

Temporal eigenfunction methods for multiscale analysis of community composition and other multivariate data

Pierre Legendre

Département de sciences biologiques

Université de Montréal

Pierre.Legendre@umontreal.ca

<http://www.bio.umontreal.ca/legendre/>

Time-series in marine science and applications for industry

A Europole Mer Gordon-like conference at the crossroad of marine sciences

Centre nautique de Moulin Mer, Brittany, September 17-21, 2012

Outline of the talk

1. Motivation
2. Variation partitioning
3. Multi-scale analysis
 - The dbMEM method
 - Simulation study (summary)
4. Applications to real ecological data
5. Recent developments: MEM, AEM
6. R packages and functions

Setting the stage

Ecologists want to understand and model temporal community structures through the analysis of **species assemblages**.

- Species assemblages are the best response variable available to estimate the impact of [anthropogenic] changes in ecosystems.
- Difficulty: species assemblages form **multivariate data tables** (sites x species).

Setting the stage

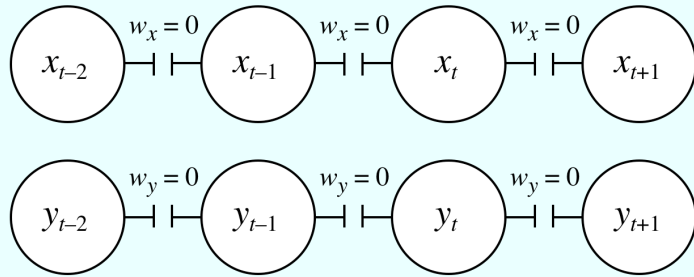
Ecologists want to understand and model temporal community structures through the analysis of **species assemblages**.

- Species assemblages are the best response variable available to estimate the impact of [anthropogenic] changes in ecosystems.
- Difficulty: species assemblages form **multivariate data tables** (sites x species).

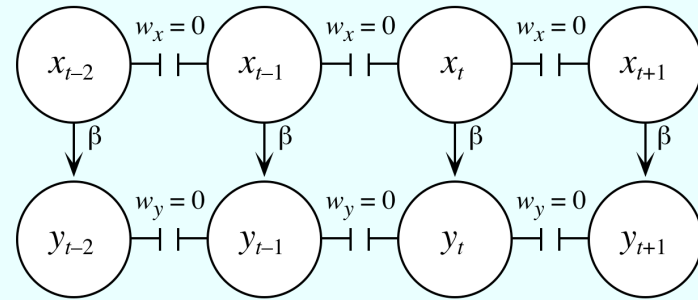
Temporal structures in communities indicate that some process has been at work to create them. Two families of mechanisms can generate temporal structures in response data, e.g. communities.

Illustration =>

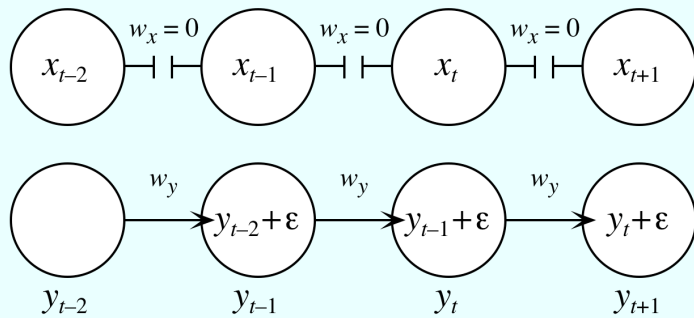
Case 1: Null situation



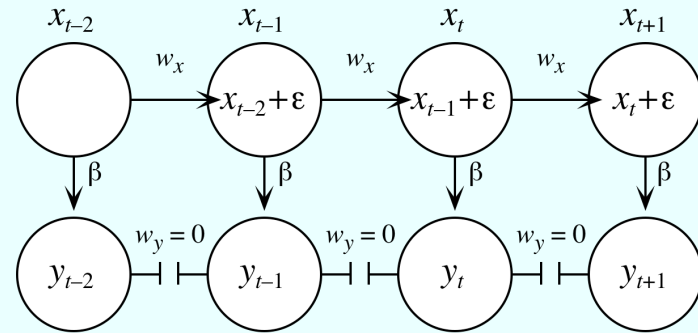
Case 2: y depends on x



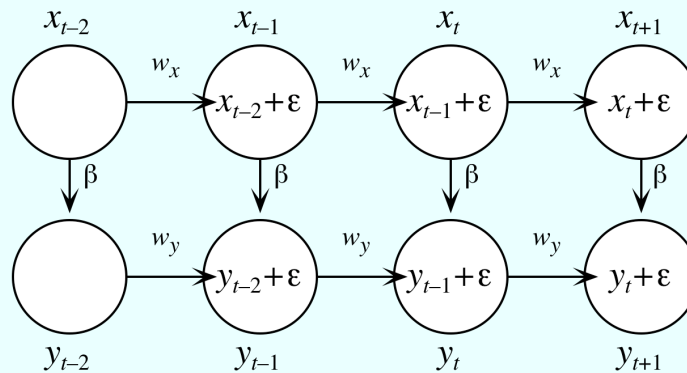
Case 3: Temporal autocorrelation (TA) in y



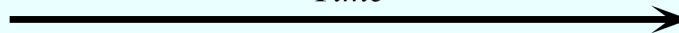
Case 4: Induced temporal dependence



Case 5: TA in x and y , y depends on x



Time



Temporal structures in communities indicate that some process has been at work to create them. Two families of mechanisms can generate temporal structures in communities:

- **Induced time dependence:** forcing (explanatory) variables are responsible for the temporal structures found in the species assemblage. They represent environmental or biotic control of the assemblages. Generally broad-scaled.
- **Community dynamics:** the temporal structures are generated by the species assemblage themselves, creating **autocorrelation**¹ in the response variables (species). Mechanisms: *neutral processes* such as reproductive dynamics, *interactions* among species, and random events. Temporal structures are generally fine-scaled.

¹ *Temporal autocorrelation* (SA) is technically defined as the dependence, due to temporal proximity, present in the residuals of a [regression-type] model of a response variable y that takes into account all deterministic effects due to forcing variables. Model: $y_i = f(\mathbf{X}_i) + SA_i + \varepsilon_i$.

Multivariate variation partitioning

Borcard & Legendre 1992 [1367 citations]

Borcard & Legendre 1994

and many published application papers



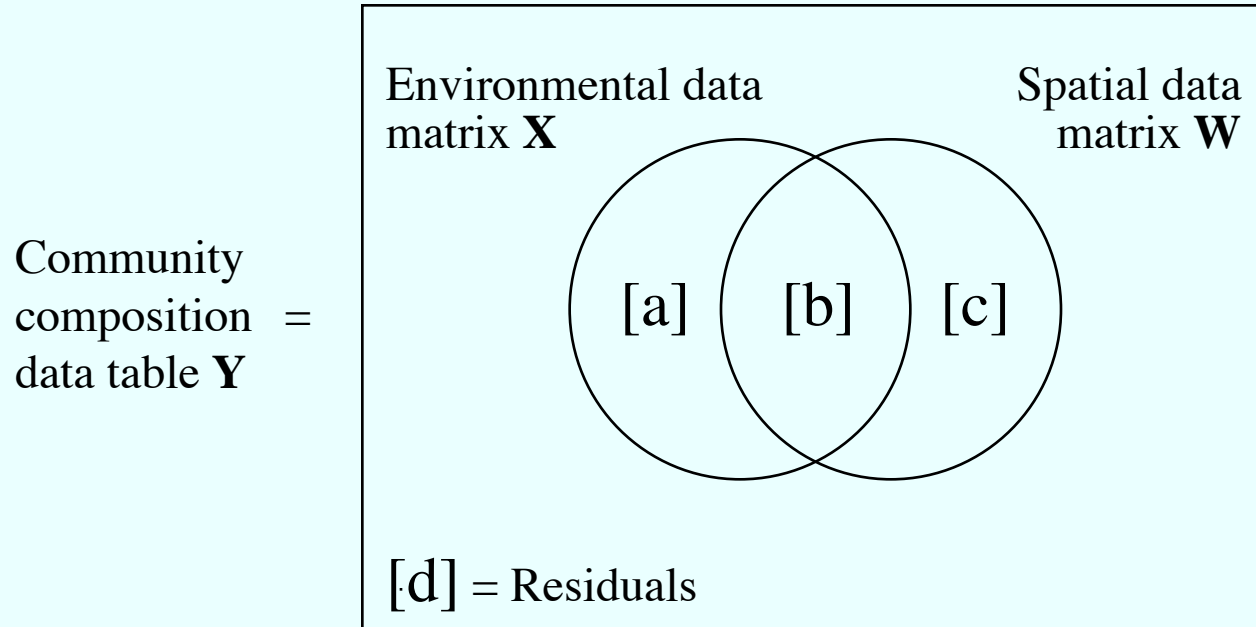


Figure – Venn diagram illustrating a partition of the variation of a response matrix \mathbf{Y} (e.g., community composition data) between environmental (matrix \mathbf{X}) and spatial (matrix \mathbf{W}) explanatory variables. The rectangle represents 100% of the variation of \mathbf{Y} .

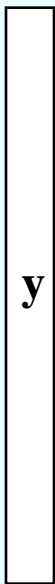
Method described in:

Borcard, Legendre & Drapeau 1992, Borcard & Legendre 1994; Legendre & Legendre 2012.

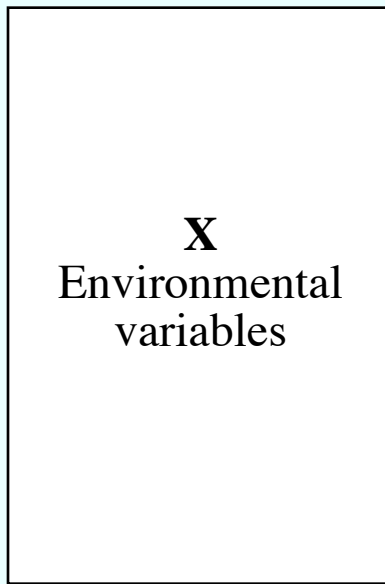
How to combine environmental and spatial variables in modeling community composition data?

- A single response variable: partial multiple regression.

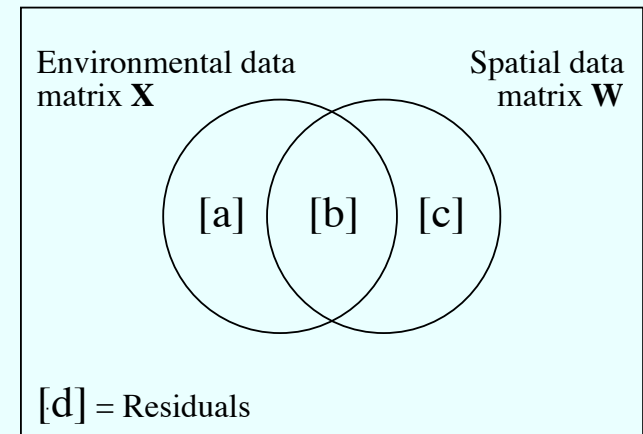
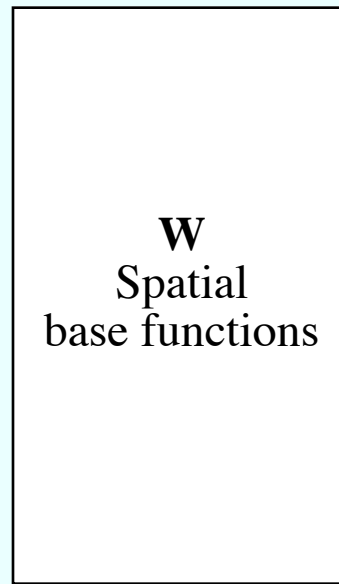
Response variable



Explanatory table



Explanatory table

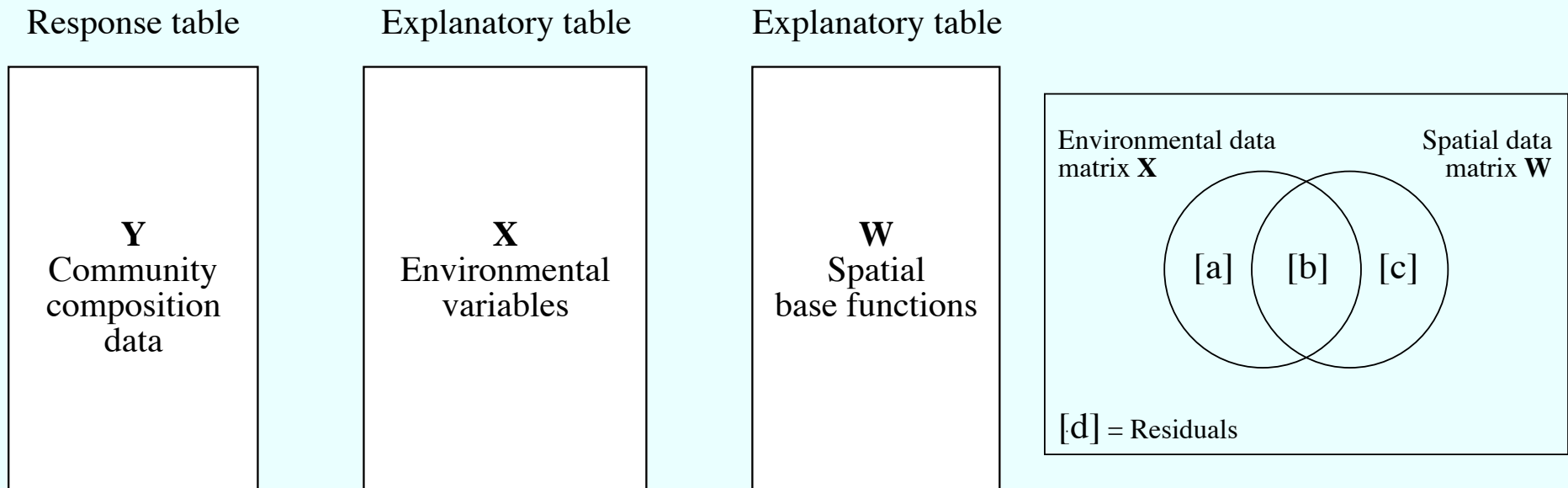


Partial multiple regression is computed as follows:

1. Compute the residuals \mathbf{X}_{res} of the regression of \mathbf{X} on \mathbf{W} : $\mathbf{X}_{\text{res}} = \mathbf{X} - [\mathbf{W} [\mathbf{W}'\mathbf{W}]^{-1} \mathbf{W}' \mathbf{X}]$
2. Regress \mathbf{y} on \mathbf{X}_{res}

How to combine environmental and spatial variables in modelling community composition data?

- Multivariate data: partial canonical analysis (RDA or CCA).



Partial canonical analysis is computed as follows:

1. Compute the residuals \mathbf{X}_{res} of the regression of \mathbf{X} on \mathbf{W} : $\mathbf{X}_{\text{res}} = \mathbf{X} - [\mathbf{W} [\mathbf{W}'\mathbf{W}]^{-1} \mathbf{W}' \mathbf{X}]$
2. Regress \mathbf{Y} on \mathbf{X}_{res} to obtain \mathbf{Y}_{fit} . Compute PCA of \mathbf{Y}_{fit} .

Geographic base functions

First (simple) representation: Polynomial function of geographic coordinates (polynomial trend-surface analysis).

Example 1: 20 sampling sites in the Thau lagoon, southern France.

	x	y	x2	xy	y2	x3	x2y	xy2	y3
Type:	Real	Real	Real	Real	Real	Real	Real	Real	Real
Source:	User...	User...	User Ent...	User Ent...	User Ent...	User Ent...	User Ent...	User Ent...	User Entered
1	3	10	9	30	100	27	90	300	1000
2	5	9	25	45	81	125	225	405	729
3	6	8	36	48	64	216	288	384	512
4	6	10	36	60	100	216	360	600	1000
5	8	9	64	72	81	512	576	648	729
6	9	10	81	90	100	729	810	900	1000
7	10	7	100	70	49	1000	700	490	343
8	11	6	121	66	36	1331	726	396	216
9	12	5	144	60	25	1728	720	300	125
10	12	7	144	84	49	1728	1008	588	343
11	12	9	144	108	81	1728	1296	972	729
12	13	8	169	104	64	2197	1352	832	512
13	15	2	225	30	4	3375	450	60	8
14	15	4	225	60	16	3375	900	240	64
15	15	5	225	75	25	3375	1125	375	125
16	16	1	256	16	1	4096	256	16	1
17	16	2	256	32	4	4096	512	64	8
18	16	4	256	64	16	4096	1024	256	64
19	17	6	289	102	36	4913	1734	612	216
20	18	4	324	72	16	5832	1296	288	64

$$\hat{z} = f(X, Y) = b_0 + b_1X + b_2Y + b_3X^2 + b_4XY + b_5Y^2 + b_6X^3 + b_7X^2Y + b_8XY^2 + b_9Y^3$$

Small textbook example:

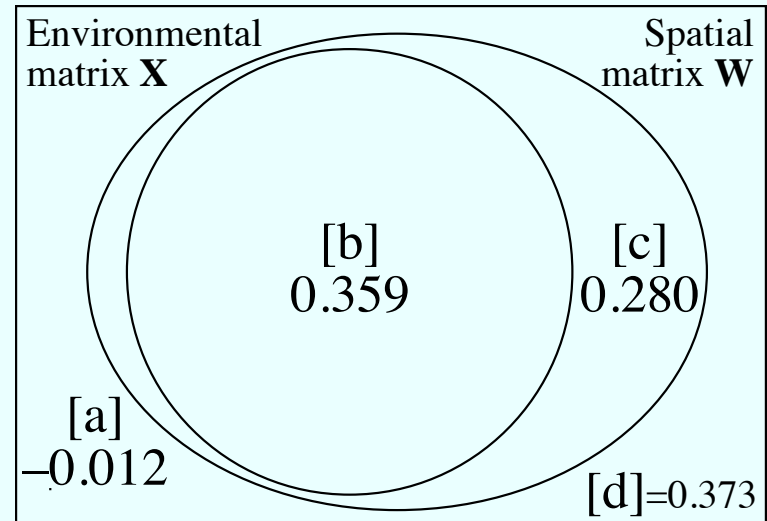
20 sampling sites in the Thau lagoon, southern France.

Response (**Y**): 2 types of aquatic heterotrophic bacteria (log-transf.)

Environmental (**X**): NH_4 , phaeopigments, bacterial production

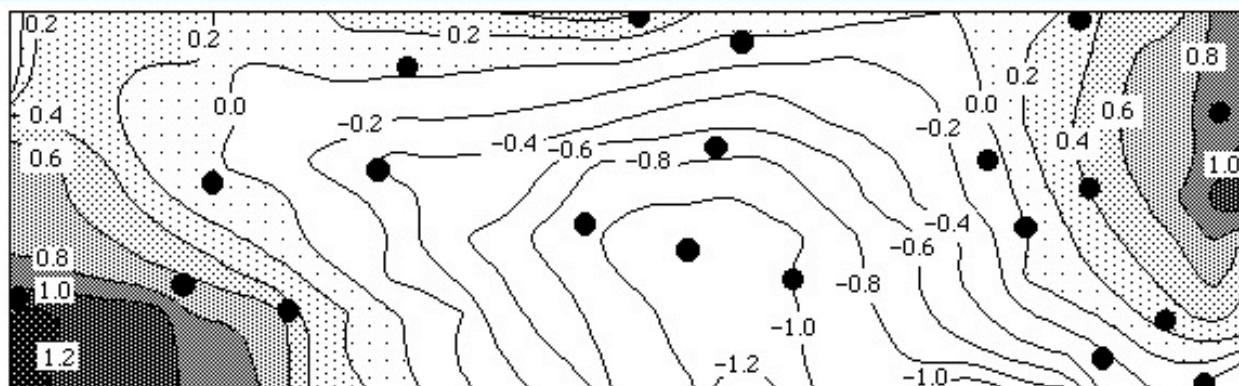
Spatial (**W**): selected geographic monomials X^2 , X^3 , X^2Y , XY^2 , Y^3

In real-life studies, partitioning is carried out on larger data sets.

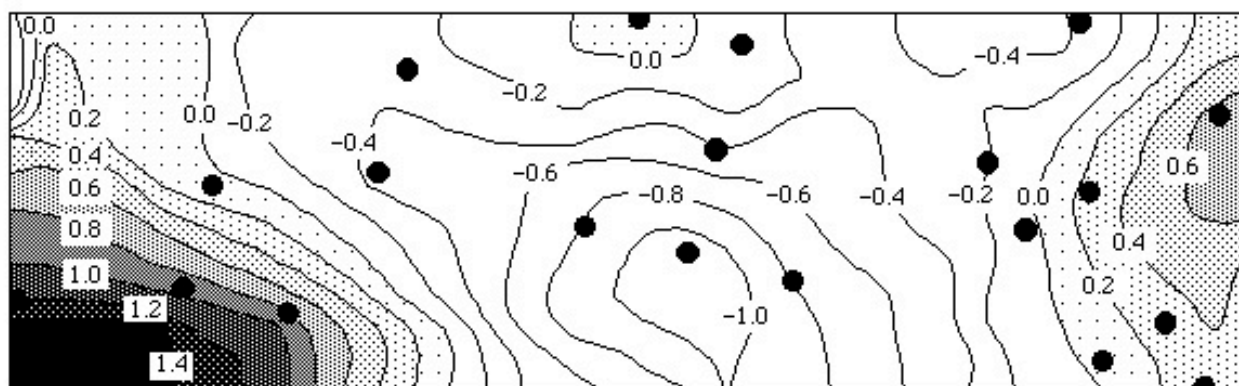


Fractions of variation	Proportion of variation of Y (R^2)	Probability (999 perm.)	Adjusted R^2
[a+b]	0.450	0.005*	0.347
[b+c]	0.734	0.001*	0.639
[a+b+c]	0.784	0.001*	0.627
[a]	0.051	0.549	-0.012
[b]	0.399	<i>Cannot be tested</i>	0.359
[c]	0.334	0.011*	0.280
Residuals = [d]	0.216		0.373
[a+b+c+d]	1.000		1.000

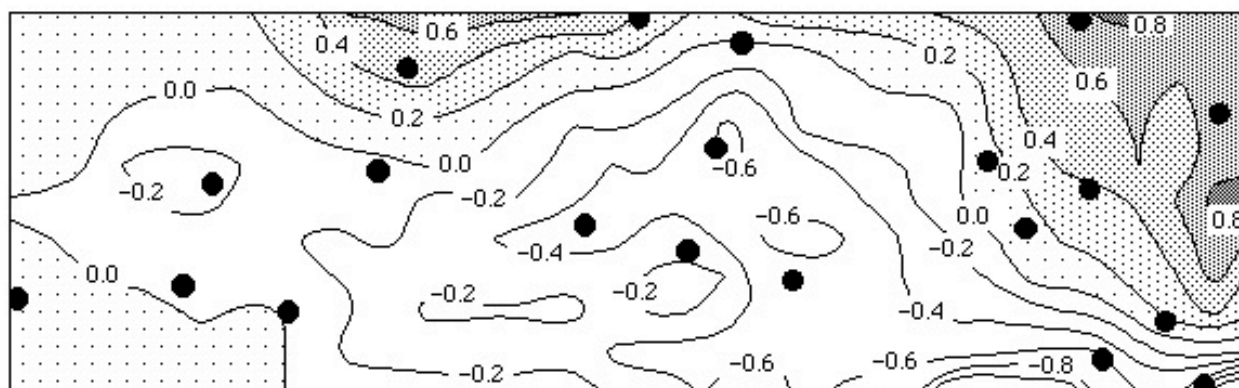
[a+b+c]



[a+b]



[c]



↑ Sea water ↑

Temporal eigenfunctions

Second representation:

Distance-based Moran's eigenvector maps (dbMEM)

alias *Principal Coordinates of Neighbor Matrices (PCNM)*

leading to multiscale analysis in variation partitioning

Borcard & Legendre 2002, 2004

Dray, Legendre & Peres-Neto 2006

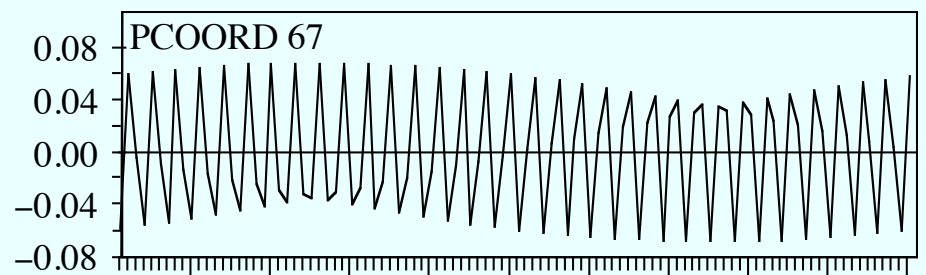
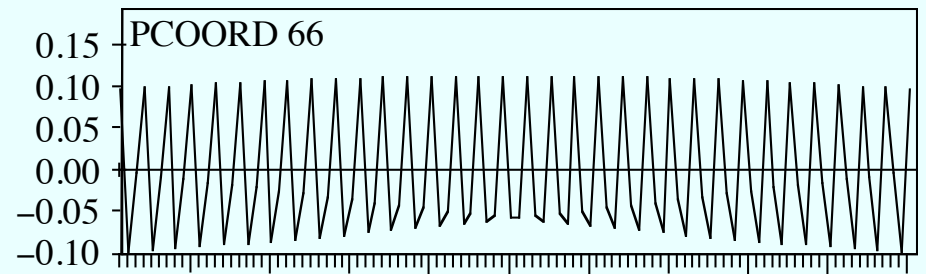
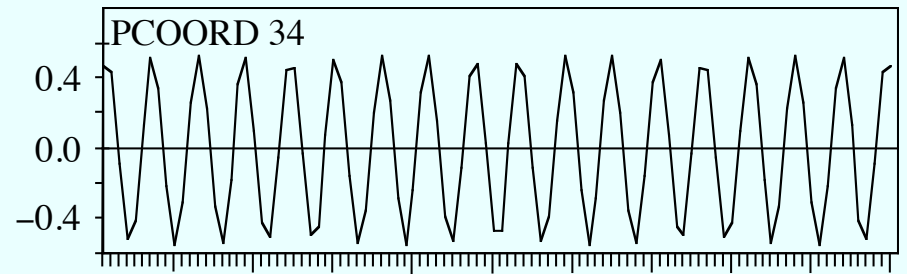
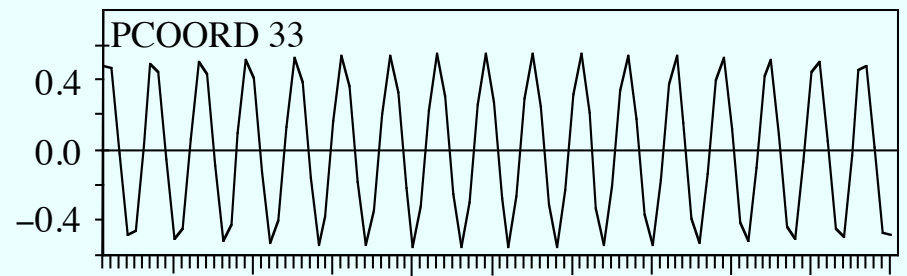
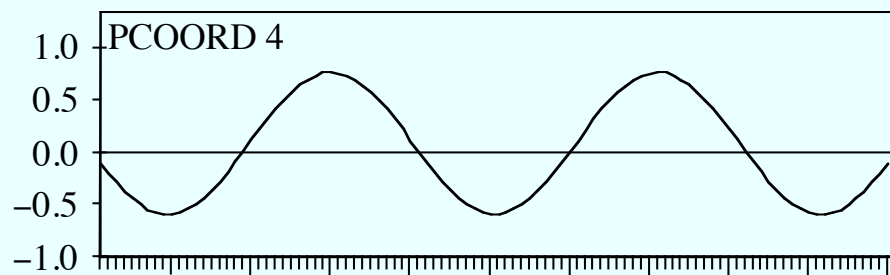
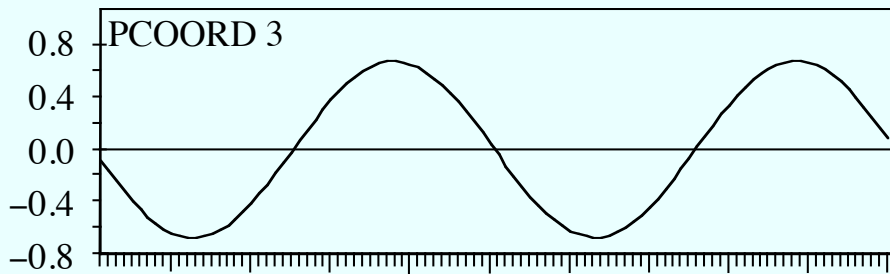
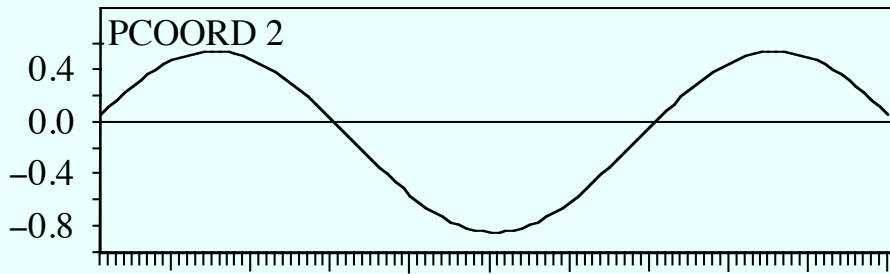
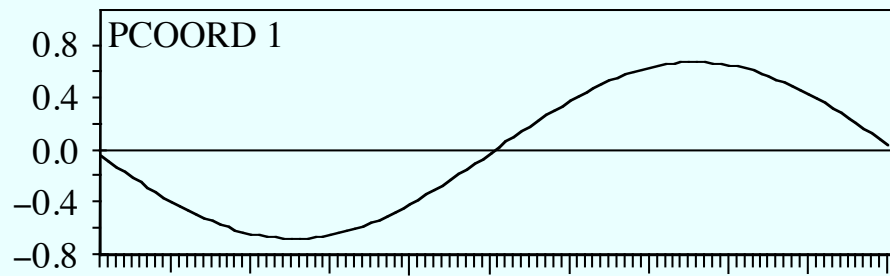


Figure – Eight of the 67 orthogonal PCNM eigenfunctions obtained for 100 equally-spaced points along a transect or time series. Truncation at first-neighbour distance.

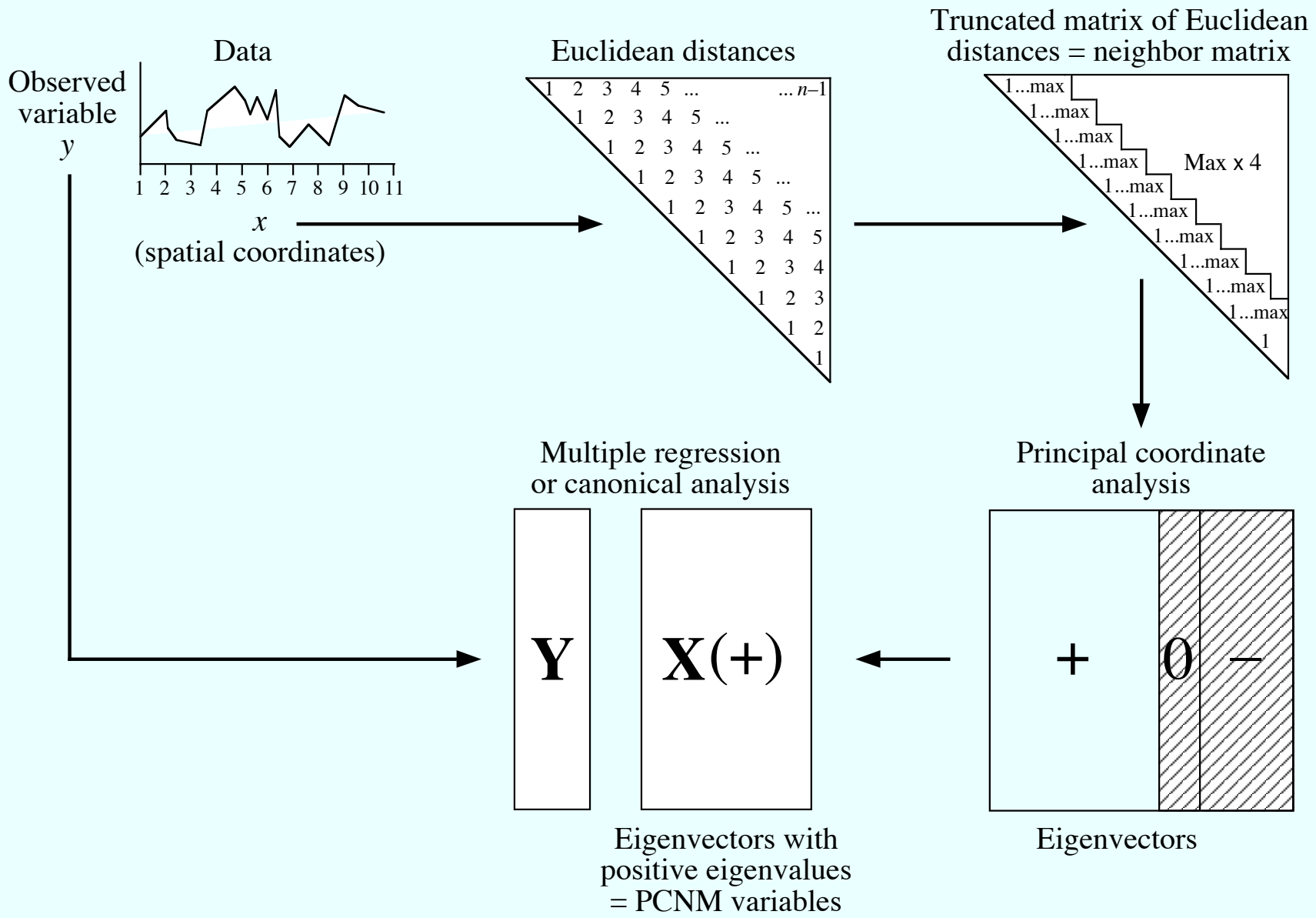
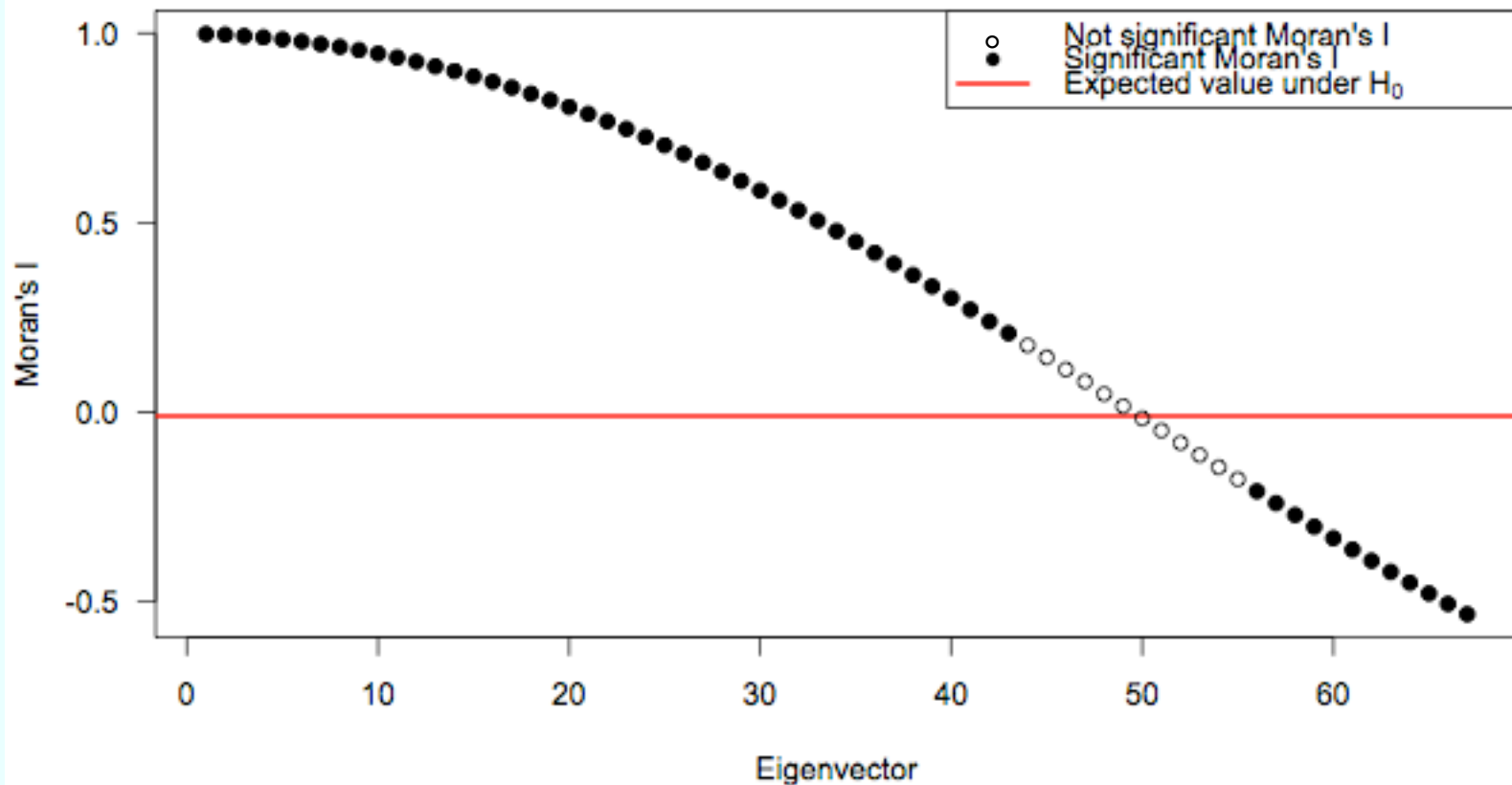


Figure – Schematic description of dbMEM analysis. The descriptors of spatial or temporal relationships (dbMEM eigenfunctions) are obtained by principal coordinate analysis of a truncated matrix of Euclidean distances among the sampling units.

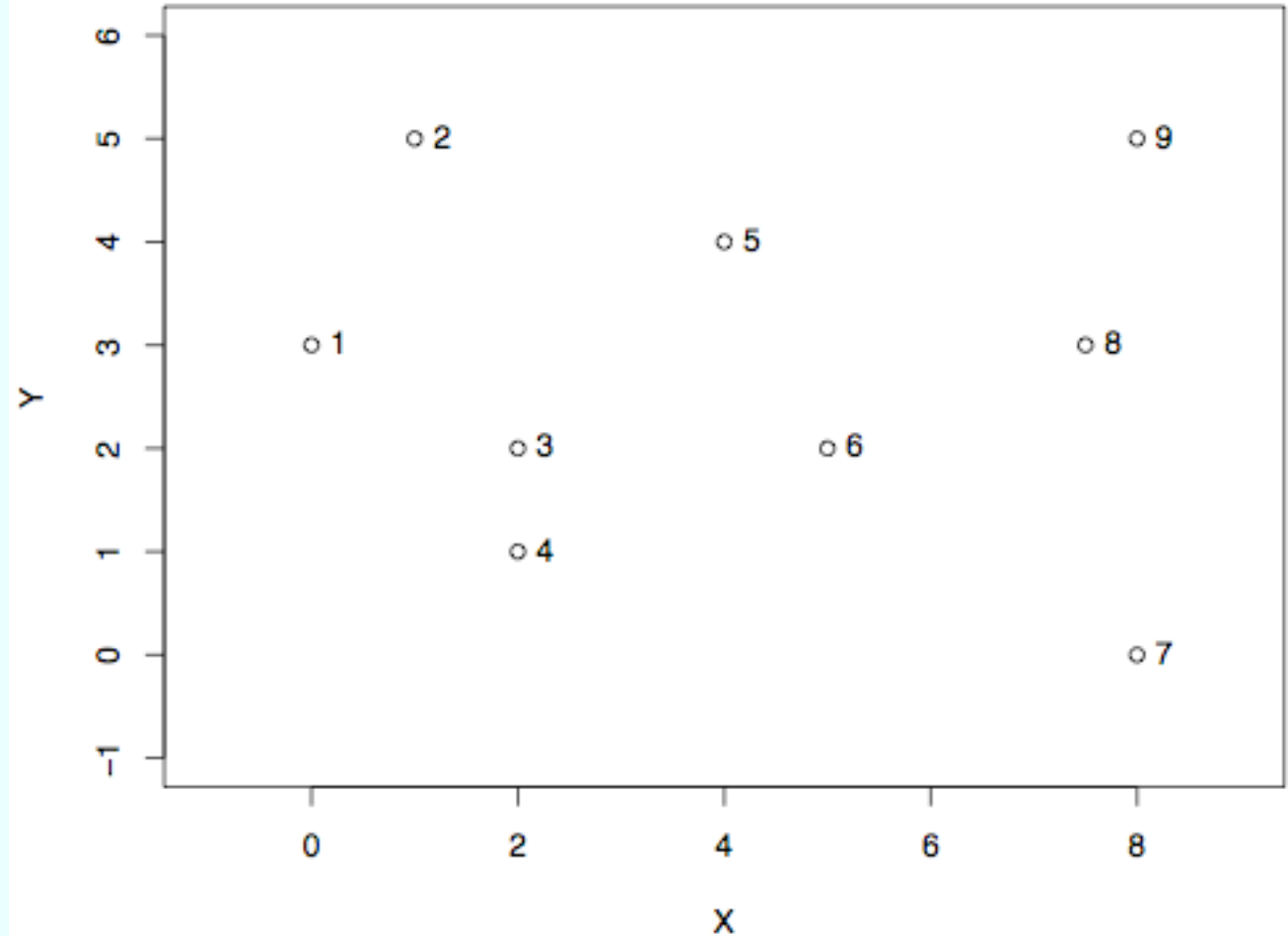
dbMEM eigenfunctions display autocorrelation

An autocorrelation coefficient (Moran's I) can be computed for each dbMEM. They indicate the sign of the autocorrelation (+ or -) modelled by the dbMEMs. Ex.100-point transect, 67 dbMEMs shown:

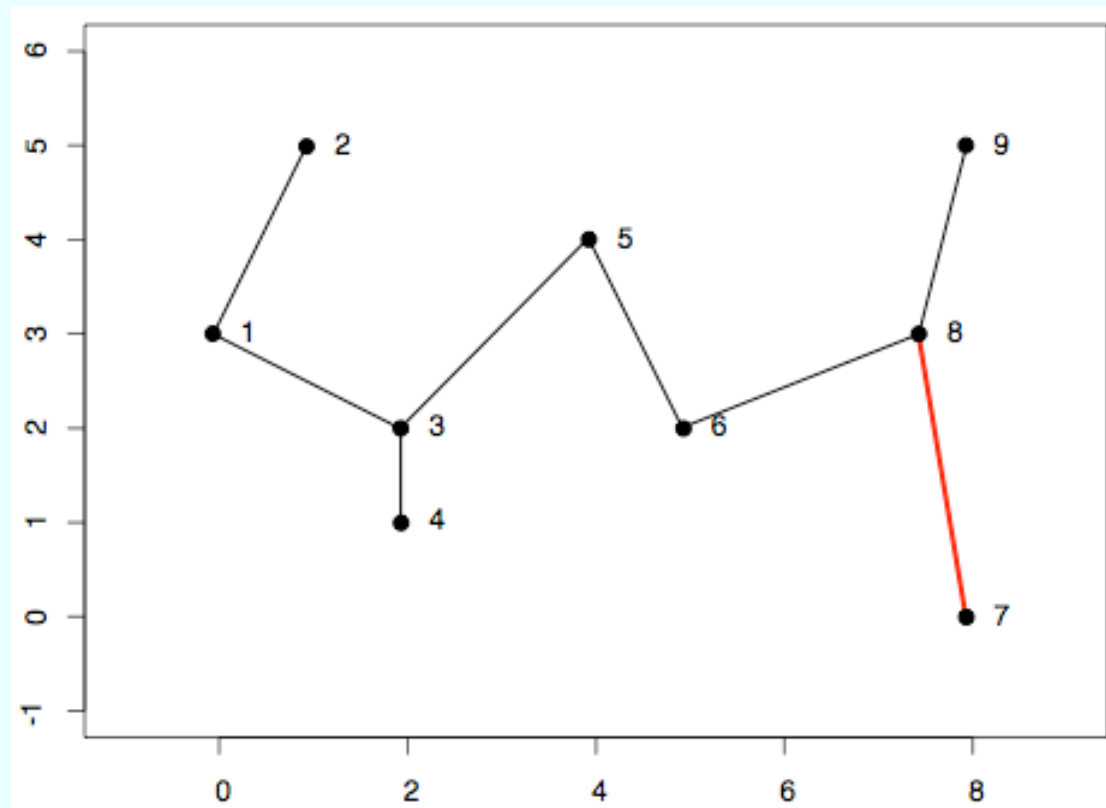


How to find the truncation distance in 2-dimensional spatial problems?

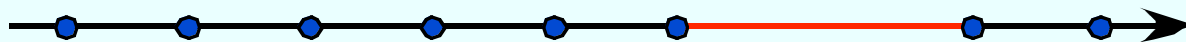
Point labels	Coordinates	
	X	Y
1	0	3
2	1	5
3	2	2
4	2	1
5	4	4
6	5	2
7	8	0
8	7.5	3
9	8	5



Spatial data: compute a minimum spanning tree



- Truncation distance \geq length of the longest edge. Longest edge here:
 $D(7, 8) = 3.0414$
- Irregular time series: truncation distance \geq length of longest edge.



Technical notes on dbMEM eigenfunctions

dbMEM variables represent a **spectral decomposition of the spatial/temporal relationships** among the study sites. They can be computed for regular or irregular sets of points in space or time.

dbMEM eigenfunctions are **orthogonal**. If the sampling design is regular, they look like sine waves; this is a property of the eigen-decomposition of the centred form of a distance matrix. If the design is irregular, the sine waves are distorted.

Simulation study

Type I error study

Simulations showed that the procedure is honest. It does not generate more significant results than it should for a given significance level α .

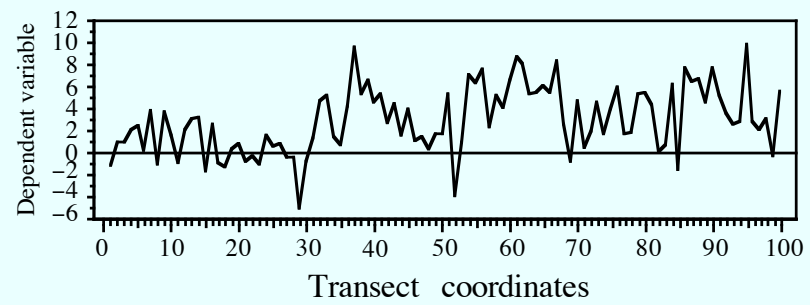
Power study

Simulations showed that dbMEM analysis is capable of detecting spatial structures of many kinds:

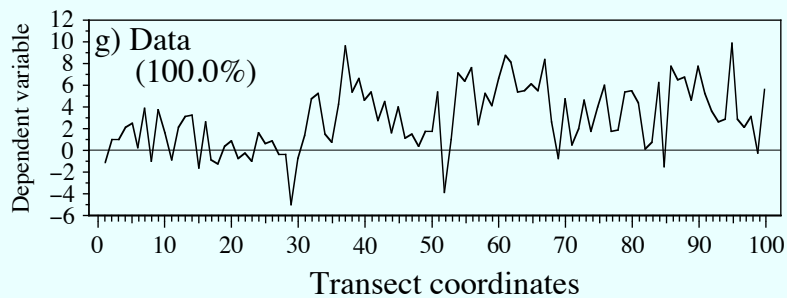
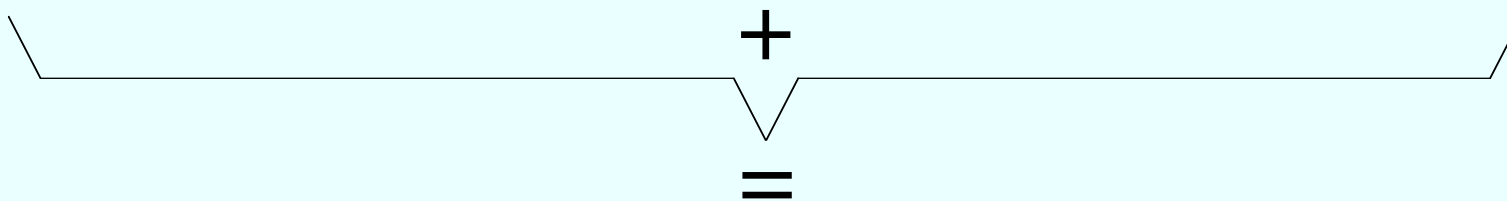
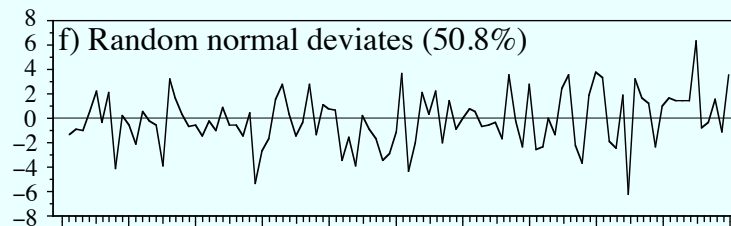
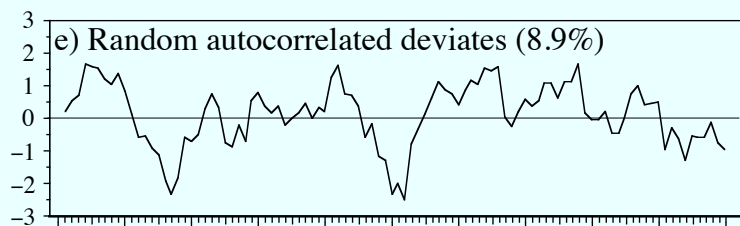
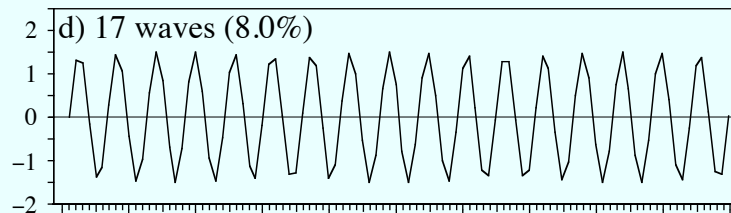
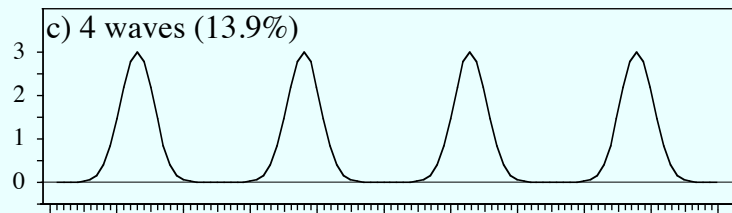
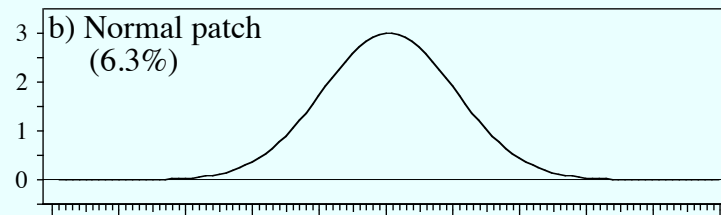
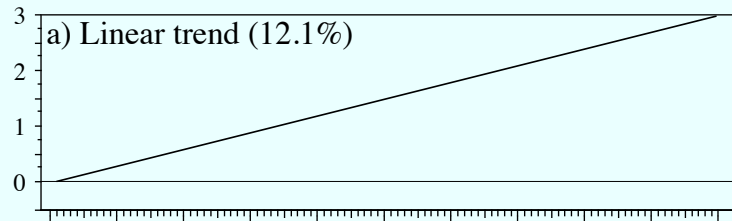
- random autocorrelated data,
 - bumps and sine waves of various sizes, without or with random noise, representing deterministic structures,
- as long as the structures are larger than the truncation value used to create the dbMEM eigenfunctions.

Detailed results are found in Borcard & Legendre 2002.

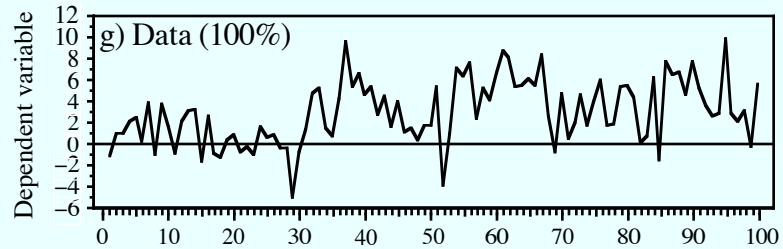
A difficult test case



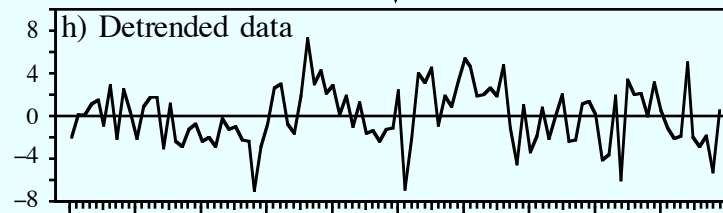
A difficult test case



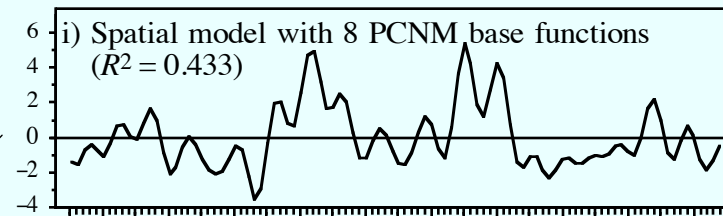
A difficult test case



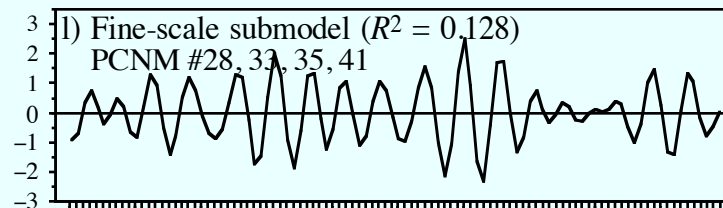
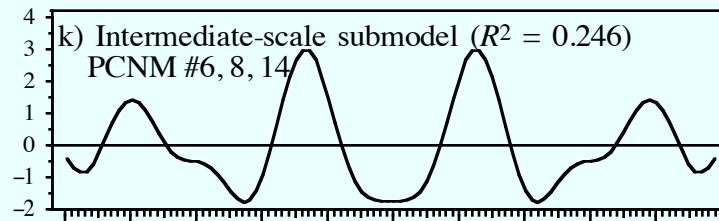
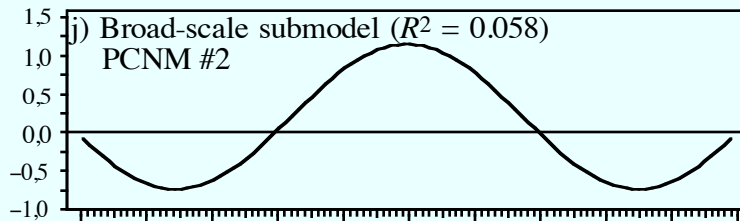
Simulated data



Step 1:
detrending



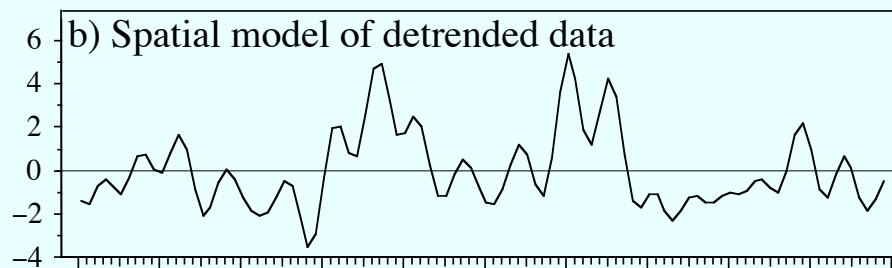
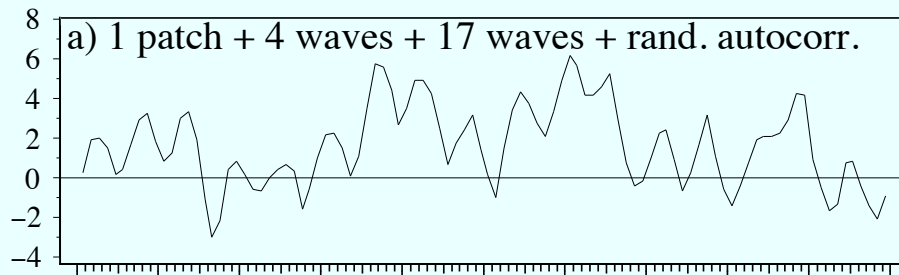
PCNM analysis
of detrended data



A difficult test case

Data

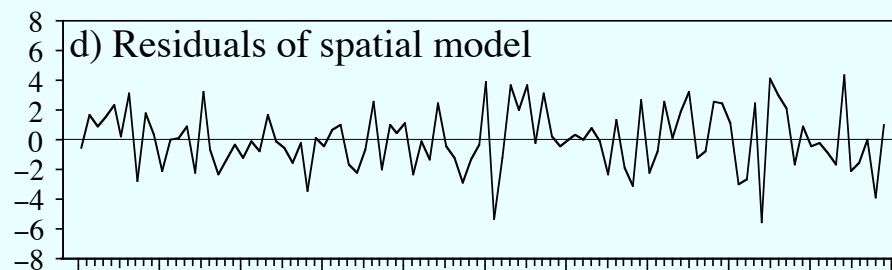
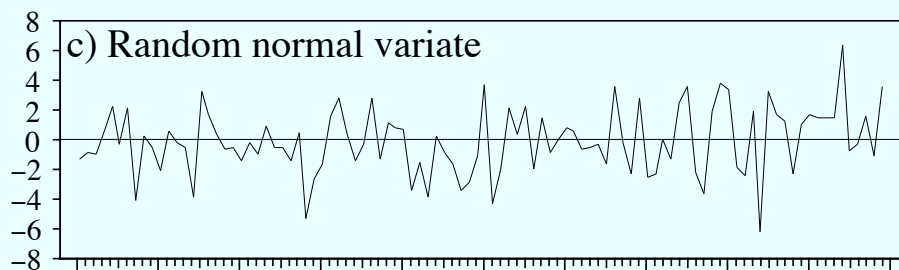
Model



+

$r = 0.775$

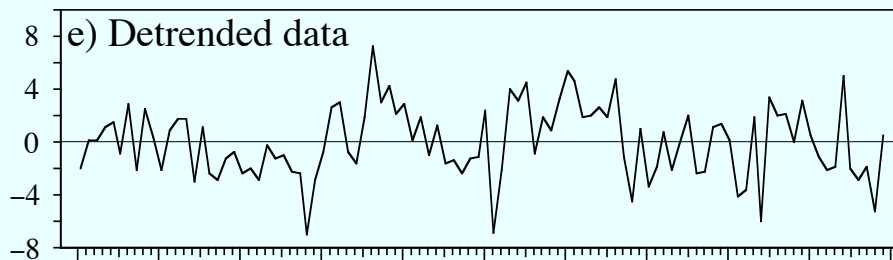
+



$r = 0.796$

=

=



Selection of explanatory variables?

Significance of the adjustment of a dbMEM model can be tested using the full set of dbMEM variables modelling positive spatial/temporal correlation, without selection of any kind.

The *adjusted R²* gives a correct estimate of the variation explained by the dbMEMs by correcting for the number of sampling units (n) and the number of explanatory variables in the model.¹

Before constructing submodels, **forward selection** of dbMEMs can be carried out by combining two criteria during model selection: the alpha significance level and the *adjusted R²* of the model containing all dbMEM eigenfunctions.²

¹ Peres-Neto, P. R., P. Legendre, S. Dray and D. Borcard. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* 87: 2614-2625.

² Blanchet F. G., P. Legendre and D. Borcard. 2008. Forward selection of explanatory variables. *Ecology* 89: 2623-2632.

Example 1

Regular one-dimensional transect in upper Amazonia¹

Data: abundance of the fern *Adiantum tomentosum* in quadrats.

Sampling design: 260 adjacent, square (5 m x 5 m) subplots forming a transect in the region of Nauta, Peru.

Questions

- At what spatial scales is the abundance of this species structured?
- Are these scales related to those of the environmental variables?

Pre-treatment

- The abundances were square-root transformed
- and detrended (significant linear trend: $R^2 = 0.102$, $p = 0.001$)

¹ Data from Tuomisto & Poulsen 2000, reanalysed in Borcard, Legendre, Avois-Jacquet & Tuomisto 2004.

Forward selection

50 PCNM eigenfunctions were selected out of 176 (permutation test, 999 permutations).

The PCNMs were arbitrarily divided into 4 submodels.

The submodels are orthogonal to one another.

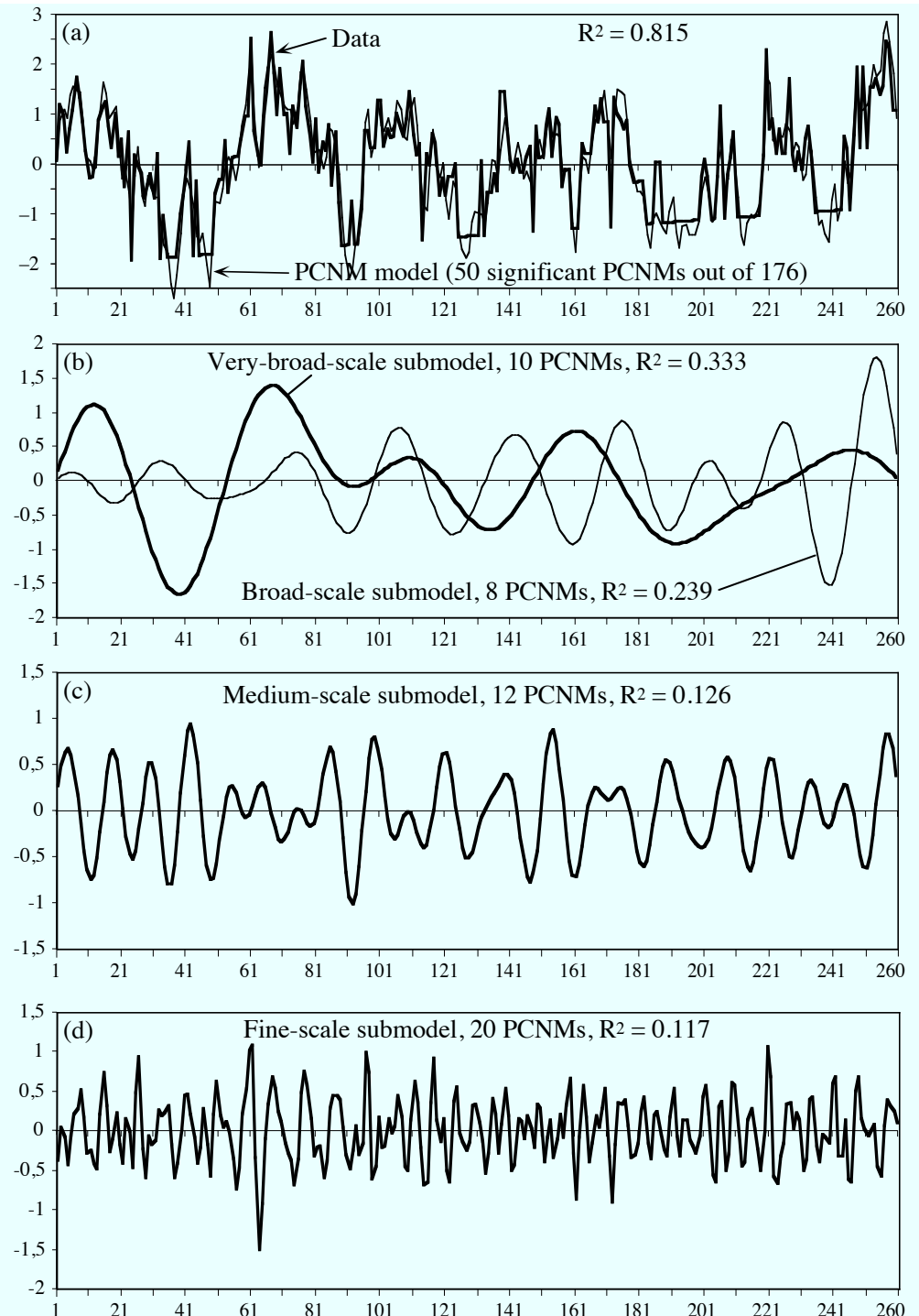
Significant wavelengths (periodogram analysis):

V-br-scale: 250, 355-440 m

Broad-scale: 180 m

Medium-scale: 90 m

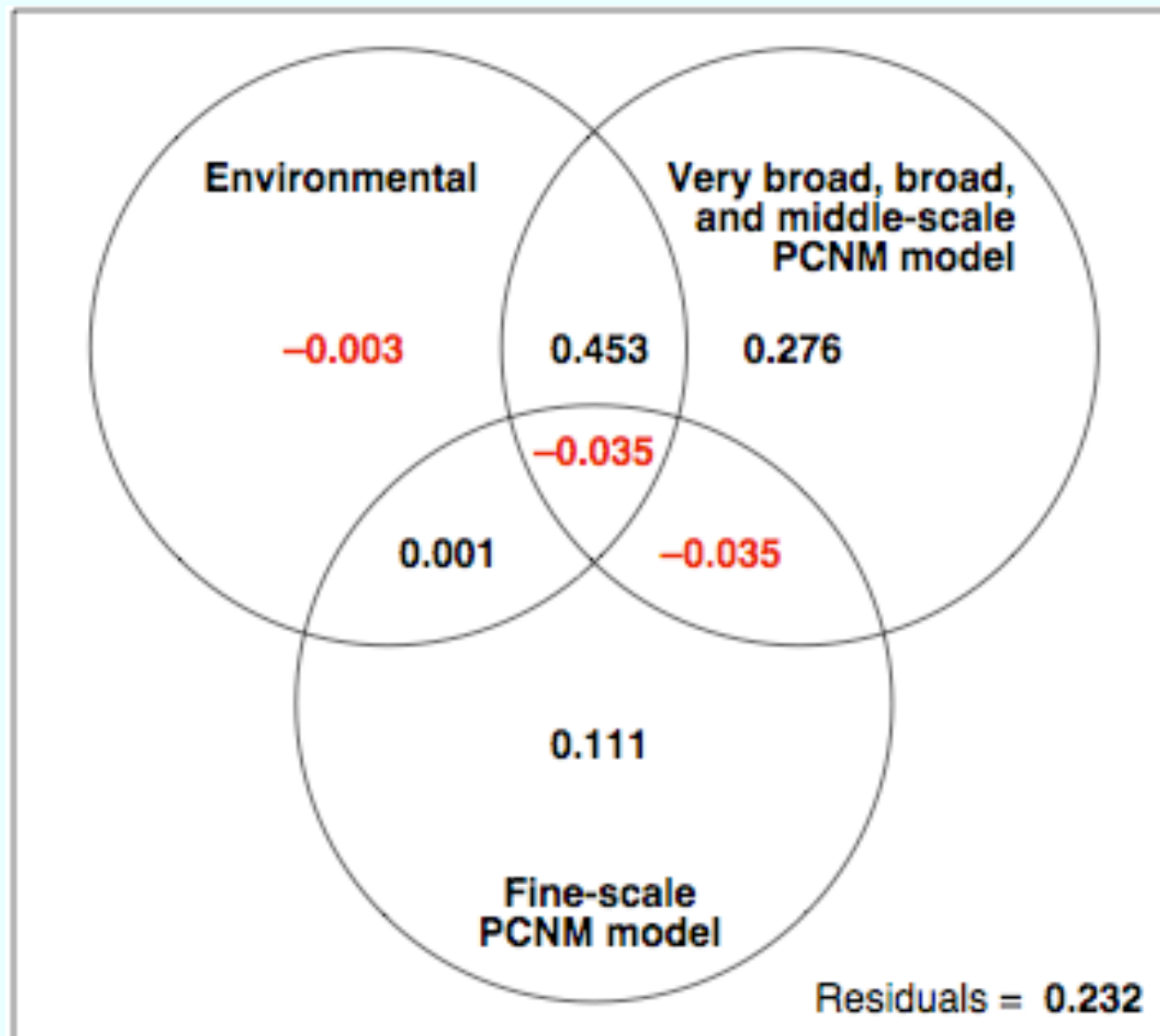
Fine-scale: 50, 65 m



Interpretation: regression on the environmental variables

	<i>Adiantum tomentosum</i>	Very broad	Broad	Medium	Fine
R^2 of PCNM submodel on <i>A. tomentosum</i>		0.333	0.239	0.126	0.117
R^2 of envir. on submodel		0.347	0.334	0.157	
R^2 of envir. on <i>A. tomentosum</i>	0.436	0.116	0.080	0.020	
Elevation (m)	< 0.0001	< 0.0001	< 0.0001		
Thickness of soil organic horizon (cm)	0.0004	< 0.0001			
Waterlogging	0.0393			0.0007	
Canopy height (m)	0.0011	< 0.0001			
Canopy coverage (%)				0.0024	
Shrub coverage (%)	0.0621	0.0341		0.0001	
Herb coverage (%)		0.0002			
Trees 3 - 7.5 cm DBH	0.0083		0.0009		
Lianas 3 - 7.5 cm diameter				0.0413	
Lianas 8 - 15 cm diameter				0.0183	

Use PCNM in variation partitioning: *Adiantum tomentosum* at Nauta, Peru (R^2_a).



Example 2

Temporal structure in a multivariate sediment core¹

- Data: the Round Loch of Glenhead (RLGH) fossil data.
- 139 Holocene diatom taxa in 101 levels of a sediment core from a small lake in Galloway, S-W Scotland.
- The data series covers the past 10 000 years.
- Level no. 1 is the top (most recent), no. 101 is the oldest.

¹ Legendre, P. and H. J. B. Birks. 2012. From classical to canonical ordination. Chapter 8 in: *Tracking Environmental Change using Lake Sediments, Volume 5: Data handling and numerical techniques*. Birks, Lotter, Juggins and Smol [eds.]. Springer, The Netherlands.

Analysis of RLGH data

Linear trend in the multivariate data: $R^2 = 0.190$

Temporal model of detrended data: 50 dbMEM variables; 30 were selected for modelling.

- Broad-scale submodel (dbMEM #1-10):

$R^2 = 0.530$, 3 significant axes

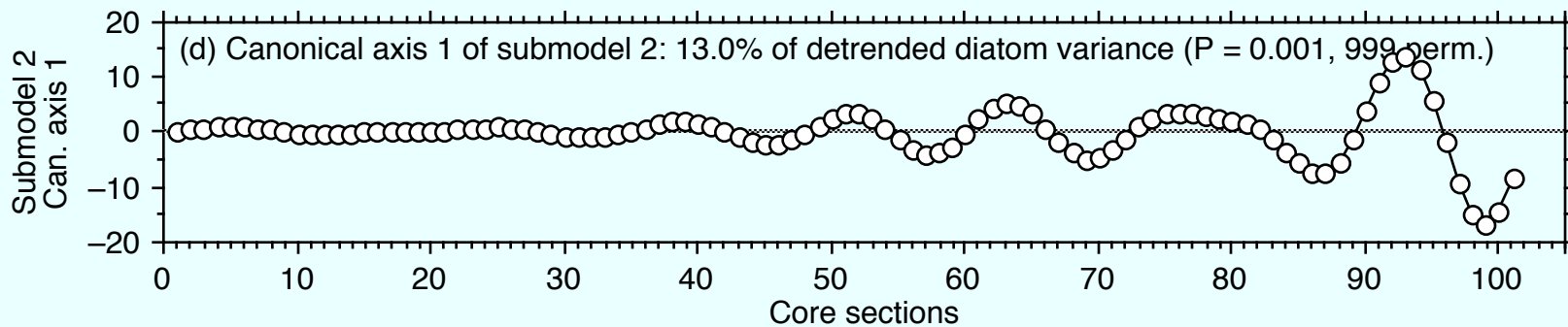
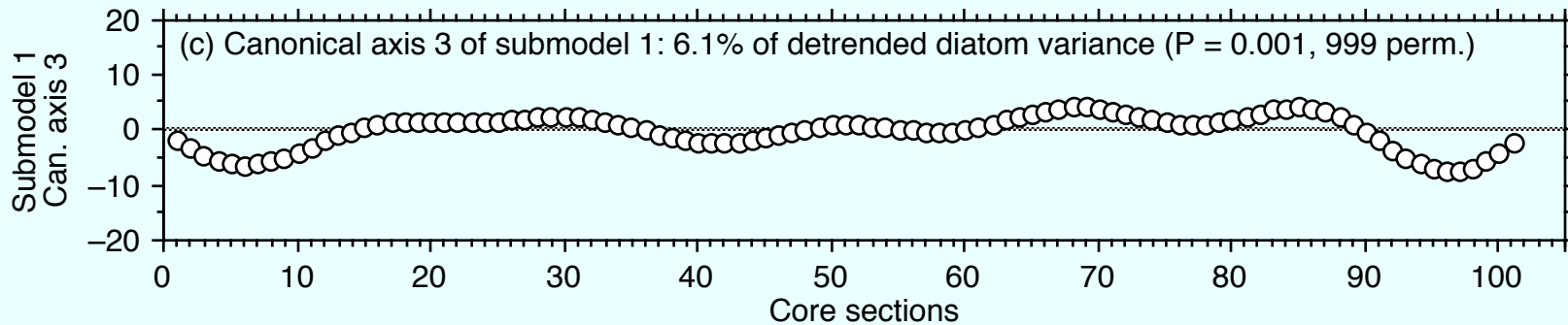
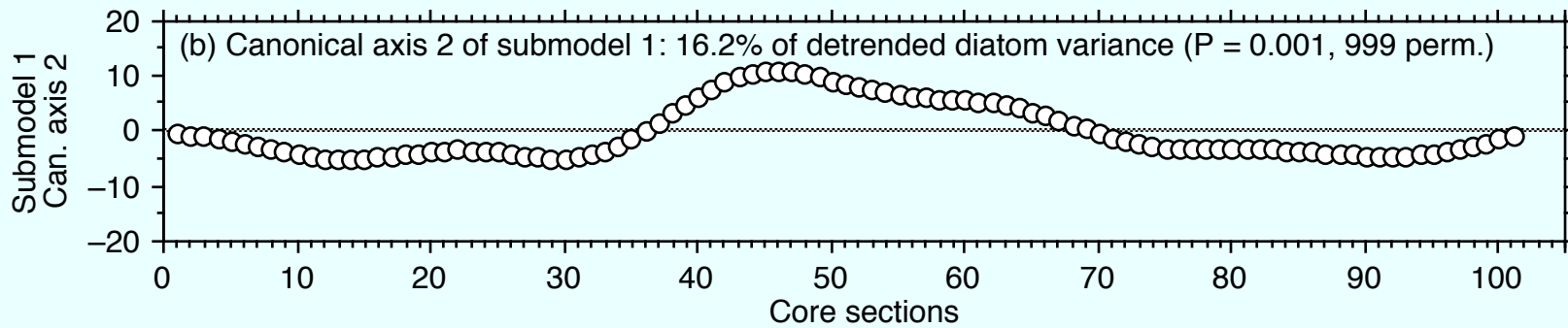
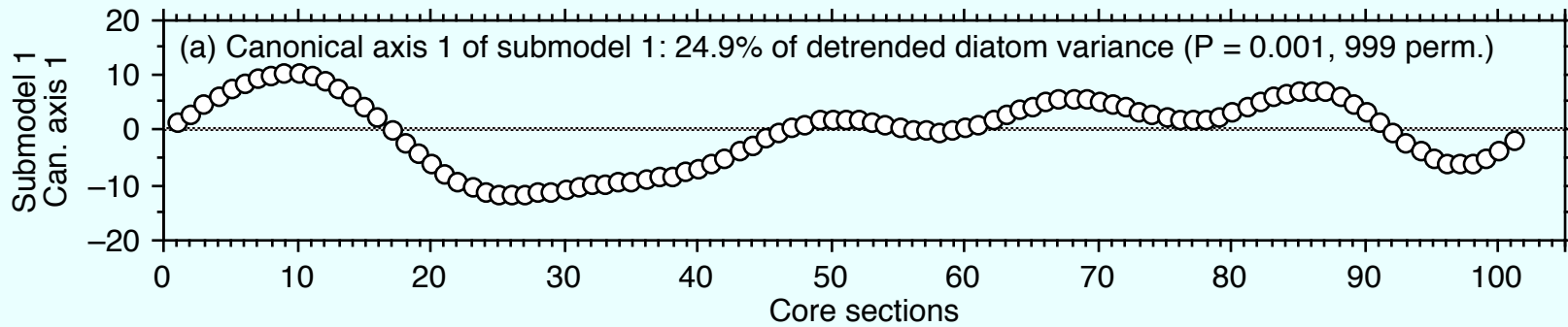
- Medium-scale submodel (dbMEM #11-20):

$R^2 = 0.182$, 1 significant axis

- Fine-scale submodel (10 dbMEM from #28-45):

$R^2 = 0.079$, model not significant

=> The diatom species that were positively or negatively correlated with each canonical axis are identified in the publication.



How to use dbMEM eigenfunctions?

a) We proceeded as follows in the first example:

- dbMEM analysis of the response table \mathbf{Y} ;
- Division of the significant eigenfunctions into submodels;
- Interpretation of the submodels using explanatory variables.

The objective was to divide the variation of \mathbf{Y} into submodels and relate those to explanatory environmental variables.

b) dbMEM eigenfunctions can also be used in the framework of variation partitioning. The variation of \mathbf{Y} is then partitioned with respect to a table of explanatory variables \mathbf{X} and (for example) several tables $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$, containing dbMEM submodels.

How to use dbMEM eigenfunctions?

c) dbMEMs can efficiently model spatial/temporal structures in data. They can be used to control for autocorrelation in tests of significance of the species-environment relationship (fraction [a]).

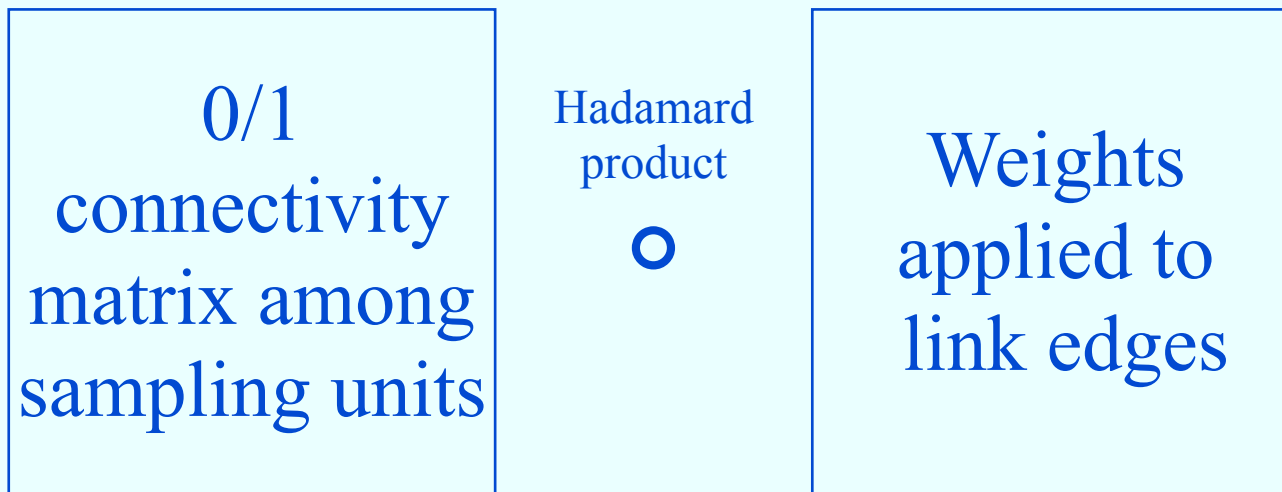
Peres-Neto, P. R. and P. Legendre. 2010. Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography* 19: 174-184.

Stéphane Dray: generalized MEM analysis

The Moran's eigenvector maps (MEM) method is a generalization of dbMEMs to different types of spatial weights. The result is a set of spatial eigenfunctions, as in dbMEM analysis.



Eigen-decomposition
of a spatial/temporal weighting matrix (SWM)



¹ Dray, S., P. Legendre, and P. R. Peres-Neto. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling* 196: 483-493.

$$\mathbf{D}'_{\text{PCNM}} =$$

0	d_{12}	$4t$	$4t$
d_{21}	0		
$4t$		0	
$4t$			0

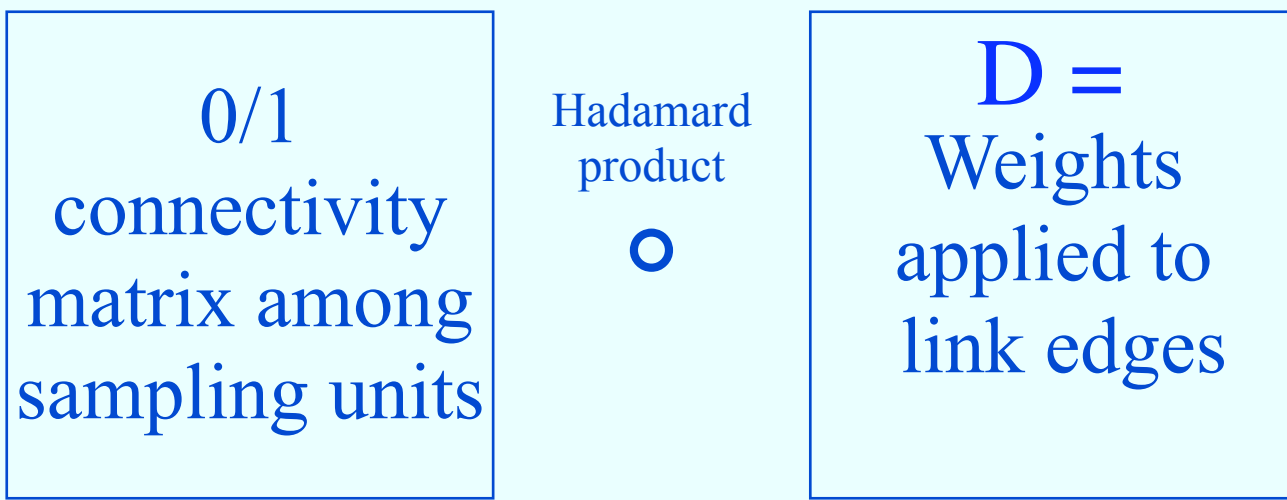
$$\mathbf{D}'_{\text{MEM}} =$$

$4t$	d_{12}	$4t$	$4t$
d_{21}	$4t$		
$4t$		$4t$	
$4t$			$4t$

- In classical PCNM, a point is at distance 0 of itself (diagonal values), meaning that it is connected to itself.
- In MEM, a point is not connected to itself, as in spatial/temporal correlation analysis. The eigenvalues are proportional to Moran's I (the proportionality factor is the sum of the eigenvalues).

Other forms of Moran's Eigenvector Maps (MEM) can be created besides dbMEM (Dray et al. 2006):

- Binary MEMs: double-centre matrix \mathbf{A} , then compute its eigenvalues and eigenvectors.
- Replace matrix \mathbf{D} by some function of the distances.
- Replace \mathbf{D} by some other weights, e.g. resistance of the landscape in landscape ecology.



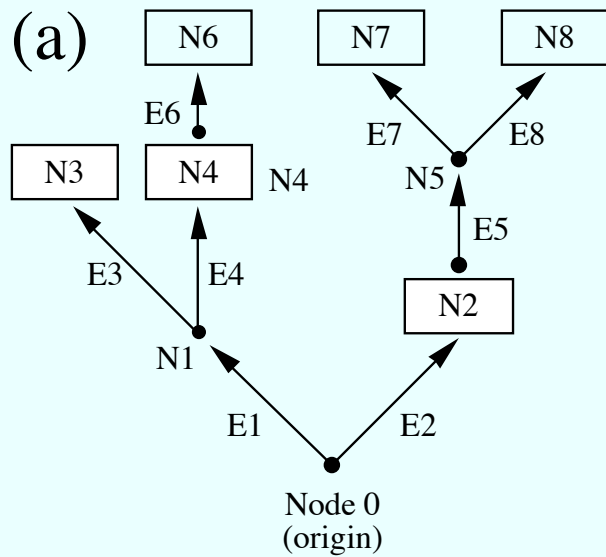
Guillaume Blanchet: AEM analysis

Asymmetric eigenvector maps (AEM) is a spatial eigenfunction method developed to model multivariate (e.g. species) spatial distributions generated by an asymmetric, directional physical process.¹

AEM can also be applied to time series.



¹ Blanchet, F.G., P. Legendre and D. Borcard. 2008. Modelling directional spatial processes in ecological data. *Ecological Modelling* 215: 325-336.



(b)

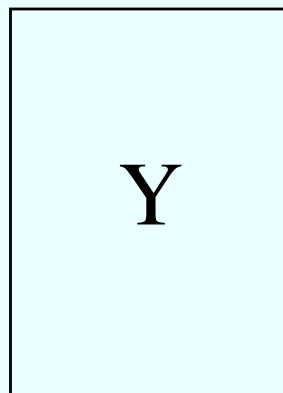
Nodes-by-edges matrix \mathbf{E}

	Edges							
	E1	E2	E3	E4	E5	E6	E7	E8
N1	1	0	0	0	0	0	0	0
N2 (Lake 6)	0	1	0	0	0	0	0	0
N3 (Lake 3)	1	0	1	0	0	0	0	0
N4 (Lake 2)	1	0	0	1	0	0	0	0
N5	0	1	0	0	1	0	0	0
N6 (Lake 1)	1	0	0	1	0	1	0	0
N7 (Lake 4)	0	1	0	0	1	0	1	0
N8 (Lake 5)	0	1	0	0	1	0	0	1

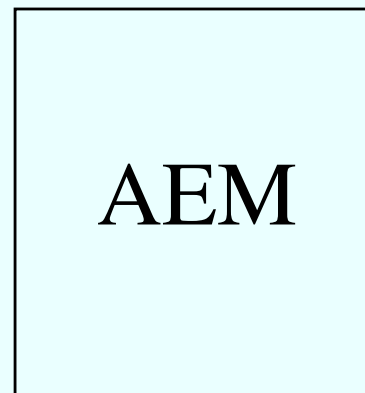
\mathbf{w} = vector of weights

PCA, SVD or PCoA

(c)



Response data

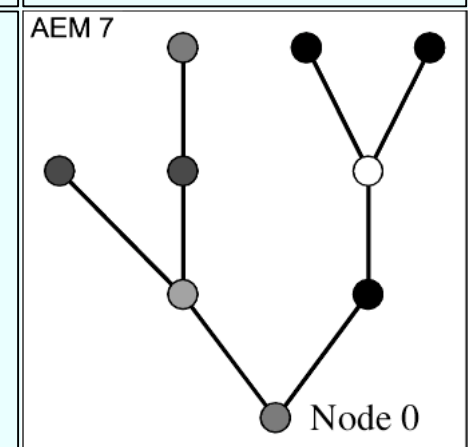
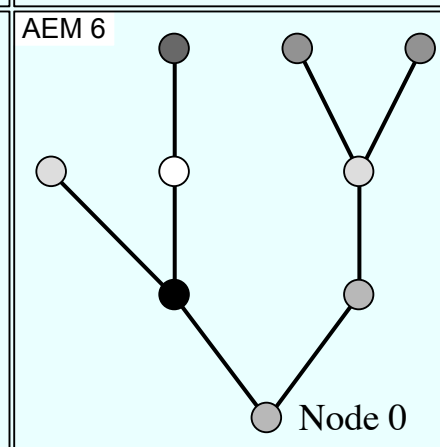
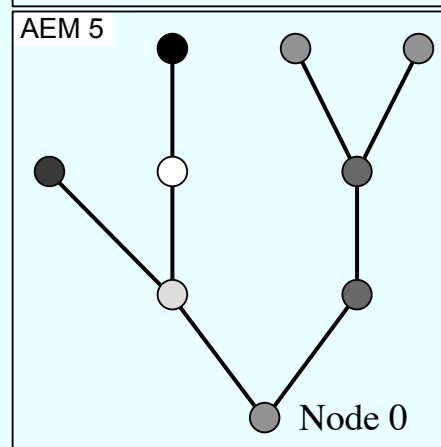
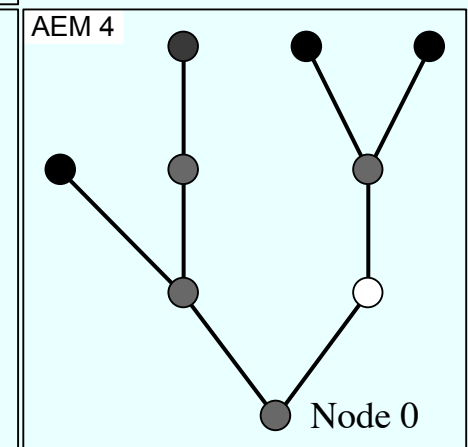
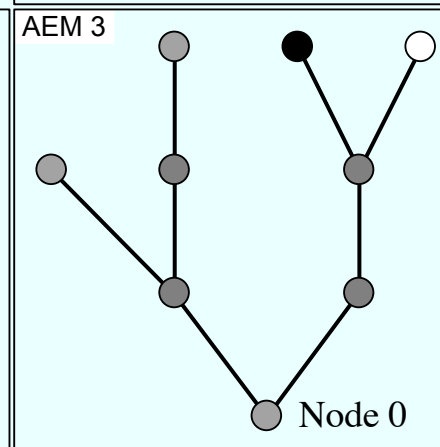
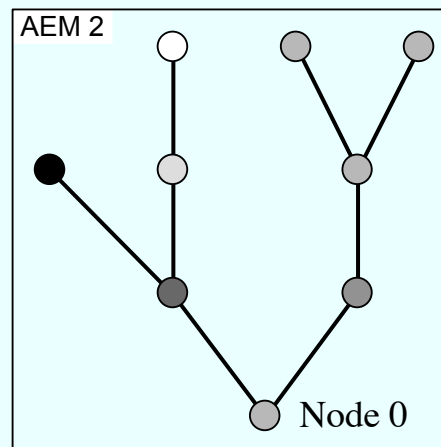
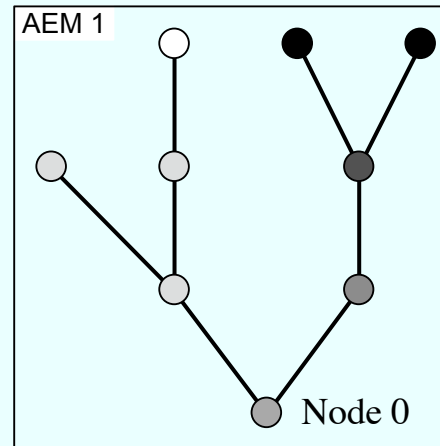


Spatial eigenfunctions

Figures 14.11-14.12 from Legendre & Legendre 2012

For 8 nodes, 7 AEM eigenfunctions are produced.

Shades of grey represent the values in each eigenvector, from white (largest negative value) to black (largest positive value). Signs are arbitrary; they may be reverted with no consequence for the analysis.

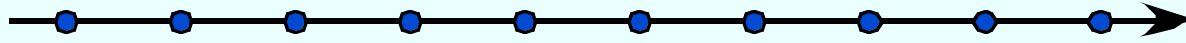


Three applications of AEM analysis to spatial data –

Blanchet, F. G., P. Legendre, R. Maranger, D. Monti, and P. Pepin.
2011. Modeling the effect of directional spatial ecological processes at
different scales. *Oecologia* 166: 357-368.

AEM analysis for a time series

Example: 10 sampling units along time, equal spacing –



Nodes-by-edges matrix \mathbf{E} :

	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9
Time.1	1	0	0	0	0	0	0	0	0	0
Time.2	1	1	0	0	0	0	0	0	0	0
Time.3	1	1	1	0	0	0	0	0	0	0
Time.4	1	1	1	1	0	0	0	0	0	0
Time.5	1	1	1	1	1	0	0	0	0	0
Time.6	1	1	1	1	1	1	0	0	0	0
Time.7	1	1	1	1	1	1	1	0	0	0
Time.8	1	1	1	1	1	1	1	1	0	0
Time.9	1	1	1	1	1	1	1	1	1	0
Time.10	1	1	1	1	1	1	1	1	1	1

Nine AEM eigenfunctions are produced by PCA of matrix \mathbf{E} –

- 4 AEM model positive temporal correlation
- 5 AEM model negative temporal correlation

Time series represent a form of directional stochastic process.

- To emphasize the directional nature of the process influencing the data, **AEM analysis**, which was designed to take trends into account, should be applied to the non-detrended series.
- **MEM analysis** can be applied to data series that were detrended to remove the directional component.

=> By applying both methods to a time series, one can differentiate the directional and non-directional components of variation in a [multivariate] series.

=> Detrended palaeoecological sediment core data, for example, could be studied by dbMEM analysis, and undetrended by AEM analysis.

Computer programs

in the R statistical language

On R-Forge: http://r-forge.r-project.org/R/?group_id=195

PCNM package for classical PCNM and dbMEM (P. Legendre)

AEM package (F. G. Blanchet)

spacemakeR: an R package to compute dbMEM and generalized MEM (D. Dray)

packfor: R package for forward selection of explanatory variables (S. Dray)

On the CRAN page: <http://cran.r-project.org>

vegan package : function **varpart** for multivariate variation partitioning;

pcnm(): construction of classical PCNM eigenfunctions for RDA and CCA;

ordistep() and **ordiR2step()**: forward selection of explanatory variables.

References, p. 1

Legendre, P. and L. Legendre. 2012. Multiscale analysis: spatial eigenfunctions. Chapter 14 in: *Numerical ecology, 3rd English edition*. Elsevier Science BV, Amsterdam.

Borcard, D., F. Gillet, and P. Legendre. 2011. *Numerical ecology with R*. Use R! series, Springer Science, New York.

Papers available in PDF at <http://numericecology.com/>

Blanchet, F. G., P. Legendre, and D. Borcard. 2008. Modelling directional spatial processes in ecological data. *Ecological Modelling* 215: 325-336.

Blanchet, F. G., P. Legendre, R. Maranger, D. Monti, and P. Pepin. 2011. Modelling the effect of directional spatial ecological processes at different scales. *Oecologia* 166: 357-368.

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. *Ecology* 73: 1045-1055.

Borcard, D., and P. Legendre. 1994. Environmental control and spatial structure in ecological communities: an example using Oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics* 1: 37-61.

References, p. 2

Available in PDF at <http://numericecology.com/>

Borcard, D., and P. Legendre. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling* 153: 51-68.

Borcard, D., P. Legendre, C. Avois-Jacquet, and H. Tuomisto. 2004. Dissecting the spatial structure of ecological data at multiple scales. *Ecology* 85: 1826-1832.

Dray, S., P. Legendre, and P. Peres-Neto. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling* 196: 483-493.

Guénard, G., P. Legendre, D. Boisclair, and M. Bilodeau. 2010. Assessment of scale-dependent correlations between variables. *Ecology* 91: 2952-2964.

Peres-Neto, P. R., and P. Legendre. 2010. Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography* 19: 174-184.

Peres-Neto, P. R., P. Legendre, S. Dray, and D. Borcard. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* 87: 2614-2625.

The End