Community ecology with multiple data tables: the interface between data management and analysis

Steve C. Walker, Guillaume Guénard, and Pierre Legendre

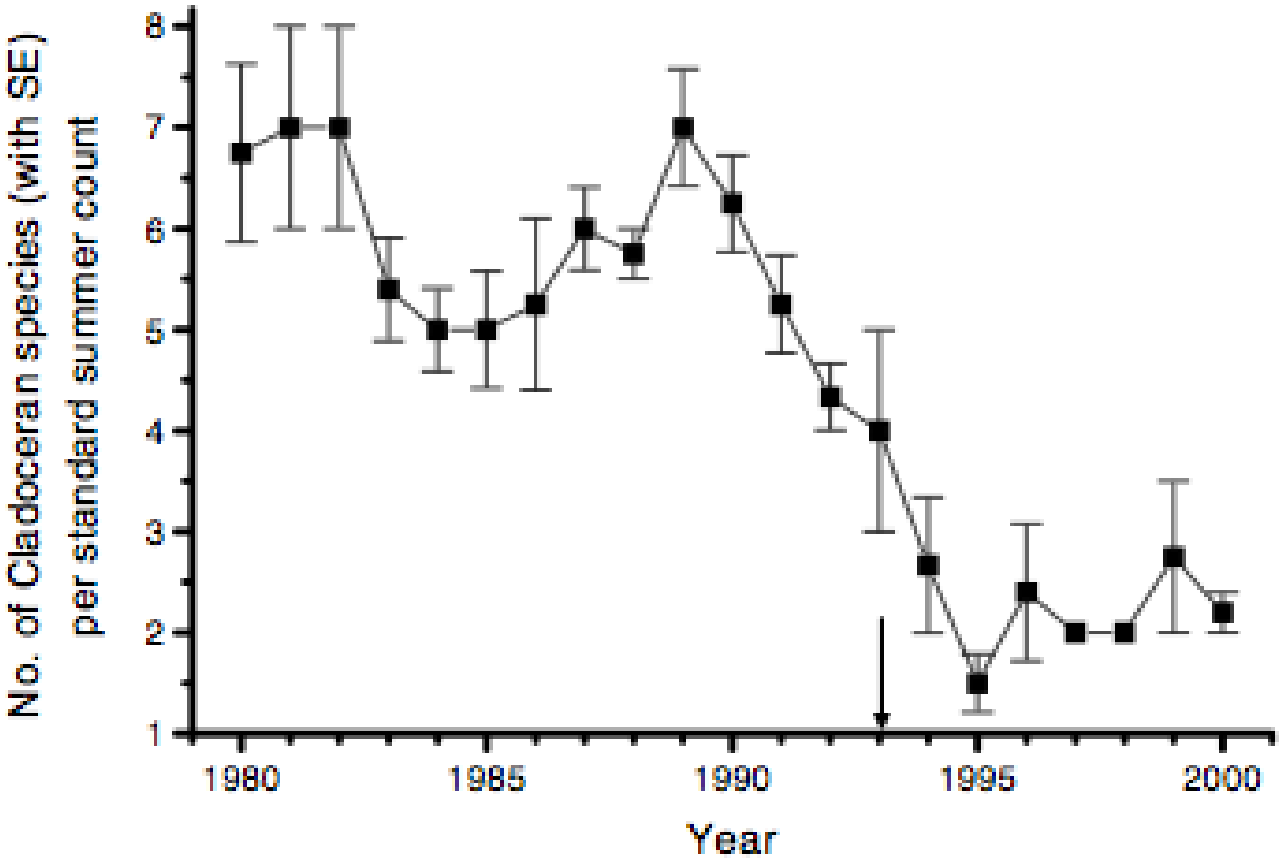
Université de Montréal
Département de Sciences Biologiques

August 12, 2011
Ecological Society of America
Austin, Texas

Bythotrephes longimanus

Wisconsin Department of Natural Resources

Bythotrephes longimanus



Yan et al. (2002)

Fourth-corner

	sp 1	sp 2	sp 3	sp 4
site 1	0.1	2.1	0.1	1.5
site 2	0.7	-0.9	1.8	3.7
site 3	1.1	0.5	1.5	2.8
site 4	1.3	-2.0	3.0	-0.2
site 5	1.7	2.0	1.3	1.2
site 6	0.8	-0.1	2.0	1.1
site 7	-2.6	-1.4	1.8	4.1
site 8	-0.0	1.5	2.3	2.3

Fourth-corner

	sp 1	sp 2	sp 3	sp 4	environment
site 1	0.1	2.1	0.1	1.5	-0.3
site 2	0.7	-0.9	1.8	3.7	1.4
site 3	1.1	0.5	1.5	2.8	-0.1
site 4	1.3	-2.0	3.0	-0.2	0.4
site 5	1.7	2.0	1.3	1.2	-0.3
site 6	0.8	-0.1	2.0	1.1	-0.6
site 7	-2.6	-1.4	1.8	4.1	2.0
site 8	-0.0	1.5	2.3	2.3	0.7

Fourth-corner

	sp 1	sp 2	sp 3	sp 4	environment
site 1	0.1	2.1	0.1	1.5	-0.3
site 2	0.7	-0.9	1.8	3.7	1.4
site 3	1.1	0.5	1.5	2.8	-0.1
site 4	1.3	-2.0	3.0	-0.2	0.4
site 5	1.7	2.0	1.3	1.2	-0.3
site 6	0.8	-0.1	2.0	1.1	-0.6
site 7	-2.6	-1.4	1.8	4.1	2.0
site 8	-0.0	1.5	2.3	2.3	0.7
trait	-1.0	-1.0	1.0	1.0	

Fourth-corner

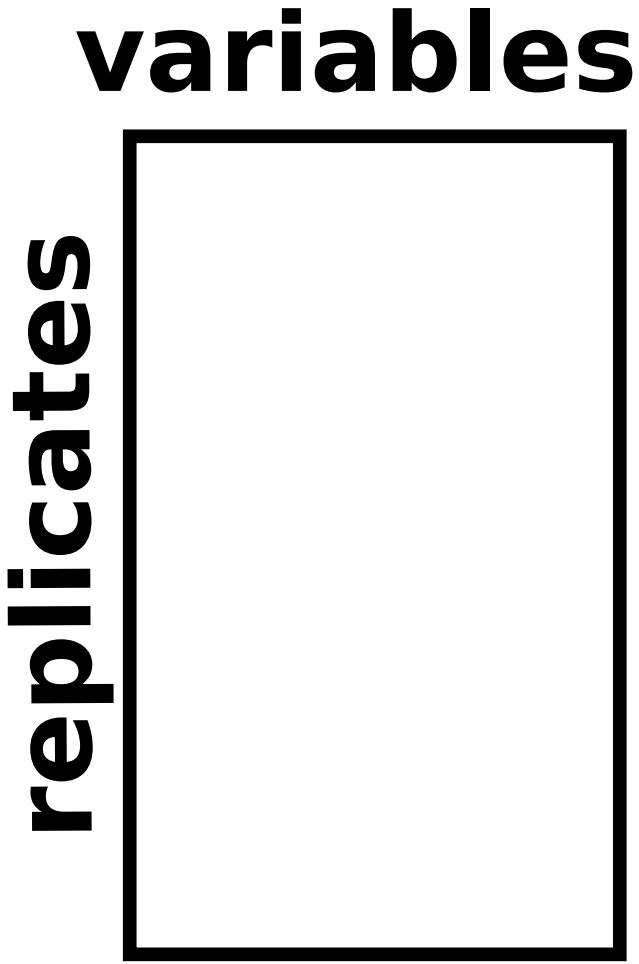
	sp 1	sp 2	sp 3	sp 4	environment
site 1	0.1	2.1	0.1	1.5	-0.3
site 2	0.7	-0.9	1.8	3.7	1.4
site 3	1.1	0.5	1.5	2.8	-0.1
site 4	1.3	-2.0	3.0	-0.2	0.4
site 5	1.7	2.0	1.3	1.2	-0.3
site 6	0.8	-0.1	2.0	1.1	-0.6
site 7	-2.6	-1.4	1.8	4.1	2.0
site 8	-0.0	1.5	2.3	2.3	0.7
trait	-1.0	-1.0	1.0	1.0	→ **??** ←

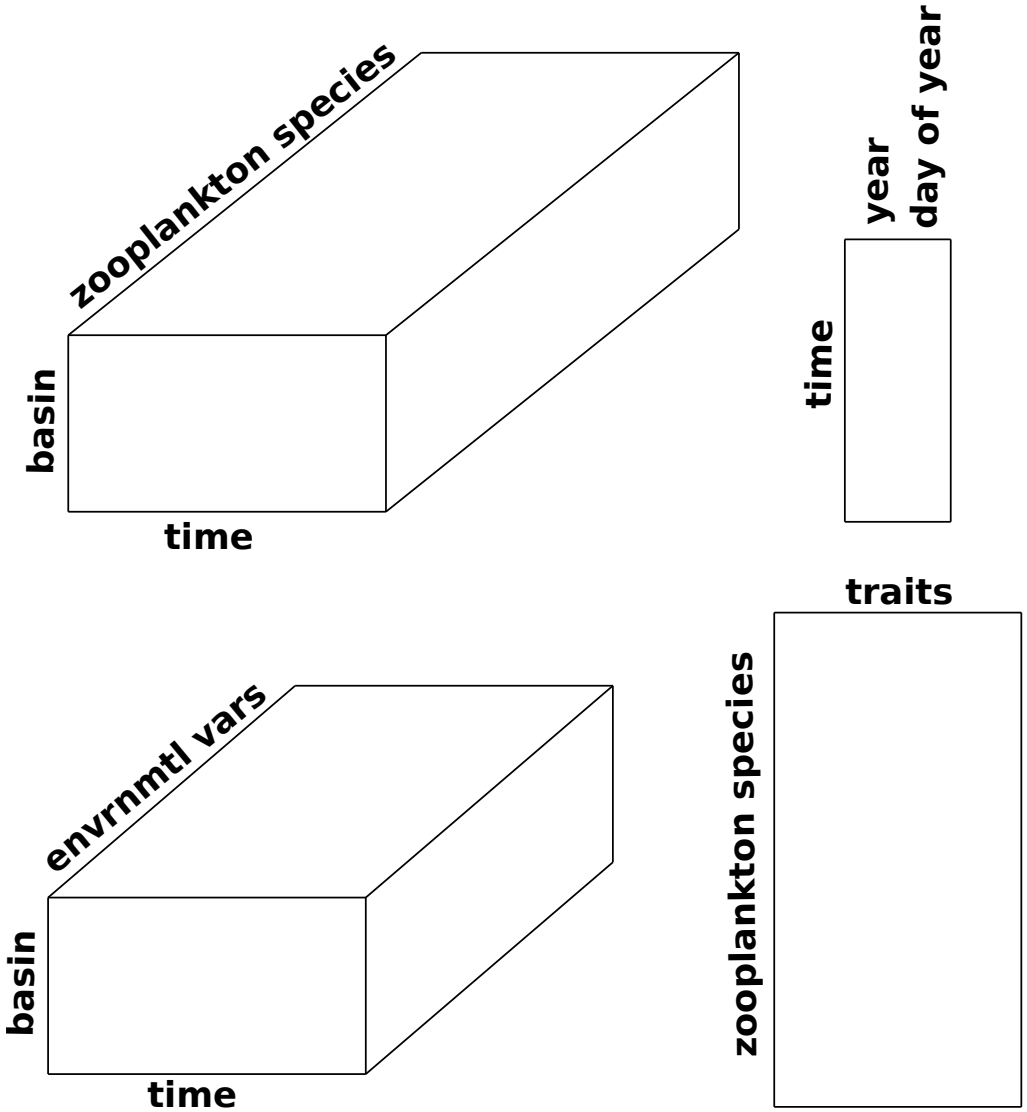
↑

Statistical methods for analyzing 'fourth-corner'-esque data

- Chessel et al. (1996) — RLQ analysis
- Legendre et al. (1997) — coined term 'fourth-corner'
- Ives and Godfray (2006) — mixed models of phylogenetically-structured foodwebs
- Dray and Legendre (2008) — extends Legendre et al.
- Pillar and Duarte (2010) — phylogenetic null models
- Leibold et al. (2010) — semi-partial correlations
- Ives and Helmus (in press) — phylogenetic generalized linear mixed models (PGLMMs)

The data frame — replicates-by-variables





Cantin et al. 2011 – Lac Croche, Québec, Canada

Linear algebra as data management

Solve for the b's

$$\begin{array}{ccccccccc} y_1 & = & b_1 x_{11} & + & b_2 x_{12} & + & \dots & + & b_m x_{1m} \\ y_2 & = & b_1 x_{21} & + & b_2 x_{22} & + & \dots & + & b_m x_{2m} \\ \vdots & & \vdots & & \vdots & & \ddots & & \vdots \\ y_n & = & b_1 x_{n1} & + & b_2 x_{n2} & + & \dots & + & b_m x_{nm} \end{array} \quad (1)$$

Linear algebra as data management

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{2}$$

Linear algebra as data management

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

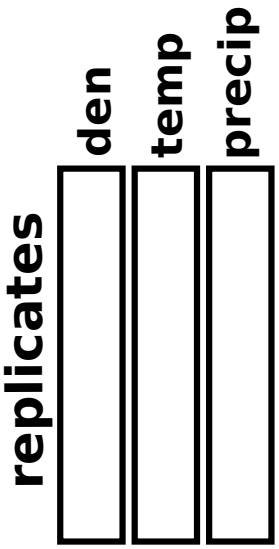
$$\mathbf{y} = \mathbf{X}\mathbf{b} \tag{2}$$

Linear algebra as data management

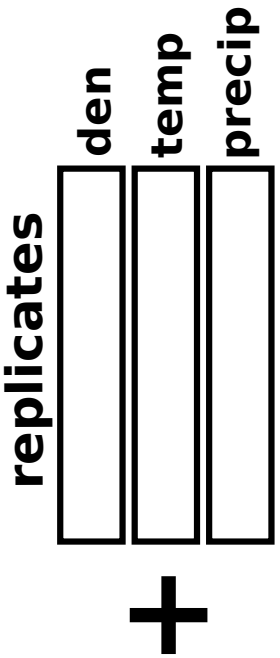
$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\mathbf{b} \\ \mathbf{X}^{\mathsf{T}}\mathbf{y} &= \mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{b} \\ (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} &= \mathbf{b} \end{aligned} \tag{2}$$

The R framework for data management

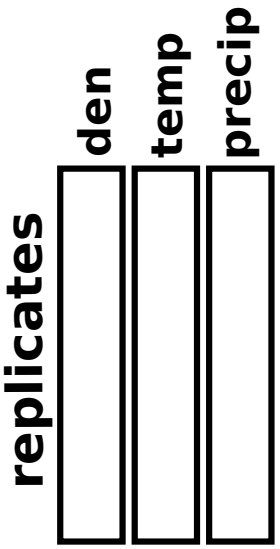


The R framework for data management



den ~ temp + precip

The R framework for data management



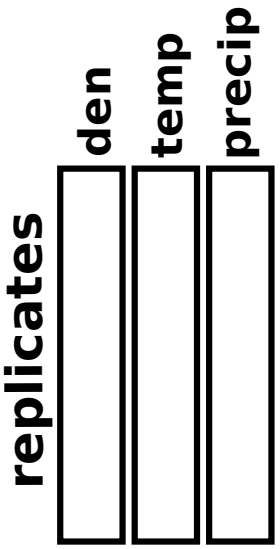
+

den ~ temp + precip

+

lm / glmer / plot / xyplot

The R framework for data management



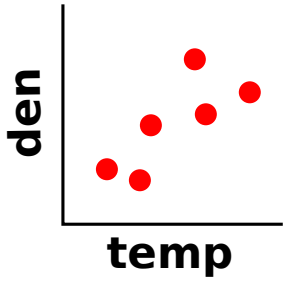
+

den ~ temp + precip

+

lm / glmer / plot / xyplot

=



p < 0.0001

	coef	s.e.
(intcpt)	-1.2	0.4
temp	2.1	0.1
precip	-0.1	0.1

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables —

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables — without explicit reference to complex mathematical and algorithmic details.

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables — without explicit reference to complex mathematical and algorithmic details.
- It also provides access to those details,

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables — without explicit reference to complex mathematical and algorithmic details.
- It also provides access to those details, which are required (1)

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables — without explicit reference to complex mathematical and algorithmic details.
- It also provides access to those details, which are required (1) for more effective analyses and (2)

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables — without explicit reference to complex mathematical and algorithmic details.
- It also provides access to those details, which are required (1) for more effective analyses and (2) to develop new methods of analysis within the framework.

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables — without explicit reference to complex mathematical and algorithmic details.
- It also provides access to those details, which are required (1) for more effective analyses and (2) to develop new methods of analysis within the framework.
- As new methods are developed,

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables — without explicit reference to complex mathematical and algorithmic details.
- It also provides access to those details, which are required (1) for more effective analyses and (2) to develop new methods of analysis within the framework.
- As new methods are developed, researchers simply pass their data frames to new functions in much the same way they would pass them to older functions.

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables — without explicit reference to complex mathematical and algorithmic details.
- It also provides access to those details, which are required (1) for more effective analyses and (2) to develop new methods of analysis within the framework.
- As new methods are developed, researchers simply pass their data frames to new functions in much the same way they would pass them to older functions.
- Thus, by separating low-level methods development from high-level data analysis,

The R framework for data management

- This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables — without explicit reference to complex mathematical and algorithmic details.
- It also provides access to those details, which are required (1) for more effective analyses and (2) to develop new methods of analysis within the framework.
- As new methods are developed, researchers simply pass their data frames to new functions in much the same way they would pass them to older functions.
- Thus, by separating low-level methods development from high-level data analysis, R fosters the formation of a community of researchers where both methodologists and analysts can have mutually beneficial interactions.

Goal Analyze multiple table data sets using this framework

Goal Analyze multiple table data sets using this framework

Problem R doesn't do multiple tables 'out-of-the-box'

Goal Analyze multiple table data sets using this framework

Problem R doesn't do multiple tables 'out-of-the-box'

Strategy Develop some theory to better understand multiple table data management and then use that theory to extend the R framework to allow multiple table data sets

Goal Analyze multiple table data sets using this framework

Problem R doesn't do multiple tables 'out-of-the-box'

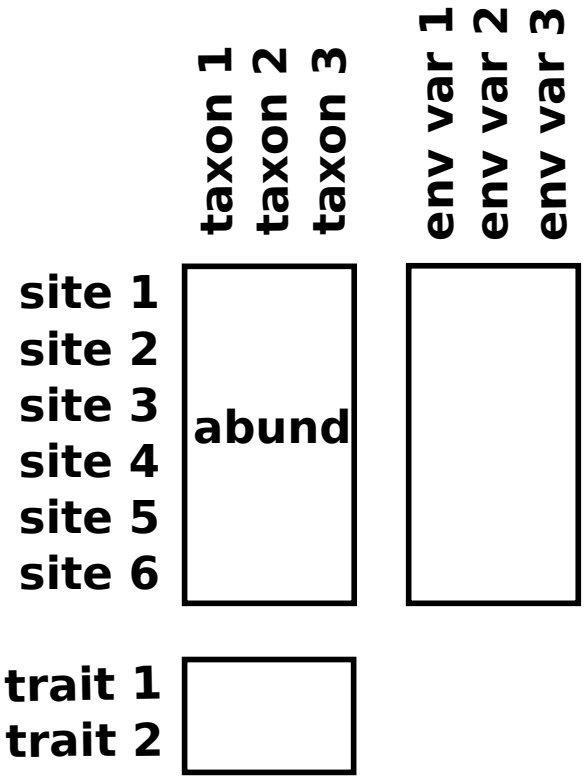
Strategy Develop some theory to better understand multiple table data management and then use that theory to extend the R framework to allow multiple table data sets

$$\text{DATA FRAME} + \text{FORMULA} + \text{FUNCTION} = \text{ANALYSIS}$$

Goal	Analyze multiple table data sets using this framework
Problem	R doesn't do multiple tables 'out-of-the-box'
Strategy	Develop some theory to better understand multiple table data management and then use that theory to extend the R framework to allow multiple table data sets

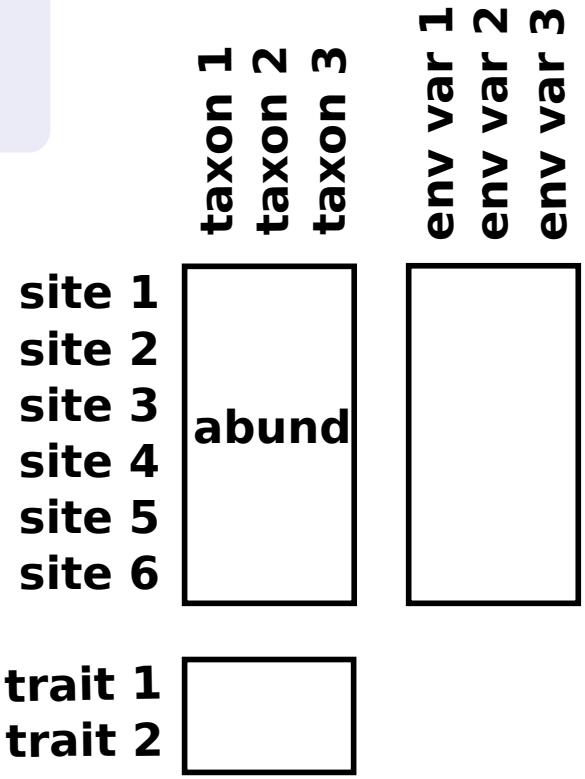
$$\begin{array}{c} \text{DATA LIST} \\ \downarrow \\ \text{DATA FRAME} + \text{FORMULA} + \text{FUNCTION} = \text{ANALYSIS} \end{array}$$

How can we convert this to a data frame?

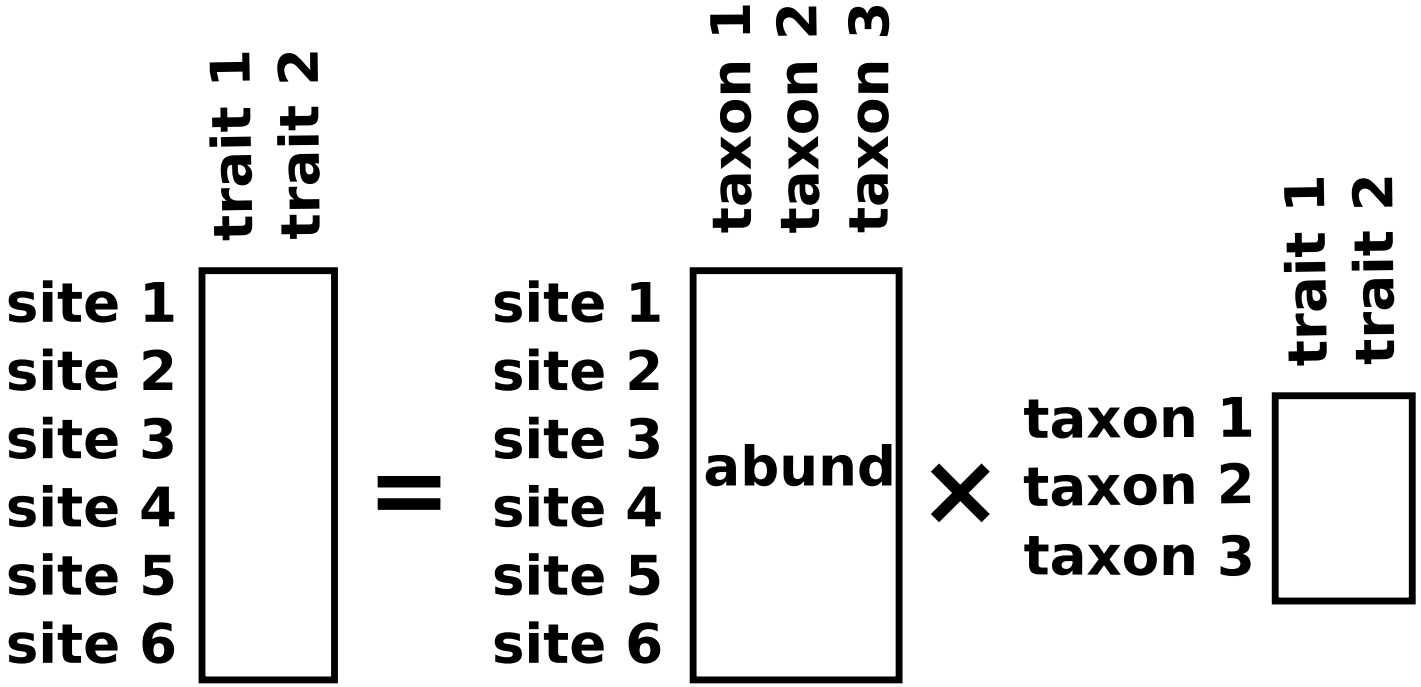


How can we convert this to a data frame?

- Lost information
- Redundant information

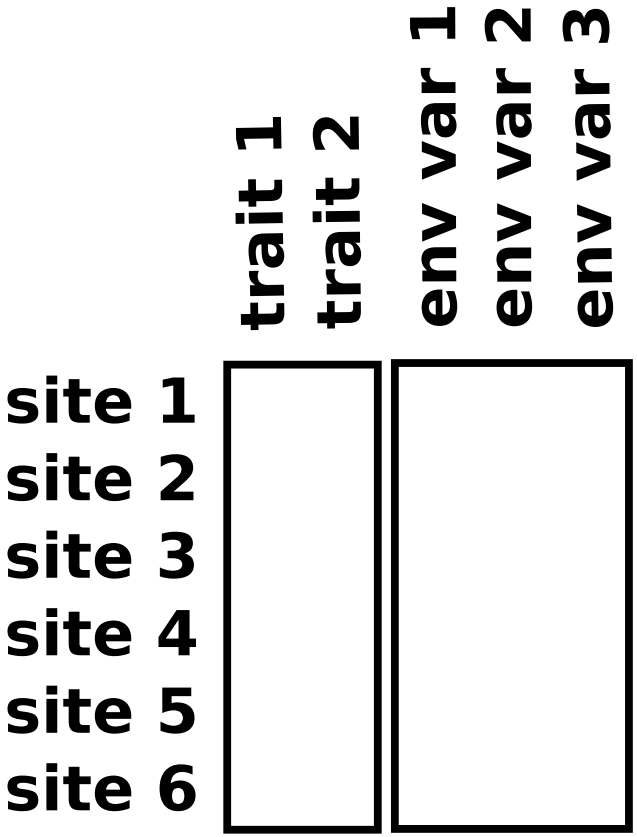


Lost information



e.g. Leibold et al. (2010)

Lost information



e.g. Leibold et al. (2010)

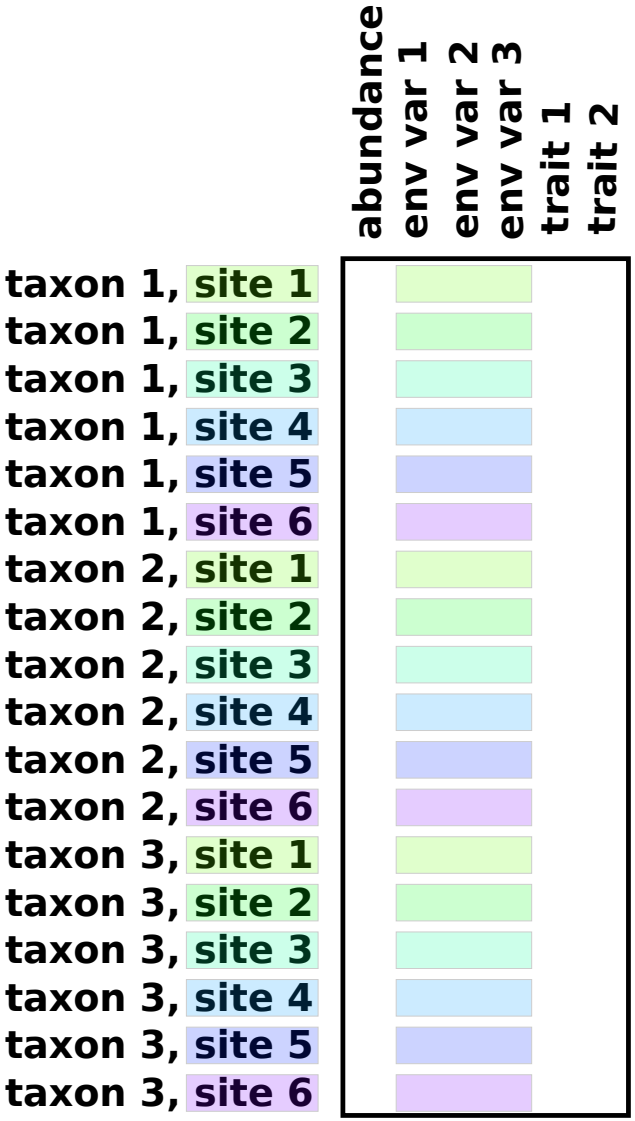
Redundant information

	abundance	env var 1	env var 2	env var 3	trait 1	trait 2
taxon 1, site 1						
taxon 1, site 2						
taxon 1, site 3						
taxon 1, site 4						
taxon 1, site 5						
taxon 1, site 6						
taxon 2, site 1						
taxon 2, site 2						
taxon 2, site 3						
taxon 2, site 4						
taxon 2, site 5						
taxon 2, site 6						
taxon 3, site 1						
taxon 3, site 2						
taxon 3, site 3						
taxon 3, site 4						
taxon 3, site 5						
taxon 3, site 6						

Redundant information

	abundance	env var 1	env var 2	env var 3	trait 1	trait 2
taxon 1, site 1						
taxon 1, site 2						
taxon 1, site 3						
taxon 1, site 4						
taxon 1, site 5						
taxon 1, site 6						
taxon 2, site 1						
taxon 2, site 2						
taxon 2, site 3						
taxon 2, site 4						
taxon 2, site 5						
taxon 2, site 6						
taxon 3, site 1						
taxon 3, site 2						
taxon 3, site 3						
taxon 3, site 4						
taxon 3, site 5						
taxon 3, site 6						

Redundant information



Redundant information

	abundance	env var 1	env var 2	env var 3	trait 1	trait 2
taxon 1, site 1						
taxon 1, site 2						
taxon 1, site 3						
taxon 1, site 4						
taxon 1, site 5						
taxon 1, site 6						
taxon 2, site 1						
taxon 2, site 2						
taxon 2, site 3						
taxon 2, site 4						
taxon 2, site 5						
taxon 2, site 6						
taxon 3, site 1						
taxon 3, site 2						
taxon 3, site 3						
taxon 3, site 4						
taxon 3, site 5						
taxon 3, site 6						

Classifying the fourth-corner dimensions

Rules

- Dimensions that can not grow with more sampling represent groups of variables

Classifying the fourth-corner dimensions

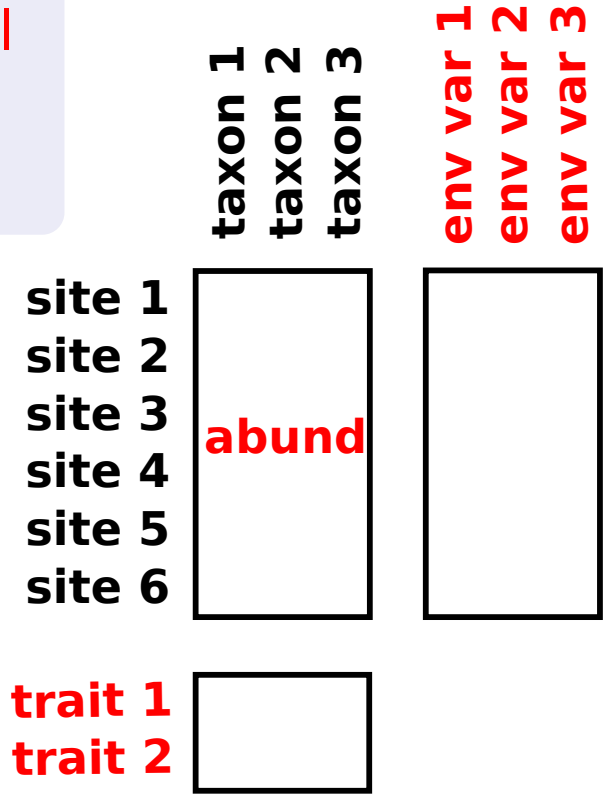
Rules

- Dimensions that can not grow with more sampling represent groups of variables
- Dimensions that can grow with more sampling represent replication

Classifying the fourth-corner dimensions

Variables

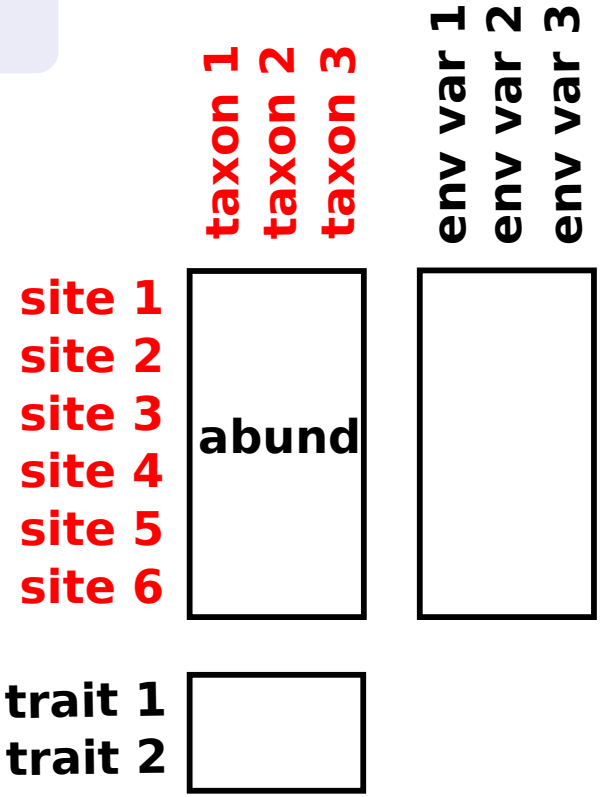
- abundance
- environmental variables
- traits



Classifying the fourth-corner dimensions

Replicates

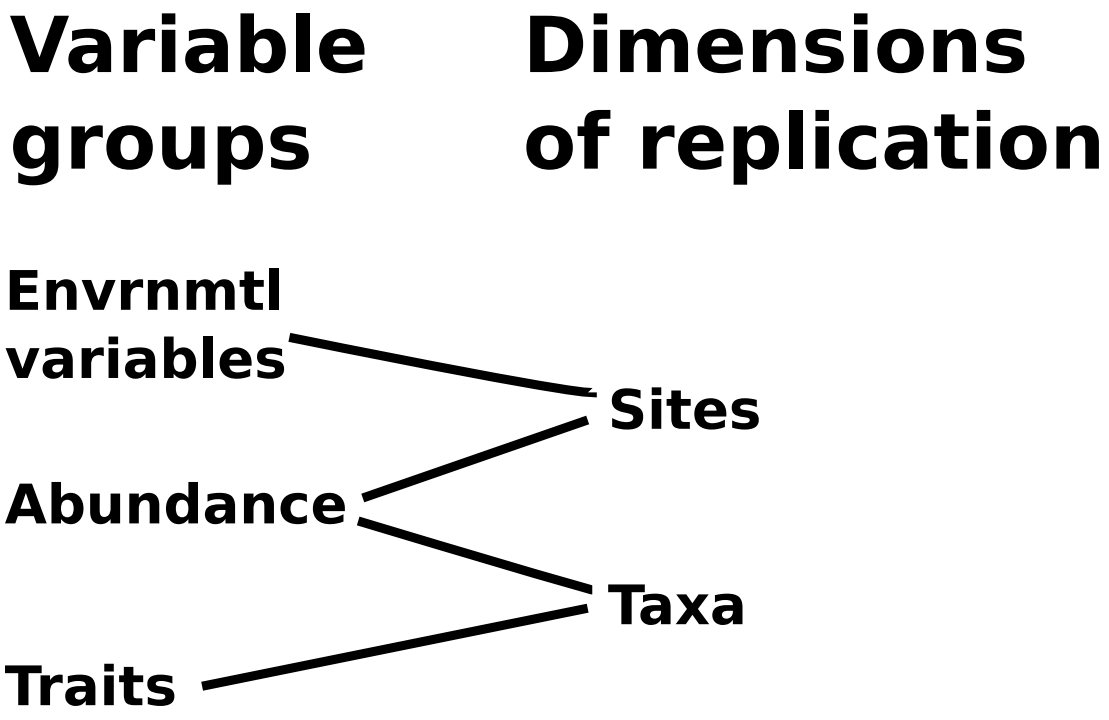
- sites
- taxa



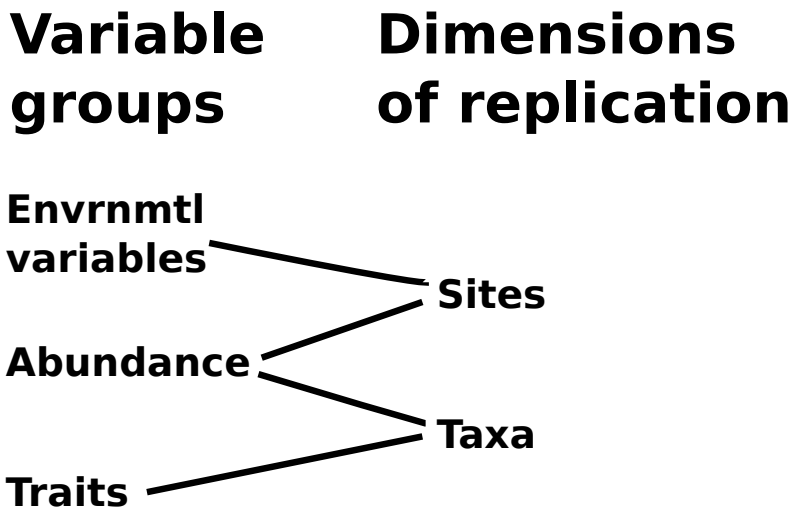
Classifying the fourth-corner dimensions

	abundance	env var 1	env var 2	env var 3	trait 1	trait 2
taxon 1, site 1						
taxon 1, site 2						
taxon 1, site 3						
taxon 1, site 4						
taxon 1, site 5						
taxon 1, site 6						
taxon 2, site 1						
taxon 2, site 2						
taxon 2, site 3						
taxon 2, site 4						
taxon 2, site 5						
taxon 2, site 6						
taxon 3, site 1						
taxon 3, site 2						
taxon 3, site 3						
taxon 3, site 4						
taxon 3, site 5						
taxon 3, site 6						

Bipartite graphs



Biadjacency matrices



	abund.	env.	traits
sites	1	1	0
taxa	1	0	1

Biadjacency matrices

Identifying data sets that are not multiple-table

If a data set has a biadjacency matrix with at least one row of all ones,

Biadjacency matrices

Identifying data sets that are not multiple-table

If a data set has a biadjacency matrix with at least one row of all ones,

Example

	abund.	env.	geog.	traits
space	1	1	1	1
time	1	1	0	0
taxa	1	0	0	1

Biadjacency matrices

Identifying data sets that are not multiple-table

If a data set has a biadjacency matrix with at least one row of all ones, then that data set can be expressed as a single table

Example

	abund.	env.	geog.	traits
space	1	1	1	1
time	1	1	0	0
taxa	1	0	0	1

Biadjacency matrices

Identifying data sets that are not multiple-table

If a data set has a biadjacency matrix with at least one row of all ones, then that data set can be expressed as a single table without redundant or lost information.

Example

	abund.	env.	geog.	traits
space	1	1	1	1
time	1	1	0	0
taxa	1	0	0	1

Biadjacency matrices

Necessarily un-correlated variables

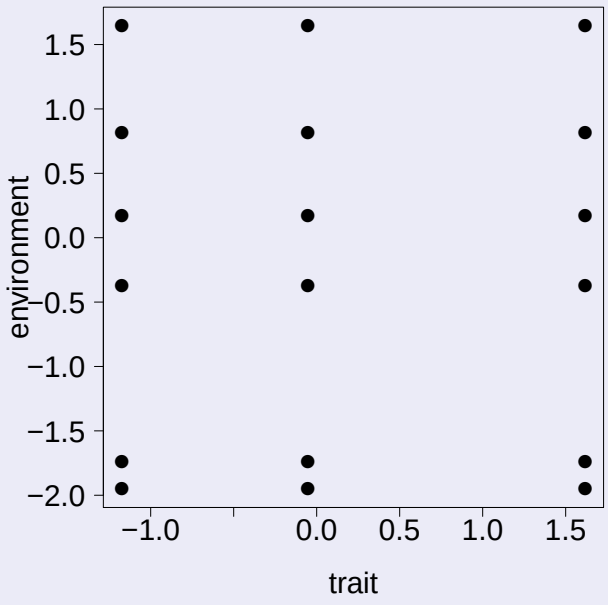
If two columns in a biadjacency matrix are orthogonal (i.e. have zero dot product)

Biadjacency matrices

Necessarily un-correlated variables

If two columns in a biadjacency matrix are orthogonal (i.e. have zero dot product)

Example



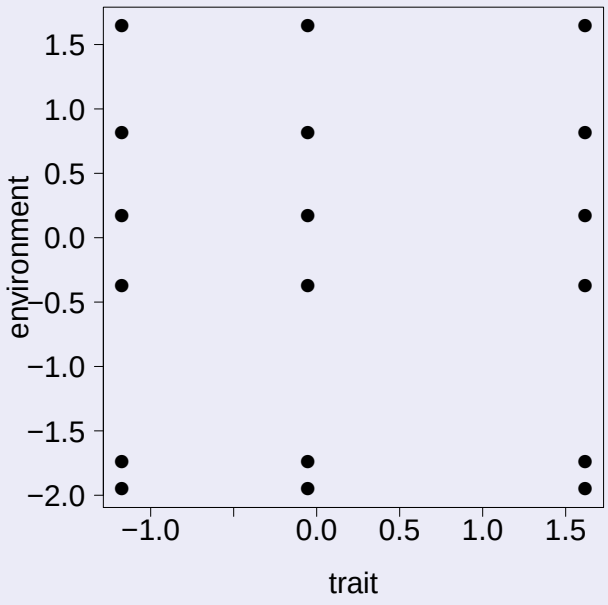
	abund.	env.	traits
sites	1	1	0
taxa	1	0	1

Biadjacency matrices

Necessarily un-correlated variables

If two columns in a biadjacency matrix are orthogonal (i.e. have zero dot product) then the associated variable groups are also orthogonal (i.e. uncorrelated),

Example



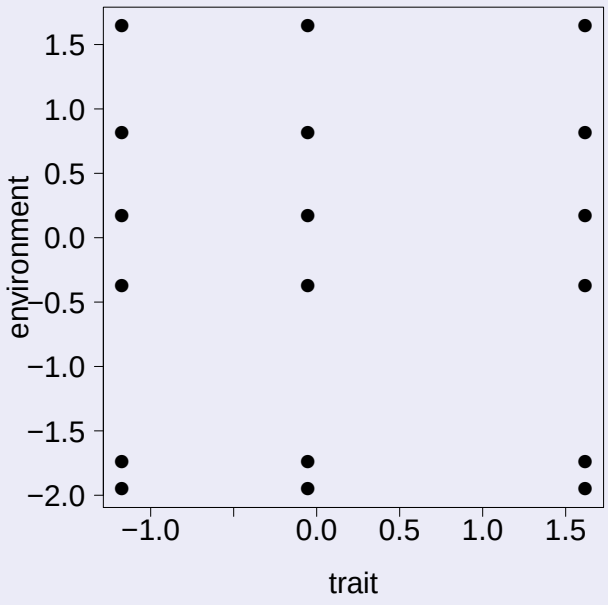
	abund.	env.	traits
sites	1	1	0
taxa	1	0	1

Biadjacency matrices

Necessarily un-correlated variables

If two columns in a biadjacency matrix are orthogonal (i.e. have zero dot product) then the associated variable groups are also orthogonal (i.e. uncorrelated), after the data set has been coerced to a data frame by the method of repetition.

Example



	abund.	env.	traits
sites	1	1	0
taxa	1	0	1

Biadjacency matrices

The meaning of zeros

A variable with a zero for a particular dimension of replication, is assumed (statistically) constant across that dimension.

Biadjacency matrices

The meaning of zeros

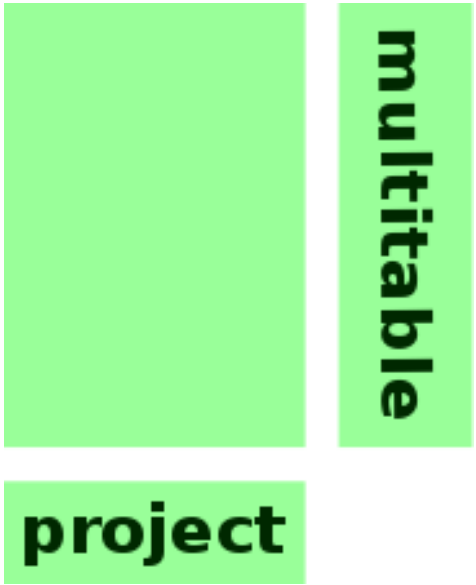
A variable with a zero for a particular dimension of replication, is assumed (statistically) constant across that dimension.

Example

	abund.	env.	traits
sites	1	1	0
taxa	1	0	1

```
> library(multitable)
```

http://multitable.r-forge.r-project.org/



```
> dl <- data.list(abundance = ab, temperature = tp,
+ bodysize = bs, dnames = c("sites", "species"))
```

```
> dl
abundance:
---------
       sppA  sppB  sppC
siteA  1.17 -0.04  0.85
siteB  0.65 -0.06 -0.37
siteC  0.51 -2.73  1.07
siteD -1.19  2.81  0.17
siteE -0.69 -0.21  0.38
Replicated along:  || sites || || species ||

temperature:
-----------
siteA siteB siteC siteD siteE
-1.04  0.77  0.82 -0.38 -0.06
Replicated along:  || sites ||
```

continued...

```
bodysize:
--------
 sppA  sppB  sppC
-0.45 -0.07  1.48
Replicated along:  || species ||


REPLICATION DIMENSIONS:
  sites species
      5       3
```

```
> dl[1:3, ]
```

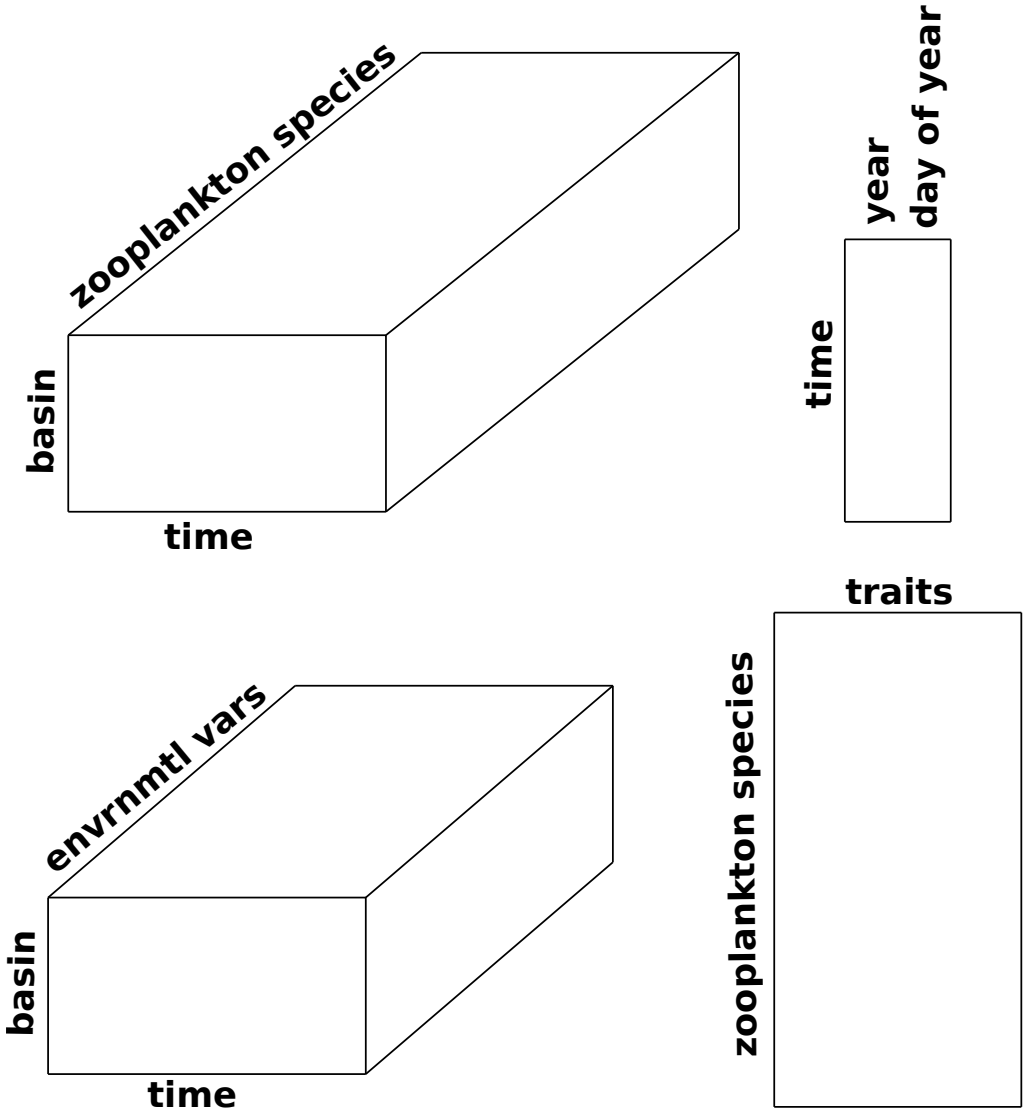
```
abundance:
---------
      sppA  sppB  sppC
siteA 1.17 -0.04  0.85
siteB 0.65 -0.06 -0.37
siteC 0.51 -2.73  1.07
Replicated along:  || sites || species ||


temperature:
-----------
siteA siteB siteC
-1.04  0.77  0.82
Replicated along:  || sites ||
```

continued...

```
bodysize:
--------
 sppA  sppB  sppC
-0.45 -0.07  1.48
Replicated along:  || species ||


REPLICATION DIMENSIONS:
  sites species
      3       3
```

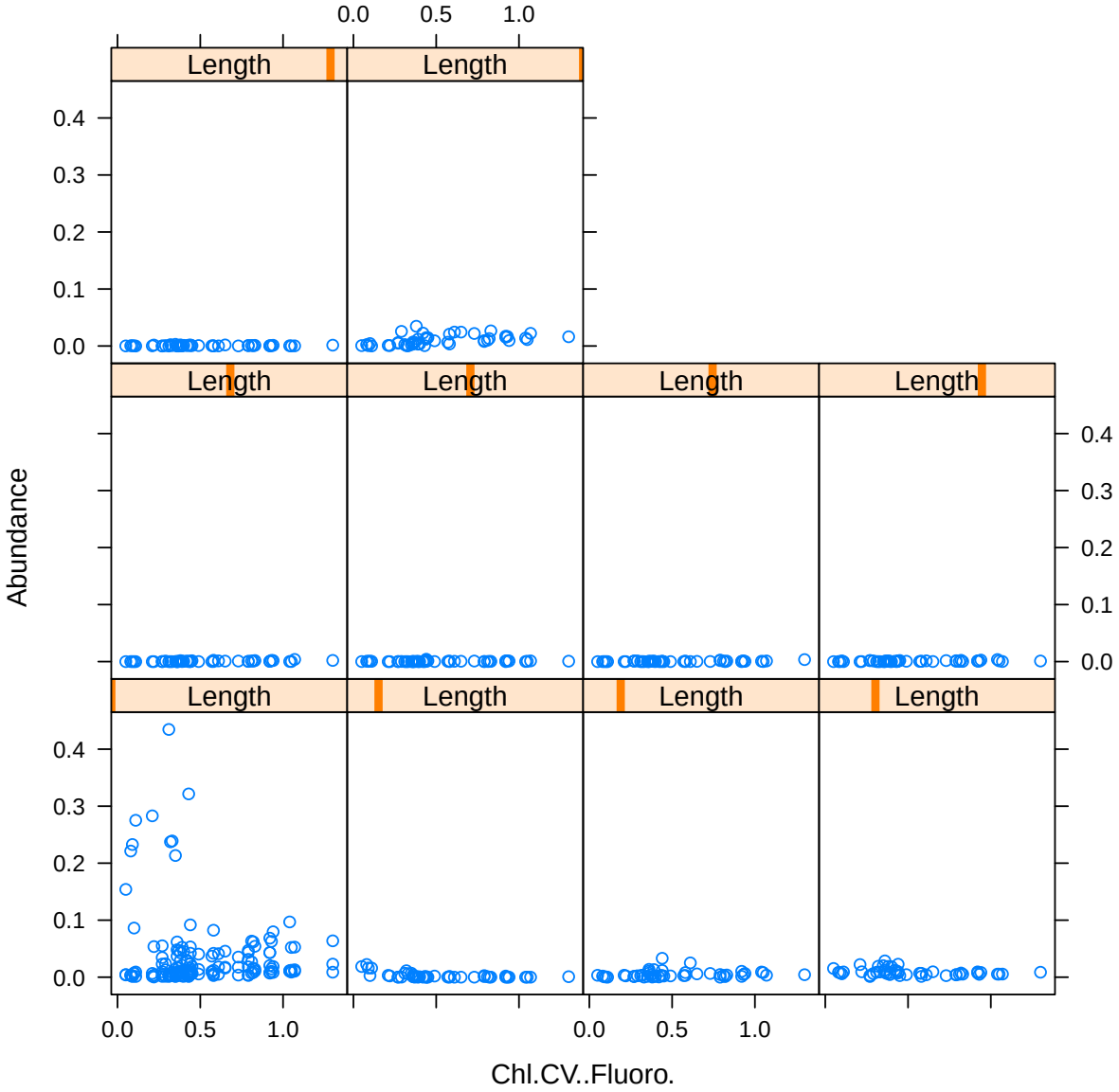
```
> lm(abundance ~ temperature * bodysize, dl)

Call:
lm(formula = abundance ~ temperature * bodysize,
data = dl)

Coefficients:
(Intercept)     temperature     bodysize
    0.08795        -0.40439      0.20848

temperature:bodysize
             0.09822
```

```
> as.data.frame(dl)
   abundance temperature bodysize
1       1.17  -1.3453605    -0.45
2       0.65   0.9475797    -0.45
3       0.51   1.0109206    -0.45
4      -1.19  -0.5092608    -0.45
5      -0.69  -0.1038791    -0.45
6      -0.04  -1.3453605    -0.07
7      -0.06   0.9475797    -0.07
8      -2.73   1.0109206    -0.07
9       2.81  -0.5092608    -0.07
10     -0.21  -0.1038791    -0.07
11      0.85  -1.3453605     1.48
12     -0.37   0.9475797     1.48
13      1.07   1.0109206     1.48
14      0.17  -0.5092608     1.48
15      0.38  -0.1038791     1.48
```



Cantin et al. 2011 – Lac Croche, Québec, Canada

```
> xyplot(Abundance ~ Chl.CV..Fluoro. | Length,
+ data = as.data.frame(dl))
```



Take-home message

Don't be scared of multiple table data sets.

Take-home message

Don't be scared of multiple table data sets. Collect more of them!

Take-home message

Don't be scared of multiple table data sets. Collect more of them! With the right data management framework,

Take-home message

Don't be scared of multiple table data sets. Collect more of them! With the right data management framework, multiple table data can be modeled in much the same way that we model single table data.

Acknowledgements

- Natural Sciences and Engineering Research Council of Canada
- Laura Timms (McGill University)
- Beatrix Beisner (Université du Québec à Montréal)
- Ben Bolker (McMaster University)
- The many people who gave their time to develop free software: R and LaTeX

http://multitable.r-forge.r-project.org/