Neotoma paper

Simon Goring

29 July, 2014

# neotoma: A Programmatic Interface to the Neotoma Paleoecological Database

## Abstract:

Paleoecological data is an integral part of ecological analysis. It provides an opportunity to study vegetation and climate interactions at time scales that cannot be observed through modern field studies, and allows us to observe changes in ecological processes that operate at centennial and millennial scales. Paleoecological inference also can allow us to understand ecological processes in the absence of widespread antrhopogenic influence.

Here we describe the R package neotoma, to be used to obtain and manipulate paleoecological data from the Neotoma Paleoecological Database (). neotoma searches the Neotoma Database for datasets associated with location, taxa or dataset types using the database's Application Programming Interface. The package can return full datasets or metadata associated with sites and provides the ability to standardize taxonomies using one of several recognized standard taxonomies from the published literature.

To assist with the use of the package we provide examples of key functions using examples from the published literature, for both plant and mammal taxa.

## Introduction

Paleoecological data is increasingly used to understand patterns of biogeographical, climatic and evolutionary change at multiple spatial and temporal scales. Paleoecoinformatics ([@brewer2012paleo; @uhen2013card]) is increasingly providing tools to researchers across disciplines to access and use large datasets spanning thousands of years. These datasets may be used to provide better insight into patterns of biomass burning (Blarquez et al, 2013; Power et al.), regional vegetation change ([@blois2013modeling; @blarquez2014disentangling]) or changes in physical processes over time ([@goring2012depo]). Critically, paleoecological data lags behind modern ecological cyber-infrastructure in regards to accessibility and extent. The increasing interest in uniting ecological and paleoecological data to understand modern ecological patterns and future responses [@fritz2013diversity; @behrensmeyer2012building; @dietl2011conservation] means that efforts to unite these two, seemingly independednt data-streams will rely, in part, on more robust tools to access and synthesize paleoecological data.

The statistical software R ([@RCoreTeam2014]) is commonly used for analysis of paleoecological data and several packages in R exist for analysis (analogue: [@analogue2013; @analogue2007]; rioja: [@rioja2013], Bchron: [@bchron2014], paleofire: [@paleofire2014]). Notwithstanding these packages, the use of extensive paleoecological resources within R has traditionally relied on *ad hoc* methods of obtaining and importing data. This has meant reliance on static online datasets such as the NOAA Paleoclimate repository or North American Modern Pollen Database, and on the distribution of individual datasets from author to analyst.

With an increasing push to provide paleoecological publications that include numerically reproducible results (e.g., [@goring2012depo; @gill2013linking; @goring2013pollen]) it is important to provide tools that allow analysts to directly access dynamic datasets, and to provide tools to support reproducible workflows. The rOpenSci project has provided a number of tools that can directly interact with Application Programmatic Interfaces (APIs) to access data from a number of databases including rfishbase (FishBase: [@boettiger2012rfishbase]) and taxize (Encyclopedia of Life, iPlant/Taxosaurus and others: [@chamberlain2013taxize]) among others.

To illustrate use cases for the neotoma package we present examples drawn from the paleoecological literature to illustrate how neotoma provides the tools to perform research that is critical to understanding paleoecological change in the Pleistocene in an open and reproducible manner.

## The neotoma package

Here we describe neotoma, an R package that acts as an interface between a large dynamic database (the Neotoma Paleoecological Database; http://neotomadb.org) and statistical tools in R. The neotoma package uses a programmatic interface (an API) to send data requests to Neotoma, and then forms data objects that can interact with existing packages such as analogue and rioja. The neotoma package also includes tools to standardize pollen data across sample sites using a set of commonly accepted pollen taxa.

## Examples

Macdonald and Cwynar ([@macdonald1991post]) used pollen percentage data for *Pinus* to map the northward migration of lodgepole pine (*Pinus contorta* var *latifolia*) following glaciation. In their study a cutoff of 15% Pinus pollen is associated with presence at pollen sample sites. Recent work by Strong and Hills ([@strong2013holocene]) has remapped the migration front using a lower pollen proportion (5%) and more sites. Here we attempt to replicate the analysis as an example both of the strengths of the package and limitations of paleoinformatic approaches.

To begin we must define a spatial bounding box and a set of taxa of interest. Strong and Hills ([@strong2013holocene]) use a region approximately bounded by 54^oN to the south and 65^oN to the North, and from 110^oW to 130^oW. The command get\_site is used to find all sites within a bounding box:

library(neotoma)  
library(ggmap)  
library(ggplot2)  
library(reshape2)  
library(plyr)  
library(Bchron)  
library(gridExtra)  
  
all.sites <- get\_site(loc = c(-140, 50, -110, 65))

#> The API call was successful, you have returned 97 records.

The get\_sites command returns a site data.frame, with siteID, latitude, longitude, altitude, SiteName, and SiteDescription. Each row represents a unique site.

We can see that this returns a total of R nrow(all.sites) sites. Sites are effectively containers for datasets though. Generally it's better to search for datasets. When you search for a dataset you can limit the type of dataset, either by looking for specific taxa, or by describing the dataset type. Here we will look for all taxa beginning with *Pinus* in pollen dataset. We use the \* wildcard to indicate any and all taxa with *Pinus* in their name:

all.datasets <- get\_dataset(loc = c(-140, 50, -110, 65), datasettype = "pollen",   
 taxonname = "Pinus\*")

A dataset is a larger data object. The dataset has site information, but it also has information about the specific dataset.

Here the API tells us we now have only 42 records of the original 97. Many of the samples are pollen surface samples, or vertebrate fauna, meaning pollen core data comprises less than half of the records. Regardless, we now know that there is pollen core data from 42 sites and we can plot those sites over our original 97.