Importing and Analyzing Pb Isotope and Concentration Data

Marc Los Huertos

12/3/2017

## Purpose of Document

I have written this to help you read csv file and manipulate Pb data into R. Then provide some tools in the analysis. Please do not just copy and paste commands, or you will be terribly frustrated, becuase this was not designed for that purpose.

However, if you use this document as a "basis" for your lab report, customizing the headers, r code, and text as you go, you will have the word document that you can use as a template for your lab.

### Work Flow suggestions

I suggest using the following work flow:

* Create a new Rmd File
* Read each section in order and adjust the text and code as you go.
* Revise the Rmd to create a nice word document with figures while removing the R code that the reader does not need to see in the final report.
* Create a word document and edit the document to create a lab report using the report guidelines on Sakai.

## Finding the Data

First, we need to convert the excel file into comma separated values, which we can do by saving the file as a csv. This will give a few warnings, but don't worry about that.

Now you'll need to upload the data into Rstudio.

Next we need is to find the data -- the name and the directory path. With this, we can then tell R where the data reside and create an "object" that refers to the path and filename.

I like using the **file.choose()** function, because it provides a popup window that we can use to search for the file. Use this function in the console and copy the path and file name to insert into your Rmd and create an object that references the path and filename. I have done this below for **my** file and path, yours will be different!!

file = "/home/CAMPUS/mwl04747/github/Environmental\_Hazards/Data/Pb\_Data/171116\_EA30\_MLH4\_revised.csv"

Notice that in the Rblock above the the r block as a label "filename". This helps me figure out where I am when I find a knitting problem. You can leave that blank or put in a label. If you do put in a label, they need to be unique -- repeated ones will give an error.

## Reading Data into R

To read data into R we use **read.csv()** function

import = read.csv(file)

Whenever I read data, I check to make sure the object I created is what I had expected and intended.

### Checking the file

1. What are the variable names?

names(import)

## [1] "Sample.Name" "Park" "Replicate" "Analyte"   
## [5] "Mass" "Concentration"

What is the structure?

1. What is the structure of the object?

str(import)

## 'data.frame': 480 obs. of 6 variables:  
## $ Sample.Name : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Park : Factor w/ 7 levels "B","Blank","C",..: 6 6 6 6 6 6 6 6 6 6 ...  
## $ Replicate : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Analyte : Factor w/ 14 levels "Ag","As","Ba",..: 8 12 5 6 2 13 11 1 7 3 ...  
## $ Mass : int 6 45 52 72 75 78 103 107 115 137 ...  
## $ Concentration: num 0 0 0.02 0 -77.02 ...

A dataframe is the most common data structure in R. Notice the dollar signs -- these can be used to select specific variables within the dataframe.

## Preparing Data for Analysis

### Tranform data mg/kg of soil

The concentration from the ICP-MS was based on the extraction concentration, so we need to divide by the soil weight and multipy by the amount of dilution solution (water and nitric acid). I have created a new variable name, "Result", so I don't get mixed up and remove the old variable name.

import$Result = 25/5\*import$Concentration  
drops <- c("Concentration")  
import = import[ , !(names(import) %in% drops)]

Again, it's a good idea to check the data:

head(import)

## Sample.Name Park Replicate Analyte Mass Result  
## 1 1 R 1 Li 6 0.000  
## 2 1 R 1 Sc 45 0.000  
## 3 1 R 1 Cr 52 0.100  
## 4 1 R 1 Ge 72 0.000  
## 5 1 R 1 As 75 -385.095  
## 6 1 R 1 Se 78 0.000

### Subset Pb Analytes

Let's subset the data -- for now, we will only look at Pb. NOTE: Lab reports that also analyze other analytes will be awarded more points.

We can subset the rows where the Analyte = Pb. Unfortunately, R's method to do this is a bit obtuse:

Pb\_data = import[import$Analyte=="Pb",]  
head(Pb\_data)

## Sample.Name Park Replicate Analyte Mass Result  
## 13 1 R 1 Pb 206 0.035  
## 14 1 R 1 Pb 207 0.030  
## 15 1 R 1 Pb 208 0.070  
## 29 2 R 2 Pb 206 0.035  
## 30 2 R 2 Pb 207 0.030  
## 31 2 R 2 Pb 208 0.070

### Remove Blank

It turns out removing our blank is trickier than I had anticipated. we need to "Install" a library or package call *gdata*. NOTE: A report discussion about the role of the Blank and our results will be given more points.

Pb\_data = Pb\_data[!Pb\_data$Park=="Blank", ]  
levels(Pb\_data$Park)

## [1] "B" "Blank" "C" "G" "P" "R" "W"

library(gdata)

## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

##

## gdata: Unable to load perl libaries needed by read.xls()  
## gdata: to support 'XLSX' (Excel 2007+) files.

##

## gdata: Run the function 'installXLSXsupport()'  
## gdata: to automatically download and install the perl  
## gdata: libaries needed to support Excel XLS and XLSX formats.

##   
## Attaching package: 'gdata'

## The following object is masked from 'package:stats':  
##   
## nobs

## The following object is masked from 'package:utils':  
##   
## object.size

## The following object is masked from 'package:base':  
##   
## startsWith

drop.levels(Pb\_data$Park)

## [1] R R R R R R R R R B B B B B B P P P W W W P P P P P P P P P W W W C C  
## [36] C P P P G G G B B B B B B W W W C C C W W W W W W G G G R R R C C C R  
## [71] R R C C C C C C G G G G G G G G G  
## Levels: B C G P R W

levels(Pb\_data$Park)

## [1] "B" "Blank" "C" "G" "P" "R" "W"

## Wrangle Data Format

Wrangling data is reformats the data -- moving columns into rows based on a 'key'. In our case, we need to separate the 3 isotopes so we can add them together and make a new variable, TotalPb!

To do this, I'll create a new file name. The *spread()* function is a bit complicated, so I can't go into it, but check to see how the data change!

library(tidyr)  
  
Pb = Pb\_data %>% spread(Mass, Result)  
  
Pb$Pb67 = round(Pb$'206'/Pb$'207',3)  
Pb$Pb68 = round(Pb$'206'/Pb$'208',3)  
Pb$Total = Pb$'206'+ Pb$'207'+ Pb$'208'  
Pb

## Sample.Name Park Replicate Analyte 206 207 208 Pb67 Pb68 Total  
## 1 1 R 1 Pb 0.035 0.030 0.070 1.167 0.500 0.135  
## 2 2 R 2 Pb 0.035 0.030 0.070 1.167 0.500 0.135  
## 3 3 R 3 Pb 0.085 0.065 0.165 1.308 0.515 0.315  
## 4 4 B 4 Pb 0.075 0.060 0.150 1.250 0.500 0.285  
## 5 5 B 5 Pb 0.060 0.045 0.115 1.333 0.522 0.220  
## 6 6 P 3 Pb 0.025 0.020 0.045 1.250 0.556 0.090  
## 7 7 W 1 Pb 0.090 0.070 0.180 1.286 0.500 0.340  
## 8 8 P 5 Pb 0.020 0.015 0.035 1.333 0.571 0.070  
## 9 9 P 1 Pb 0.025 0.020 0.050 1.250 0.500 0.095  
## 10 10 P 2 Pb 0.025 0.020 0.050 1.250 0.500 0.095  
## 11 11 W 3 Pb 0.045 0.040 0.090 1.125 0.500 0.175  
## 12 12 C 3 Pb 0.030 0.025 0.060 1.200 0.500 0.115  
## 13 13 P 4 Pb 0.025 0.020 0.045 1.250 0.556 0.090  
## 14 14 G 3 Pb 0.100 0.080 0.200 1.250 0.500 0.380  
## 15 15 B 2 Pb 0.035 0.025 0.065 1.400 0.538 0.125  
## 16 16 B 1 Pb 0.050 0.040 0.090 1.250 0.556 0.180  
## 17 17 W 2 Pb 0.065 0.050 0.130 1.300 0.500 0.245  
## 18 18 C 2 Pb 0.045 0.035 0.090 1.286 0.500 0.170  
## 19 19 W 4 Pb 0.095 0.080 0.200 1.188 0.475 0.375  
## 20 20 W 5 Pb 0.085 0.070 0.175 1.214 0.486 0.330  
## 21 21 G 4 Pb 0.100 0.085 0.205 1.176 0.488 0.390  
## 22 22 R 4 Pb 0.045 0.040 0.090 1.125 0.500 0.175  
## 23 24 C 1 Pb 0.045 0.035 0.085 1.286 0.529 0.165  
## 24 25 R 4 Pb 0.075 0.060 0.155 1.250 0.484 0.290  
## 25 26 C 5 Pb 0.040 0.030 0.070 1.333 0.571 0.140  
## 26 27 C 4 Pb 0.045 0.035 0.090 1.286 0.500 0.170  
## 27 28 G 5 Pb 0.060 0.050 0.125 1.200 0.480 0.235  
## 28 29 G 1 Pb 0.090 0.070 0.170 1.286 0.529 0.330  
## 29 30 G 2 Pb 0.075 0.060 0.155 1.250 0.484 0.290

## Creating Hypotheses to Test

There are numerous hyptheses that one can test:

* Are [Pb] elevated?
* Is [Pb] above an USEPA threshold?
* What is the source of the Pb?
* Do the parks have different [Pb]?
* If the parks differ, does the distance from a source explain the differences?
* If the parks differ, does some other mechanism explain the differences (pH, moisture, etc.)?

I will give you two examples and some ideas of how to set up the other hypotheses.

## Analyze Basic Hypothesis

Are the park [Pb] different? Our null hypothesis is that the means of each park are the same, or = \_{...} = \_n$.

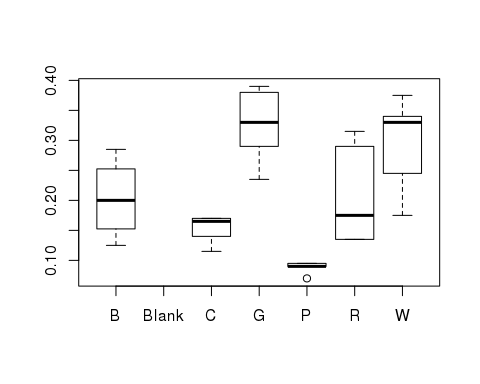
We can analyze this with a ANOVA (in R using *aov()*):

Pb.aov = aov(Total ~ Park, data=Pb)  
summary(Pb.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Park 5 0.19191 0.03838 9.873 3.84e-05 \*\*\*  
## Residuals 23 0.08942 0.00389   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

To visualize this we might use a boxplot:

boxplot(Total ~ Park, data=Pb)



## Analyze Data relative to "anthropogenic sources"

exceed = sum(Pb$Pb67 > 1.14); exceed

## [1] 27

prop.test(exceed, length(Pb$Pb67))

##   
## 1-sample proportions test with continuity correction  
##   
## data: exceed out of length(Pb$Pb67), null probability 0.5  
## X-squared = 19.862, df = 1, p-value = 8.324e-06  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.7578883 0.9879640  
## sample estimates:  
## p   
## 0.9310345

## Distance to Park from I10

Need to create a value for each park...

Pb$Distance = NA  
Pb$Distance[Pb$Park=="C"] <- 3.2  
head(Pb)

## Sample.Name Park Replicate Analyte 206 207 208 Pb67 Pb68 Total  
## 1 1 R 1 Pb 0.035 0.030 0.070 1.167 0.500 0.135  
## 2 2 R 2 Pb 0.035 0.030 0.070 1.167 0.500 0.135  
## 3 3 R 3 Pb 0.085 0.065 0.165 1.308 0.515 0.315  
## 4 4 B 4 Pb 0.075 0.060 0.150 1.250 0.500 0.285  
## 5 5 B 5 Pb 0.060 0.045 0.115 1.333 0.522 0.220  
## 6 6 P 3 Pb 0.025 0.020 0.045 1.250 0.556 0.090  
## Distance  
## 1 NA  
## 2 NA  
## 3 NA  
## 4 NA  
## 5 NA  
## 6 NA

## Hypotheses Testing

1. Is Pb Correlated to Distance?

Null Hypothesis: There is no relationship between Distance and Pb concentrations.

#Pb.lm <- lm(Lead ~ Distance10Bogus, data=import)  
#summary(Pb.lm)

## Hypothesis Testing

Null Hypothesis: There is no relationship between Nitrate and Salinity.

#pH.lm <- lm(pH ~ Salinity, data=import)  
#summary(pH.lm)

## Ploting Best Fit Line

What is r-squared? It tells us how much of the variation in the response is explained by the predictor. See the variation around the line? Only 48.6% of the data is explained by thes data, while some 51% is unexplained.

#plot(Lead ~ Distance10Bogus, data=import)  
#abline(coef(Pb.lm))

## Are the park Pb concentratins different?

#boxplot(Lead ~ Location, data = import, xlab="Park", ylab="[Pb] (mg/L)")

## Testing the Hypothsis

Null Hypothesis: There is no difference in Pb concentrations between parks.

#summary(aov(Lead ~ Location, data = import))

## Is there a relationship between Pb and Distance

#plot(Lead ~ Distance10Bogus, data=import)