Topic models

Machine Learning Journal Club

November 22nd 2017

Classifying text documents

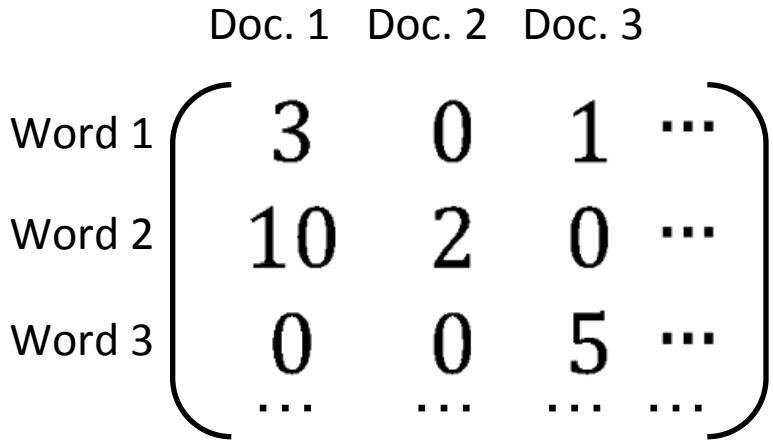
How to classify very large numbers of text documents to make them easily searchable?

→ Associate ‘topics’ to documents based on their word content

Earlier approaches: matrix factorization

Latent Semantic Analysis (1988)

- Based on the singular-value decomposition of the matrix of word counts
- Aims at a finding vectors of co-occurring words



Probabilistic topic models

Instead of a deterministic decomposition, a probabilistic model is fitted to the data

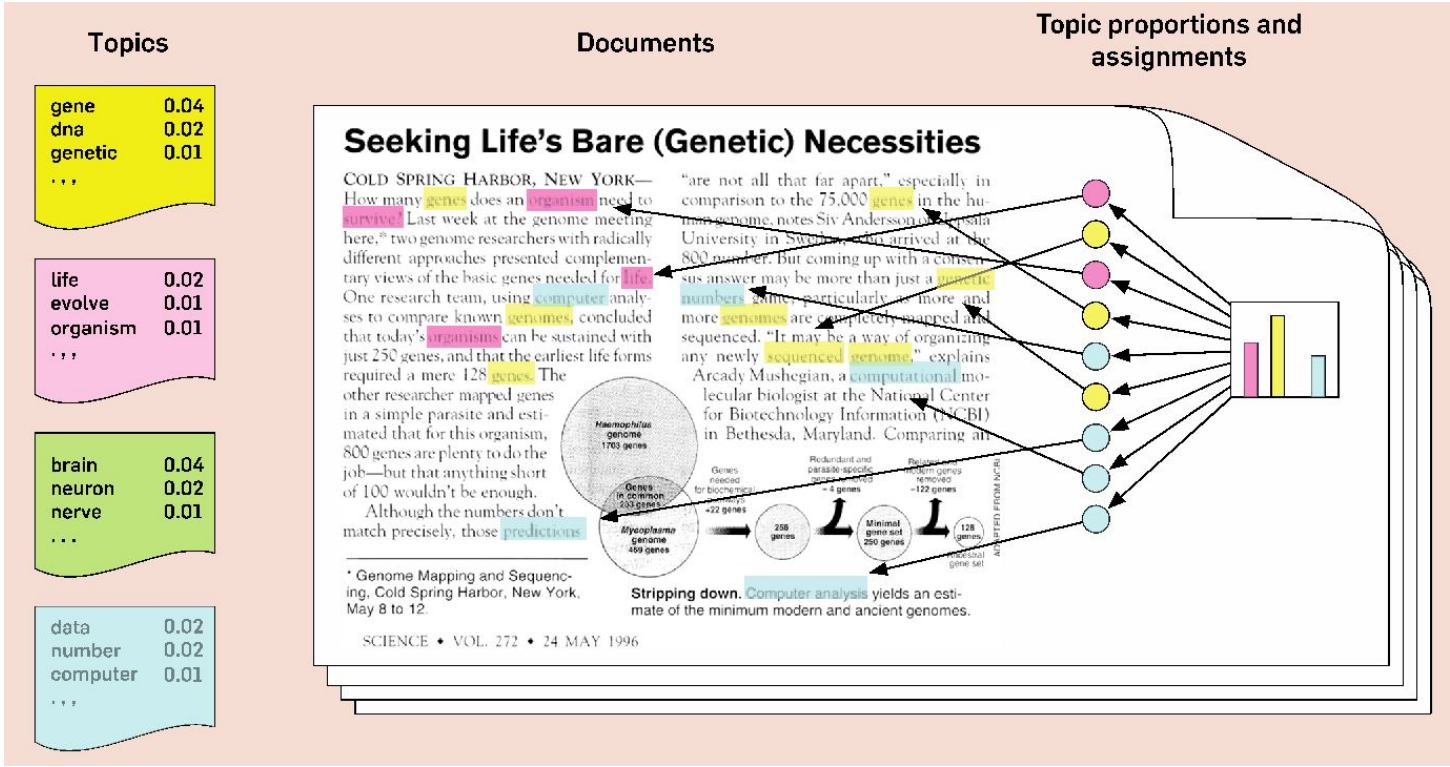
→ Fitted using a stochastic algorithm (Markov Chain Monte Carlo or Expectation-Maximization)

Latent Dirichlet Allocation

Most popular topic model (*Blei, 2003*)

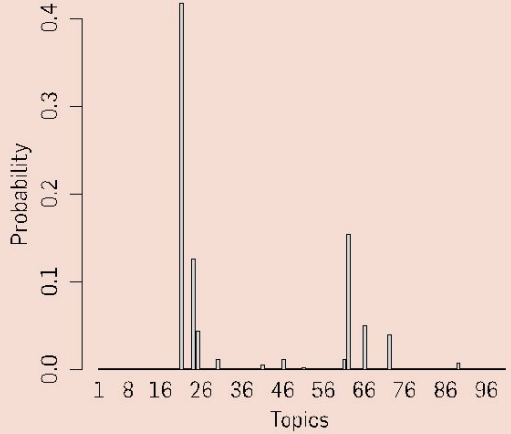
- Aims at decomposing a corpus of text documents into groups of co-occurring words
- Each word occurrence in each document is assigned to a topic (the number of topics is fixed in advance)
- Topics are described by a distribution over words
- Documents are described by a distribution over topics

Latent Dirichlet Allocation



Blei et al. 2012

Latent Dirichlet Allocation



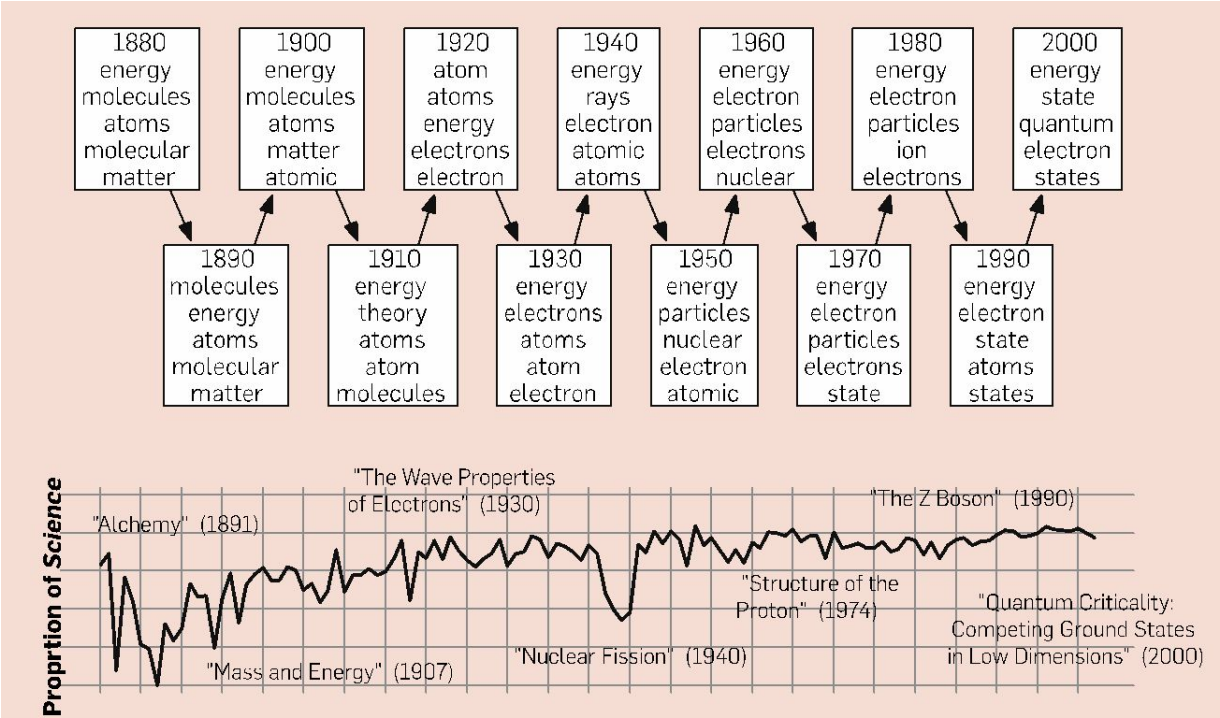
"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Blei et al. 2012

Extensions to Latent Dirichlet Allocation

- Hierarchical Dirichlet Processes (*Teh et al. 2006*)
  - → Fits the number of topics as part of the probabistic model
- Author-topic models (*Rosen-Zvi et al. 2004*)
  - → Accounts for metadata, such as documents authors
- Time-dependent topic models
  - → Allows for time-varying topic composition

Extensions to Latent Dirichlet Allocation

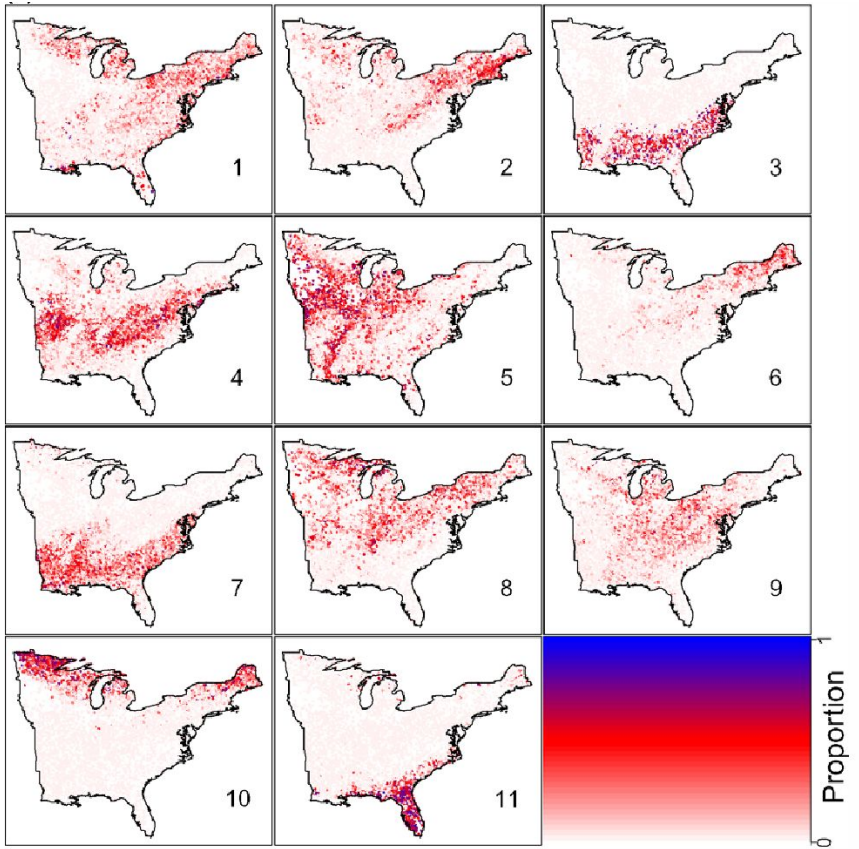


Blei et al. 2012

Use beyond natural language processing

- Bioinformatics
- Image classification
- Music classification
- Fraud detection in telecommunications …

Biodiversity data



Valle et al. 2014

DNA-based biodiversity survey: ‘metabarcoding’

Environmental sample (e.g. soil) → DNA extraction → PCR amplification of a DNA barcode → High-throughput Illumina sequencing

DNA-based biodiversity surveys generate:

- abundance data (number of sequence reads)
- for a large number of OTUs
- in a large number of sampled locations.

	Sample 1	Sample 2	Sample 3	…
OTU 1 ■ OTU 2 ★ OTU 3 ◆ …	■■■■ ★	■■ ★★ ◆	★★★ ◆◆◆	

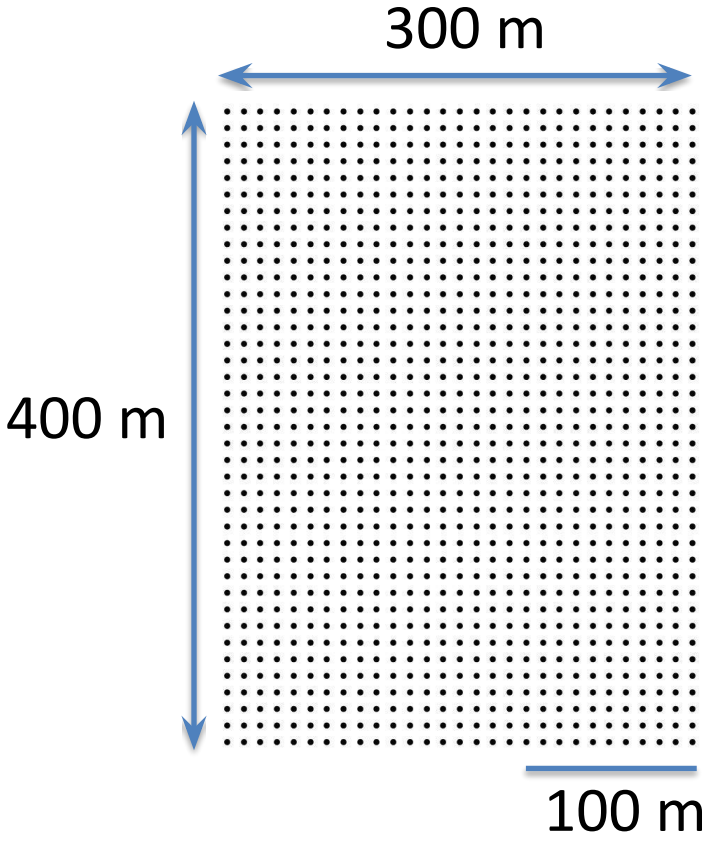
All the information is contained in the co-occurrence of OTUs and in the covariance of their abundances among samples

➔ How to retrieve and represent this information?

Metabarcoding dataset

Sampling scheme:

- 12 ha of tropical forest
- 1 soil sample every 10 m
- 1,136 samples



Metabarcoding dataset

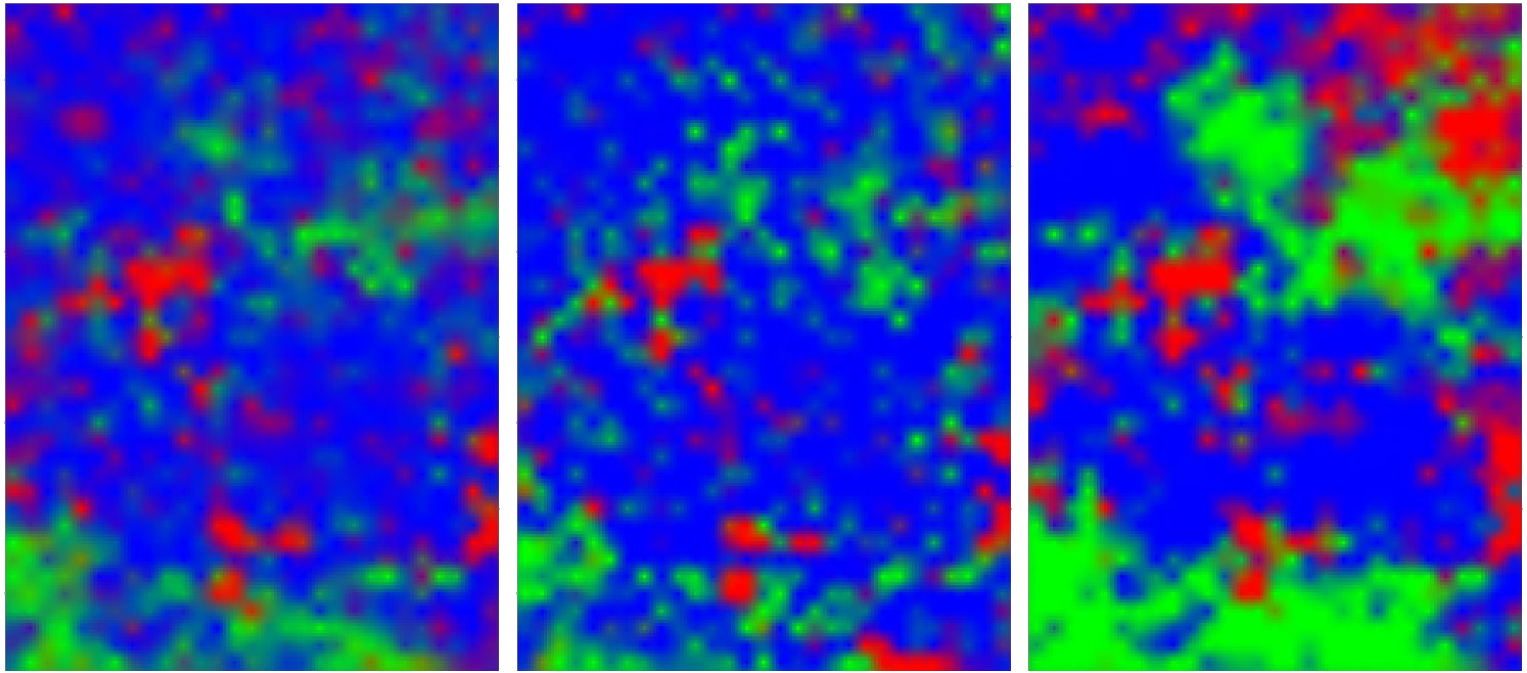
5 barcodes:

Bacteria (16S): 20,162 OTUs
Archaea (16S): 4,101 OTUs
Fungi (ITS1): 9,855 OTUs
Plants (trnL): 1,360 OTUs

Protists (18S): 1,648 OTUs
Arthropods (18S): 1,881 OTUs
Annelids (18S): 51 OTUs
Nematodes (18S): 378 OTUs
Platyhelminthes (18S): 126 OTUs

LDA decomposition

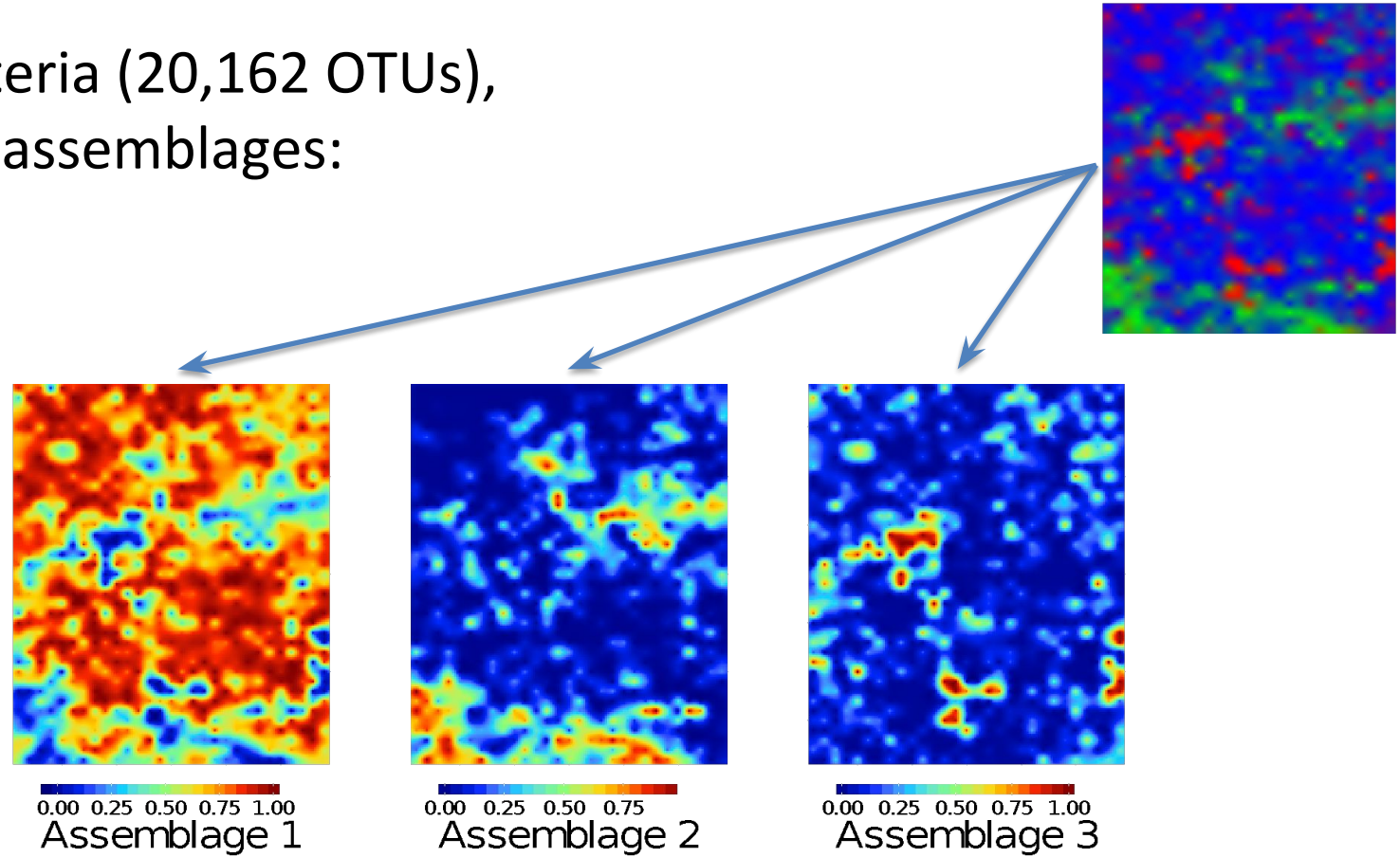
K=3 assemblages:



Bacteria Protists Fungi

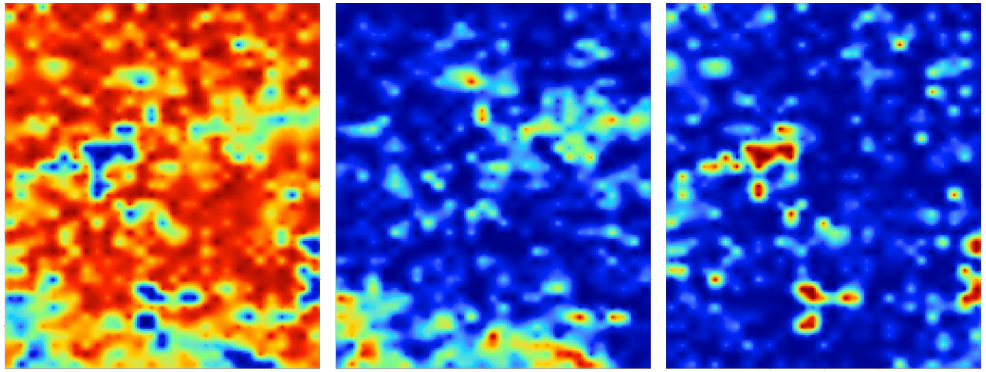
LDA decomposition

Bacteria (20,162 OTUs),
K=3 assemblages:



LDA decomposition

Bacteria (20,162 OTUs),
K=3 assemblages:



LiDAR measurements	Terra firme	Hydromorphic	Exposed rock
Topography	0.36*	-0.43*	0.00
Slope	-0.26*	0.31*	0.01
Wetness	-0.27*	0.40*	-0.08*

Stability

How strongly structured are the data?

➔ Stability of the taxonomic composition over 100 realizations with random initial conditions