Data ingestion for 2017

K Todd-Brown (ktoddbrown@gmail.com)

11/13/2017

Contents


```
library(SoilDataR) #library(devtools); install_github("ktoddbrown/soilDataR")
library(ggplot2) #make pretty plots
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(knitr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(tidyr)

#mapping librarys to help with global/regional plots
library(ggmap)
library(maps)
library(mapdata)
library(fiftystater)
```

Datasets are: 1) Treat: Peat properties synthesis dataset (2MB, XLSX format, download only; ISCNtemplate_Treat_peatProps_v2): This dataset is a synthesis of literature and site-level data on peat properties, C, N, 14C, and vegetation from 366 sites worldwide. Data are available for nearly 16,000 layers from 659 profiles. Data contributed by Claire Treat. 2) Alamos soil C stocks (<1MB, XLSX format, download only; ISCNtemplate_Alamos): This site-level dataset comes courtesy of Kris Johnson and collaborators at ITSON (Obregon, MX). It contains 30 profiles sampled by quantitative pit as part of a NASA-supported C monitoring

study. 3) Berhe et al 2012. Fractionation example from the Power Center Working Group, manuscript DOI: 10.1029/2011JG001790

```
##source('../../SoilDataR/R/processData_Templet.R') ##Uncomment to debug template files
ingestFiles <- list(
  filename = c('../repoData/Treat_2015/ISCNtemplate_Treat_peatProps_v2.xlsx',
               '../repoData/Alamos/ISCNtemplate_Alamos.xlsx',
               '../repoData/Berhe2012/Berhe_2012.xlsx'),
  keyFile = c(rep('../templates/ISCNtemplate_2016Key.xlsx', 2),
              '../templates/PowellCenterKey.xlsx'),
  verticalSheets = c(rep('metadata', 2), ''),
  skip=list(c(1:2), 1:2, NA))

data.ls <- list(study=data.frame(), field=data.frame(), sample=data.frame(), treatment=data.frame())
for(ii in 1:length(ingestFiles$filename)){
  temp <- processData_Templet(
    filename=ingestFiles$filename[[ii]],
    key.df=readxl::read_excel(path=ingestFiles$keyFile[[ii]], sheet='headerKey'),
    skip=ingestFiles$skip[[ii]],
    verticalSheets=ingestFiles$verticalSheets[[ii]])

  ##append data set name
  datasetName <- unique(temp$study$dataset_name[!is.na(temp$study$dataset_name)])
  temp$field$dataset_name <- datasetName
  temp$sample$dataset_name <- datasetName
  if(nrow(temp$treatment) > 0) temp$treatment$dataset_name <- datasetName

  ##append units to sample
  temp$sample <- temp$sample %>%
    mutate(unit = as.character(unit)) %>%
    left_join(select(temp$key %>% filter(type == 'value'), var, hardUnit)) %>%
    mutate(var=as.factor(var))

  data.ls$study <- bind_rows(data.ls$study,
                             temp$study)
  data.ls$field <- bind_rows(data.ls$field,
                             temp$field)
  data.ls$sample <- bind_rows(data.ls$sample,
                              temp$sample)
  data.ls$treatment <- bind_rows(data.ls$treatment,
                                 temp$treatment)
}
```

```
## Joining, by = "site_name"
```

```
## Joining, by = c("site_name", "profile_name")
```

```
## Joining, by = "site_name"
```

```
## Joining, by = "var"
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## character vector and factor, coercing into character vector
```

```
## Joining, by = "site_name"
```

```
## Joining, by = c("site_name", "profile_name")
```

```
## Joining, by = "site_name"
```

```
## Joining, by = "var"
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## character vector and factor, coercing into character vector
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
## Joining, by = c("site_name", "dataset_name")
```

```
## Joining, by = c("site_name", "dataset_name")
```

```
## Joining, by = c("site_name", "dataset_name", "profile_name")
```

```
## Joining, by = c("site_name", "dataset_name", "layer_name", "profile_name")
```

```
## Joining, by = "var"
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## character vector and factor, coercing into character vector
```

```
##Filter the messy study names
data.ls$study <- data.ls$study %>%
  filter(!is.na(dataset_name) & !is.na(curator_email)) %>%
  arrange(dataset_name)

data.ls$sample <- data.ls$sample %>%
  mutate(unit = if_else(is.na(unit), hardUnit, unit))
```

Observation time

```
location.df <- data.ls$field %>%
  select(lat, long, observation_date, state, country, dataset_name) %>%
  unique() %>%
  ###cetch any N|S or E|W notations
  mutate(lat = if_else(grepl('S', as.character(lat)), -1*as.numeric(gsub('S', '', as.character(lat))),
                                        as.numeric(gsub('N', '', as.character(lat)))),
         long = if_else(grepl('W', as.character(long)), -1*as.numeric(gsub('W', '', as.character(long))),
                                      as.numeric(gsub('E', '', as.character(long))))) %>%
  ###convert the observation dates
  mutate(observation_date = as.character(observation_date)) %>%
  separate(observation_date, c('monthStr', 'dayStr', 'yearStr'),
           remove=FALSE, fill='left') %>%
  mutate(year=if_else(is.na(as.numeric(yearStr)), as.numeric(yearStr),
                      if_else(as.numeric(yearStr) < 20, as.numeric(yearStr) + 2000,
                              if_else(as.numeric(yearStr) < 100, as.numeric(yearStr) + 1900,
                                      as.numeric(yearStr)))),
         month=if_else(is.na(as.numeric(monthStr)), 1,  as.numeric(monthStr)),
```

```
        day=if_else(is.na(as.numeric(dayStr)), 1, as.numeric(dayStr))) %>%
  select(-contains('Str')) %>%
  mutate(obsDate = ymd(paste(year, month, day, sep='-'))) %>%
  arrange(lat, long, obsDate) %>%
  ###segment everything
  mutate(yrCut = cut(year, seq(from = floor(min(year, na.rm=TRUE)/10)*10,
                               to = ceiling(max(year, na.rm=TRUE)/10)*10, by=10),
                     dig.lab=4),
         latCut = cut(lat, seq(-90, 90, by=0.05)),
         longCut = cut(long, seq(-180, 180, by=0.05))) %>%
  ###replace common country names
  mutate(country = if_else(grepl('USA', country), 'United States', country))
```

```
## Warning in if_else(grepl("S", as.character(c(NA, "56.883333329999999",
## "56.633333329999999", : NAs introduced by coercion
```

```
## Warning in replace_with(out, !condition & !is.na(condition), false,
## "`false`"): NAs introduced by coercion
```

```
## Warning: 13 failed to parse.
```

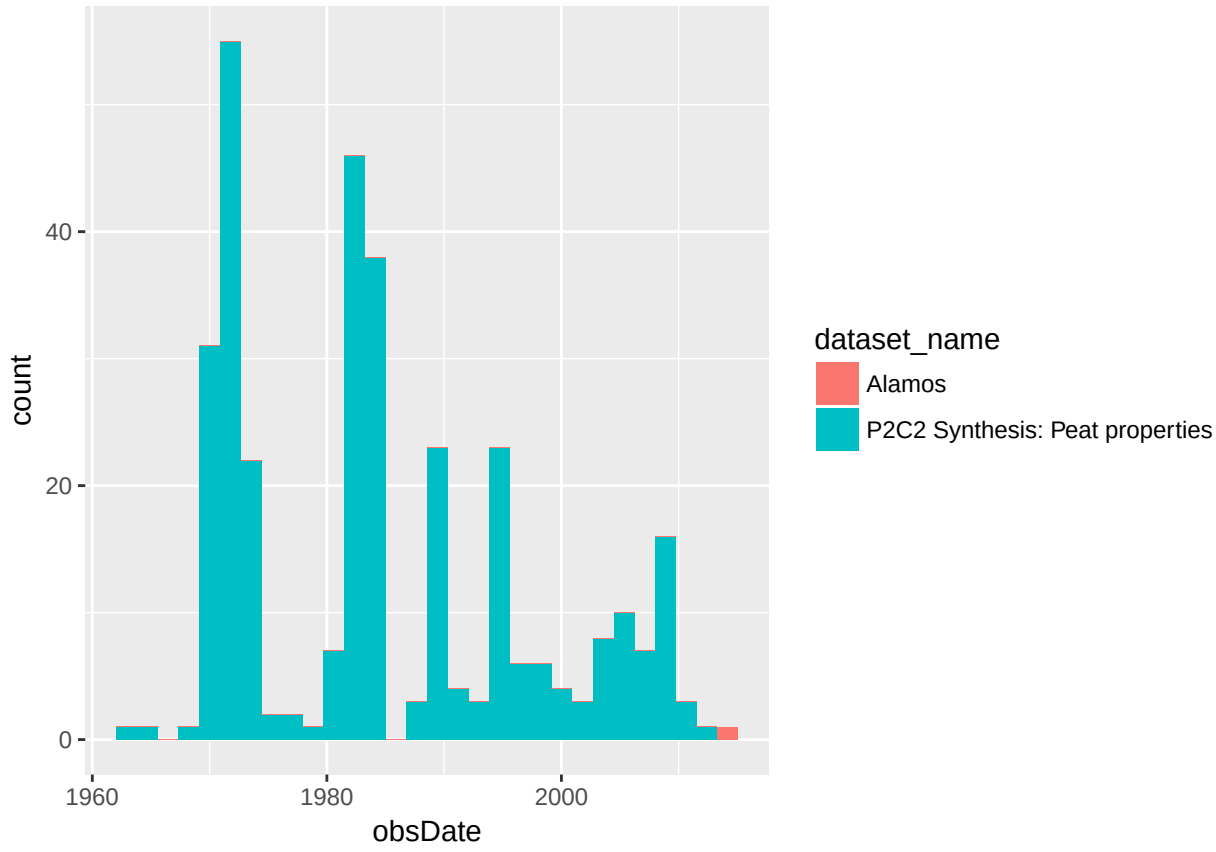
```
timeSpaceCounts <- location.df %>%
  group_by(yrCut,  country) %>%
  tally

print(timeSpaceCounts)
```

```
## Source: local data frame [22 x 3]
## Groups: yrCut [?]
##
## # A tibble: 22 x 3
##          yrCut       country     n
##         <fctr>         <chr> <int>
##  1 (1960,1970]        Canada    34
##  2 (1970,1980]        Canada    83
##  3 (1970,1980]        Russia     1
##  4 (1970,1980]          <NA>     1
##  5 (1980,1990]        Canada   108
##  6 (1980,1990]        Russia     5
##  7 (1980,1990] United States     1
##  8 (1990,2000]        Canada    31
##  9 (1990,2000]        Russia    12
## 10 (1990,2000]        Sweden     1
## # ... with 12 more rows
```

```
ggplot(location.df, aes(x=obsDate, fill=dataset_name)) + geom_histogram()
```

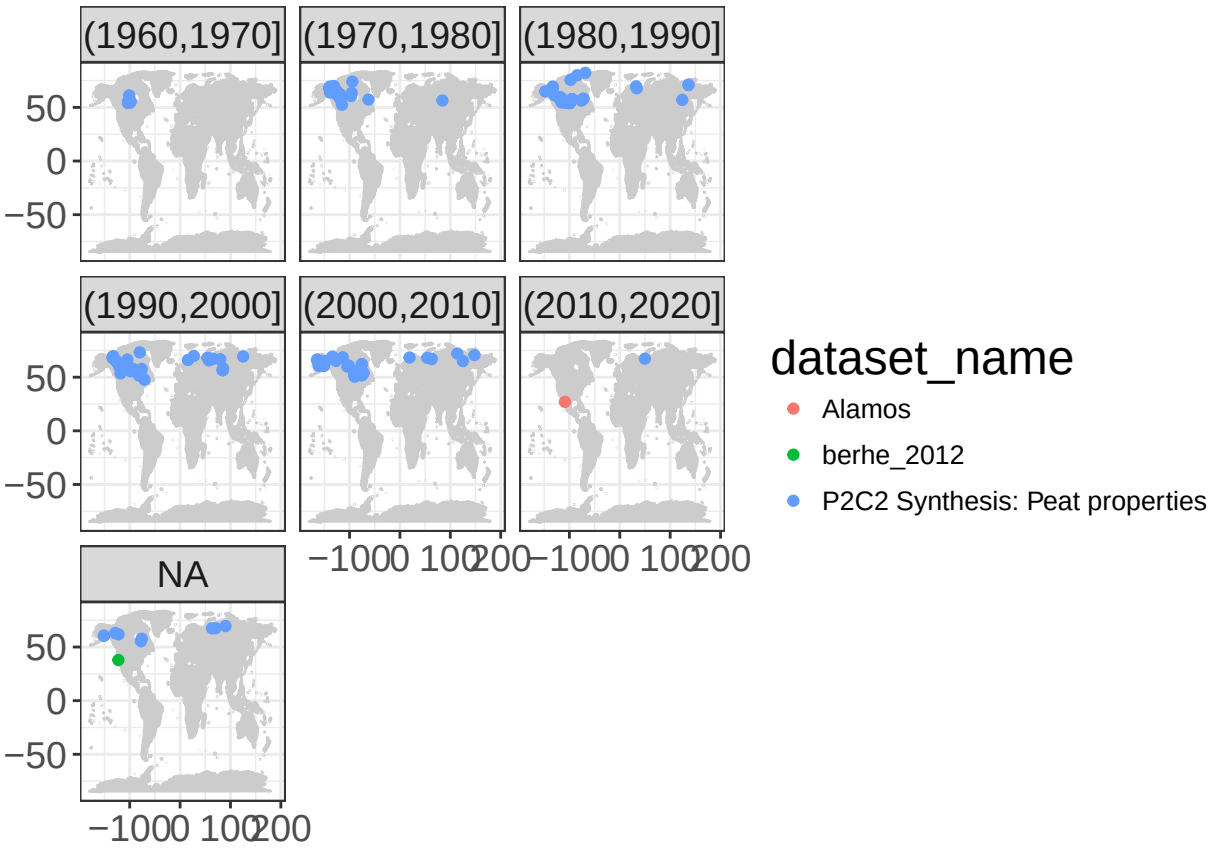
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 13 rows containing non-finite values (stat_bin).
```




```
mapWorld <- borders("world", colour="gray80", fill="gray80") # create a layer of borders
#ggplot() + mapWorld
ggplot(location.df) +
  mapWorld +
  #geom_hex(aes(x=long, y=lat), bins=200) +
  geom_point(aes(x=long, y=lat, color=dataset_name)) +
  scale_fill_gradient(trans='log10') +
  theme_bw() +
  theme(text=element_text(size=18),
        legend.text=element_text(size=10),
        axis.title=element_blank()) +
  #ylim(45, 90) +
  #coord_map(projection='azequidistant') +
  facet_wrap(~yrCut)
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```

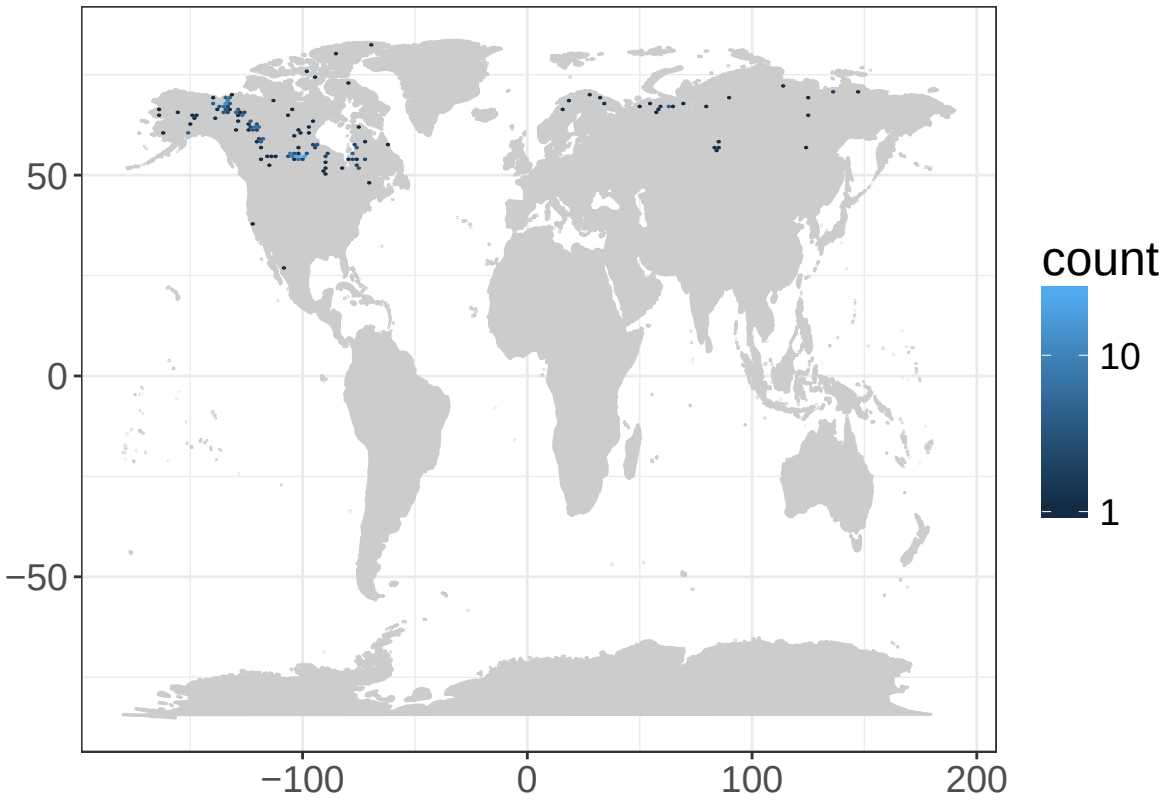


Site locations

Lat-lon map

```
mapWorld <- borders("world", colour="gray80", fill="gray80") # create a layer of borders
#ggplot() + mapWorld
ggplot(unique(location.df[, c('lat', 'long')])) +
  mapWorld +
  geom_hex(aes(x=long, y=lat), bins=200) +
  scale_fill_gradient(trans='log10') +
  theme_bw() +
  theme(text=element_text(size=18)) +
  labs(x='', y='')
```

```
## Warning: Removed 1 rows containing non-finite values (stat_binhex).
```

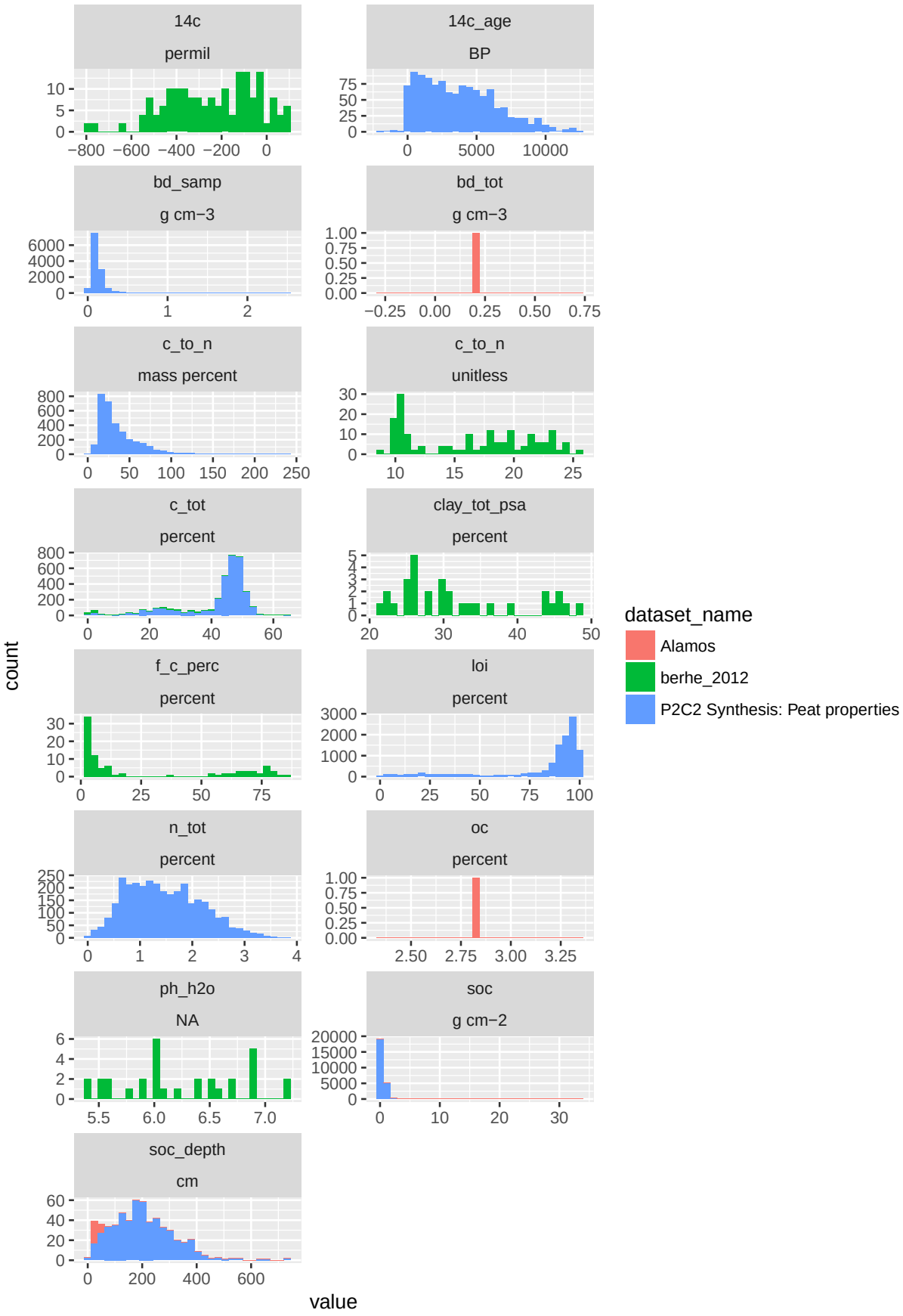


Measruements distribution

```
ggplot(data.ls$sample) +
  geom_histogram(aes(x=value, fill=dataset_name)) +
  facet_wrap(var~unit, scales='free', ncol=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```


14c
permil
14c_age
BP
bd_samp
g cm−3
bd_tot
g cm−3
c_to_n
mass percent
c_to_n
unitless
c_tot
percent
clay_tot_psa
percent
f_c_perc
percent
loi
percent
n_tot
percent
oc
percent
ph_h2o
NA
soc
g cm−2
soc_depth
cm
count
value
dataset_name
Alamos
berhe_2012
P2C2 Synthesis: Peat properties

Comparison with ISCN3

```
#source('../../SoilDataR/R/processData_ISCN3.R')
ISCN <- processData_ISCN3(layersDir='../repoData/ISCN_3/Layers/', metaDir='../repoData/ISCN_3/Meta/',
                                    keyFile='../repoData/ISCN_3/ISCNKey.xlsx',
                          loadVars=as.character(unique(data.ls$sample$var)))
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factor and character vector, coercing into character vector
```

```
## Joining, by = c("ISCN 1-1 (2015-12-10)", "dataset_name", "dataset_type (dataset_type)", "curator_name
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

```
## Joining, by = "dataset_name"
```

```
ISCN$field <- ISCN$field %>% select(ends_with('_name'), lat, lon, state, country, observation_date)

ISCNLocation <- ISCN$field %>%
  select(lat, lon, observation_date, state, country, dataset_name) %>%
  unique() %>%
  ##Convert to numeric
  mutate(lon=as.numeric(lon), lat=as.numeric(lat)) %>%
  mutate(observation_date = as.character(observation_date)) %>%
  separate(observation_date, c('monthStr', 'dayStr', 'yearStr'),
           remove=FALSE, fill='left') %>%
  mutate(year=if_else(is.na(as.numeric(yearStr)), as.numeric(yearStr),
                      if_else(as.numeric(yearStr) < 20, as.numeric(yearStr) + 2000,
                              if_else(as.numeric(yearStr) < 100, as.numeric(yearStr) + 1900,
                                      as.numeric(yearStr)))),
         month=if_else(is.na(as.numeric(monthStr)), 1,  as.numeric(monthStr)),
         day=if_else(is.na(as.numeric(dayStr)), 1, as.numeric(dayStr))) %>%
  select(-contains('Str')) %>%
  mutate(obsDate = ymd(paste(year, month, day, sep='-'))) %>%
  arrange(lat, lon, obsDate) %>%
  mutate(yrCut = cut(year, seq(from = floor(min(year, na.rm=TRUE)/10)*10,
                               to = ceiling(max(year, na.rm=TRUE)/10)*10, by=10),
                     dig.lab=4),
         latCut = cut(lat, seq(-90, 90, by=0.05)),
         lonCut = cut(lon, seq(-180, 180, by=0.05)))
```

```
## Warning: 905 failed to parse.
```

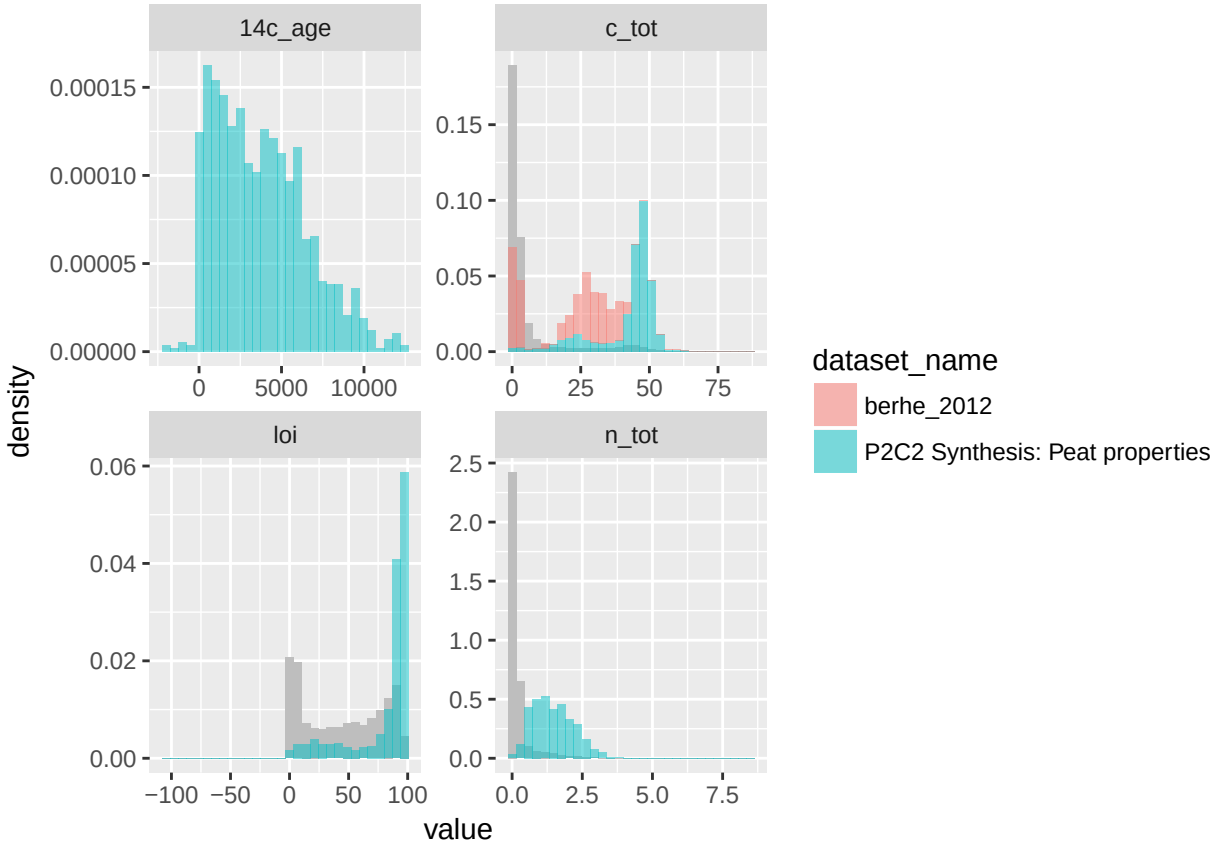
```
refData <- ISCN$measure %>%
  filter(var %in% c('14c_age', 'n_tot', 'c_tot', 'loi')) %>%
  left_join(ISCN$sample)
```

```
## Joining, by = "measureID"
```

```
ggplot(data.ls$sample %>% filter(var %in% c('14c_age', 'n_tot', 'c_tot', 'loi'))) +
  geom_histogram(data=refData, aes(x=value, y=..density..), fill='grey') +
  geom_histogram(aes(x=value, y=..density.., fill=dataset_name), alpha=0.5) +
  facet_wrap(~var, scales='free')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

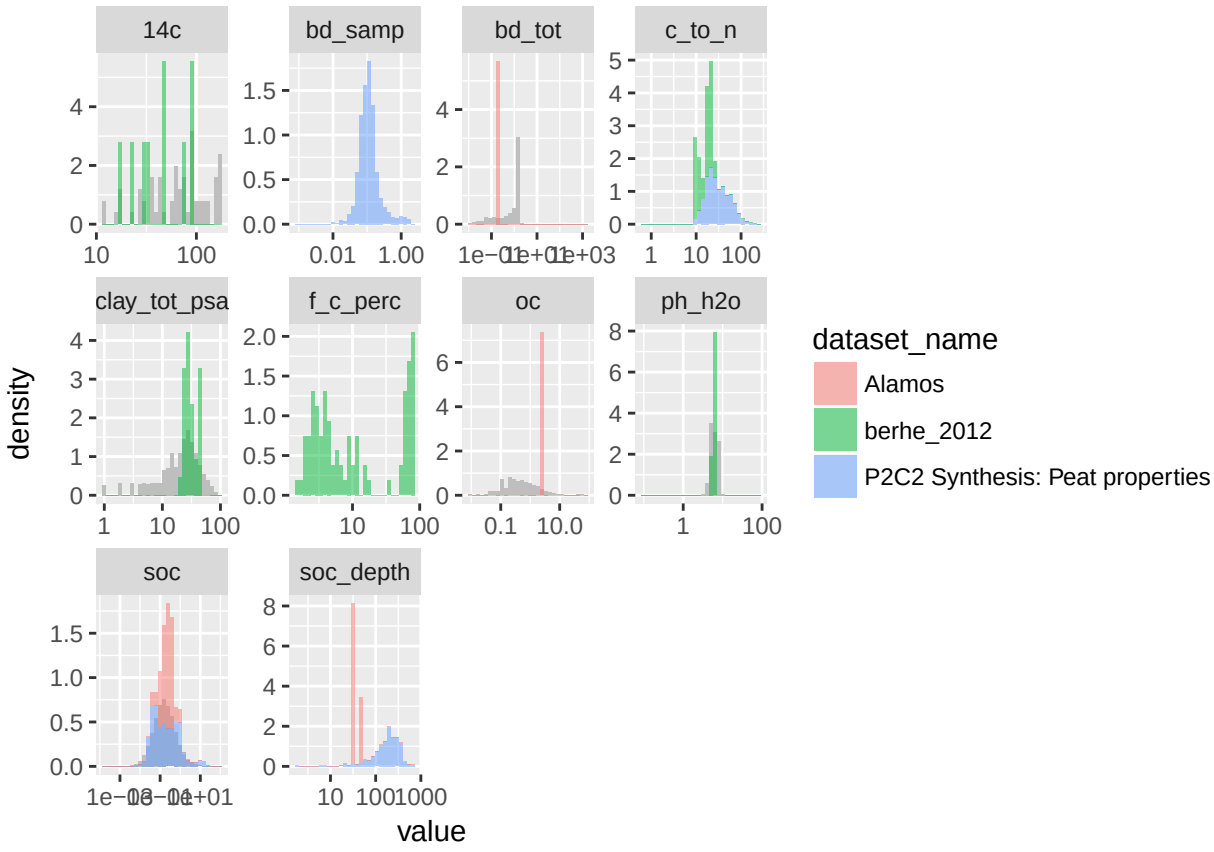



```
refData <- ISCN$measure %>%
  filter(!var %in% c('14c_age', 'n_tot', 'c_tot', 'loi')) %>%
  left_join(ISCN$sample)
```

```
## Joining, by = "measureID"
```

```
ggplot(data.ls$sample %>% filter(!var %in% c('14c_age', 'n_tot', 'c_tot', 'loi'))) +
  geom_histogram(data=refData, aes(x=value, y=..density..), fill='grey') +
  geom_histogram(aes(x=value, y=..density.., fill=dataset_name), alpha=0.5) +
  scale_x_log10() +
  facet_wrap(~var, scales='free')
```

```
## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous x-axis

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 6849 rows containing non-finite values (stat_bin).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 165 rows containing non-finite values (stat_bin).
```


14c
bd_samp
bd_tot
c_to_n
clay_tot_psa
f_c_perc
oc
ph_h2o
soc
soc_depth
density
value
dataset_name
Alamos
berhe_2012
P2C2 Synthesis: Peat properties