UNIVERSITY OF TWENTE.

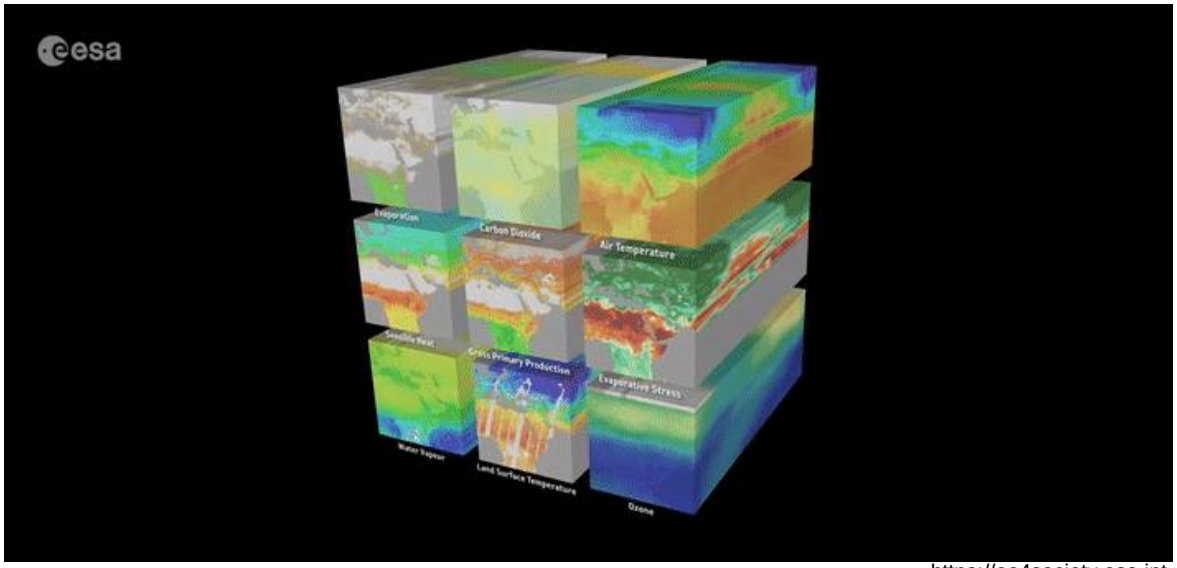
ON THE MAPPING OF PHENOLOGICAL REGIONS VIA ADVANCED CLUSTERING

RAÚL ZURITA-MILLA[1], EMMA IZQUIERDO-VERDIGUIER[2], FRANCESCO NATTINO[3], OU KU[3], MEIERT W. GROOTES[3] AND SERKAN GIRGIN[1]

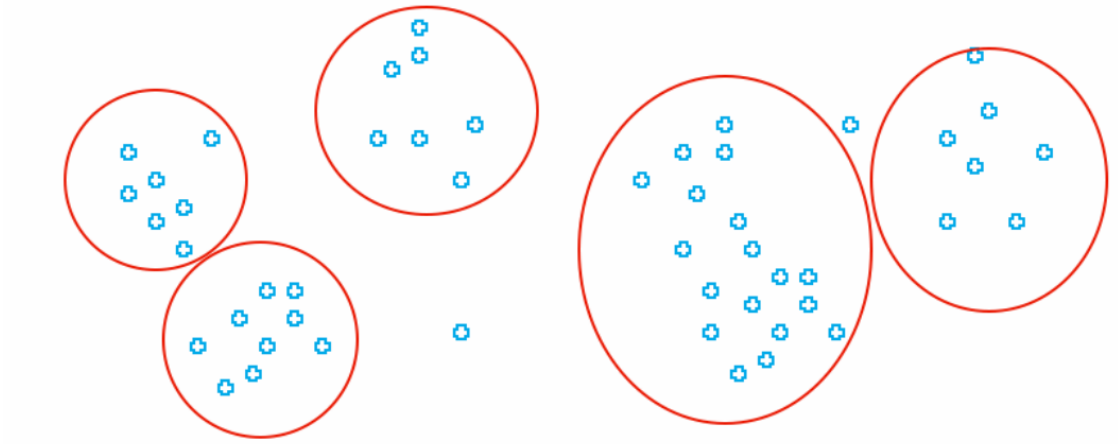
(1) FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION (ITC), UTWENTE, NL
(2) UNIVERSITY OF NATURAL RESOURCES AND LIFE SCIENCES (BOKU) AT.
(3) NETHERLANDS ESCIENCE CENTER (NLESC), NL.

ITC FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

EXPLORING DATA CUBES



https://eo4society.esa.int



SPATIO-TEMPORAL CLUSTERING

Spatial clustering

Time / Space	T1	T2	T3	T4	…	Tn
S1	213	221	63	194	…	216
S2	100	97	16	179	…	186
S3	33	67	239	15	…	111
S4	161	46	237	52	…	142
…	…	…	…	…	…	…
Sn	17	123	216	174	…	210

SPATIO-TEMPORAL CLUSTERING

Temporal clustering

Time / Space	T1	T2	T3	T4	…	Tn
S1	213	221	63	194	…	216
S2	100	97	16	179	…	186
S3	33	67	239	15	…	111
S4	161	46	237	52	…	142
…	…	…	…	…	…	…
Sn	17	123	216	174	…	210

SPATIO-TEMPORAL CLUSTERING

S-T clustering

Space \ Time	T1	T2	T3	T4	…	Tn
S1	213	221	63	194	…	216
S2	190	97	16	179	…	186
S3	32	57	239	75	…	111
S4	161	46	237	52	…	142
…	…	…	…	…	…	…
Sn	17	123	216	174	…	210

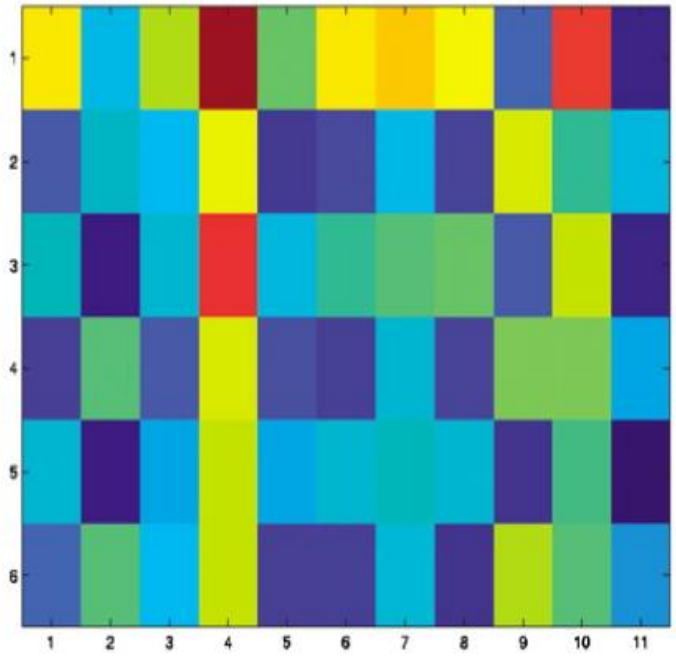
Co-clustering → ‘blocks’ of rows and columns

Hartigan, 1972 & Banerjee et al., 2007

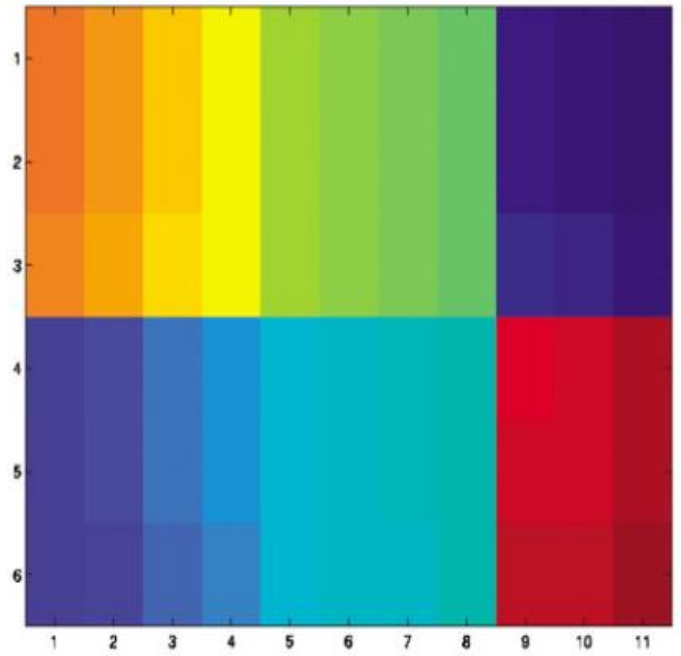
Wu, Zurita-Milla et al. 2015; 2016

Wu, Cheng, Zurita-Milla & Song, 2020

CO-CLUSTERING



Input data matrix



Co-clustered data matrix

PHENOLOGICAL REGIONS

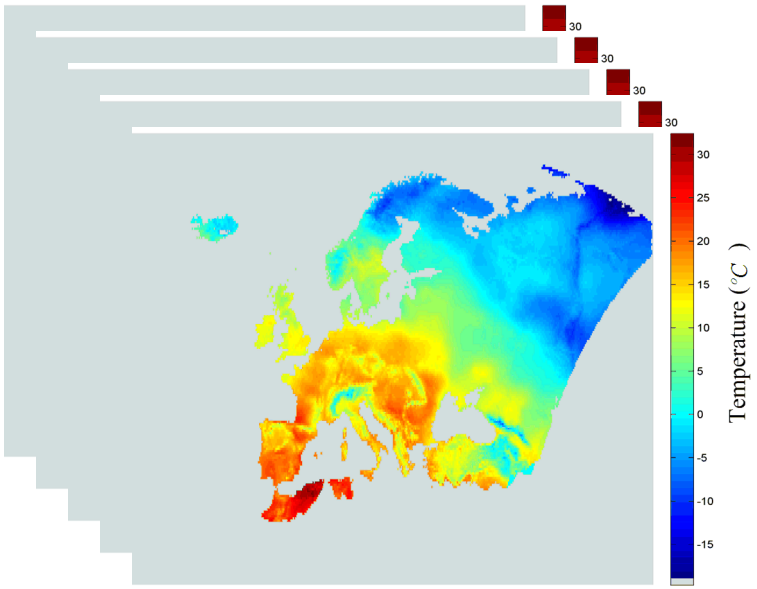
- OBJECTIVE: To identify phenoregions and study their temporal dynamics

- DATA: ECA → daily Tmin, Tmax since 1950
- PRE-PROCESSING: SI-x models used to transform temperature and latitudinal information into a biologically meaningful product: first leaf day (FLD) for key indicator species (lilac and honeysuckle).

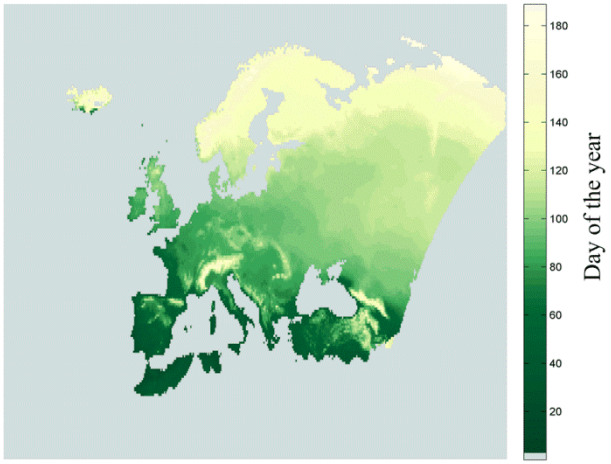
- METHODS: Co-clustering (refined by using k-means)

SPRING INDICES

FLD = f (TN, TX, Lat)



Daily TN and TX

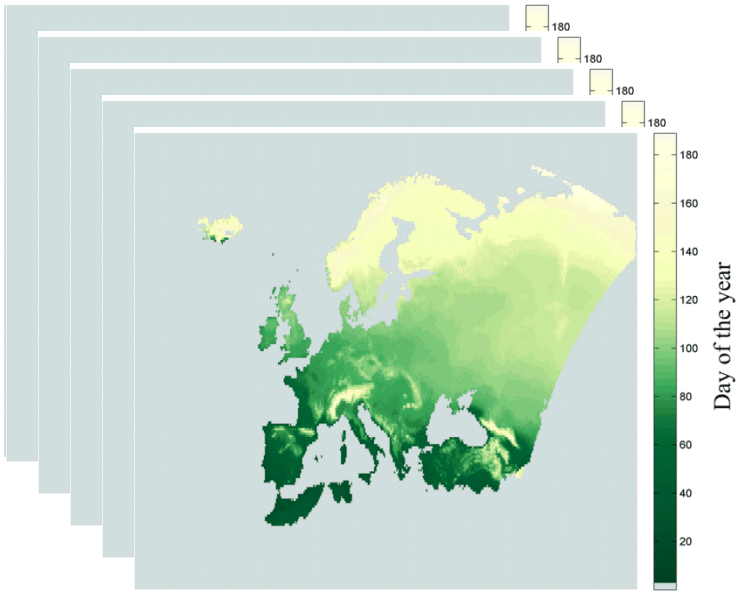


FLD

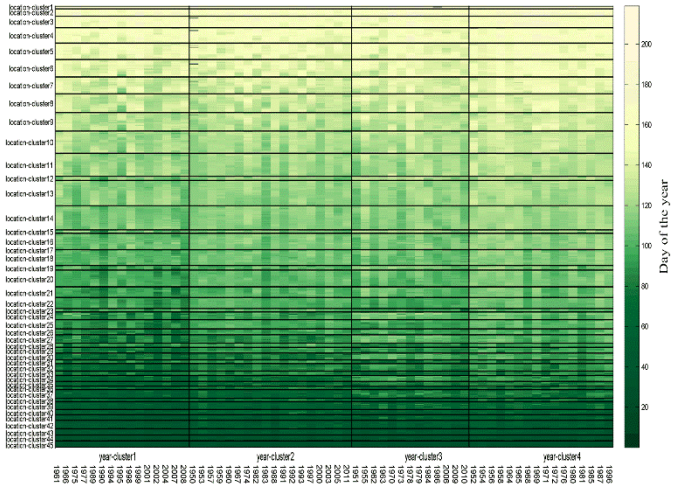
Spatial resolution = 0.25° (28500 grid cells)
Temporal res= 1950 – 2012

CO-CLUSTERING

k=45; l = 4
$\varepsilon=10^{-6}$
N=200



Yearly FLD maps
1950-2012



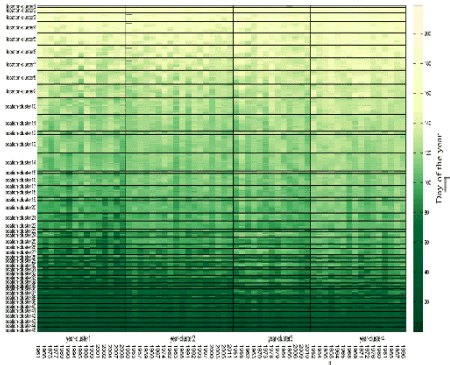
Re-ordered FLD matrix

CO-CLUSTERS

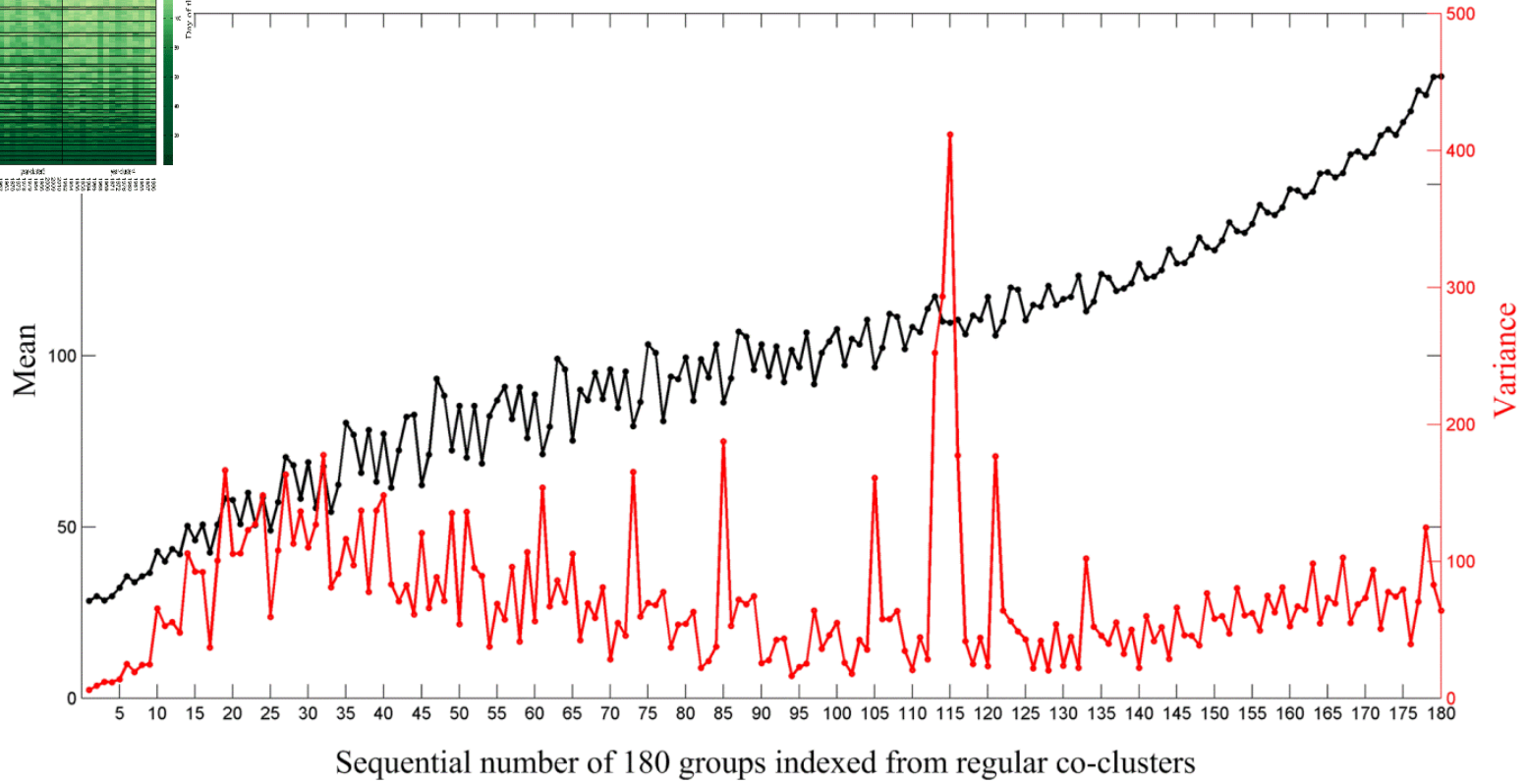
location-cluster45

https://doi.org/10.1002/2015JG003308

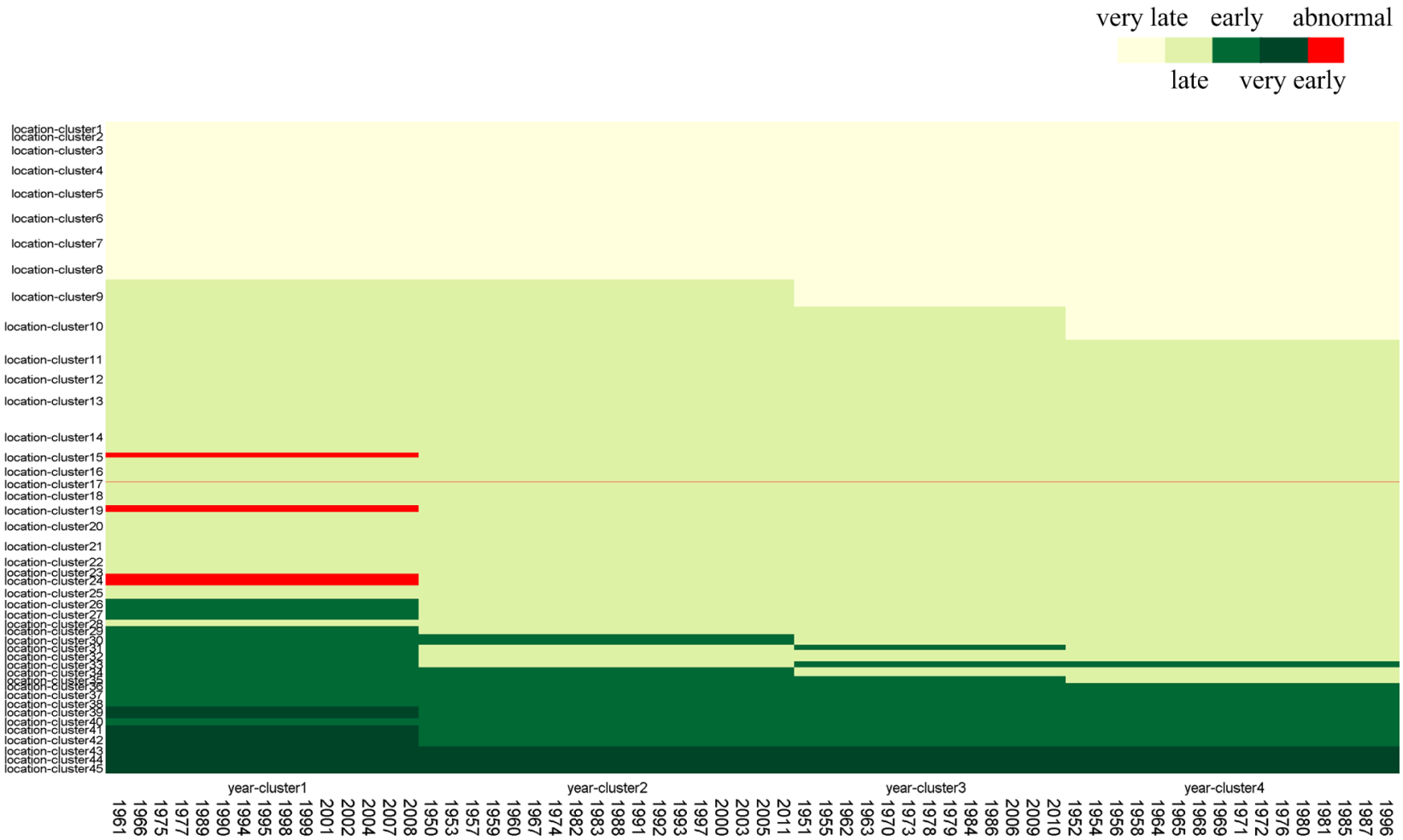
FEATURES FOR KMEANS



Kmeans optimized using the silhouette method (Rousseeuw, 1987)

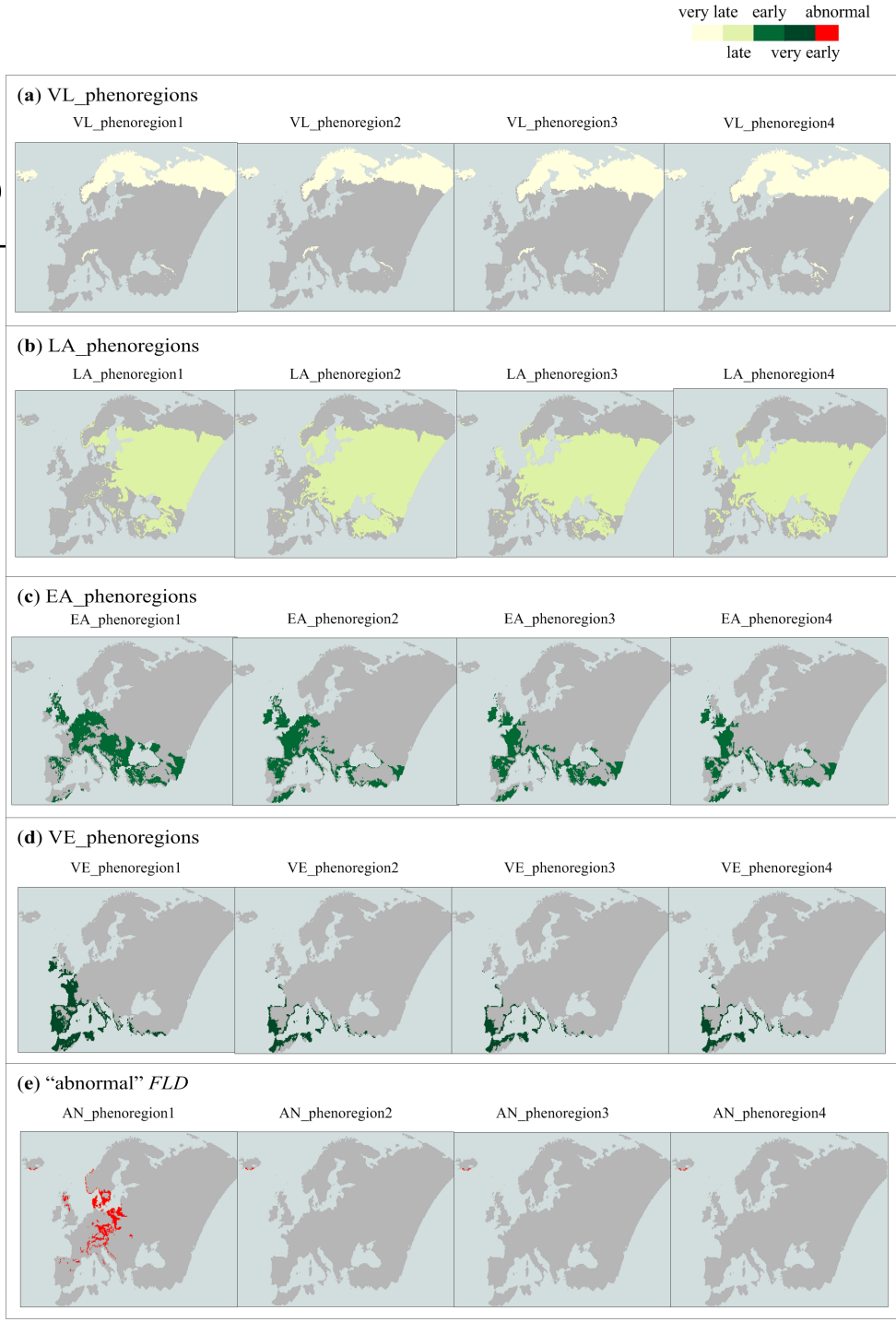


REFINED CO-CLUSTERS

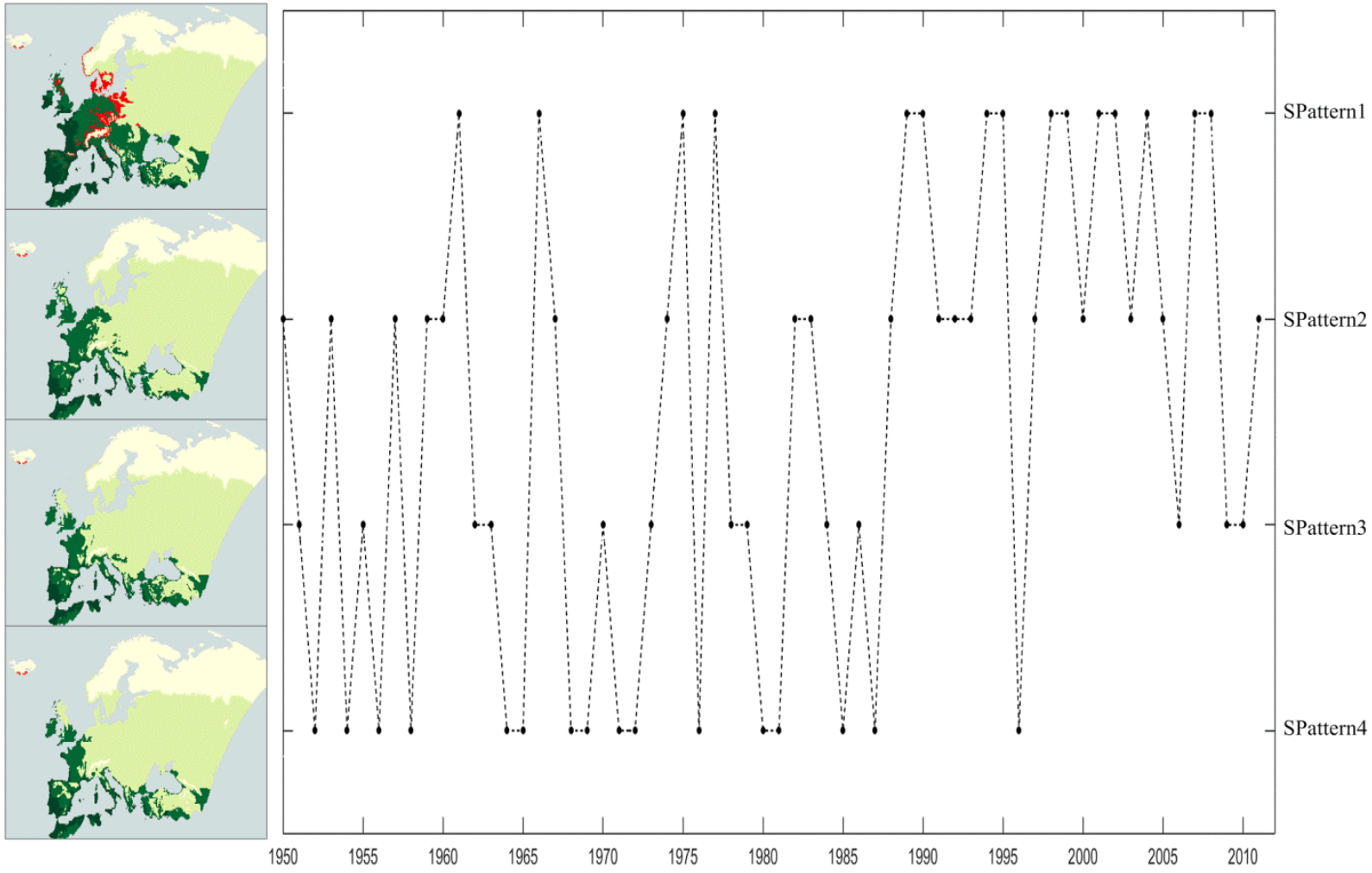


LOCATION CO-CLUSTERS

- Phenoregions
- 4 temporal clusters
- 5 spatial categories

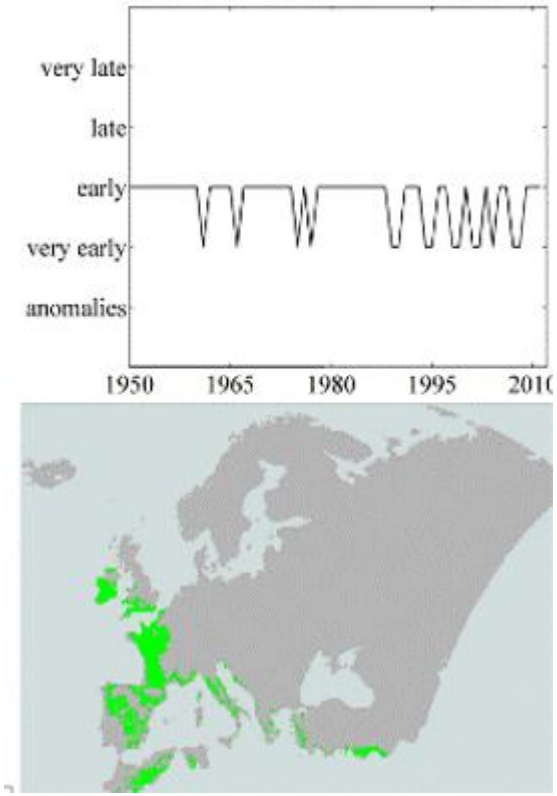


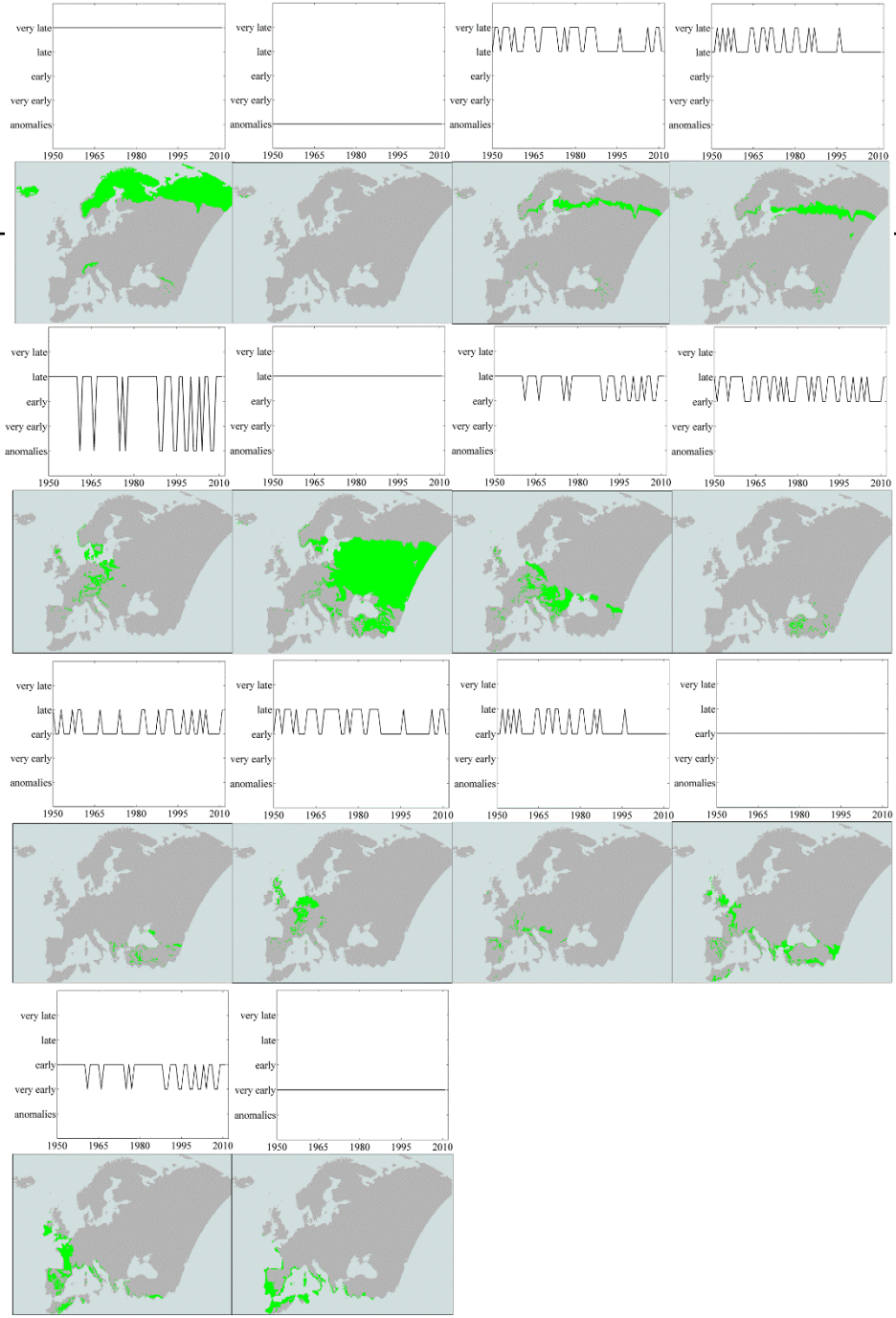
SPATIAL PATTERNS AND THEIR TIMELINE



https://doi.org/10.1002/2015JG003308

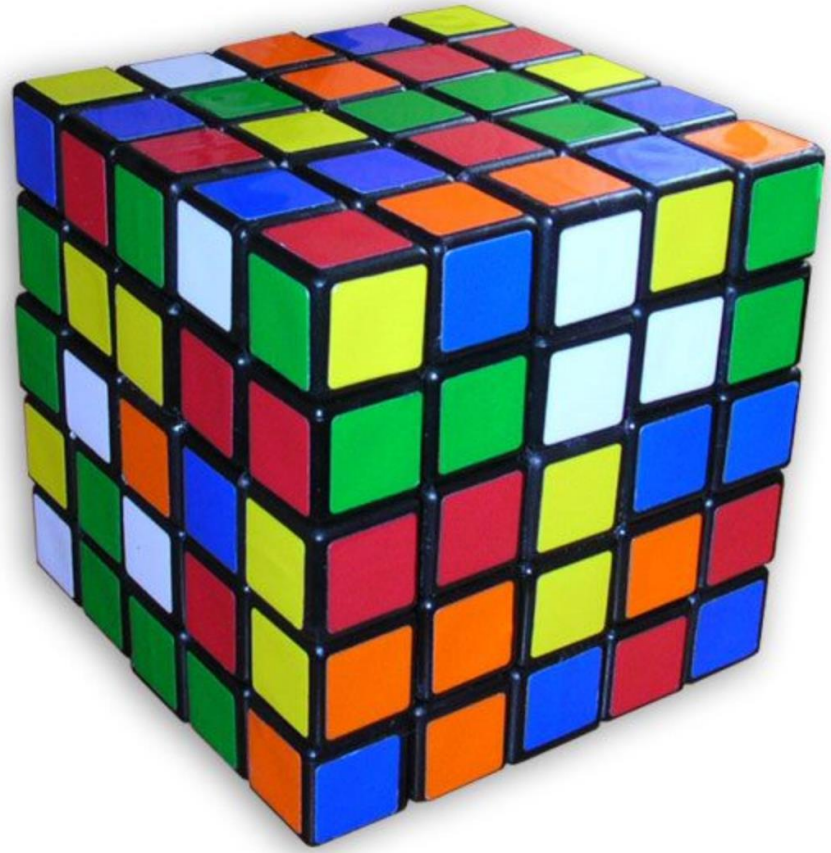
STABLE/CHANGEABLE





UNIVERSITY OF TWENTE.

TRI-CLUSTERING



Data cubes

Georeferenced time series

CGC PACKAGE

github.com/phenology/cgc

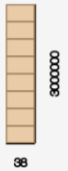
Clustering **G**eo-Data **C**ubes

- Co-clustering
- Tri-clustering
- K-means refinement

```
from cgc.coclustering import Coclustering
```

```
matrix
```

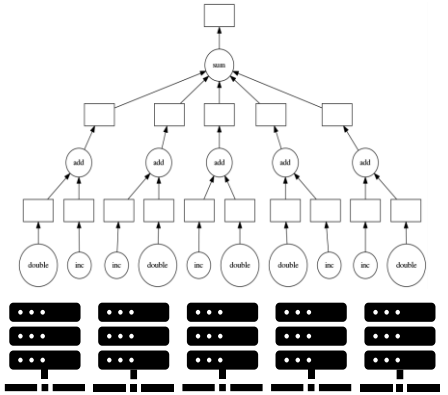
	Array	Chunk
Bytes	912.00 MB	114.00 MB
Shape	(3000000, 38)	(375000, 38)
Count	9 Tasks	8 Chunks
Type	float64	numpy.ndarray




```
cc = Coclustering(matrix, nclusters_row=30, nclusters_col=5, max_iterations=100)
```

```
cc.run_with_dask(client, low_memory=True)
```

```
2020-08-27 23:01:18,032 INFO  - Run 0 ..
2020-08-27 23:01:19,471 DEBUG - Iteration # 0 ..
2020-08-27 23:01:26,916 DEBUG - Error = -4.481409657333588e+10, dE = -4.481409657333590e+10
2020-08-27 23:01:26,917 DEBUG - Iteration # 1 ..
2020-08-27 23:01:39,800 DEBUG - Error = -4.490170942544089e+10, dE = -8.761285210501099e+07
2020-08-27 23:01:39.801 DEBUG - Iteration # 2 ..
```



The Journal of Open Source Software

CGC: a Scalable Python Package for Co- and Tri-Clustering of Geodata Cubes

Francesco Nattino[1], Ou Ku[1], Meiert W. Grootes[1], Emma Izquierdo-Verdiguier[2], Serkan Girgin[3], and Raul Zurita-Milla[3]

1 Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam, The Netherlands **2** Institute of Geomatics, University of Natural Resources and Life Science (BOKU), 1190, Vienna, Austria **3** Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, PO Box 217, 7500 AE, Enschede, the Netherlands

DOI: 10.21105/joss.04032

Software

- Review
- Repository
- Archive

Editor: Hugo Ledoux

Reviewers:

- @Subho07
- @Narayana-Rao

Submitted: 17 December 2021
Published: 10 April 2022

Summary

Multidimensional data cubes are increasingly ubiquitous, in particular in the geosciences. Clustering techniques encompassing their full dimensionality are necessary to identify patterns "hidden" within these cubes. Clustering Geodata Cubes (CGC) is a Python package designed for partitional clustering, which identifies groups of similar data across two (e.g., spatial and temporal) or three (e.g., spatial, temporal, and thematic) dimensions. CGC provides efficient and scalable co- and tri-clustering functionality appropriate to analyze both small and large datasets as well as a cluster refinement functionality that supports users in their quest to make sense of complex datasets.

Introduction

Faced with the increasing ubiquity of large datasets, data mining techniques have become essential to extracting patterns and generating insights. In this regard, clustering techniques, which aim to identify groups or subgroups with similar properties within a larger dataset, are becoming ever more popular.

Traditional clustering techniques focus on a single dimension and may therefore obfuscate relevant groups (Cheng & Church, 2000; Hartigan, 1972). Hence, clustering techniques capable

SUMMARY & OUTLOOK

- CGC helps to ”mine” phenological patterns
- Not only data cubes!
  - E.g., years when species have similar phenology
  - Open invitation to collaborate

- CGC package is still in development (and testing)
  - Currently analysing high spatial resolution versions of the SI-x over Europe and USA

r.zurita-milla@utwente.nl