title: “cosore-working-with-data”
author: “Ben Bond-Lamberty”
date: “2023-02-23”
output: rmarkdown::html_vignette
vignette: >
%
%
%

Introduction

Soil respiration–the flux of CO2 from the soil surface to the atmosphere–has been measured by ‘continuous’ (e.g. half-hourly) measurement systems in many places over the last 20 years. The goal of the COSORE database is to collect many of these data into a single open community resource to support and speed up data synthesis and to fight the file drawer problem (http://dx.doi.org/10.1037/0033-2909.86.3.638).

The database is distributed as an R package. Use the `remotes` or `devtools` packages to install COSORE, e.g. `devtools::install_github("bpbond/cosore")`.

(For non-R users, the database is also distributed as a zip file of comma-separated files, available through the package’s release on GitHub (https://github.com/bpbond/cosore/releases). When you download and extract the file, there’s a README, the data in two separate formats, and a number of other files, including a version of this vignette.)

But how do we work with this database, exactly? Let’s start by loading it into R:

```
library(cosore)
#> Type citation("cosore") for the main COSORE database reference.
```

The database is comprised of a collection of *datasets*, each converted to a standard format and units. A dataset is one or more files of continuous (automated) soil respiration data, with accompanying metadata, with all measurements (i) taken at a single site (although different chambers can have individual geographic coordinates) and (ii) with constant treatment assignments. In R, each dataset is a list (http://www.r-tutor.com/r-introduction/list) of data frames (http://www.r-tutor.com/r-introduction/data-frame):

```
dataset 1
|- description table (a data.frame)
|- contributors table (ditto...)
|- ports table
|- columns table
|- ancillary table
|- data table
|- diagnostics table

dataset 2
|   |- etc.
```

For most analyses we want to extract one or more of these pieces and combine them–for example, to get a single dataset, a table of contributors, or an overview of the entire database.

The package provides a useful function that gives an overview of the entire database:

```
db_info <- csr_database()
tibble::glimpse(db_info)
#> Rows: 91
#> Columns: 11
#> $ CSR_DATASET     <chr> "d20190409_ANJILELI", "d20190409_ZOU", "d20190415_VARN…
#> $ CSR_LONGITUDE   <dbl> -117.8490, -7.2500, -72.1700, -86.4131, -86.4131, -90.…
#> $ CSR_LATITUDE    <dbl> 33.65800, 52.95000, 42.54000, 39.32320, 39.32320, 45.8…
#> $ CSR_ELEVATION   <dbl> 2.0, 260.0, 340.0, 275.0, 275.0, 520.0, 352.5, 352.5, …
#> $ CSR_IGBP        <chr> "Wetland", "Evergreen needleleaf forest", "Deciduous b…
#> $ CSR_PRIMARY_PUB <chr> "10.1029/2018JG004640", NA, "10.1029/2008JG000858", "1…
#> $ CSR_RECORDS     <int> 46271, 82314, 34641, 56701, 59181, 53886, 43656, 12446…
#> $ CSR_GASES       <chr> "CO2", "CO2", "CO2", "CO2", "CO2", "CO2", "CO2", "CO2"…
#> $ CSR_DATE_BEGIN  <date> 2016-02-05, 2013-11-22, 2003-04-20, 2012-01-01, 2011-…
#> $ CSR_DATE_END    <date> 2017-04-21, 2015-02-17, 2006-12-12, 2013-11-15, 2012-…
#> $ CSR_MSMT_VAR    <chr> "Rs", "Rs", "Rs", "Rh, Rs", "Rh, Rs", "Rs, Rh", "Rs", …
```

There's lots of information here, one row per dataset, including dataset name; geographic location; number of records; vegetation types; and gases, fluxes, and dates measured. Much of this is also summarized in the `Report-all.html` file included with the data in each release.

Exploring a single dataset

To begin, we pick a single dataset (`d20190415_VARNER`), get some information about it, and plot it.

```
varner <- csr_dataset("d20190415_VARNER")
#> d20190415_VARNER Reading standardized data
tibble::glimpse(varner$description)
#> Rows: 1
#> Columns: 20
#> $ CSR_DATASET          <chr> "d20190415_VARNER"
#> $ CSR_SITE_NAME        <chr> "Prospect Hill Tract (Harvard Forest)"
#> $ CSR_LONGITUDE        <dbl> -72.17
#> $ CSR_LATITUDE         <dbl> 42.54
#> $ CSR_ELEVATION        <dbl> 340
#> $ CSR_TIMEZONE         <chr> "Etc/GMT+5"
#> $ CSR_IGBP             <chr> "Deciduous broadleaf forest"
#> $ CSR_NETWORK          <chr> "Ameriflux"
#> $ CSR_SITE_ID          <chr> "US-Ha1"
#> $ CSR_INSTRUMENT       <chr> "LI-820"
#> $ CSR_MSMT_LENGTH      <dbl> 480
#> $ CSR_FILE_FORMAT      <chr> "Processed_csv"
#> $ CSR_TIMESTAMP_FORMAT <chr> "%Y-%m-%dT%H:%M"
#> $ CSR_TIMESTAMP_TZ     <chr> "Etc/GMT+5"
#> $ CSR_PRIMARY_PUB      <chr> "10.1029/2008JG000858"
#> $ CSR_OTHER_PUBS       <chr> NA
#> $ CSR_DATA_URL         <chr> "10.6073/pasta/29aae9def8e977d8ee67f1ca2f54b632"
#> $ CSR_ACKNOWLEDGMENT   <chr> "Varner R. 2008. Soil Respiration Along a Hydrolo…
#> $ CSR_NOTES            <chr> NA
#> $ CSR_EMBARGO          <chr> NA
```

The `description` table gives the basic information about this dataset: where it was measured, the time zone that the `data` timestamps are in, instrument used, and citation and acknowledgment information.

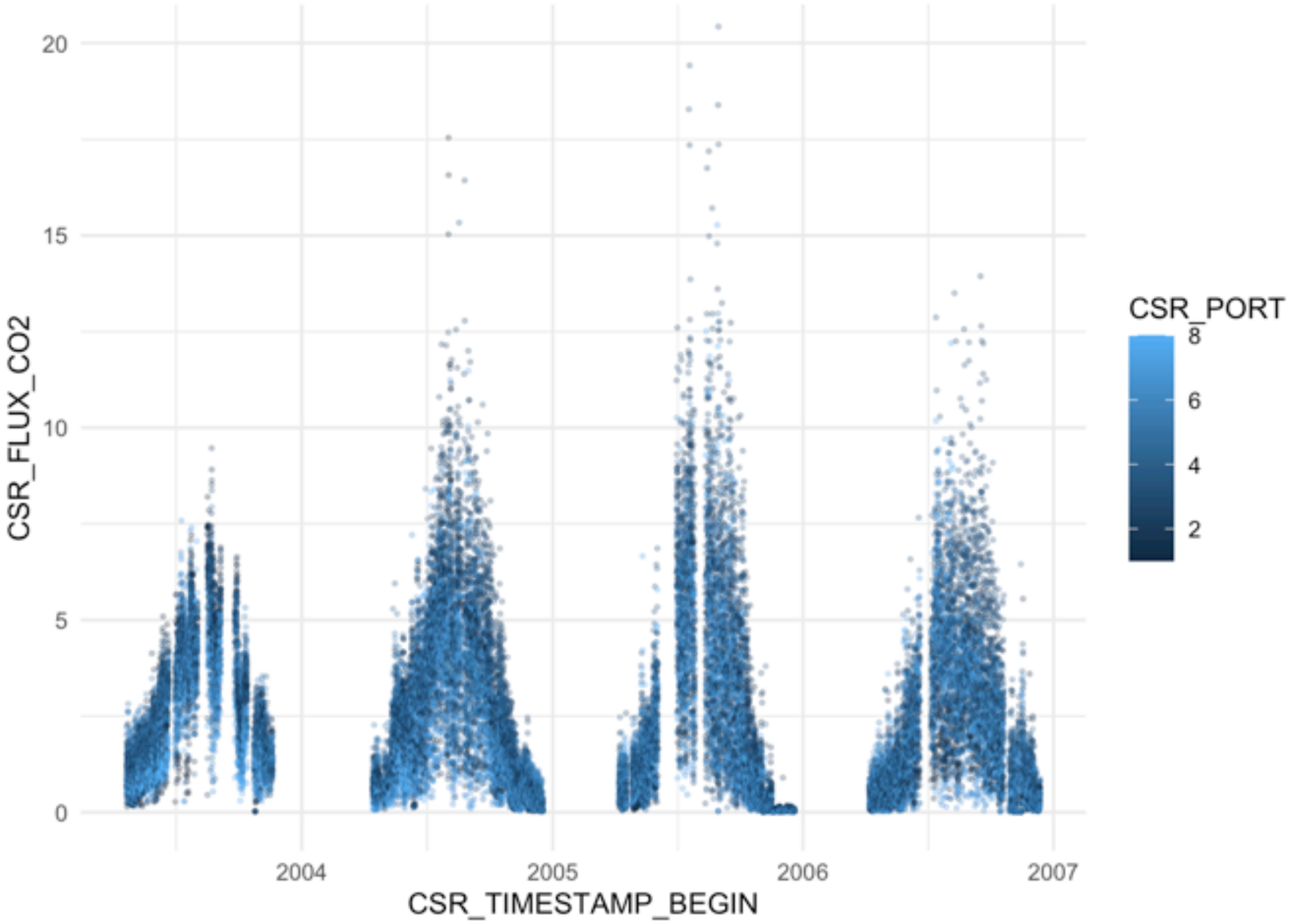
If you have questions about what COSORE fields contain, their units, etc., the `csr_metadata()` function returns a table with full information about this. This information is also packaged into each flat-file release.

Next, we want to look at the actual data:

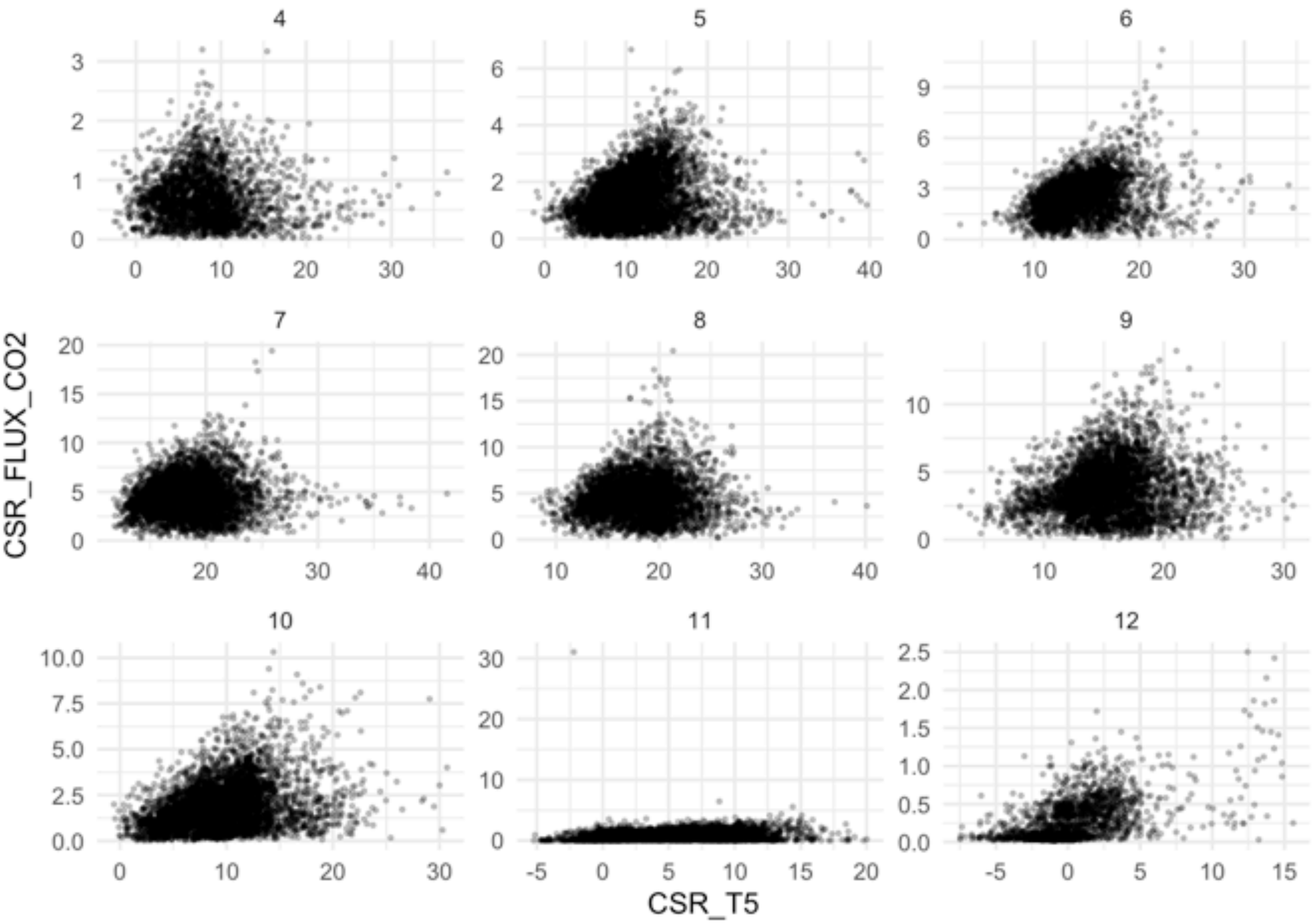
```
sr <- varner$data
nrow(sr)
#> [1] 34641
summary(sr)
#>     CSR_PORT     CSR_TIMESTAMP_BEGIN           CSR_TIMESTAMP_END
#>  Min.   :1.000   Min.   :2003-04-20 14:05:00   Min.   :2003-04-20 14:13:00
#>  1st Qu.:2.000   1st Qu.:2004-05-25 15:32:00   1st Qu.:2004-05-25 15:40:00
#>  Median :4.000   Median :2004-12-01 06:10:00   Median :2004-12-01 06:18:00
#>  Mean   :4.478   Mean   :2005-02-08 10:16:50   Mean   :2005-02-08 10:24:50
#>  3rd Qu.:6.000   3rd Qu.:2005-11-22 18:10:00   3rd Qu.:2005-11-22 18:18:00
#>  Max.   :8.000   Max.   :2006-12-12 12:39:00   Max.   :2006-12-12 12:47:00
#>
#>   CSR_FLUX_CO2       CSR_T5       CSR_TAIR_AMB
#>  Min.   : 0.01   Min.   :-7.49   Min.   :-17.08
#>  1st Qu.: 0.91   1st Qu.: 7.91   1st Qu.:  7.16
#>  Median : 2.05   Median :12.97   Median : 13.07
#>  Mean   : 2.59   Mean   :12.35   Mean   : 12.63
#>  3rd Qu.: 3.85   3rd Qu.:17.04   3rd Qu.: 18.30
#>  Max.   :31.03   Max.   :41.55   Max.   : 47.86
#>                  NA's   :7
```

This dataset has 34,641 observations; extends from April 2003 to December 2006; and soil respiration was measured using eight chambers, along with air and 5 cm soil temperature. Visualizing it:

```
library(ggplot2)
theme_set(theme_minimal()) # so much nicer
ggplot(sr, aes(CSR_TIMESTAMP_BEGIN, CSR_FLUX_CO2, color = CSR_PORT)) +
  geom_point(size = 0.5, alpha = 0.25) +
  coord_cartesian(ylim = c(0, 20))
```



```
library(lubridate, warn.conflicts = FALSE)
ggplot(sr, aes(CSR_T5, CSR_FLUX_CO2)) +
  facet_wrap(~month(CSR_TIMESTAMP_BEGIN), scales = "free") +
  geom_point(size = 0.5, alpha = 0.25)
#> Warning: Removed 7 rows containing missing values (`geom_point()`).
```



The November data include one very large flux that we’d probably want to exclude.

Did these eight different ports (chambers) represent different treatments? We might want to exclude treatment collars, or color them differently in the plots above. The `ports` table holds this information.

```
tibble::glimpse(varner$ports)
#> Rows: 1
#> Columns: 8
#> $ CSR_PORT           <int> 0
#> $ CSR_MSMT_VAR       <chr> "Rs"
#> $ CSR_TREATMENT      <chr> "None"
#> $ CSR_AREA           <int> 1886
#> $ CSR_DEPTH          <int> 2
#> $ CSR_SPECIES        <lgl> NA
#> $ CSR_OPAQUE         <lgl> TRUE
#> $ CSR_PLANTS_REMOVED <lgl> TRUE
```

From this we see that the only `CSR_PORT` entry is zero (meaning that this information applies to *all* ports), has has a `CSR_TREATMENT` of “None”. Also, the collars were 1,886 cm2.

Finally, we can use the `description` table information to get a full reference (which you will **definitely** cite in your published analysis, right?):

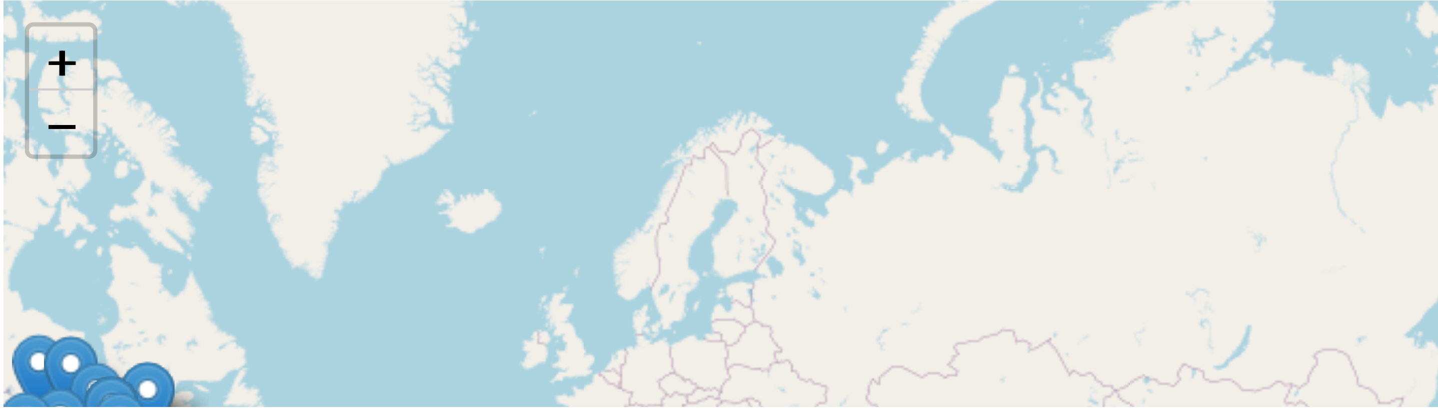
```
doi <- varner$description$CSR_PRIMARY_PUB
print(doi)
#> [1] "10.1029/2008JG000858"
try({  # in case you don't have 'rcrossref' installed...
  library(rcrossref)
  cr_cn(dois = doi, format = "text")
})
#> [1] "Phillips, S. C., Varner, R. K., Frolking, S., Munger, J. W., Bubier, J. L., W
ofsy, S. C., & Crill, P. M. (2010). Interannual, seasonal, and diel variation in soil
respiration relative to ecosystem respiration at a wetland to upland slope at Harvard
Forest. Journal of Geophysical Research: Biogeosciences, 115(G2), n/a-n/a. Portico. h
ttps://doi.org/10.1029/2008jg000858"
```

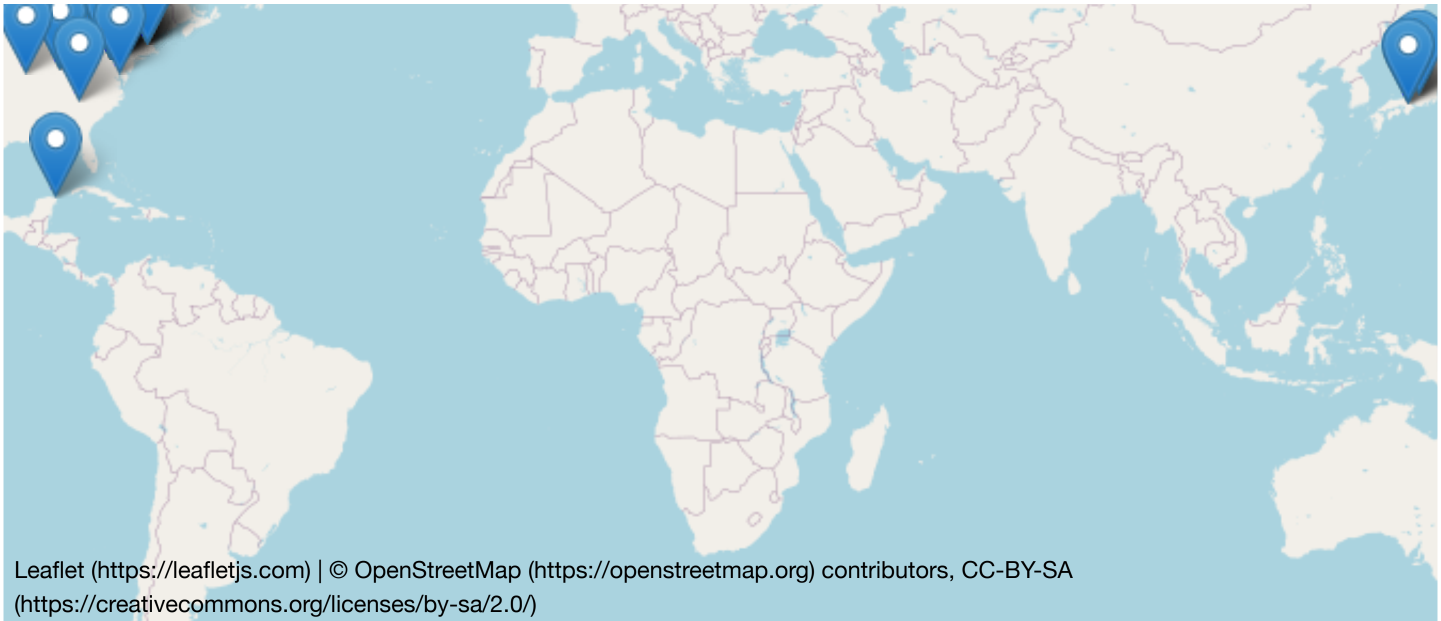
Selecting and combining multiple datasets

Time for something more ambitious: let's examine how soil respiration varies over the course of the day in temperate deciduous forests. For this we use the `csr_table()` function, which combined data across multiple datasets.

```
dbf_datasets <- subset(db_info, CSR_IGBP == "Deciduous broadleaf forest")$CSR_DATASET
tdf <- csr_table("description", dbf_datasets)
```

```
# Make a map of these datasets
library(sp)
library(leaflet)
map <- data.frame(lon = tdf$CSR_LONGITUDE, lat = tdf$CSR_LATITUDE)
coordinates(map) <- ~lon + lat
leaflet(map) %>%
  addMarkers() %>%
  addTiles()
```





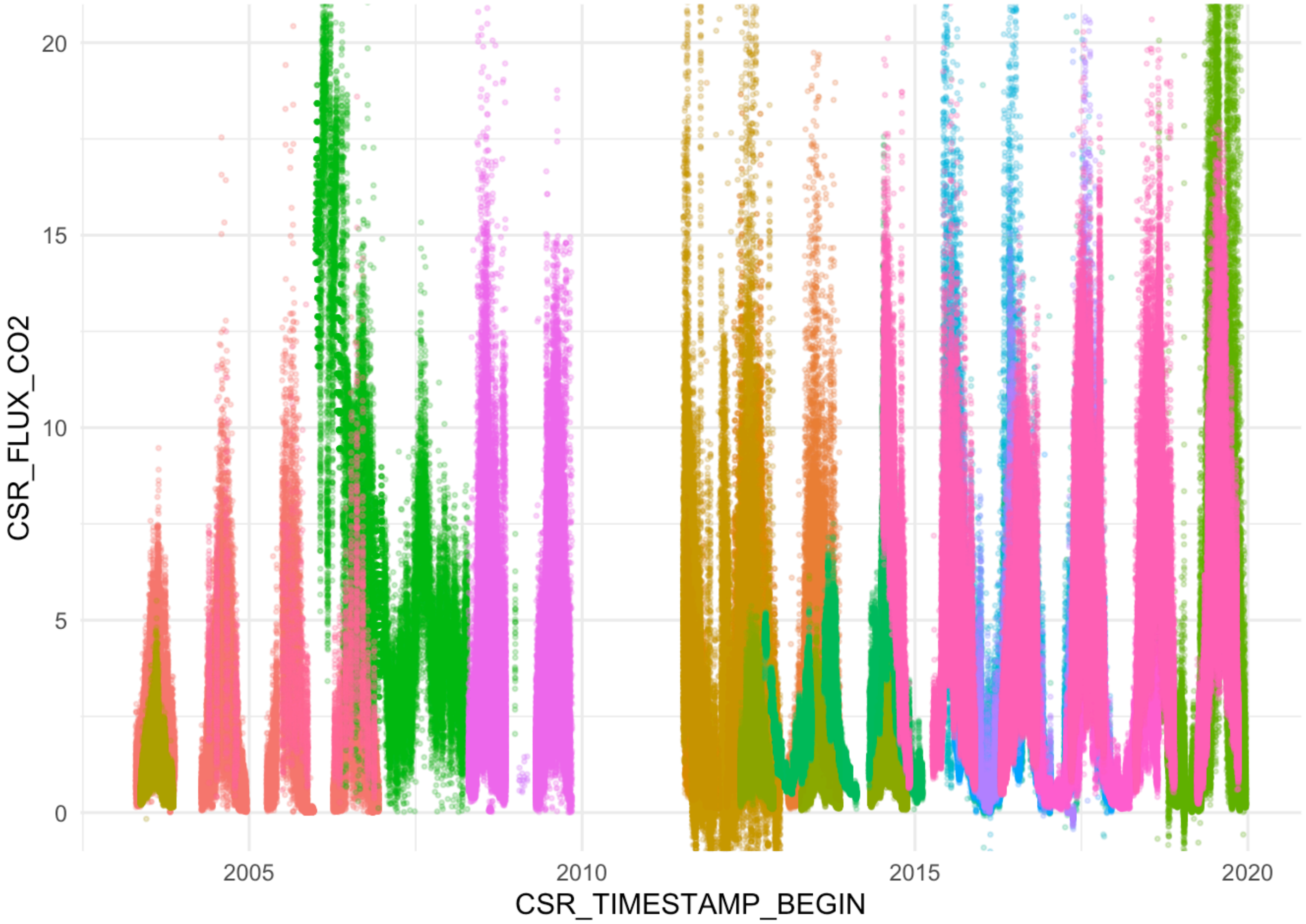
Leaflet (https://leafletjs.com) | © OpenStreetMap (https://openstreetmap.org) contributors, CC-BY-SA (https://creativecommons.org/licenses/by-sa/2.0/)

There are 24 datasets here. Extract and visualize their data:

```
tdf_dat <- csr_table("data", dbf_datasets, quiet = TRUE)
```

992,365 rows of data!

```
ggplot(tdf_dat, aes(CSR_TIMESTAMP_BEGIN, CSR_FLUX_CO2, color = CSR_DATASET)) +
  geom_point(size = 0.5, alpha = 0.25) +
  scale_color_discrete(guide = FALSE) +
  coord_cartesian(ylim = c(0, 20))
#> Warning: The `guide` argument in `scale_*()` cannot be `FALSE`. This was deprecate
d in
#> ggplot2 3.3.4.
#> ℹ Please use "none" instead.
```

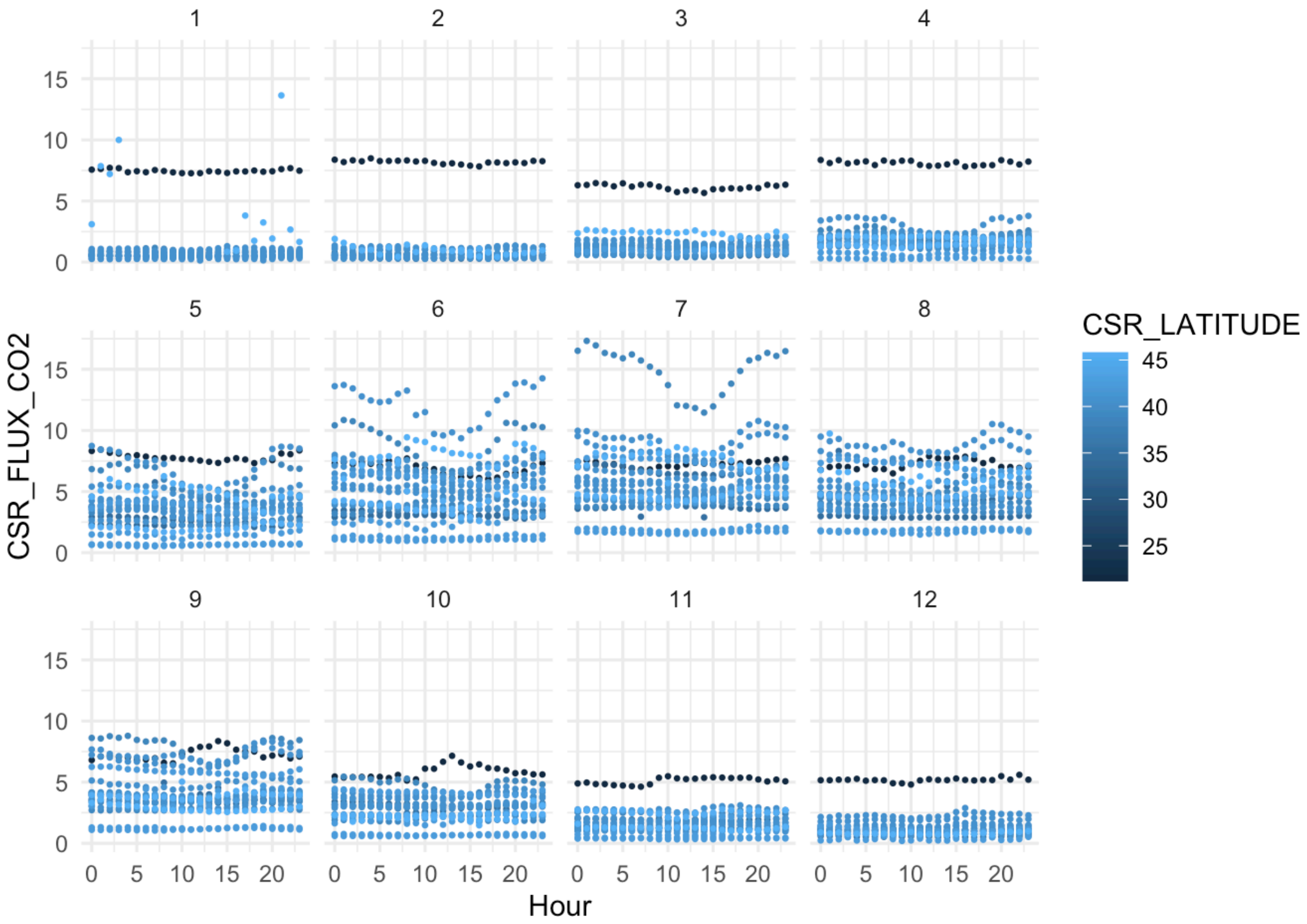


The original question we were interested in was how respiration varies over the course of the day. Say we're also interested in site latitude as a possible covariate, requiring us to join together two of the tables we've extracted.

```
site_info <- tdf[c("CSR_DATASET", "CSR_SITE_NAME", "CSR_LATITUDE")]
# join two tables together
tdf_combined <- merge(tdf_dat, site_info, by = "CSR_DATASET")
# add some new fields
tdf_combined$Hour <- hour(tdf_combined$CSR_TIMESTAMP_BEGIN)
tdf_combined$Month <- month(tdf_combined$CSR_TIMESTAMP_BEGIN)

# for each month, compute mean flux for each hour of the day
tdf_smry <- aggregate(CSR_FLUX_CO2 ~ CSR_DATASET + CSR_LATITUDE + Month + Hour,
                      FUN = mean, data = tdf_combined)

ggplot(tdf_smry, aes(Hour, CSR_FLUX_CO2, color = CSR_LATITUDE)) +
  geom_point(size = 0.5) + facet_wrap(~Month)
```



Note that in COSORE all timestamps are in the site's *local*, *standard* (https://en.wikipedia.org/wiki/Standard_time) time.

Other tables and data

This vignette has mostly focused on the `description` and `data` tables, and briefly mentioned `ports` and `contributors` . Others include:

- The `contributors` table: dataset contributors. The first person listed should be considered the primary point of contact for questions about the data or offers of co-authorship.
- `ancillary` : ancillary site-level data such as leaf area index, net primary production, soil texture, etc. All optional.

Two additional tables provide information about the processing of raw (contributed) data into standardized COSORE datasets:

- `columns` : describes how columns in the original dataset (i.e. as contributed) were mapped to the COSORE standard fields, including any unit changes or transformations.
- `diagnostics` : metadata about the data ingestion process: rows and columns removed, errors, etc.

These can all be extracted using the `csr_table()` function shown above.

Useful package functions for users

- `csr_database()` , demonstrated above, provides a database overview
- `csr_dataset()` , demonstrated above, loads a single dataset
- `csr_table()` , demonstrated above, load a single *table* across multiple datasets
- `csr_report_database()` generates a HTML summary of the entire database
- `csr_report_dataset()` generates an HTML summary of a single dataset
- `csr_metadata()` returns an informational table describing columns in all tables

Feedback and contributions

Feedback is welcome on any aspects of the database design, strengths, limitations, formats, documentation…please open a GitHub issue (https://github.com/bpbond/cosore/issues/new) or email; see the README.

Interested in contributing data to COSORE? Pleae contact the maintainer (mailto:bondlamberty@pnnl.gov) or open a GitHub issue.