Partitioning beta diversity in landscape ecology and genetics

Pierre Legendre

Département de sciences biologiques

Université de Montréal

Workshop on *Mathematics for an evolving biodiversity*

Centre de recherches mathématiques, Université de Montréal, September 19, 2013

Outline of talk

1. Whittaker's alpha, beta and gamma diversities
2. Measuring beta by a single number: different approaches
3. BD_{Total}, SCBD and LCBD
4. Landscape ecology example
5. Compute BD from a dissimilarity matrix
6. Calculation summmary
7. Properties of **D** matrices for beta assessment
8. Multiple ways of partitioning BD_{Total}
9. Landscape genetics example
10. Conclusion

1. Whittaker's alpha, beta and gamma diversities

• ***Alpha diversity*** is local diversity –

or species diversity at a site. Estimated by species richness or by one of the alpha diversity indices.

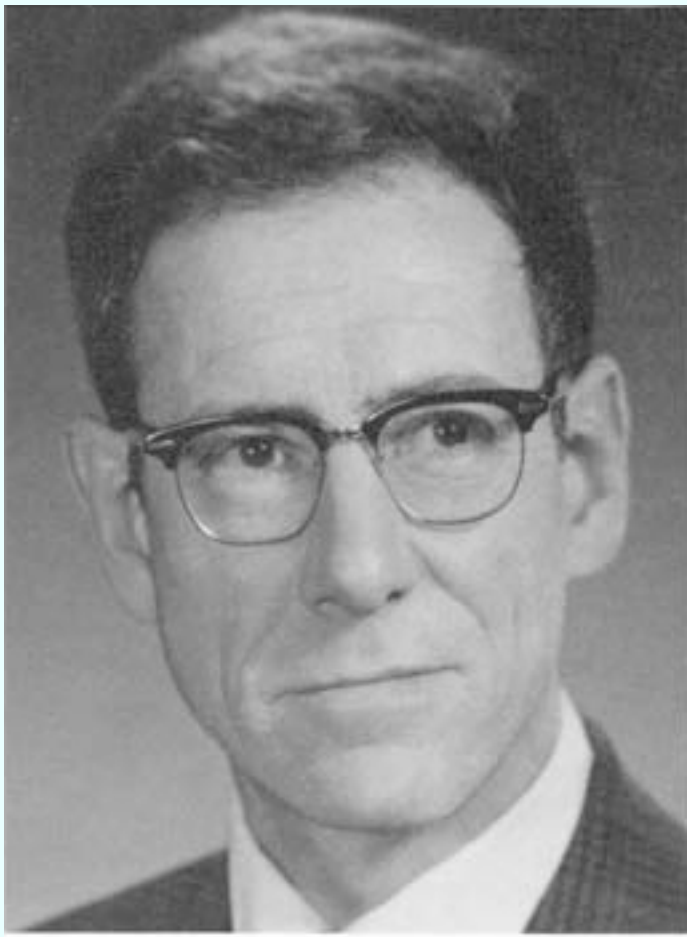
• ***Beta diversity*** is spatial differentiation –

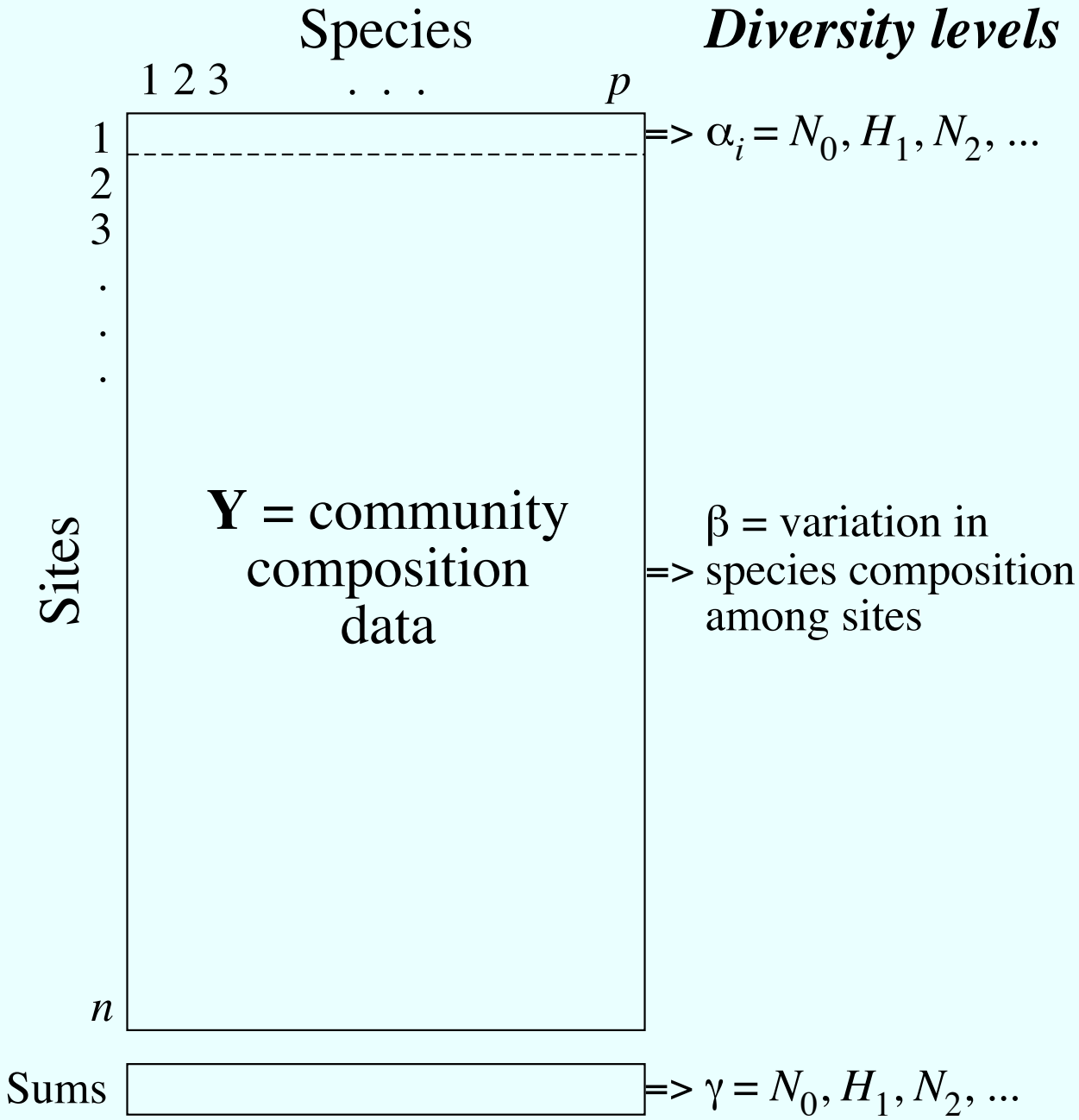
or the variation in species composition among sites within a region of interest.

• ***Gamma diversity*** is regional diversity –

or species diversity in a region of interest. Estimated by pooling observations from a large number of sites in the area and computing an alpha diversity index.

Robert Whittaker (1960, 1972).





Legendre & Legendre *Numerical ecology* (2012, Fig. 6.3).

2. Measuring beta by a single number: different approaches

Studies of beta diversity may focus on two aspects of community structure, distinguishing two types of beta diversity –

- The first is *turnover*, or the directional change in community composition from one sampling unit to another along a predefined spatial, temporal, or environmental gradient. Measure dissimilarities between neighbouring points along the gradient and relate the changes to the gradient values (positions in space, time, or other).

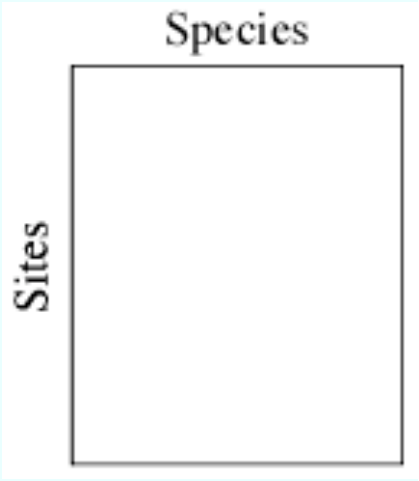
- The second is a *non-directional approach* to the study of community variation through space. It does not refer to any explicit gradient but simply focuses on the variation in community composition among the sampling units.

Vellend (2001), Legendre *et al*. (2005), Anderson *et al*. (2011).

Non-directional **beta diversity** (Whittaker 1960, 1972) can be summarized by a single number –

- Computed as $\beta = S/\overline{\alpha}$ or $\log(\beta) = \log(S) - \log(\overline{\alpha})$

where S = number of species in the larger area of interest (γ diversity) and $\overline{\alpha}$ is the mean number of species at the sampling sites. β indicates how many more species are present in the region than at an average site within the region.



- Or from the Sites × Species data table **Y:**
- Total sum of squares in the community composition table, SS(**Y**)
- Total variance in the data table: BD_{Total} = Var(**Y**) = SS(**Y**)/(n–1)

Many other beta diversity indices are reviewed in Koleff *et al*. (2003), Anderson *et al*. (2011), and other papers.

3. BD_{Total}, SCBD and LCBD

1. Centre data table **Y** by columns, then square the values:

$$\mathbf{S} = \left[s_{ij} \right] = \left[\left(y_{ij} - \bar{y}_j \right)^2 \right]$$

$$\mathbf{Y} = \begin{array}{c} \\ \text{Sites} \end{array} \overbrace{\boxed{[y_{ij}]}}^{\text{Species}} \quad \Rightarrow \quad \mathbf{S} = \begin{array}{c} \\ \text{Sites} \end{array} \overbrace{\boxed{\begin{array}{c}[s_{ij}] = \\ [y_{c.ij}{}^2]\end{array}}}^{\text{Species}}$$

2. Sum all values in matrix **S** to obtain SS(**Y**):

$$\text{SS}_{\text{Total}} = \text{SS}(\mathbf{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{p} s_{ij}$$

3. Divide by the degrees of freedom $(n - 1)$ to obtain Var(**Y**):

$$\text{BD}_{\text{Total}} = \text{Var}(\mathbf{Y}) = \text{SS}_{\text{Total}} / (n-1)$$

Note 1 – These equations should not be computed directly on raw species abundance or biomass data.

Reason: this calculation assumes that the Euclidean distance correctly represents the relationships among sites. However, the Euclidean distance is inappropriate and should not be used for beta diversity assessment (Section 7 of the talk).

Note 2 –

Community composition data should be transformed in some ecologically meaningful way before BD_{Total} is calculated. The chord and Hellinger transformations[1] are appropriate because –

• chord tranformation: $y'_{ij} = y_{ij} \Big/ \sqrt{\sum_{j=1}^{p} y_{ij}^2}$

chord-tranformed data + Euclidean distance = chord distance

• Hellinger transformation: $y'_{ij} = \sqrt{y_{ij} / y_{i+}}$

Hellinger-tranformed data + Euclidean distance = Hellinger distance

Section 7 of the talk will show that the chord and Hellinger distances are appropriate for beta diversity assessment.

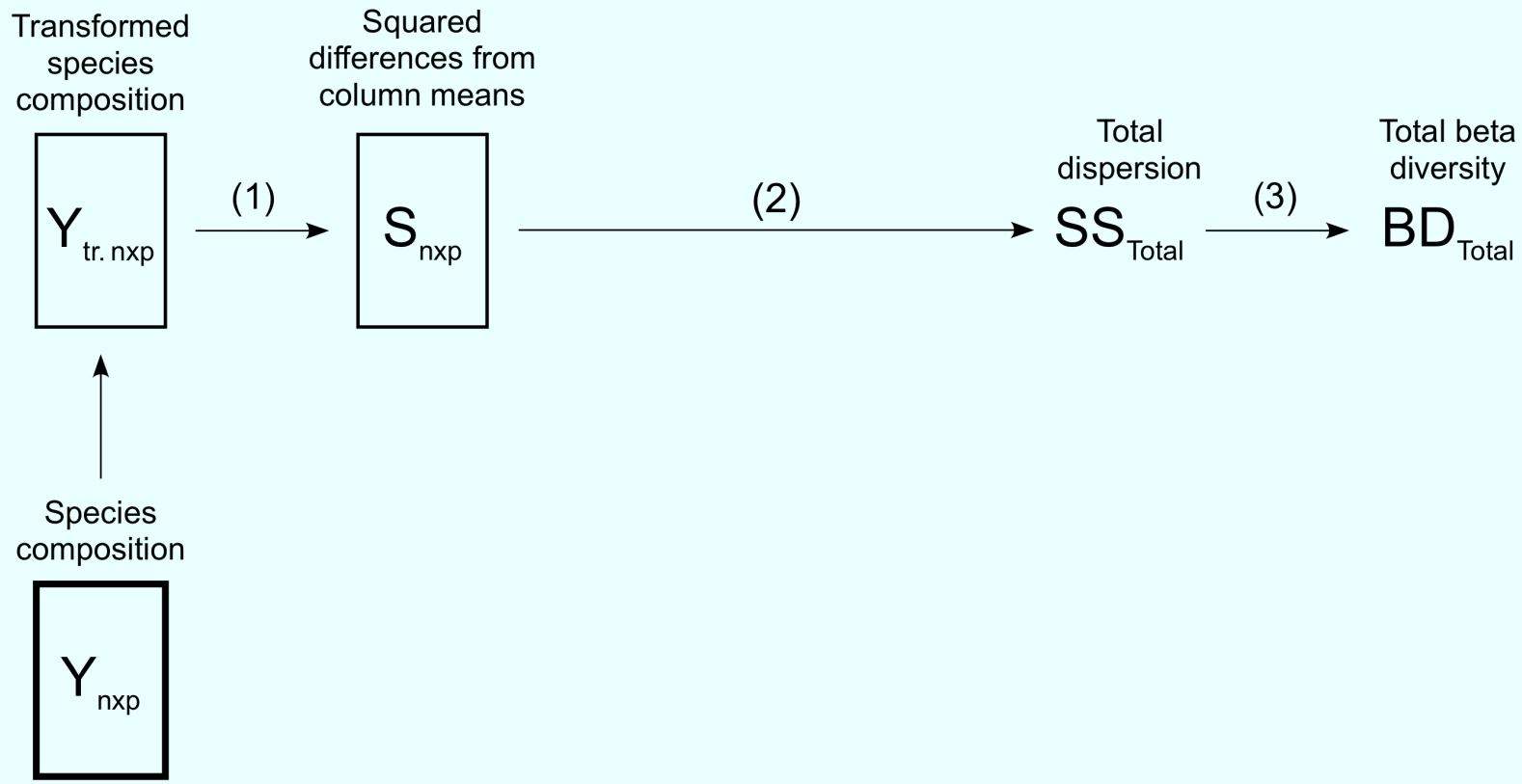
[1] Legendre & Gallagher (*Oecologia* 2001)

Note 3 – Are BD_{Total} statistics comparable?

BD_{Total} statistics computed with the same index are comparable –

- among taxonomic groups observed at the same sites in a geographic area of interest,

- among study areas represented by data sets having the same or different numbers of sampling units (n), for a given taxonomic group, provided that the sampling units are of the same size or represent the same sampling effort.


Transformed species composition
$Y_{tr.\ nxp}$
(1)
Squared differences from column means
S_{nxp}
(2)
Total dispersion
SS_{Total}
(3)
Total beta diversity
BD_{Total}
Species composition
Y_{nxp}

An advantage of conceiving beta as the total variation in **Y** is that SS_{Total} can be decomposed into species and site contributions.

1. *Local Contributions to Beta Diversity* (LCBD) are computed as the sum of the values of **S** in each row *i* :

$$LCBD_i = \sum_{j=1}^{p} s_{ij} \Big/ SS_{Total}$$

=> LCBD values represent *the degree of uniqueness of the sampling units in terms of community composition*.

An advantage of conceiving beta as the total variation in **Y** is that SS_{Total} can be decomposed into species and site contributions.

1. *Local Contributions to Beta Diversity* (LCBD) are computed as the sum of the values of **S** in each row *i* :

$$LCBD_i = \sum_{j=1}^{p} s_{ij} \Big/ SS_{Total}$$

=> LCBD values represent *the degree of uniqueness of the sampling units in terms of community composition*.

2. *Species Contributions to Beta Diversity* (SCBD) are computed as the sum of the values of **S** in each column *j* :

$$SCBD_j = \sum_{i=1}^{n} s_{ij} \Big/ SS_{Total}$$

=> Species with high SCBD values have high abundances at a few sites, hence high variance.

Small numerical example – 7 fish species at 11 sites along a river

	TRU	VAI	LOC	CAR	TAN	GAR	ABL
1	3	0	0	0	0	0	0
2	5	4	3	0	0	0	0
3	5	5	5	0	0	0	0
9	0	1	3	0	1	4	0
18	1	3	3	1	1	2	2
19	0	3	5	1	2	5	3
20	0	1	2	1	4	5	5
21	0	1	1	2	4	5	5
23	0	0	0	0	0	1	2
24	0	0	0	0	0	2	5
25	0	0	0	0	0	1	3

Small numerical example – 7 fish species at 11 sites along a river

	TRU	VAI	LOC	CAR	TAN	GAR	ABL
1	3	0	0	0	0	0	0
2	5	4	3	0	0	0	0
3	5	5	5	0	0	0	0
9	0	1	3	0	1	4	0
18	1	3	3	1	1	2	2
19	0	3	5	1	2	5	3
20	0	1	2	1	4	5	5
21	0	1	1	2	4	5	5
23	0	0	0	0	0	1	2
24	0	0	0	0	0	2	5
25	0	0	0	0	0	1	3

$$SS_{Total} = SS(\mathbf{Y}) = 5.30^{1}$$

$$BD_{Total} = Var(\mathbf{Y}) = 0.530^{1}$$

[1] After chord transformation of the abundance data.

Small numerical example – 7 fish species at 11 sites along a river

	TRU	VAI	LOC	CAR	TAN	GAR	ABL	LCBD
1	3	0	0	0	0	0	0	
2	5	4	3	0	0	0	0	
3	5	5	5	0	0	0	0	
9	0	1	3	0	1	4	0	
18	1	3	3	1	1	2	2	
19	0	3	5	1	2	5	3	
20	0	1	2	1	4	5	5	
21	0	1	1	2	4	5	5	
23	0	0	0	0	0	1	2	
24	0	0	0	0	0	2	5	
25	0	0	0	0	0	1	3	
SCBD								

$SS_{Total} = SS(\mathbf{Y}) = 5.30$

$BD_{Total} = Var(\mathbf{Y}) = 0.530$

• LCBD indices can be tested for significance by random, independent permutations of the columns of **Y**. Example of a permutation of **Y**:

Matrix **Y**

	Sp.1	Sp.2	Sp.3	Sp.4	Sp.5
Site.1	0	10	20	30	40
Site.2	1	11	21	31	41
Site.3	2	12	22	32	42
Site.4	3	13	23	33	43
Site.5	4	14	24	34	44
Site.6	5	15	25	35	45
Site.7	6	16	26	36	46
Site.8	7	17	27	37	47
Site.9	8	18	28	38	48
Site.10	9	19	29	39	49

Matrix **Y** permuted

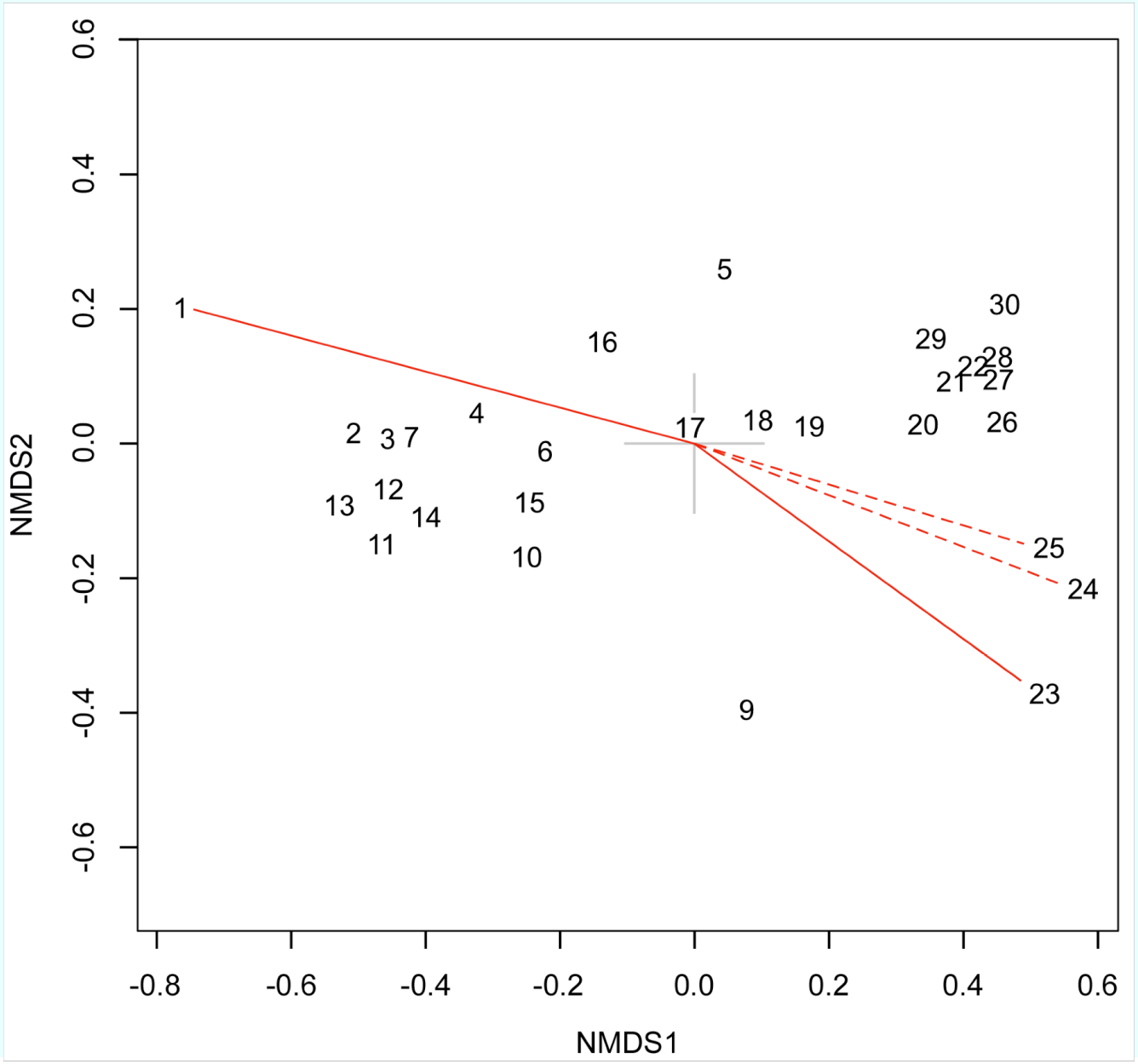
	Sp.1	Sp.2	Sp.3	Sp.4	Sp.5
Site.1	7	12	26	33	41
Site.2	4	15	28	36	40
Site.3	2	18	20	35	44
Site.4	0	19	22	32	46
Site.5	9	17	21	34	49
Site.6	8	10	27	30	48
Site.7	3	14	25	37	45
Site.8	5	13	24	31	42
Site.9	1	11	23	39	43
Site.10	6	16	29	38	47

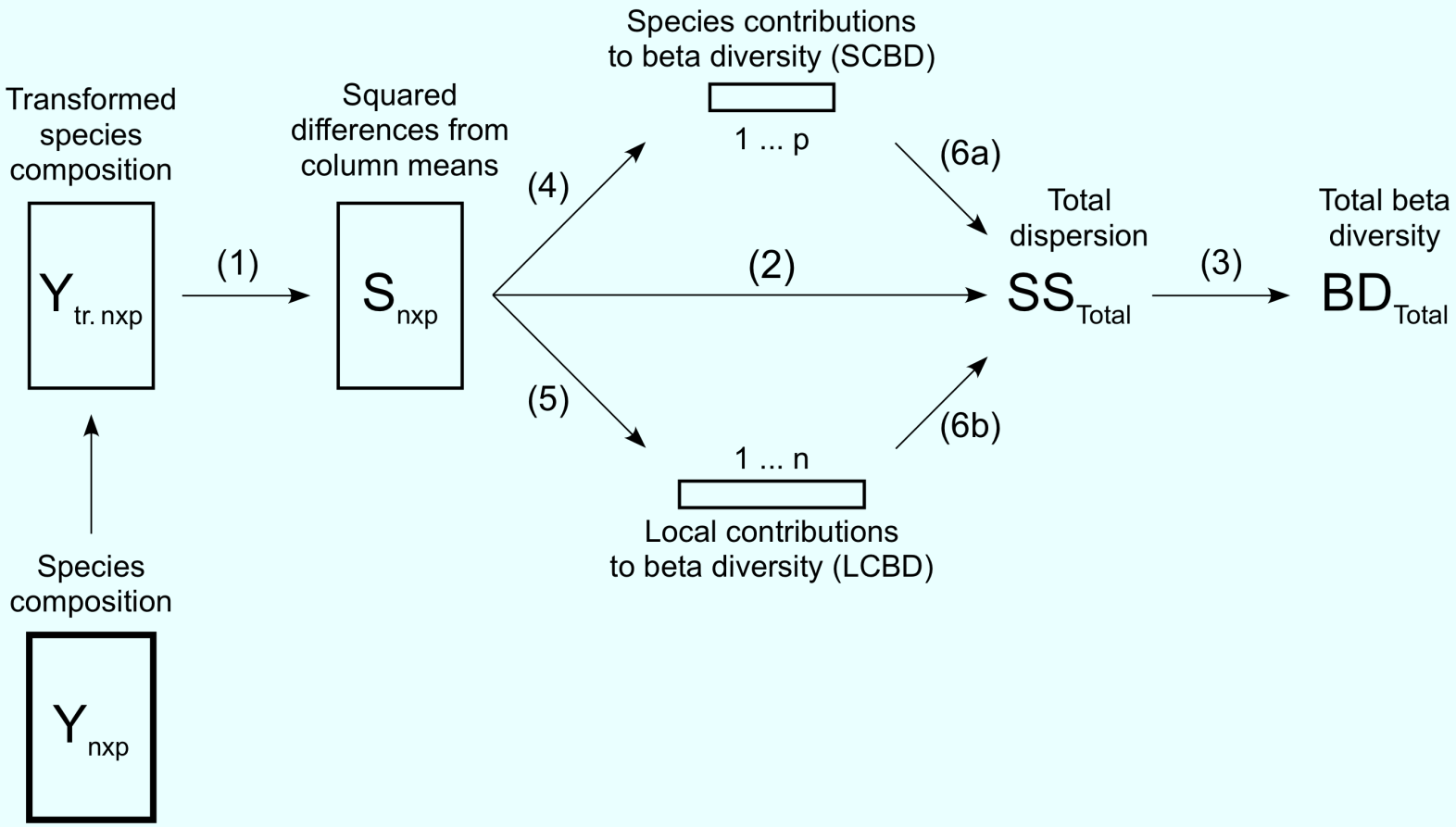
R command to produce a random permutation:

```
Y.perm <- apply(Y, 2, sample)
```

Or use a permutation method that preserves spatial correlation.

LCBD: squared distance to the centroid in an ordination diagram. The sites near the centre are not exceptional in species combination.




Species contributions
to beta diversity (SCBD)
1 ... p
Transformed
species
composition
Squared
differences from
column means
(4)
(6a)
Total
dispersion
Total beta
diversity
(1)
(2)
(3)
$Y_{tr.\ nxp}$
S_{nxp}
SS_{Total}
BD_{Total}
(5)
(6b)
1 ... n
Local contributions
to beta diversity (LCBD)
Species
composition
Y_{nxp}

4. Full landscape ecology example

Fish observed at 29 sites along the Doubs river, a tributary of the Saône running near the France-Switzerland border in the Jura Mountains, eastern France.

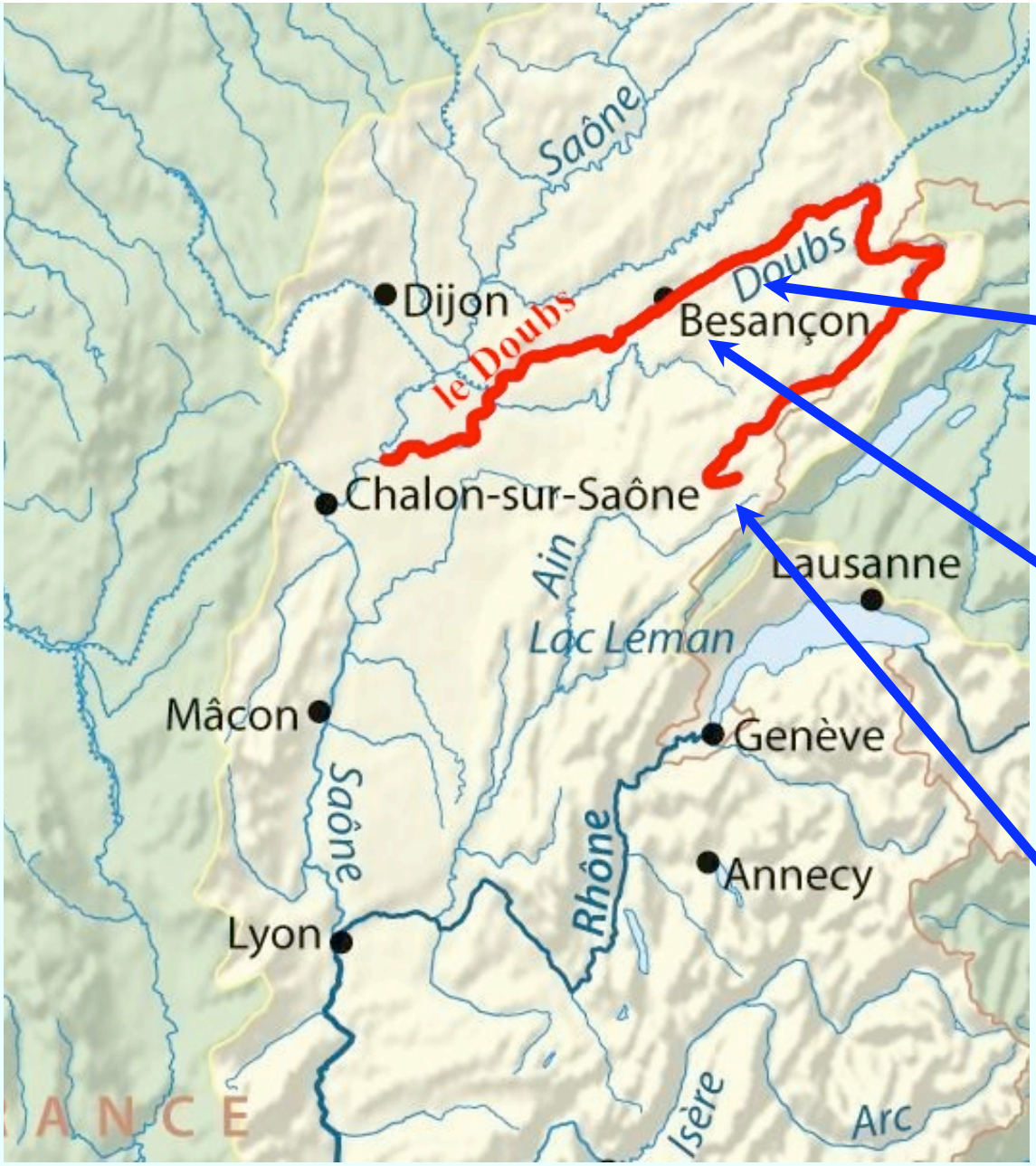
- Data from Verneaux (1973), available at

http://adn.biol.umontreal.ca/~numericalecology/numecolR/

the Web page of *Numerical ecology with R* (Borcard *et al*. 2011).

- Analysis of the chord-transformed fish abundance data:

$$SS_{Total} = SS(\mathbf{Y}) = 15.243 \qquad BD_{Total} = Var(\mathbf{Y}) = 0.544$$

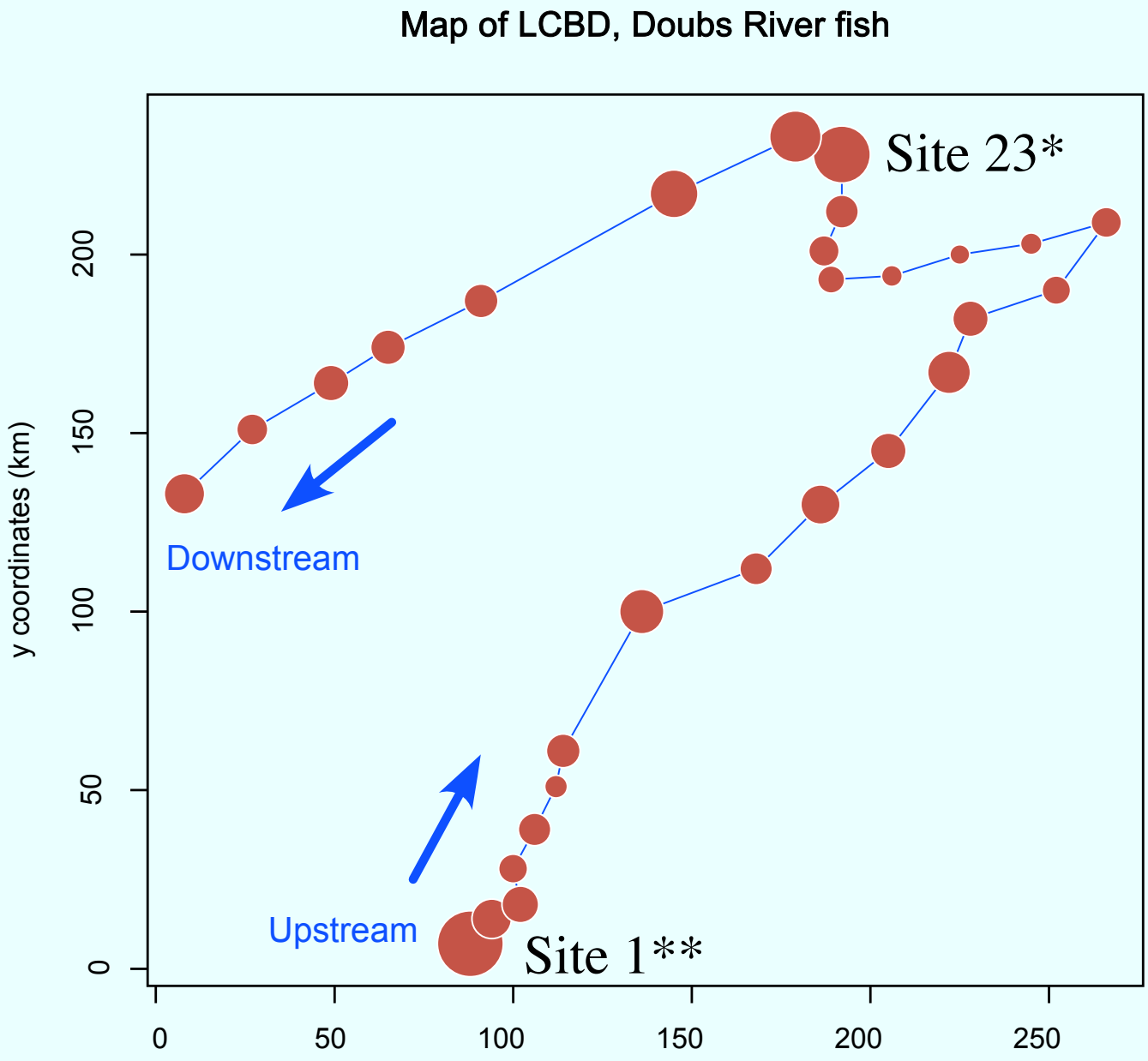


Entre Laissey et Deluz, peu avant Besançon

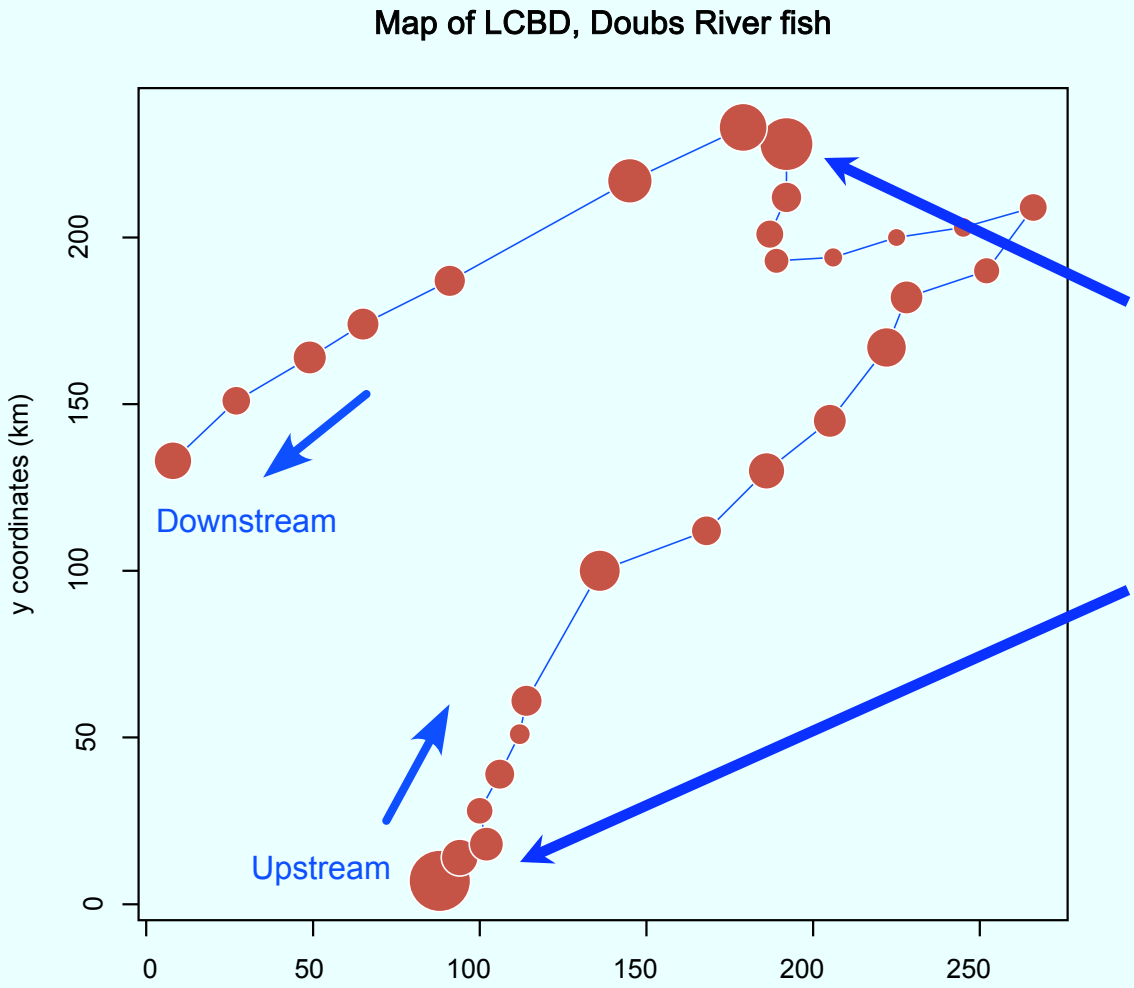
Besançon

Les sources du Doubs à Mouthe

http://fr.wikipedia.org/wiki/Doubs_(rivière)



LCBD: uniqueness of community composition at each site.



Species with high SCBD:

Common common bleak/Ablette (*Alburnus alburnus*) abundant in eutrophic sites mid-river

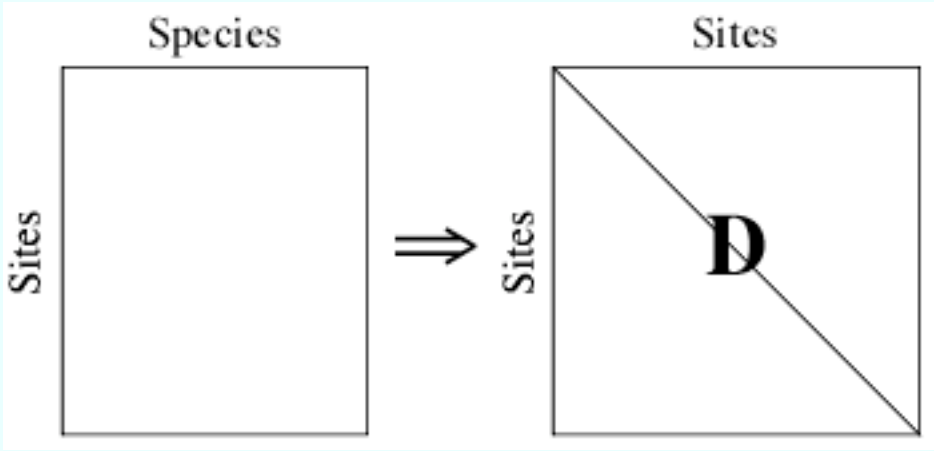
Brown trout/Truite brune (*Salmo trutta*), Eurasian minnow/Vairon (*Phoxinus phoxinus*) and stone loach/Loche franche (*Nemacheilus barbatulus*) in oligotrophic sites upriver

Two signif. LCBD (sites 1, 23) after correction for multiple testing.

Regression of LCBD on environmental variables: LCBD positively related to slope of the riverbed and BOD; adjusted $R^2 = 0.58$.

5. Compute BD from a dissimilarity matrix

Whittaker (1972) –
Beta biversity can be computed from a dissimilarity matrix



• Presence-absence data:
Jaccard or Sørensen coefficient, or β computed for pairs of sites.

• Quantitative community composition data: Odum percentage difference, Hellinger, chord or chi-square distance, etc.

=> Whittaker (1972): the mean of the dissimilarities is another single-number index of beta diversity.

Compute SS_{Total} from the upper triangular portion of a dissimilarity matrix[1] :

$$SS_{Total} = SS(\mathbf{Y}) = \frac{1}{n}\sum_{h=1}^{n-1}\sum_{i=h+1}^{n} D_{hi}^2$$

For **D** that are not Euclidean but $\mathbf{D}^{(0.5)} = \left[\sqrt{D_{hi}^2}\right]$ is Euclidean:

$$SS_{Total} = SS(\mathbf{Y}) = \frac{1}{n}\sum_{h=1}^{n-1}\sum_{i=h+1}^{n} D_{hi}$$

Then, compute BD_{Total} :

$$BD_{Total} = Var(\mathbf{Y}) = \frac{SS_{Total}}{n-1}$$

[1] Proof in Legendre & Fortin (2010, Appendix 1).

Compute LCBD from a dissimilarity matrix –

Compute Gower-centred matrix **G** containing the centred dissimilarities in principal coordinate analysis (PCoA; classical or metric scaling)[1]:

$$\mathbf{G} = \left(\mathbf{I} - \frac{\mathbf{11}'}{n}\right)\left[-0.5 D_{hi}^2\right]\left(\mathbf{I} - \frac{\mathbf{11}'}{n}\right)$$

=> The LCBD values are the diagonal elements of **G** divided by SS_{Total}

$$\left[\mathrm{LCBD}_i\right] = \frac{\mathrm{diag}(\mathbf{G})}{SS_{Total}}$$

LCBD indices can be computed and tested for significance.

☹ SCBD cannot be computed from a dissimilarity matrix.

[1] Legendre & Legendre, *Numerical ecology* (2012), eq. 9.42.

Range of values of BD_{Total}

All dissimilarity functions used to analyse beta diversity have a maximum value (D_{max}), reached when two sites have completely different community compositions.

- For example, the Hellinger and chord distances have a minimum value of 0 and a maximum of $\sqrt{2}$.
- If all sites in **Y** have the exact same species composition, all distances in **D** are 0 and

$$\text{Var}(\mathbf{Y}) = \frac{1}{n(n-1)} \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} D_{hi}^2 = 0$$

- If all sites in **Y** have entirely different species compositions, all $n(n-1)/2$ distances in **D** are $\sqrt{2}$ and

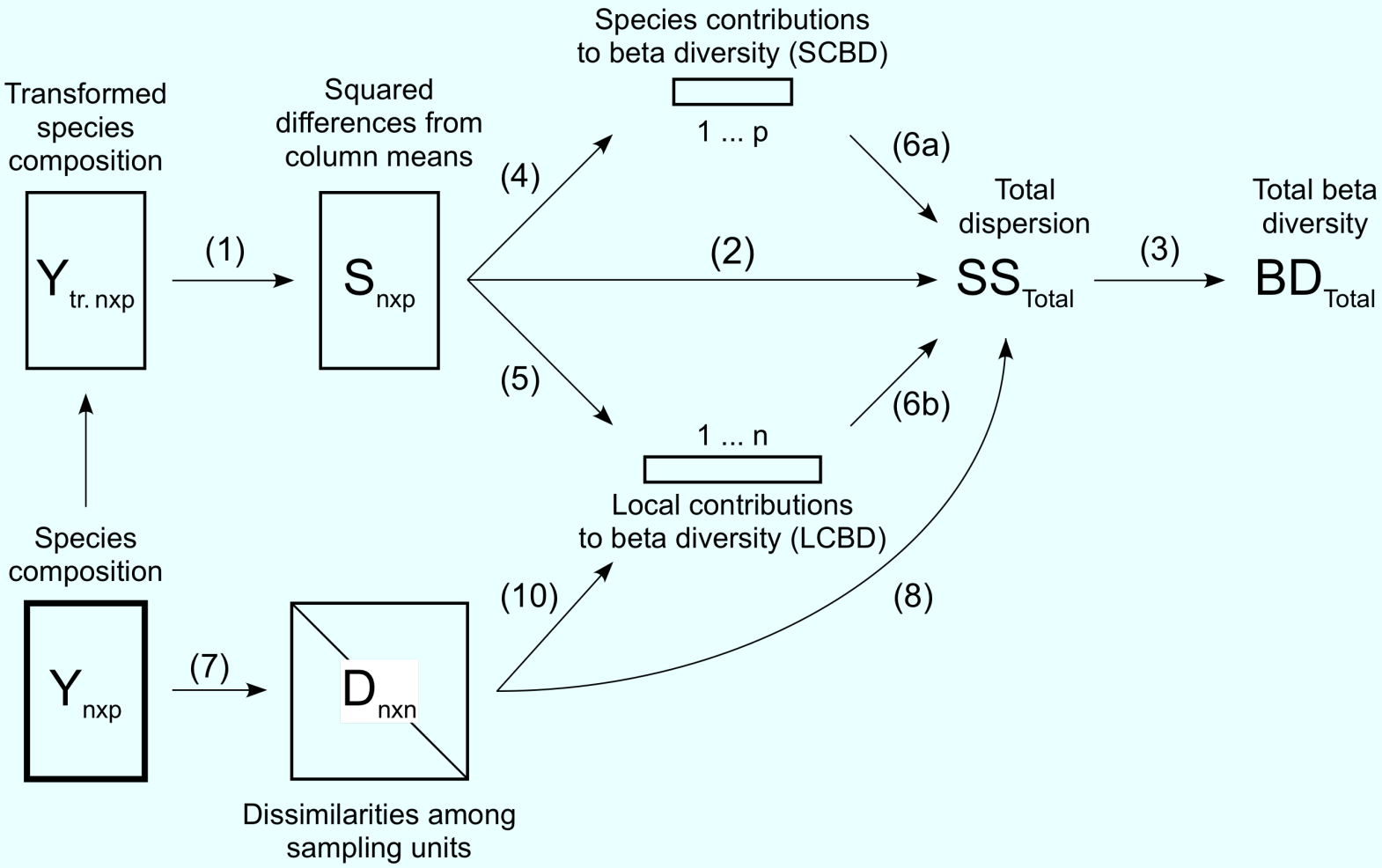
$$\text{Var}(\mathbf{Y}) = \frac{1}{n(n-1)} \left(\frac{n(n-1)}{2} \sqrt{2}^2 \right) = 1$$

For these distances, BD_{Total} is in the range [0, 1].

• Dissimilarity indices with $D_{max} = 1$ have maximum BD = 0.5 when all sites have different species compositions. Hence the range of their BD_{Total} values is $[0, 0.5]$.

For these distances, multiply BD by 2 to produce normalized BD values in the range [0, 1].

6. Calculation summary



7. Properties of **D** *matrices for beta assessment*

Basic necessary properties

P1 – Minimum of zero, positiveness: $D \geq 0$.

P2 – Symmetry: $D(\mathbf{x}_1,\mathbf{x}_2) = D(\mathbf{x}_2,\mathbf{x}_1)$.

P3 – Monotonicity to changes in abundance: D increases when differences in abundance increase.

P4 – Double-zero asymmetry: D does not change when adding double-zeros but D changes when double-X are added where $X > 0$.

P5 – Sites without species in common have the largest D.

P6 – D does not decrease in series of nested species assemblages.

Comparability between data sets

P7 – Species replication invariance.

P8 – Invariance to measurement units, e.g. for biomass data.

P9 – Existence of a fixed upper bound, D_{max}.

Additional properties useful in some studies.

Sampling issues

P10 – Invariance to the number of species in each sampling unit.

P11 – Invariance to the total abundance in each sampling unit.

P12 – Coefficients with corrections for undersampling.

Ordination-related properties

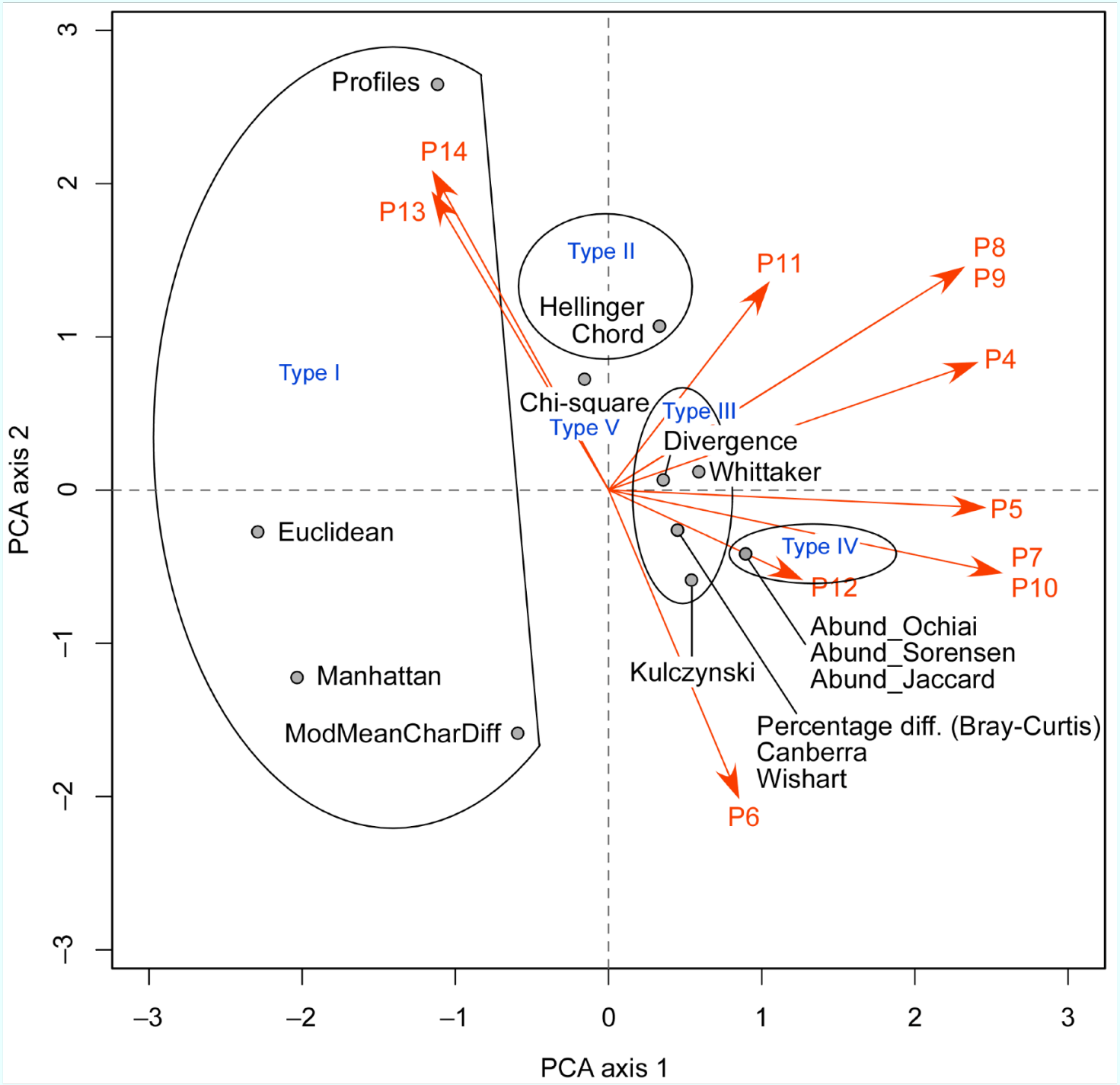
P13 – $\mathbf{D}$ or $\mathbf{D}^{(0.5)} = [D^{0.5}]$ is Euclidean. PCoA ordinations without negative eigenvalues and complex axes.

P14 – Dissimilarity function emulated by transformation of the raw frequency data followed by Euclidean distance.

Example: the chord distance can be computed by applying the chord transformation to the community composition data, followed by calculation of the Euclidean distance. The Hellinger, chord, profile and chi-square distances have that property.

#1–9: Necessary properties for beta assessment

Dissimilarity	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	D_{max}
Euclidean distance	0	0	1	0	0	0	0	0	0	2	1	—
Manhattan distance	0	0	1	0	0	0	0	0	0	1	0	—
Modified mean character diff.	1	0	1	1	0	0	1	0	0	0	0	—
Species profile distance	1	0	0	0	1	1	0	1	0	2	1	$\sqrt{2}$
Hellinger distance	1	1	1	1	1	1	1	1	0	2	1	$\sqrt{2}$
Chord distance	1	1	1	1	1	1	1	1	0	2	1	$\sqrt{2}$
Chi-square distance	1	0	1	1	1	1	NA	0	0	2	1	$\sqrt{2y_{++}}$
Coefficient of divergence	1	1	1	1	1	1	1	0	0	2	0	1
Canberra metric	1	1	1	1	1	1	1	0	0	1	0	1
Whittaker's index	1	1	1	1	1	1	1	1	0	1	0	1
Percentage difference (B-C)	1	1	1	1	1	1	1	0	0	1	0	1
Wishart coefficient	1	1	1	1	1	1	1	0	0	1	0	1
D = (1–Kulczynski coefficient)	1	1	1	1	1	1	1	0	0	0	0	1
Abundance-based Jaccard	1	1	1	1	1	1	1	1	1	0	0	1
Abundance-based Sørensen	1	1	1	1	1	1	1	1	1	0	0	1
Abundance-based Ochiai	1	1	1	1	1	1	1	1	1	0	0	1


PCA axis 2
PCA axis 1
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Profiles
P14
P13
Type II
Hellinger
Chord
P11
P8
P9
P4
Type I
Chi-square
Type V
Type III
Divergence
Whittaker
P5
Euclidean
Type IV
P7
P10
P12
Abund_Ochiai
Abund_Sorensen
Abund_Jaccard
Kulczynski
Manhattan
ModMeanCharDiff
Percentage diff. (Bray-Curtis)
Canberra
Wishart
P6

The following 11 coefficients are appropriate for BD studies

Type II: Hellinger and chord distances. They justify the application of the Hellinger and chord transformations to raw abundance data and direct calculation of BD_{Total}, followed by partition analysis *(next slide)*.

Type III: Canberra, Whittaker, divergence, percentage difference (*alias* B-C), Wishart, Kulczyinki.

Type IV: Abundance-based quantitative forms of Jaccard, Sørensen and Ochiai coefficients with corrections for undersampling.

The following 5 coefficients are inappropriate

Type I: Euclidean, Manhattan, modified mean character difference, species profiles.

Type V: The chi-square distance.

8. Multiple ways of partitioning BD_{Total}

1. Partition BD_{Total} among species (SCBD) and among sites (LCBD).

2. Multivariate analysis of variance (MANOVA) using a single factor partitions the SS_{Total} into within- and among-group sums of squares.

In MANOVA involving two or several crossed factors, SS_{Total} is partitioned SS_{Total} among the factors and their interaction.

3. Partition SS_{Total} by simple and canonical ordinations, e.g. PCA and redundancy analysis (RDA).

4. SS_{Total} can be partitioned with respect to two or more matrices of explanatory variables by variation partitioning (Borcard & Legendre 1992, 1994).

5. SS_{Total} can be partitioned as a function of different spatial scales by spatial eigenfunction analysis (MEM, AEM), multivariate variogram, and multiscale ordination analysis (Wagner 2003, 2004).

Reference

Legendre, P. and M. De Cáceres. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology Letters* 16: 951-963.

PDF available on:

http://adn.biol.umontreal.ca/~numericalecology/Reprints/

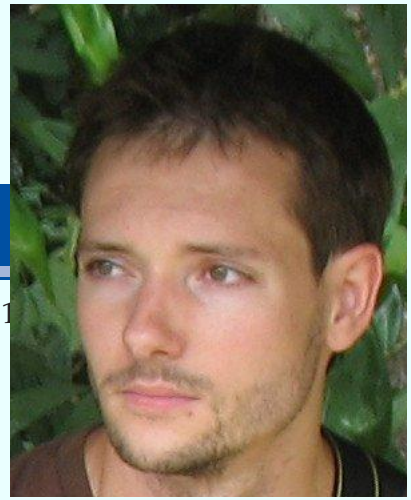
9. Landscape genetics example

Freshwater snail *Drepanotrema depressissimum* in a fragmented landscape of tropical ponds in Grande-Terre, Guadeloupe.

Photo: Jean-Pierre Pointier

- Microsatellite data from Lamy *et al*. (2012)

See also http://amnat.org/an/newpapers/AprLamy.html



MOLECULAR ECOLOGY

Molecular Ecology (2012) 21, 1394–1410 doi: 10.111

Testing metapopulation dynamics using genetic, demographic and ecological data

T. LAMY,* J. P. POINTIER,† P. JARNE* and P. DAVID*

**UMR 5175 CEFE, campus CNRS, 1919 route de Mende, 34293 Montpellier cedex 5, France, †USR 3278 CNRS-EPHE, 52 avenue Paul Alduy, 66860 Perpignan cedex, France*

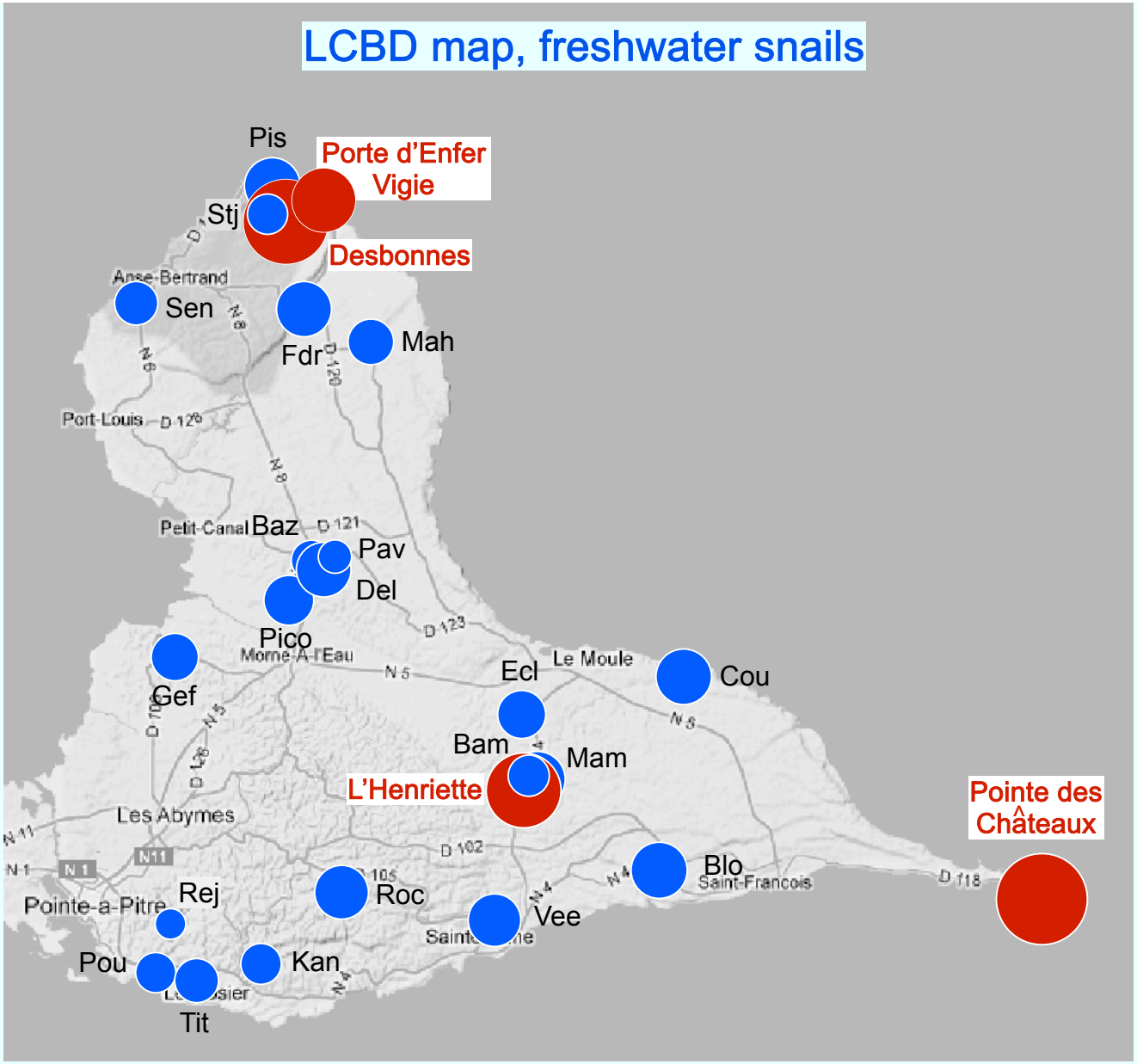
Landscape genetics example

Drepanotrema depressissimum (Gastropoda, Planorbidae)

Microsatellite data from Lamy *et al*. (2012)

- 25 populations in ponds, rivulets, & swamp grasslands, Guadeloupe
- 749 individual snails were genotyped (diploids)
- 10 microsatellite loci, with a mean of 34 alleles per locus
- LCBD analysis through the genetic chord distance:

$$BD_{Total} = Var(\mathbf{Y}) = 0.197$$



Four sites have significant LCBD indices, indicating the most genetically unique populations.

Sites with high LCBD values indicate the most genetically unique populations. Something happened to create exceptional allele combinations. What was it?

Regression tree analysis of LCBD values on environmental variables (pond size, vegetation cover, connectivity, and temporal stability) showed that the four sites with high LCBD are ponds –>

• where temporal stability is the lowest (sites regularly dry up), causing loss of alleles through random sampling (genetic drift),

• and connectivity is low with neighbouring ponds (no connection at all), preventing migration of snails from adjacent areas.

Snails can survive in dessicated ponds by aestivating in the sediment.

These mechanisms reduced the gene pool of these four populations to a few alleles per locus.

10. Conclusion

Beta diversity (BD) is the spatial variation in community – or genetic composition – among sites in a geographic region of interest.

• BD can be estimated in various ways. The estimator described and used in this talk is the variance of the community composition data, Var(**Y**). BD_{Total} = Var(**Y**) is a general, flexible index of beta diversity.

• BD_{Total} can be computed either from the [transformed] raw data or from a dissimilarity matrix.

• At least 11 dissimilarity coefficients are appropriate for beta diversity studies.

• BD_{Total} can be decomposed into SCBD and LCBD (–> maps).

• BD_{Total} = Var(**Y**) links beta diversity to all well-known methods of multivariate analysis of community composition data.

Do you have questions?