
## position \*\*references\*\* phd   
##   
## \*\*faculty\*\* applications university   
##   
## department \*\*names\*\* ecology   
##   
## \*\*curriculum\*\* \*\*postdoc\*\* project   
##   
## research \*\*associate\*\* \*\*motivated\*\*   
##   
## \*\*successful\*\* project department   
##   
## \*\*candidates\*\* develop \*\*students\*\*   
##   
## \*\*vitae\*\* \*\*conduct\*\* will   
##   
## \*\*search\*\* \*\*year\*\* interests   
##   
## \*\*affirmative\*\* \*\*dynamics\*\* \*\*study\*\*   
##   
## \*\*candidate\*\* \*\*begin\*\* \*\*fall\*\*   
##   
## interests \*\*laboratory\*\* \*\*opportunity\*\*   
##   
## graduate ecology \*\*can\*\*   
##   
## application \*\*doctoral\*\* \*\*tuition\*\*   
## ---------------------------------------------------  
##   
## Table: Table continues below  
##   
##   
## ---------------------  
## Interdisciplinary   
## ---------------------  
## \*\*interdisciplinary\*\*  
##   
## research   
##   
## position   
##   
## project   
##   
## applications   
##   
## graduate   
##   
## develop   
##   
## university   
##   
## \*\*environmental\*\*   
##   
## modeling   
##   
## postdoctoral   
##   
## \*\*projects\*\*   
##   
## phd   
##   
## ecology   
##   
## application   
##   
## available   
##   
## \*\*applicants\*\*   
##   
## will   
##   
## \*\*employer\*\*   
##   
## interests   
## ---------------------

The random forest model to explicitly identify Tenure Track positions indicates that terms of high importance include "tenuretrack" (indicating the removal of a hyphen), "assistant", "professor" and "undergraduate" (Table 2). For postdoctoral positions, it is again, postdoctoral, position, and then research that are most important. The key term in this case is "interdisciplinary". We're really looking at interdisciplinary jobs here, so it's no surprise that terms like "position", "project", "applications" show up, but the importance rank of "postdoctoral" (rank = 5) is much higher than the importance of "tenuretrack" (rank = 215) or "professor" (rank = 215), indicating a much higher incidence of interdisciplinary advertisments for graduate and postdoctoral positions than for professors.

rep.row<-function(x,n){  
 matrix(rep(x,each=n),nrow=n)  
}  
  
weeks <- do.call(rbind.data.frame, lapply(1:length(all.messages),function(x)rep.row(myw2[x,],length(all.messages[[x]]))))  
  
r\_preds$time <- unlist(weeks[,3]) + (unlist(weeks[,2]) + unlist(weeks[,1])/4)/12  
  
r\_preds <- subset(r\_preds, time <= 2015)  
  
interdiscip\_gam <- gam(as.logical(predict\_in)~s(time, k = 120),   
 data = r\_preds, family='binomial')  
inter\_tt\_gam <- gam(I(as.logical(predict\_in)&as.logical(predict\_tt))~s(time, k = 120),   
 data = r\_preds, family='binomial')  
inter\_gg\_gam <- gam(I(as.logical(predict\_in)&as.logical(predict\_gr))~s(time, k = 120),   
 data = r\_preds, family='binomial')  
inter\_pd\_gam <- gam(I(as.logical(predict\_in)&as.logical(predict\_pd))~s(time, k = 120),  
 data = r\_preds, family='binomial')  
  
inter\_models <- list(inter\_gg\_gam,  
 inter\_pd\_gam,  
 inter\_tt\_gam)  
  
plot\_gam <- do.call(rbind.data.frame,   
 lapply(1:3, function(x){  
 data.frame(time = r\_preds$time,  
 pred = exp(predict(inter\_models[[x]], type = 'response')),  
 class = x)}))  
  
# plot\_gam <- data.frame(time = r\_preds$time,  
# value= c(predict(inter\_gg\_gam, type = 'response'),  
# predict(inter\_pd\_gam, type = 'response'),  
# predict(inter\_tt\_gam, type = 'response')),  
# class = rep(c('Interdisciplinary',  
# 'Graduate',  
# 'Postdoc',  
# 'Tenure Track'),   
# each = length(predict(inter\_gg\_gam))))  
  
ggplot(subset(plot\_gam, time <= 2015),   
 aes(x = time, y = pred, color = factor(class))) +  
 geom\_path() +  
 theme\_bw() +  
 #coord\_cartesian(xlim=c(2000, 2015), ylim=c(0, 0.6)) +  
 xlab('Year of posting') +  
 ylab('Proportional Postings') +  
 theme(axis.title = element\_text(family = 'serif', face = 'bold', size = 18))

 ##Discussion

## References

Feinerer, I., and K. Hornik. 2012. Tm: Text mining package. R package version 0.5-7.1.

Meyer, D., K. Hornik, and I. Feinerer. 2008. Text mining infrastructure in r. Journal of Statistical Software 25:1–54.