A Vignette on Laminations and Climate

Christina Knudson, PhD

03/26/2018

Overview

Very generally, this vignette describes our process of finding associations between climate regimes and laminated intervals. In particular, our research questions are:

- are glacial measurements unusually likely/unlikely during laminations?
- are interglacial measurements unusually likely/unlikely during laminations?

In this vignette, we first establish definitions. Second, we cover the hypotheses, the test statistic, and the intuition behind the tests. Next, we demonstrate the `Bering` functions and methods necessary to conduct the test. Then, we demonstrate visualizing the results using the `Bering` output. Finally, we discuss the computational cost of this procedure.

Definitions

The *climate* regimes are:

1. Glacial (high d18O values)
2. Interglacial (low d18O values)
3. Transition

We define a measurement as "glacial" if it is 1 standard deviation above the 100 kyr moving average of d18O. We define a measurement as "interglacial" if it is 1 standard deviation below the 100 kyr moving average of d18O. All other measurements are defined as "transition."

We define *"unusually likely/unlikely"* using measurements from laminated and non-laminated intervals. If there is no relationship between climate and laminations, then the proportion of laminated intervals with a glacial measurement will not change when we randomly reassign each measurement to laminated or nonlaminated. Similarly, if there is no relationship between climate and laminations, then the proportion of laminated intervals with an interglacial measurement will not change when we reassign each measurement to laminated or nonlaminated.

Hypotheses, test statistics, and intuition

We will conduct one test for the glacial climate regime and another test for the interglacial climate regime. Because the hypotheses and statistical analyses are almost identical, we list these for the glacial climate regime. For the interglacial regime, simply substitute "glacial" for "interglacial".

The null hypothesis is that the occurrence of glacial measurements is NOT related to an interval being laminated. The alternative hypothesis is that the occurrence of glacial measurements is related to an interval being laminated. Our test statistic is the proportion of laminated intervals with a glacial measurement.

Using Monte Carlo, we simulate 10000 data-sets and test statistics under the null hypothesis. To simulate each data-set, we randomly assign each measurement to one of two states (laminated or non-laminated) using a Markov chain with empirical transition probabilities calculated using the original data set.

If the original data's test statistic is relatively large compared to the distribution of simulated test statistics, we will reject the null hypothesis and conclude that the occurrence of glacial measurements is significantly related to an interval being laminated. In particular, our p-value is the proportion of test statistics as extreme or more extreme than the original data's test statistic.

This vignette uses `mockd18O`, which has d18O values, real measurement ages, and and simulated laminations. We will update the `Bering` package and this vignette to include the true lamination data if/when our manuscript is published.

Exploring the data

Begin by invoking the `Bering` package and `mockd18O` data.

```
library(Bering)
data(mockd18O)
```

At this point, I highly recommend conducting exploratory data analysis to become acquainted with the data. In particular, I recommend using the `summary` command (in base R) to understand the data.

```
summary(mockd18O)
```

```
##       age            d18Osmooth      windowmean       windowsd
##  Min.   :  12.08   Min.   :3.440   Min.   :3.746   Min.   :0.07343
##  1st Qu.: 307.40   1st Qu.:3.809   1st Qu.:3.894   1st Qu.:0.13027
##  Median : 602.71   Median :3.956   Median :3.938   Median :0.16239
##  Mean   : 602.71   Mean   :3.951   Mean   :3.950   Mean   :0.16572
##  3rd Qu.: 898.03   3rd Qu.:4.074   3rd Qu.:4.009   3rd Qu.:0.20305
##  Max.   :1193.30   Max.   :4.521   Max.   :4.293   Max.   :0.30615
##  mocklam
##  0:660
##  1:193
##
##
##
##
```

Our data frame `mockd18O` has three variables: `age`, `d18Osmooth`, `windowmean`, `windowsd`, and `mocklam`. Information on these variables can be found by typing the following into the `R` console:

```
?mockd18O
```

After completing exploratory data analysis, we ensure that the measurements are in order by age. The code below demonstrates one way to perform this check. TRUE indicates the measurements are indeed in order.

```
sum(mockd18O$age == sort(mockd18O$age)) == length(mockd18O$age)
```

```
## [1] TRUE
```

Summarizing the original data

Before we can conduct the test, we need to categorize each measurement's climate as glacial, interglacial, or transition using each measurement's z-score. We denote these with `G`, `IG`, and `Tr`, respectively.

```
mockd18O$zscore <- rep(-100, nrow(mockd18O))
mockd18O$climate <- rep("Tr", nrow(mockd18O))
for(i in 1:nrow(mockd18O)){
  mockd18O$zscore[i] <- (mockd18O$d18Osmooth[i]-mockd18O$windowmean[i])/mockd18O$windowsd[i]
  if(mockd18O$zscore[i]>=1) {mockd18O$climate[i] <- "G"}
  if(mockd18O$zscore[i]<=-1) {mockd18O$climate[i] <- "IG"}
}
```

Next, we use our original data set to calculate two empirical probabilities:

- the empirical probability of a laminated measurement following a nonlaminated measurement
- the empirical probability of a laminated measurement following another laminated measurement.

The `peas` function from `Bering` calculates these probabilities using a single input: the vector indicating which measurements are laminated.

```
mypea <- peas(mockd18O$mocklam)
(pswitchL <- mypea$pswitchL)
```

```
## [1] 0.2090909
```

```
(pstayL <- mypea$pstayL)
```

```
## [1] 0.2901554
```

Next, we calculate our two test statistics: the proportion of laminated intervals containing a glacial measurement and the the proportion of laminated intervals containing an interglacial measurement.

```
out <- countGIG(mockd18O$climate, mockd18O$mocklam)
(Gteststat <- out$Gcount/out$lamintcount)
```

```
## [1] 0.1654676
```

```
(IGteststat <- out$IGcount/out$lamintcount)
```

```
## [1] 0.1726619
```

Conducting the test

To begin our test, we specify a Monte Carlo sample size of 10000. For reproducible results, we set the seed. `IGdistrib` and `Gdistrib` are vectors of length 10000 that will store the test statistics simulated under the null hypothesis.

```
m <-10^4
set.seed(1234)
nobs <- length(mockd18O$climate)
IGdistrib <- Gdistrib <-rep(0,m)
```

Next, we use a loop to simulate our test statistics. We randomly assign each measurement as laminated or non-laminated and then calculate the proportion of laminated intervals that contain a glacial/interglacial measurement.

```
for(i in 1:m){
     # randomly assign measurements as lam or nonlam
     states <- assignlam(nobs, pswitchL, pstayL)

     # create sampling distribution under null hypothesis by
     # calculating prop of lams with a G (or IG) measurement
     out <- countGIG(mockd18O$climate, states)
     Gdistrib[i] <- out$Gcount/out$lamintcount
     IGdistrib[i] <- out$IGcount/out$lamintcount
}
```

We now have the information necessary to calculate p-values. Let's calculate the p-value for the glacial climate regime first. We add one to the numerator and one to the denominator to include our original data's test statistic. Because we are conducting a two-sided test and will reject the null if we see an extreme enough test statistic in either direction (either tail), we calculate the p-value by doubling the proportion of test statistics in the smaller tail.

```
Gtop <- min(sum(Gdistrib >= Gteststat) + 1, sum(Gdistrib <= Gteststat) + 1)
Gbottom <- m+1
(Gpvalue <- 2*Gtop/Gbottom)
```

```
## [1] 0.1931807
```

Now we calculate the p-value for interglacial using the same logic.

```
IGtop <- min(sum(IGdistrib >= IGteststat) + 1, sum(IGdistrib <= IGteststat) + 1)
IGbottom <- m+1
(IGpvalue <- 2*IGtop/IGbottom)
```

```
## [1] 0.4665533
```

Because the p-values are large (.193 and .467), we do not reject the null hypothesis. Based on the mock laminations data, we do not have evidence to say that climate is related to laminated intervals.

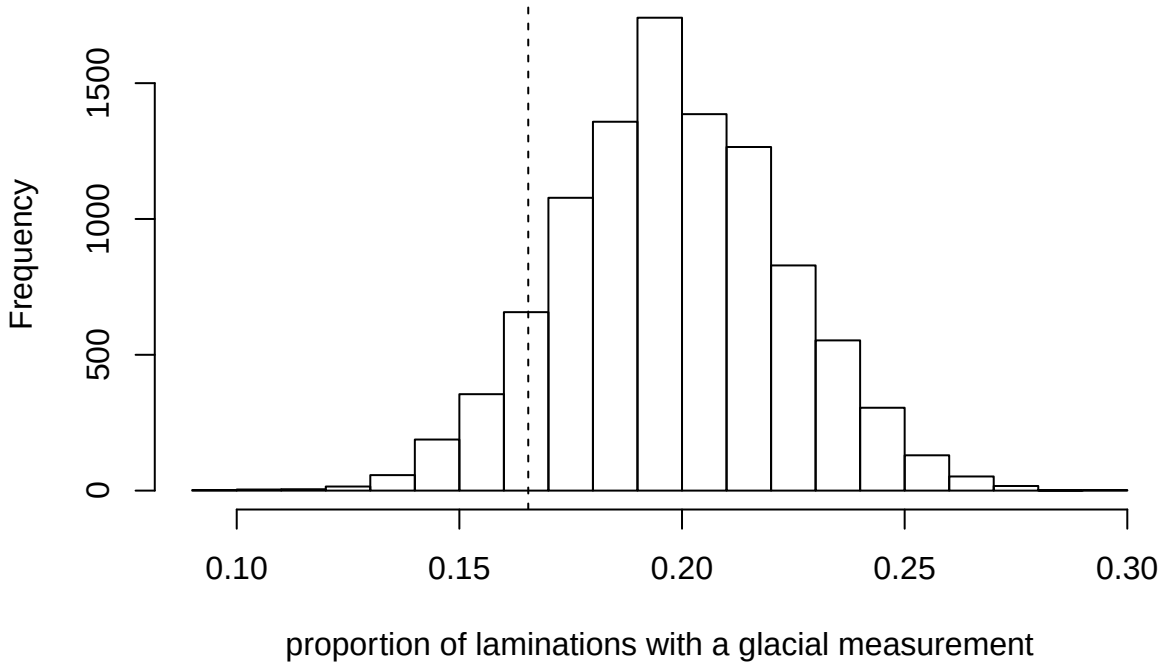
- The occurrence of glacial measurements is unrelated to laminated intervals.
- The occurrence of interglacial measurements is unrelated to laminated intervals.

Visualizing the results

To better understand our results, we can visualize the simulated test statistics and compare them to the original data's test statistic. We first create a histogram using the simulated test statistics and then we draw a vertical dashed line at our original data's test statistic. Because we conducted two tests (one for glacial, the other for interglacial), we produce two plots.

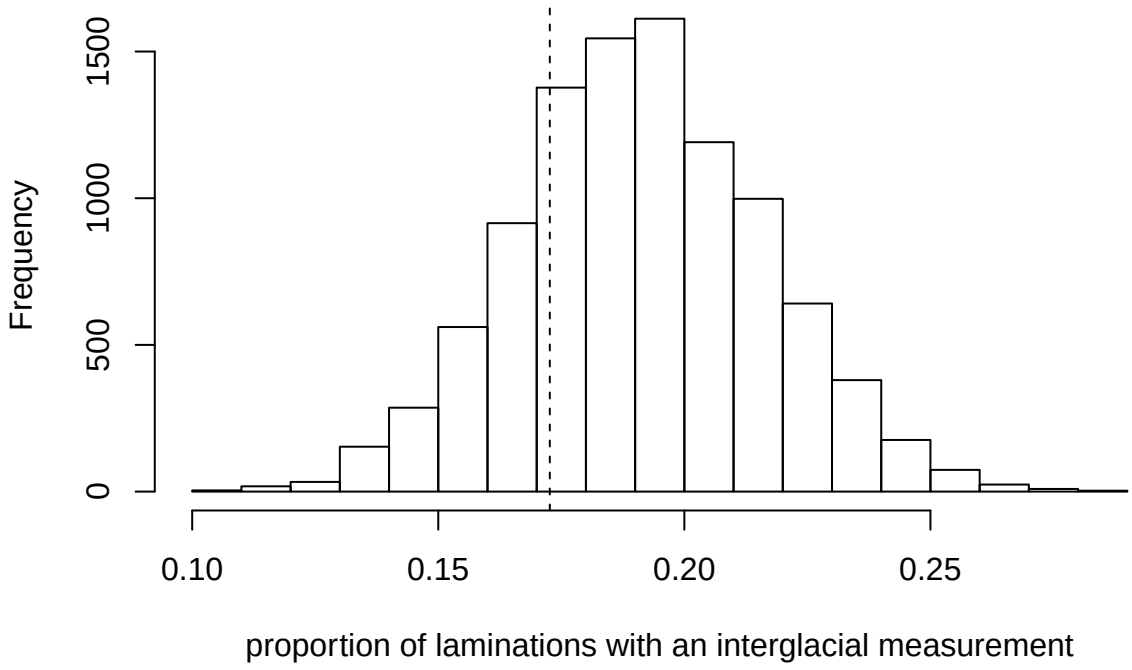
The following plot displays the sampling distribution (under the null hypothesis) for the proportion of lamnated intervals containing a glacial measurement. The dotted line shows the test statistic from our original data set. You can see our test statistic is not unusual extreme compared to those built under the null hypothesis (no relationship between climate and laminations).

```
hist(Gdistrib, main=NULL, xlab ="proportion of laminations with a glacial measurement")
abline(v=Gteststat, lty=2)
```



The following plot displays the sampling distribution (under the null hypothesis) for the proportion of lamnated intervals containing an interglacial measurement. The dotted line shows the test statistic from our original data set. You can see our test statistic is not unusual extreme compared to those built under the null hypothesis (no relationship between climate and laminations).

```
hist(IGdistrib, main=NULL, xlab ="proportion of laminations with an interglacial measurement")
abline(v=IGteststat, lty=2)
```



Computational cost

Compiling this vignette on a 'normal' computer (Windows 10 with 8 GB of RAM) took about 30 seconds.