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 The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Statistical issues

Data management

neory

 $\begin{array}{c} \mathsf{Multiple} \longrightarrow \mathsf{single} \\ \mathsf{Bipartite} \ \mathsf{graphs} \end{array}$

/lethods

Subscripting

Coercion

Real data

Introduction

Traits and the ecology of communities

Statistical issues
The data management-analysis interface

Theory

Converting multiple tables to one single table Analyzing data structure with bipartite graphs

Computational methods

Multiple tables in one R object: the data list Subscripting multiple tables simultaneously Using the simplest functions (e.g. 1m) Coercing data lists to data frames Real complex zooplankton community data

Conclusion

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Traits and ecology Statistical issues

Data management

Theory

Multiple → single Bipartite graphs

/lethods

Data lists Subscripting

Coercion

Real data

Bythotrephes longimanus

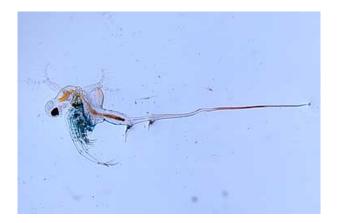


Steve C. Walker Guillaume Guénard Pierre Legendre

Traits and ecology

Multiple → single

Real data



Wisconsin Department of Natural Resources



Bythotrephes longimanus

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre



Traits and ecology

Data management

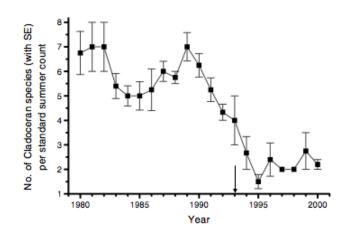
Theory

Multiple → single Bipartite graphs

Methods

Subscripting Im

Coercion



Introduction

Traits and the ecology of communities

Statistical issues

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Statistical issues

Multiple → single

Real data

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Statistical issues

Multiple → single Bipartite graphs

lm

	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	8.0	-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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and ecology Statistical issues

.

Multiple → single Bipartite graphs

Methods

Data lists Subscripting

Im Coercion

eal data

	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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and ecology Statistical issues

a management

heory

 $\begin{array}{c} \mathsf{Multiple} \longrightarrow \mathsf{single} \\ \mathsf{Bipartite} \ \mathsf{graphs} \end{array}$

1ethods

Data lists Subscripting

lm C-----

eal data

	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	

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and ecology Statistical issues

neory

Multiple → single

/lethods

Data lists Subscripting

Im Coercion

percion eal data

	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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and ecolog

Data management

heory

Multiple → single

Methods

Data lists Subscripting

n

oercion eal 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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Statistical issues

Data management

Theory

Multiple → single
Bipartite graphs

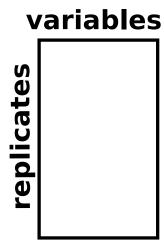
Methods

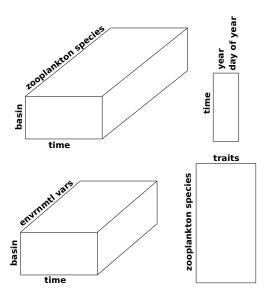
Data lists

Subscripting

Coercion

Real data





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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Statistical issues

tistical issues

Theo

Multiple → single
Bipartite graphs

Methods

Data lists Subscripting

Im

oercion

Real data

Introduction

Traits and the ecology of communities Statistical issues

The data management-analysis interface

Theory

Converting multiple tables to one single table Analyzing data structure with bipartite graphs

Computational methods

Multiple tables in one R object: the data list Subscripting multiple tables simultaneously Using the simplest functions (e.g. 1m) Coercing data lists to data frames Real complex zooplankton community data

Conclusion

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Statistical issues

Data management

Theory

Multiple → single Bipartite graphs

Methods

Data lists Subscripting

Coercion Real data

iveai uata

4□ > 4□ > 4□ > 4□ > 4□ >

Solve for the b's

$$y_{1} = b_{1}x_{11} + b_{2}x_{12} + \dots + b_{m}x_{1m}$$

$$y_{2} = b_{1}x_{21} + b_{2}x_{22} + \dots + b_{m}x_{2m}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$y_{n} = b_{1}x_{n1} + b_{2}x_{n2} + \dots + b_{m}x_{nm}$$

$$(1)$$

Traits and ecology

Data management

Multiple → single

Real data

Linear algebra as data management

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Statistical issues

Data management

Theory

Multiple → single

Methods

ata lists ubscripting

ercion

(2)

data

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

Linear algebra as data management

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Data management

Multiple → single

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

y = Xb

(2)

troduction

Statistical issues

Data management

heory

Multiple → single Bipartite graphs

Methods

Data lists Subscripting

percion

eal data

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

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

$$\mathbf{X}^{\mathsf{T}}\mathbf{y} = \mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{b}$$

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} = \mathbf{b}$$
(2)

temp precip

replicates

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and e

Data management

Theory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscriptin

lm

Real data



Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and ecology Statistical issues

Data management

Theory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

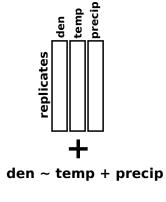
Methods

Data lists

Subscripting

Coercion

Real data





Steve C. Walker Guillaume Guénard Pierre Legendre

roduction

Traits and eco

Data management

Theory

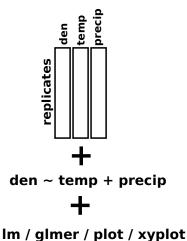
Multiple → single
Bipartite graphs

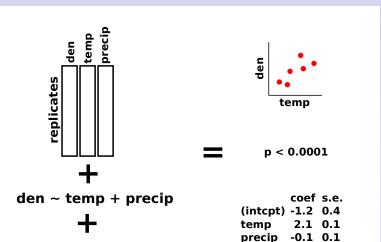
Methods

Data lists Subscripting

m

Real data





Im / glmer / plot / xyplot

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Statistical issues

Data management

Theory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

m

Real data

The interface between data management and analysis

► This framework allows ecologists to concentrate on their primary interests

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Traits and eco

Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

1

Real data

The interface between data management and analysis

► This framework allows ecologists to concentrate on their primary interests — the relationships between ecological variables — Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Traits and e Statistical is

Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

1

Real data

► 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 interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Statistical issue

Data management

Γheory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists

n

Real data

itroduction

Statistical issues

Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

/lethods

Data lists

1

Real data

- ➤ 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,

ntroduction

Statistical issues

Data management

Theory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

/lethods

Data lists

.....

Real data

- 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)

- Steve C. Walker Guillaume Guénard Pierre Legendre
 - itroduction

Statistical issues

Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

/lethods

Data lists Subscripting

ercion

Real data

- 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)

itroduction

Statistical issues

Data management

Theory

Multiple → single Bipartite graphs

Methods

Data lists Subscripting

percion

Real data

- ➤ 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.

itroduction

Statistical issues

Data management

Theory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

.

Real data

- 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,

troduction

Statistical issues

Data management

Theory

Multiple → single Bipartite graphs

Methods

Data lists Subscripting

Coercion

tear uata

- 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.

troduction

Statistical issues

Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \longrightarrow \mathsf{single} \\ \mathsf{Bipartite} \ \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

Coercion

Real data

- 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,

troduction

Statistical issues

Data management

heory

Multiple → single Bipartite graphs

Methods

Data lists Subscripting

Coercion

- 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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

meroduction

Data management

Theory

Multiple → single Bipartite graphs

Methods

Data lists

n

Coercion Real data

Goal Analyze multiple table data sets using this framework

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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Data management

Multiple → single

Real data

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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and ecolog

Data management

Theory

Multiple → single Bipartite graphs

Methods

Data lists Subscripting

Real data

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

DATA FRAME + FORMULA + FUNCTION = ANALYSIS

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Statistical issues

Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

Real data

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

DATA LIST $\downarrow \\ {\rm DATA\ FRAME\ +\ FORMULA\ +\ FUNCTION\ =\ ANALYSIS}$

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Statistical issue

Data management

heory

Multiple → single Bipartite graphs

Methods

Data lists Subscripting

oubscripting m

Coercion Real data

. . .

Introduction

Traits and the ecology of communities Statistical issues The data management-analysis interfac

Theory

Converting multiple tables to one single table

Analyzing data structure with bipartite graphs

Computational methods

Multiple tables in one R object: the data list Subscripting multiple tables simultaneously Using the simplest functions (e.g. 1m) Coercing data lists to data frames Real complex zooplankton community data

Conclusion

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Traits and ecology Statistical issues Data management

Theor

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting Im

Coercion Real data

How can we convert this to a data frame?

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and ecolog Statistical issues

Data managemer

Theory

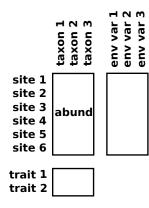
 $\begin{array}{c} \mathsf{Multiple} \longrightarrow \mathsf{single} \\ \\ \mathsf{Bipartite} \ \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

lm Coercion

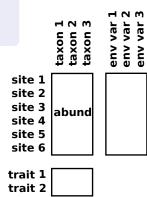
Real data



How can we convert this to a data frame?

Lost information

Redundant information



The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Traits and ecology Statistical issues

Data management

Theory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists

lm

Coercion Real data

Steve C. Walker Guillaume Guénard Pierre Legendre

roduction

Traits and ecology Statistical issues Data management

Theory

Multiple → single
Bipartite graphs

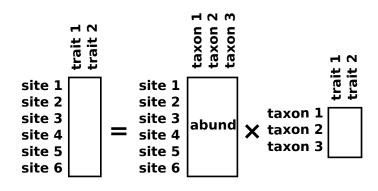
Methods

Data lists Subscripting

ercion

Real data

Conclusion



e.g. Leibold et al. (2010)

Lost information

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and ecolog Statistical issues

Data management

Theory

Multiple → single
Bipartite graphs

Methods

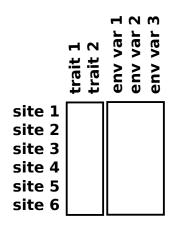
Data lists

n

Real data

Real data

Conclusion



e.g. Leibold et al. (2010)

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Statistical issues

Data management

Theory

Multiple → single Bipartite graphs

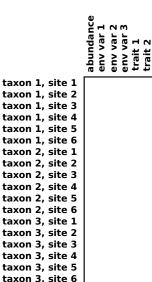
Methods

Data lists

n

Real data

Conclusion



4 □ ト 4 同 ト 4 豆 ト 4 豆 ・ り Q (へ)

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Statistical issues

Data management

Γheory

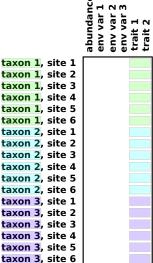
Multiple → single
Bipartite graphs

Methods

Data lists

n

Real data



The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Traits and ecology
Statistical issues
Data management

Theory

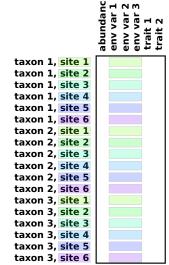
Multiple → single

Methods

Data lists

n

Real data



The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Statistical issues

Data management

Theory

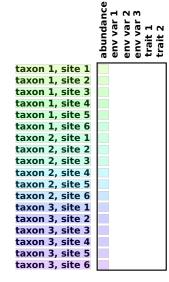
Multiple → single
Bipartite graphs

Methods

Data lists

n oercion

Real data



The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Statistical issues

Data management

Theory

Multiple → single
Bipartite graphs

Methods

Data lists Subscripting

ercion

Real data

Conclusion

Rules

▶ Dimensions that <u>can not grow</u> with more sampling represent groups of variables

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Statistical issues

Data management

Theory

Multiple → single
Bipartite graphs

Methods

Data lists Subscripting

percion

Real data

Conclusion

Rules

- Dimensions that <u>can not grow</u> with more sampling represent groups of variables
- ▶ Dimensions that <u>can grow</u> with more sampling represent replication

between data management and

analysis Steve C. Walker

The interface

Guillaume Guénard Pierre Legendre

ntroduction

Statistical issues

Data management

i neory

 $\begin{array}{c} \mathsf{Multiple} \longrightarrow \mathsf{single} \\ \\ \mathsf{Bipartite} \ \mathsf{graphs} \end{array}$

Viethods

Data lists Subscripting

> m Coercion

Real data

Conclusion

Variables

- abundance
- environmental variables
- traits



trait 1 trait 2

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Statistical issues

Theory

Multiple → single

Dipartite gr

Data lists

Subscripting

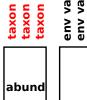
oercion

Real data

Conclusion

Replicates

- sites
- taxa



site 2 site 3 site 4 site 5 site 6

site 1

trait 1 trait 2

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Traits and ecology Statistical issues Data management

Theory

 $Multiple \rightarrow single$ Bipartite graphs

Methods

Data lists

n oercion

Real data

	abunda	env var	env var	env var	trait 1	:
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						

Introduction

Traits and the ecology of communities Statistical issues The data management-analysis interfac

Theory

Converting multiple tables to one single table Analyzing data structure with bipartite graphs

Computational methods

Multiple tables in one R object: the data list Subscripting multiple tables simultaneously Using the simplest functions (e.g. lm) Coercing data lists to data frames Real complex zooplankton community data

Conclusion

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Statistical issues

Data management

I heory

Multiple \rightarrow single

Bipartite graphs

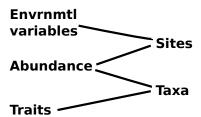
Methods

Data lists Subscripting Im

Coercion Real data

Steve C. Walker Guillaume Guénard Pierre Legendre

Variable Dimensions groups of replication



troduction

Statistical issues

Data management

heory

Multiple → single
Bipartite graphs

Methods

Data lists

m

Coercion Real data

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and ecology
Statistical issues
Data management

heory

Multiple → single
Bipartite graphs

Nethods

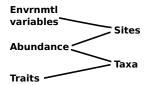
Data lists Subscripting

m

Coercion Real data

Conclusion

Variable Dimensions groups of replication



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

Biadjacency matrices

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Identifying data sets that are not multiple-table

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

ntroduction

Traits and ecology

Data management

heory

Multiple → single
Bipartite graphs

Methods

Data lists Subscripting

> n oercion

Real data

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

troduction

Statistical issues
Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting m

Coercion Real data

Keal data

Theory

Multiple → single

Multiple → single
Bipartite graphs

Data lists Subscripting

m

Coercion Real data

Conclusion

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

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

m oercion

Real data

Conclusion

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.

	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)

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Statistical issues

Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

oercion

Real data

Гһеогу

Multiple → single
Bipartite graphs

Methods

Data lists Subscripting

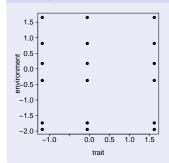
Im

Coercion Real data

Conclusion

Necessarily un-correlated variables

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



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

Traits and ecology Statistical issues Data management

I neory

Multiple → single

Bipartite graphs

Methods

Data lists Subscripting

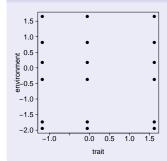
Coercion

Real data

Conclusion

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),



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

Traits and ecology Statistical issues Data management

Multiple → single
Bipartite graphs

Methods

Data lists Subscripting

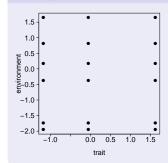
Coercion Real data

Redi data

Conclusion

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.



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

Biadjacency matrices

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

The meaning of zeros

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

Multiple → single Bipartite graphs

Real data

troduction

Statistical issues

Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Viethods

Data lists Subscripting

m

Real data

Conclusion

The meaning of zeros

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

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

> library(multitable)

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







The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Statistical issues

Data management

Theory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

n

oal data

Real data

Introduction

Traits and the ecology of communities Statistical issues The data management-analysis interfac

Theory

Converting multiple tables to one single table Analyzing data structure with bipartite graphs

Computational methods

Multiple tables in one R object: the data list

Subscripting multiple tables simultaneously Using the simplest functions (e.g. 1m)
Coercing data lists to data frames
Real complex zooplankton community data

Conclusion

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Statistical issues

Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists

Im

Real data

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Statistical issues

Data management

Theory

Multiple → single Bipartite graphs

Methods

Data lists

Subscripting

oercion

Real data

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

```
> d1
```

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...

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Statistical issues

Data management

Theory

Multiple → single Bipartite graphs

1ethods

Data lists

Subscripting

n

Real data

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Statistical issues

Multiple → single

Bipartite graphs

Methods

Data lists

Subscripting

percion

Real data

Conclusion

bodysize:

sppA sppB sppC -0.45 -0.07 1.48

Replicated along: || species ||

REPLICATION DIMENSIONS:

sites species

5

Introduction

Traits and the ecology of communities Statistical issues The data management-analysis interface

Theory

Converting multiple tables to one single table Analyzing data structure with bipartite graphs

Computational methods

Multiple tables in one R object: the data list Subscripting multiple tables simultaneously Using the simplest functions (e.g. 1m) Coercing data lists to data frames

Real complex zooplankton community data

Conclusion

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Statistical issues

Data management

heory

 $Multiple \rightarrow single$ Bipartite graphs

Methods

Data lists

Subscripting

Subscripting

Coercion Real data

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Statistical issues

Data managemen

Theor

Multiple → single
Bipartite graphs

Methods

Data lists Subscripting

Subscripting

oercion

Real data

Conclusion

◆□▶ ◆□▶ ◆■▶ ◆■▶ ■ 900

> dl[1:3,]

```
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...

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Multiple → single

Subscripting

Real data

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Statistical issues

Data manageme

Ineory

Multiple → single Bipartite graphs

Methods

Subscripting

oubscripting

n

Real data

Conclusion

bodysize:

sppA sppB sppC -0.45 -0.07 1.48

Replicated along: || species ||

REPLICATION DIMENSIONS:

sites species

3 :

Introduction

Traits and the ecology of communities Statistical issues The data management-analysis interface

Theory

Converting multiple tables to one single table Analyzing data structure with bipartite graphs

Computational methods

Multiple tables in one R object: the data list Subscripting multiple tables simultaneously

Using the simplest functions (e.g. lm)

Coercing data lists to data frames Real complex zooplankton community data

Conclusion

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Statistical issues

Data management

Multiple → single

Multiple → single Bipartite graphs

Methods

Data lists
Subscripting

Coercion

Real data

Steve C. Walker Guillaume Guénard Pierre Legendre

Multiple → single

Im

Real data

> lm(abundance ~ temperature * bodysize, dl)

Call:

lm(formula = abundance ~ temperature * bodysize, data = d1)

Coefficients:

(Intercept) bodysize temperature 0.08795 0.20848 -0.40439

temperature:bodysize

0.09822

Introduction

Traits and the ecology of communities Statistical issues The data management-analysis interface

Theory

Converting multiple tables to one single table Analyzing data structure with bipartite graphs

Computational methods

Multiple tables in one R object: the data list Subscripting multiple tables simultaneously Using the simplest functions (e.g. 1m)

Coercing data lists to data frames

Real complex zooplankton community data

Conclusion

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Traits and ecology Statistical issues Data management

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

Coercion

Real data

> as.data.frame(d1)

	${\tt abundance}$	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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

roduction

Traits and ecology Statistical issues

heory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

1ethods

ata lists ubscripting

lm Coercion

al data

Introduction

Traits and the ecology of communities Statistical issues The data management-analysis interface

Theory

Converting multiple tables to one single table Analyzing data structure with bipartite graphs

Computational methods

Multiple tables in one R object: the data list Subscripting multiple tables simultaneously Using the simplest functions (e.g. 1m) Coercing data lists to data frames

Real complex zooplankton community data

Conclusion

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction

Traits and ecology Statistical issues Data management

heory

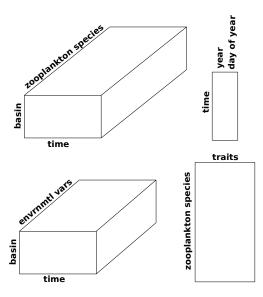
Multiple → single Bipartite graphs

Methods

Data lists Subscripting

Coercion

Real data



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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

troduction

Traits and ecology Statistical issues Data management

Theory

Multiple → single Bipartite graphs

Methods

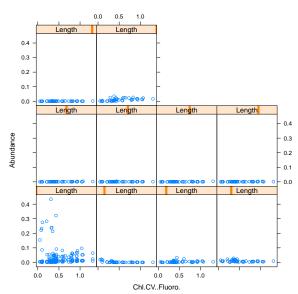
Data lists Subscripting

n abscripting

Dercion

Real data

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



The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

ntroduction

Statistical issues

Data management

Theory

Multiple → single

Methods

Data lists Subscripting

n ----:--

Real data

ricui dutu

Traits and the ecology of communities

Multiple tables in one R object: the data list

Conclusion

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Multiple → single

Real data

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Take-home message

Don't be scared of multiple table data sets.

Multiple → single

Real data

Take-home message

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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

Introduction Traits and ecology

Statistical issues

Data management

Theory

 $\begin{array}{c} \mathsf{Multiple} \, \longrightarrow \, \mathsf{single} \\ \mathsf{Bipartite} \, \, \mathsf{graphs} \end{array}$

Methods

Data lists Subscripting

oercion

Real data

Take-home message

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

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

itroduction

Statistical issues

Data managemen

Theory

 $\begin{array}{c} \mathsf{Multiple} \longrightarrow \mathsf{single} \\ \mathsf{Bipartite} \ \mathsf{graphs} \end{array}$

Methods

Data lists Subscriptin

oercion

Real data

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.

The interface between data management and analysis

Steve C. Walker Guillaume Guénard Pierre Legendre

itroduction

Statistical issues

Data management

Theory

Multiple → single Bipartite graphs

Methods

Data lists Subscripting

> n percion

Real data

 Natural Sciences and Engineering Research Council of Canada

- Laura Timms (McGill University)
- Beatrix Beisner (Université du Québec à Montréal)
- Ben Bolker (McMaster University)

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



project