



www.epa.gov/research

Safe and Sustainable Water Resources Research

Who am I and why am I here: Lakes, Linked Data, and R



Jeffrey W. Hollister

US EPA, Atlantic Ecology Division

Semantic Web/Linked Data Users Community

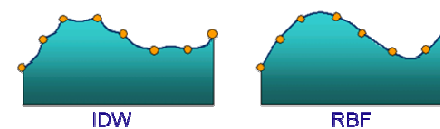
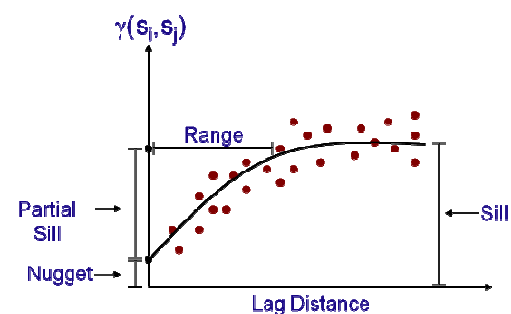
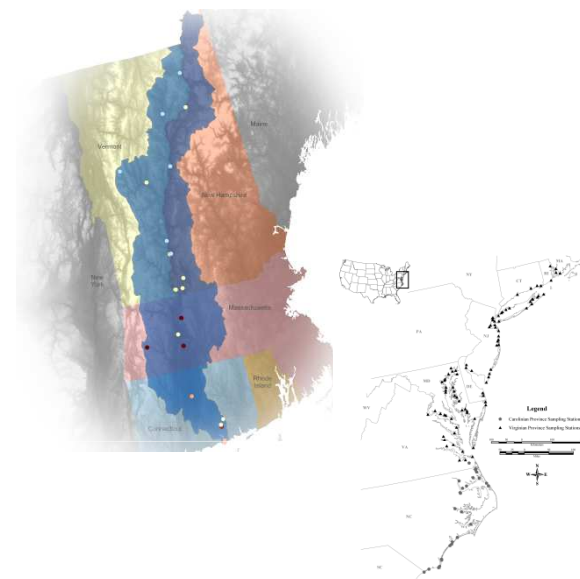
15 August 2012

Talk Outline

- Introduction
 - Who am I?
 - Why am I here?
- Lakes
 - Lake Volume
 - Lake Depth
 - Linked Data Plans
- R
 - Overview
 - How I use it

Who am I?

- Landscape Ecologist
 - GIS and Statistics
 - Research: Landscape Structure and Water Quality
- Past Work Experience
 - J.W. Jones Ecological Research Center
 - National Ecological Observatory Network (NEON)
 - Introduced to Ecoinformatics
- Current Position
 - Research Ecologist with US EPA
 - Introduced to Reproducible Research

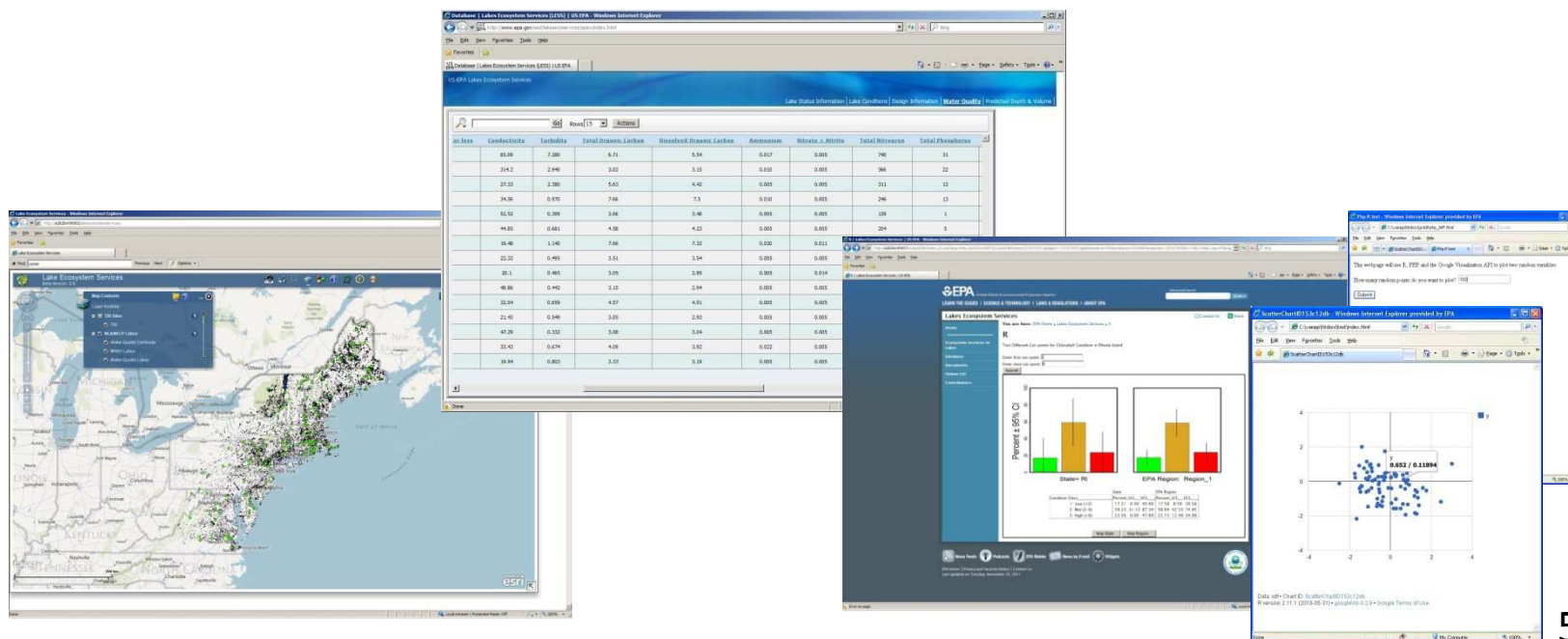


Who am I?

- My Agency and Mission
 - US Environmental Protection Agency
 - Office of Research and Development
 - National Health and Environmental Effects Research Lab,
 - » Atlantic Ecology Division
 - Monitoring and Assessment Branch
 - Research Ecologist

Why am I here?

- New Focus
 - Informatics, Decision Support, Tool Development



Northeast Lakes Projects

- Multiple Research Plans and Years
 - 2007-Present
- Common Denominators
 - Lakes
 - Nutrients
- Research Questions:
 - How do changes in nutrients change delivery of ecosystem services/risk associated with Cyanobacteria?
- Project Goals:
 - Data Sharing
 - Reproducible Research
 - Decision Support



Ecosystem Services in Lakes



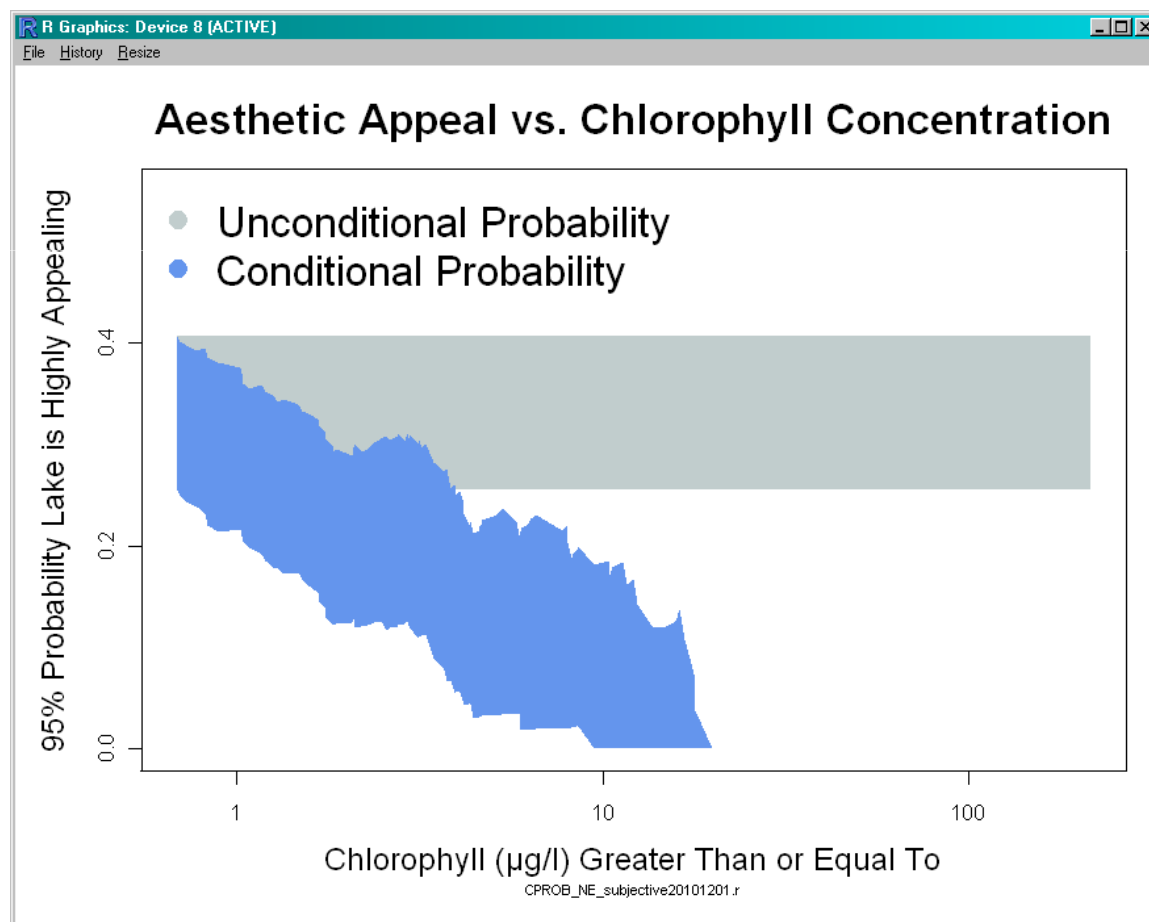
- Swimming
- Fishing
- Drinking Water
- Property Values
- Existence Value
- Aesthetics

- Aesthetic Appeal
- Disturbance
- Biotic Integrity
- Recreational Value
- Swimmability



Written Comments from Lakes in Highest Appeal Categories

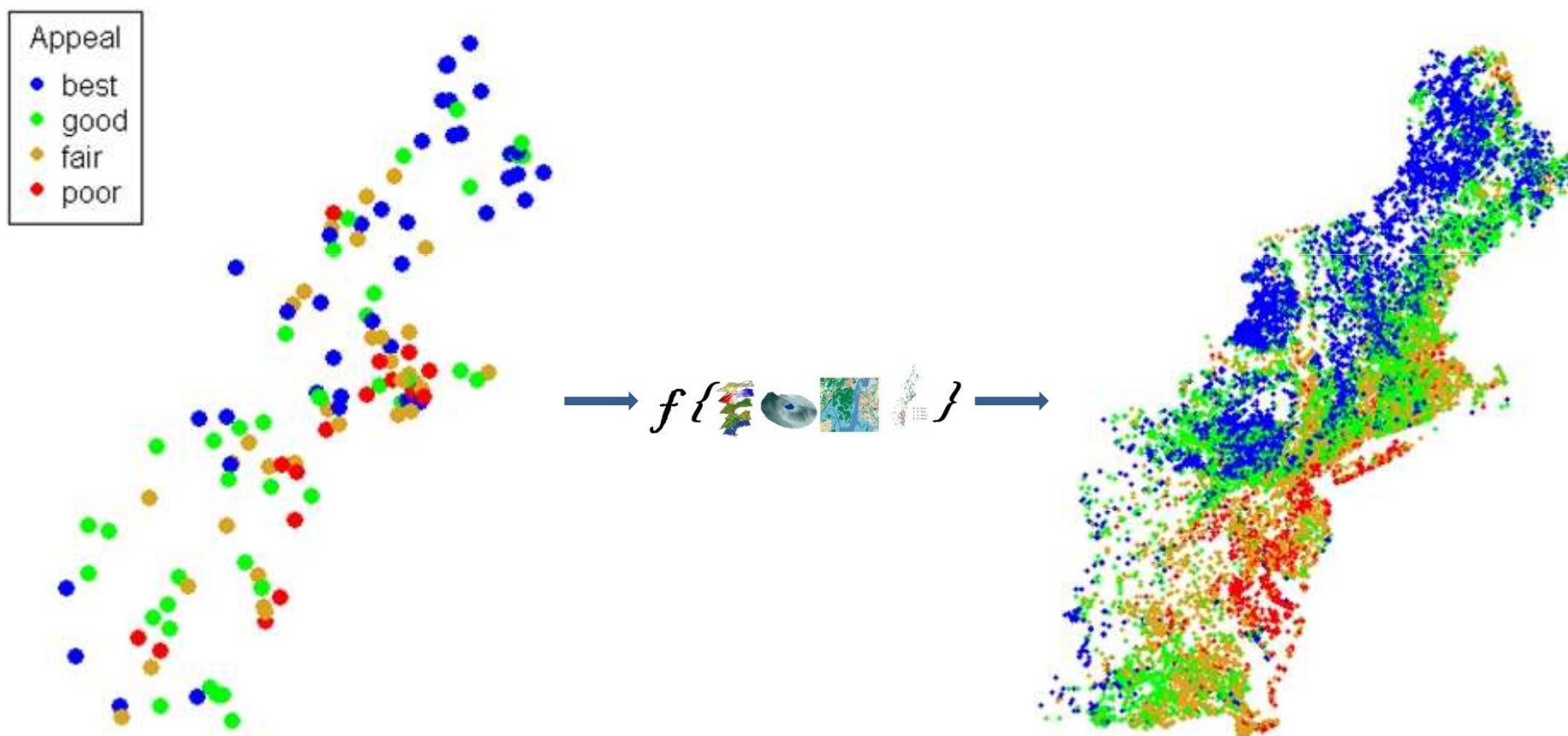
Nutrients and Ecosystem Services



1.) Start with field data

2.) Combine with landscape data in function

3.) Predict Appeal for ~18,000 Lakes



Lake Morphometry

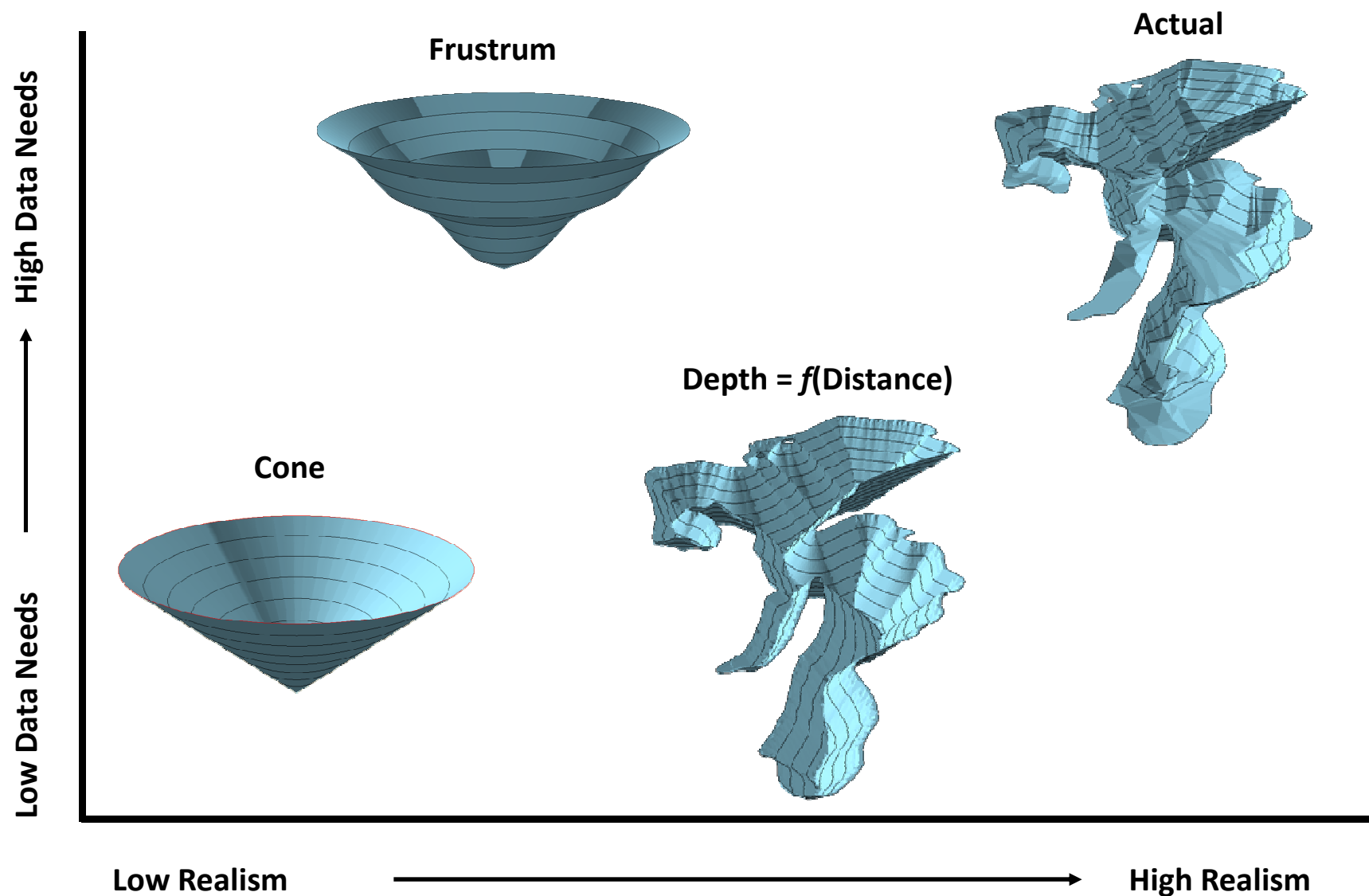
- Ordered Logit Models
 - Need Residence time
- Existing data
 - Limited resources
 - ~18,000 Lakes



Question #1

What is the best way to estimate lake volume given, lake shoreline and maximum depth?

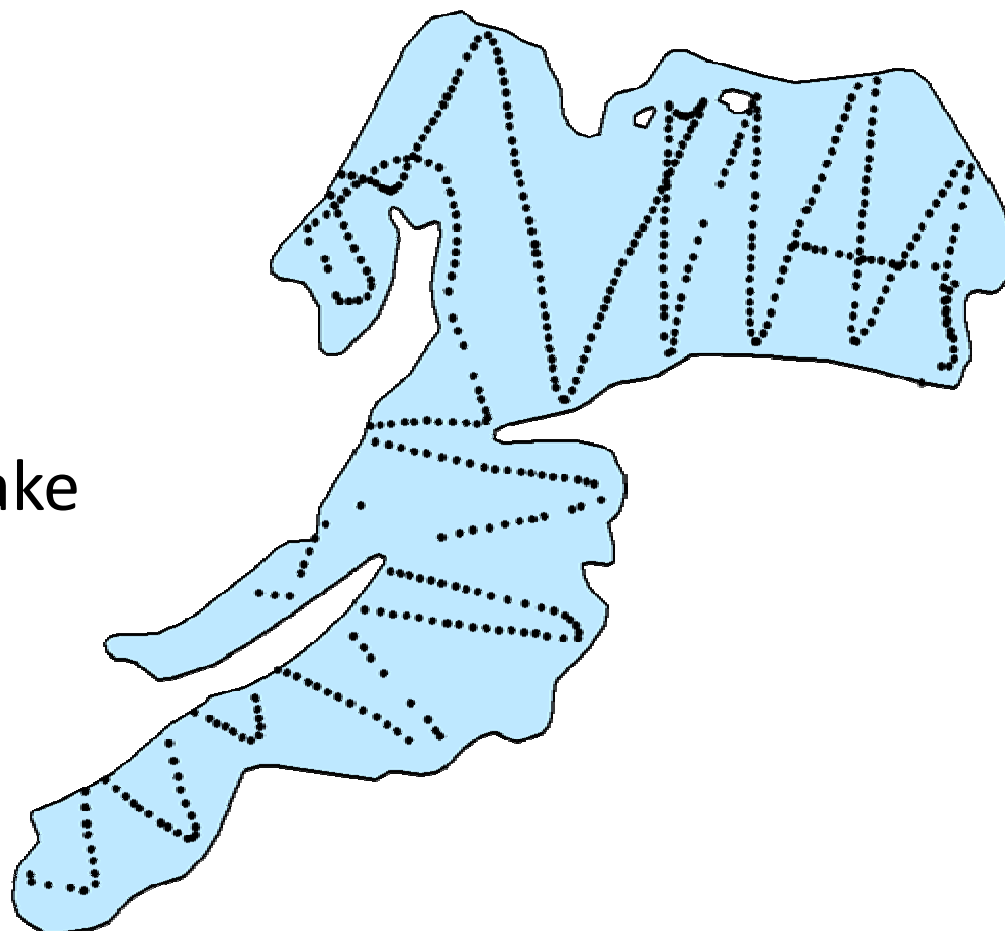
Citation: Hollister, J. W., W.B. Milstead (2010). Using GIS to Estimate Lake Volume from Limited Data. *Lake and Reservoir Management*. 26(3)194-199. Contribution no. AED-10-018.



Methods

Partridge Lake
Bathymetry Data

- Accuracy Assessment
 - Bathymetry data
 - NH DES for 132 lakes
 - Created TIN for each lake
 - Calculated volumes
 - Cone v TIN
 - GIS Method v TIN



Results - Volume Error Analysis

Method	RMSD	MD	MAD	P(Better)
GIS – All Lakes	3,287,360	8622	200734	0.59
Cone – All Lakes	6,975,740	608967	225502	0.41

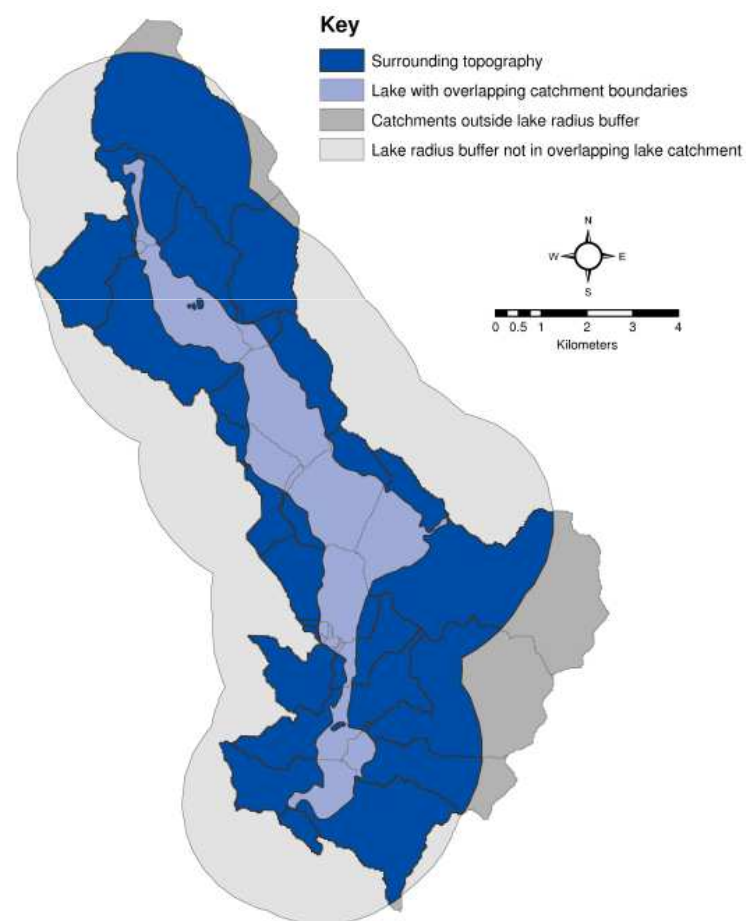
Estimating Maximum Lake Depth: Question #2

- Method in Question #1 assumes a measurement of maximum lake depth is available
- Is it possible to create a reasonable estimate of lake depth from the topography surrounding a lake?

Citation: Hollister, J. W., W.B. Milstead, M.A. Urrutia (2011). Predicting Maximum Lake Depth from Surrounding Topography. *PLoS ONE* 6(9): e25764. doi:10.1371/journal.pone.0025764. Contribution no. AED-11-013

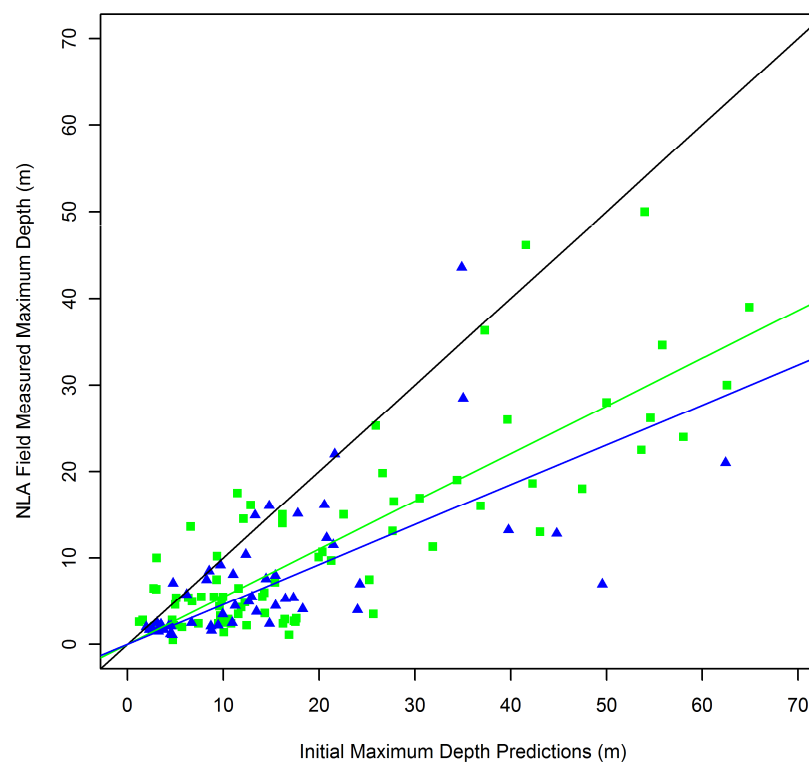
Predicting Maximum Lake Depth

- Select surrounding topography
- Determine median slope
- Determine maximum distance in lake
- Depth
 - $\text{Max.Dist} * \text{Median.Slope}$



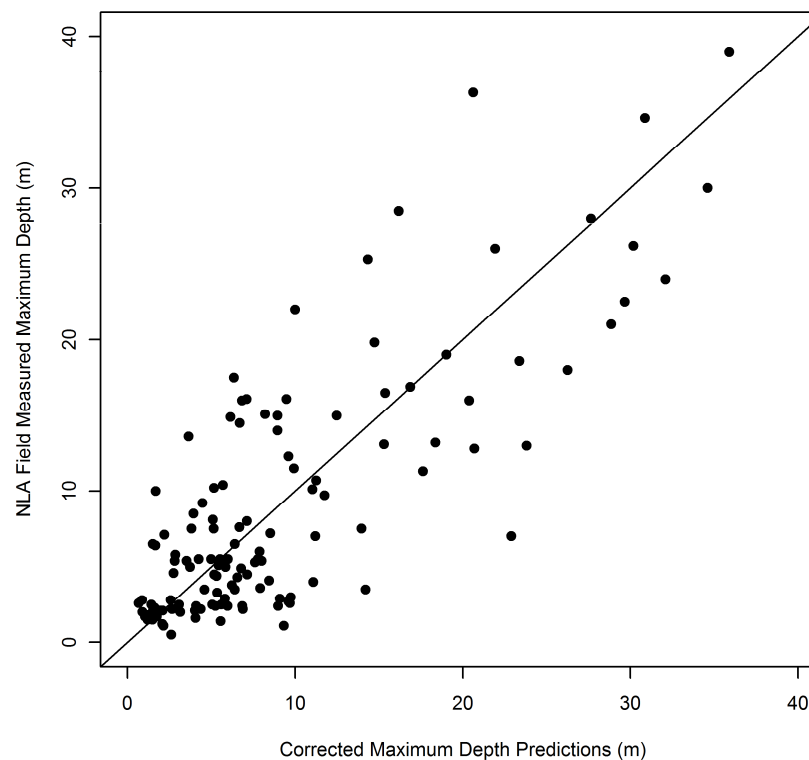
Assessing the method

- Compare to measured data
 - National Lakes Assessment Data
 - Web reported depths
- Over predicts
- Fit NLA model
- Use NLA model to correct

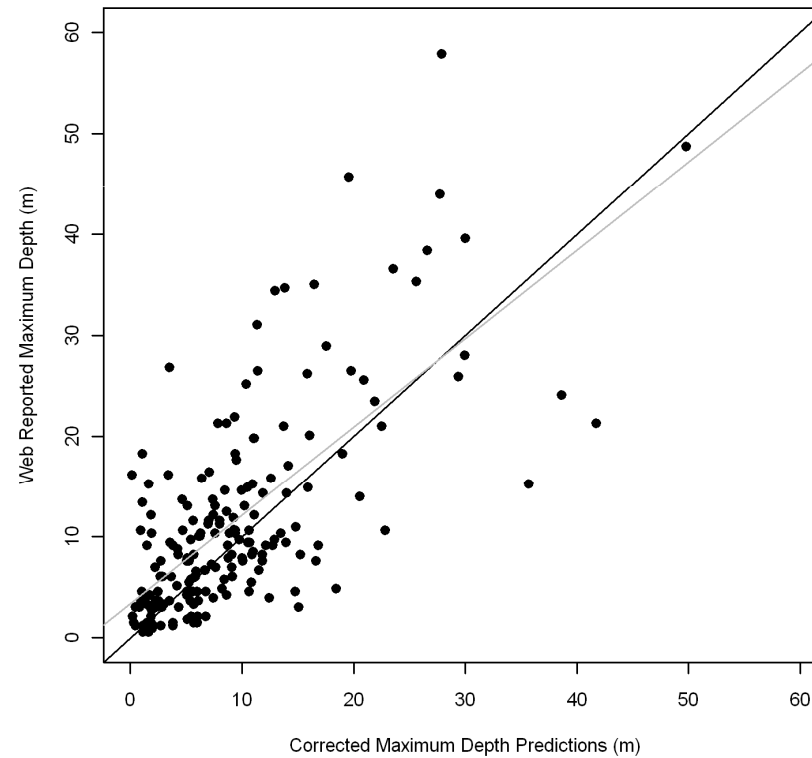


Assessing the method

- Compare to measured data
 - National Lakes Assessment Data
 - Web reported depths
- Over predicts
- Fit NLA model
- Use NLA model to correct



Web Depth Comparisons

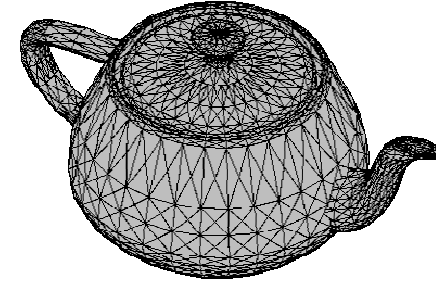


Linked Data and Lakes

- Starting a pilot
 - Convert Lake Morphometry data to Linked Data
 - Working with Mike Pendelton and David Smith
- Why?
 - Education (mine)
 - Interest
 - Potential for many more datasets
 - NLA
 - Wildlife
 - Merganser

R

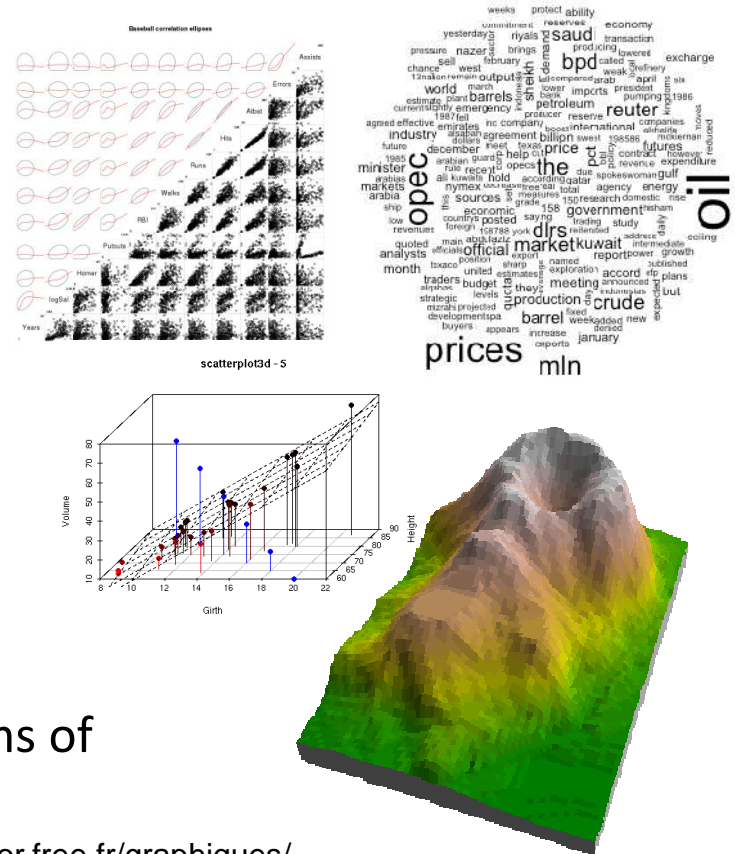
- Overview
 - Background
 - Extending R
 - Linked Data and Semantic Web
- How I use it?



- # What is R?

 - “Language and environment for statistical computing and graphics”
 - Similar to the S language
 - Extensible
- # Why use R?

 - FREE!
 - Publication quality graphics
 - Get only the results you want (i.e. no reams of output)



Examples borrowed from: <http://addictedtor.free.fr/graphiques/>

R: Overview

- Why is it named R?
 - First name of the original authors
 - Ross Ihaka and Robert Gentleman
 - Play on the S+ language
 - S- as a opposed to S+
- Who wrote/writes/contributes to R?
 - R Core Team
 - Anybody
- Is it reliable?
 - Open source = Peer review

R: Overview

- Expanding R:
 - Packages
 - Contributed collections of analytical tools
 - Extends the scope and utility of R
 - Currently 3978 packages available
 - Task Views
 - Currently 30 (e.g. Environmetrics, HighPerformanceComputing, Spatial, etc.)
 - None (yet) related directly to Semantic Web and Linked Data...
 - R, not just for stats anymore
 - GIS (sp, rgdal, raster, rgeos, ...)
 - Web and Related (Brew (+ rApache module), XML, rCurl, ...)
 - Linked Data Packages: (SPARQL , rrdf)

More Information

- R Install
 - <http://cran.r-project.org/doc/manuals/R-intro.pdf>
- Websites
 - R - <http://www.r-project.org/>
 - CRAN - <http://cran.r-project.org/>
 - simpleR - <http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>
- Listservs
 - R-help - <https://stat.ethz.ch/mailman/listinfo/r-help>
 - R-sig-ecology - <https://stat.ethz.ch/mailman/listinfo/r-sig-ecology>
- Books
 - Venables and Ripley, Modern Applied Statistics with S
 - Dalgaard, Introductory Statistics with R
 - Many, many Others!
- Journal
 - Journal of Statistical Software - <http://www.jstatsoft.org/>

R: How I use It?

- Basic Scripts and New Functions
- Demo

R: How I use It?

- Managing scientific workflow
 - From R
 - From Python



R Script

Tinn-R - [D:\DATA\EcoServices\LakeVolume\LakeVolumeManuscript\supplementals\Supp1RScript.R]

1/2

```
#####  
# Lake Volume R Script  
# Written by: Jeff W. Hollister  
# July/Aug 2009  
#  
# Function to take input bathymetry points an estimate TIN Volume (optional)  
# and Distance Based Volume  
# Function requires installed and functioning version of ArcGIS (with 3d  
# analyst and spatial analyst) and maptools,rgdal and RPyGeo packages  
# requires: bathypts of for ID, LONG, LAT, DEPTH  
# lake polygon shape file (single lake)  
# cell size  
# Coordinate Reference System of points  
# optional: Logical for calculating Volume with different methods, and  
# Python Path  
# This function assumes that units are uniform and that the Lake Polygon is  
# projected  
#####  
lrmVolume <- function(bathypts, lakepoly, cell, CRSText, PyPath="C:\\Python25",  
                      Dist=T, TIN=F, Cone=F){  
  
  #Loads required Libraries  
  require(maptools)  
  require(rgdal)  
  require(RPyGeo)  
  
  #reads in bathymetry points and creates Spatial Points Data Frame  
  xdf<-read.csv(bathypts)  
  xcoord<-coordinates(data.frame(xdf[,1:2], xdf[,4]))  
  xspdf<-SpatialPointsDataFrame(xcoord, xdf,  
                               proj4string=CRS(CRSText))  
  
  MAXDEPTH<-max(xspdf[[5]][,na.rm=T])  
  
  #Read in NH Lakes shapfiles  
  lakep<-readOGR(lakepoly, csuo(".shp", ".lakepoly"))  
  xspdf<-spTransform(xspdf, CRS(proj4string(lakep)))  
  
  #Set up rpygeo and R parameters  
  assign("rpygeo.env", rpygeo.build.env(workspace=getwd(), python.path=PyPath,  
                                         overwriteoutput=),  
        mask="lakep", cellsize=cell, extensions="3d"), envir = .GlobalEnv)  
  
  #Estimates Lake Volume based on Max and Dist  
  if(Dist==T){  
    carea<-cell*cell  
    rpygeo.geoprocessor("PolygonToLine",  
      args=list(lakepoly, "xx.shp"))  
    rpygeo.EucDistance.sa("xx.shp", "xxdist", "#", cell)  
    rpygeo.geoprocessor("ExtractByMask_sa",  
      args=list("xxdist", lakepoly, "xxdist1"))  
    assign("lakedist", readGDAL("xxdist1"))  
    MAXDIST<-max(lakedist[["band1"]], na.rm=T)  
    DistanceVolume<-sum(lakedist[["band1"]]*MAXDEPTH)/MAXDIST*carea, na.rm=T)  
  
    print(paste("DistanceVolume ", DistanceVolume))  
  }  
}
```

Tinn-R - [D:\DATA\EcoServices\LakeVolume\LakeVolumeManuscript\supplementals\Supp1RScript.R]

2/2

```
}  
  
#This section creates TIN, Add's points and hard lines, calculates and returns  
volume  
if(TIN==T){  
  SHORE<-0  
  lakep<-spCbind(lakep, SHORE)  
  writeOGR(lakep, getwd(), "lakep", driver="ESRI Shapefile")  
  writeOGR(xspdf, getwd(), "xspdf", driver="ESRI Shapefile")  
  rpygeo.env$extensions="3d"  
  rpygeo.geoprocessor("createtint",  
    args=list("temptin", "lakep.shp"), env=rpygeo.env)  
  argu<-paste("xspdf", ".shp DEPTH <none> masspoints true; ", "lakep", ".shp SHORE  
    <none> hardclip true", sep="")  
  rpygeo.geoprocessor("editint",  
    args=list("temptin", argu), env=rpygeo.env)  
  unlink("vol.txt")  
  rpygeo.geoprocessor("SurfaceVolume_3d",  
    args=list("temptin", "vol.txt", "ABOVE"), env=rpygeo.env)  
  xvcl<-read.csv("vol.txt")  
  TINVolume<-xvcl[,7]  
  
  print(paste("TINVolume ", TINVolume))  
}  
  
if(Cone==T){  
  ConeVolume<-(lakep[["AREA"]]*MAXDEPTH)/3  
  
  print(paste("ConeVolume ", ConeVolume))  
}  
}
```




Safe and Sustainable Water Resources

Python Script

TextS1.py

Page 1

```
import time
start = time.clock()
import sys, os
import arcpy
arcpy.CheckOutExtension("Spatial")
import rpy2.robjects as robjects
import math
r=robjects.r
r.library("rgdal")
#These options can be set to desired workspace and names of input elevation, catchment, and lake datasets
#Currently set to work with examples in Dataset S1 of Hollister, Milstead, and Urrutia (2011)
#Workspace is created when example dataset are extracted
arcpy.env.overwriteOutput=True
arcpy.env.workspace=r"C:\HollisterMilsteadUrrutia\DatasetS1"
inlakes = "exampleLakes.shp"
inelev = "exampleNED"
incatch = "exampleCatchment.shp"
outfile = "C:/HollisterMilsteadUrrutia/DatasetS1/exampleOutput.csv"

#Do not change the following output names, these are used in the script and would changes would
#cause script to fail
outlake = "xxlake.shp"
outbuffer = "xxlakeb.shp"
outclip = "xxlakec.shp"
outunion = "xxunion.shp"
outelev = "xxelev"
outslope = "xxslope"
outlakeline = "xxlakel.shp"
outlaker = "xxlaker"
outdist = "xxdist"
outdistl = "xxdistl"

numRows = float(arcpy.GetCount_management(inlakes)[0])
timer = time.clock()
cnt = 0
cnt2 = 0
lakes = arcpy.SearchCursor(inlakes)
if os.path.exists(outfile)==True:
    f = open(outfile,'a')
else:
    f = open(outfile, 'w')
    f.write('WB_ID,PredMaxDepth\n')
    f.close()
    f = open(outfile,'a')
for lake in lakes:
    r('test<-read.csv="'+outfile+'"')
    test=r('sum(test[,1]~1)*str(lake.getValue("WB_ID"))+') [0]
    if test > 0:
        cnt=cnt+1
        cnt2=cnt2+1
    if test == 0:
        sqlqry='WB_ID="'+str(lake.getValue("WB_ID"))+'"'
        arcpy.Select_analysis(inlakes,outlake,sqlqry)
        arcpy.PolygonToLine_management(outlake,outlakeline)
        DistOut = arcpy.sa.EucDistance(outlakeline,"#",30)
        #DistOut.save(outdist)
        arcpy.FeatureToRaster_conversion(outlake,"WB_ID",outlaker,30)
        LakeRastOut = arcpy.sa.ExtractByMask(DistOut,outlaker)
        lakebuff = float(arcpy.GetRasterProperties_management(LakeRastOut,"MAXIMUM")[0])
        if lakebuff<100:
            lakebuff=100
        arcpy.Buffer_analysis(outlake,outbuffer,lakebuff)
        arcpy.Clip_analysis(incatch,outbuffer,outclip)
        arcpy.Union_analysis(outlake+"",outbuffer+"",outclip,outunion)
        arcpy.Clip_management(inelev,"#",outelev,outbuffer,"#", "ClippingGeometry")
```

TextS1.py

Page 2

```
SlopeOut = arcpy.sa.Slope(outelev,"PERCENT_RISE")

r('memory.limit(4000)')
r.setwd(arcpy.env.workspace)
r('unionshp<-readOGR("xxunion.shp","xxunion")')
r('slopegrd<-readGDAL="'+str(SlopeOut).replace("\\","/")+'"')
r('distgrd<-readGDAL="'+str(LakeRastOut).replace("\\","/")+'"')
r('lake<-unionshp[unionshp[["FID_xxlake"]]==0,]')
r('land<-unionshp[unionshp[["FID_xxlake"]]==-1,]')
r('landovl<-overlay(slopegrd,land)')
r('lakeovl<-overlay(distgrd,lake)')
r(
'COMIDlnktblland<-data.frame(index=c(1:length(land[["COMID"]])),COMID=land[["COMID"]][1:length(land[["COMID"]])])')
r(
'COMIDlnktbllake<-data.frame(index=c(1:length(lake[["COMID"]])),COMID=lake[["COMID"]][1:length(lake[["COMID"]])])')
r(
'slopeland<-merge(data.frame(grdindex=c(1:length(landovl)),index=landovl),COMIDlnktblland,by="index")')
r(
'distlake<-merge(data.frame(grdindex=c(1:length(lakeovl)),index=lakeovl),COMIDlnktbllake,by="index")')
)
#Matches Catchments: First Catchments in lake to catchments on land
#Removes catchments in buffer but not intersecting lake
#Second matches catchments on land to those in lake
#This removes catchments entirely internal to the lake: Rare but caused by artificial flow
paths in large (usually) lakes
r('slopeland<-slopeland[slopeland$COMID~in$distlake$COMID,]')
r('distlake<-distlake[distlake$COMID~in$slopeland$COMID,]')
r(
'slopedf<-merge(data.frame(slope=slopegrd[["band1"]][slopeland$grdindex],grdindex=slopeland$grdindex),slopeland,by="grdindex")[,c(2,4)]')
r(
'distdf<-merge(data.frame(dist=distgrd[["band1"]][distlake$grdindex],grdindex=distlake$grdindex),distlake,by="grdindex")[,c(2,4)]')
r('distmax<-max(distdf[,1],na.rm=T)')
r('slopemedian<-median(slopedf[,1],na.rm=T)')
r('xdf<-data.frame(WB_ID=lake$WB_ID,predmaxdepth=distmax*(slopemedian/100)')
r('save.image()')

f.write(str(r.xdf[0][0])+"",str(r.xdf[1][0])+"\n")
cnt=cnt+1
perc = 100*(cnt/numRows)
elaps = time.clock()-timer
cnt2 = cnt2 + 1
if elaps >= 60:

    timer = time.clock()
    perlaketime = elaps/cnt2

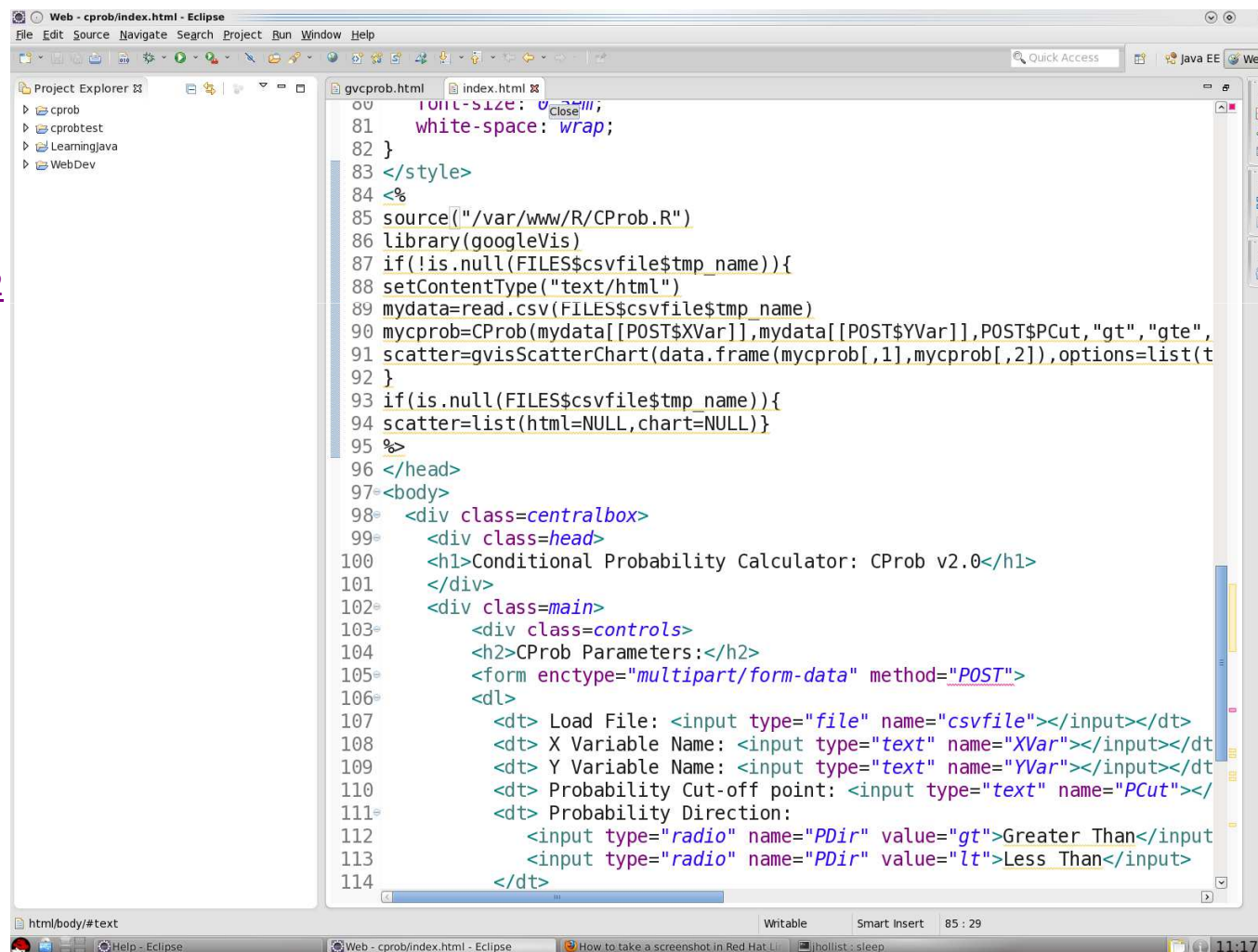
    print str(round(perc,2)) + "% of the lakes have been processed."
    print "Currently processing at approximately " + str(round(perlaketime,2)) + " seconds
    print "Approximately " + str(round(((numRows-cnt)*perlaketime)/3600,2)) + " hour(s)
    print(str(r.xdf[0][0])+"",str(r.xdf[1][0]))
    cnt2 = 0

    for feat in arcpy.ListFeatureClasses("xx*"):
        arcpy.Delete_management(feat)
    for rast in arcpy.ListRasters("xx*"):
        arcpy.Delete_management(rast)

f.close()
end = time.clock()
print (end-start)/60
```

R: How I use It?

- R for Web Apps
- Check out ggplot2
- <http://yeroon.net/ggplot2>



```
Web - cprob/index.html - Eclipse
File Edit Source Navigate Search Project Run Window Help

Project Explorer
└─ cprob
   └─ cprobtest
      └─ Learningjava
         └─ WebDev

gvcprob.html index.html
80 <!-- FONT-SIZE: 0.8em;
81 white-space: wrap;
82 }
83 </style>
84 <%
85 source("/var/www/R/CProb.R")
86 library(googleVis)
87 if(!is.null(FILE$csvfile$tmp_name)){
88 setContentType("text/html")
89 mydata=read.csv(FILE$csvfile$tmp_name)
90 mycprob=CPProb(mydata[[POST$XVar]],mydata[[POST$YVar]],POST$PCut,"gt","gte",
91 scatter=gvisScatterChart(data.frame(mycprob[,1],mycprob[,2]),options=list(t
92 }
93 if(is.null(FILE$csvfile$tmp_name)){
94 scatter=list(html=NULL,chart=NULL)}
95 %>
96 </head>
97 <body>
98 <div class=centralbox>
99 <div class=head>
100 <h1>Conditional Probability Calculator: CProb v2.0</h1>
101 </div>
102 <div class=main>
103 <div class=controls>
104 <h2>CProb Parameters:</h2>
105 <form enctype="multipart/form-data" method="POST">
106 <dl>
107 <dt> Load File: <input type="file" name="csvfile"></input></dt>
108 <dt> X Variable Name: <input type="text" name="XVar"></input></dt>
109 <dt> Y Variable Name: <input type="text" name="YVar"></input></dt>
110 <dt> Probability Cut-off point: <input type="text" name="PCut"></dt>
111 <dt> Probability Direction:
112 <input type="radio" name="PDir" value="gt">Greater Than</input>
113 <input type="radio" name="PDir" value="lt">Less Than</input>
114 </dt>
115 </dl>
116 </div>
117 </div>
118 </div>
119 </body>
120 </html>
```

R: How I use It?

- R for GIS
- Demo

Acknowledgements

- NLA Field Crews, Collaborators, & Analysis Team
- Richard Moore, USGS, MRB1 SPARROW
- Hilary Snook, Toby Stover & Carol Elliot, EPA, NELP
- John Kiddon, Jane Copeland, & the AED Aquatic Ecosystem Services Research Group
- Mike Pendleton, David Smith and George Thomas

Questions?

