MCM10 / Final Event, 22-24 Nov 2010, Vienna, Austria

Application of Affinity Propagation Clustering on Meteorological Data

Athanassios Zagouras, PhD student

Authors: Zagouras A., Argiriou A.A., Lykoudis S., Economou G. and Fotopoulos S.

Electronics Laboratory
Department of Physics
University of Patras
Greece

Affinity Propagation

- Brendan J. Frey and Delbert Dueck, *'Clustering by Passing Messages Between Data Points'*, Science 315, 972–976, 2007)
- How it works?
  - can be viewed as exchanging messages between the data points themselves
  - all data points are simultaneously considered as exemplars, but exchange deterministic messages until a good set of exemplars gradually emerges

Algorithm steps

- Input: similarity matrix of N data points, S_{NxN}, where the diagonal of the matrix is the preferences.
  - s(i,i), is the a priori suitability of point i to serve as an exemplar
- Output: exemplar of each data point.
- Initialization: $r(i, k) = 0, a(k, i) = 0 \quad \text{for all } i, k$

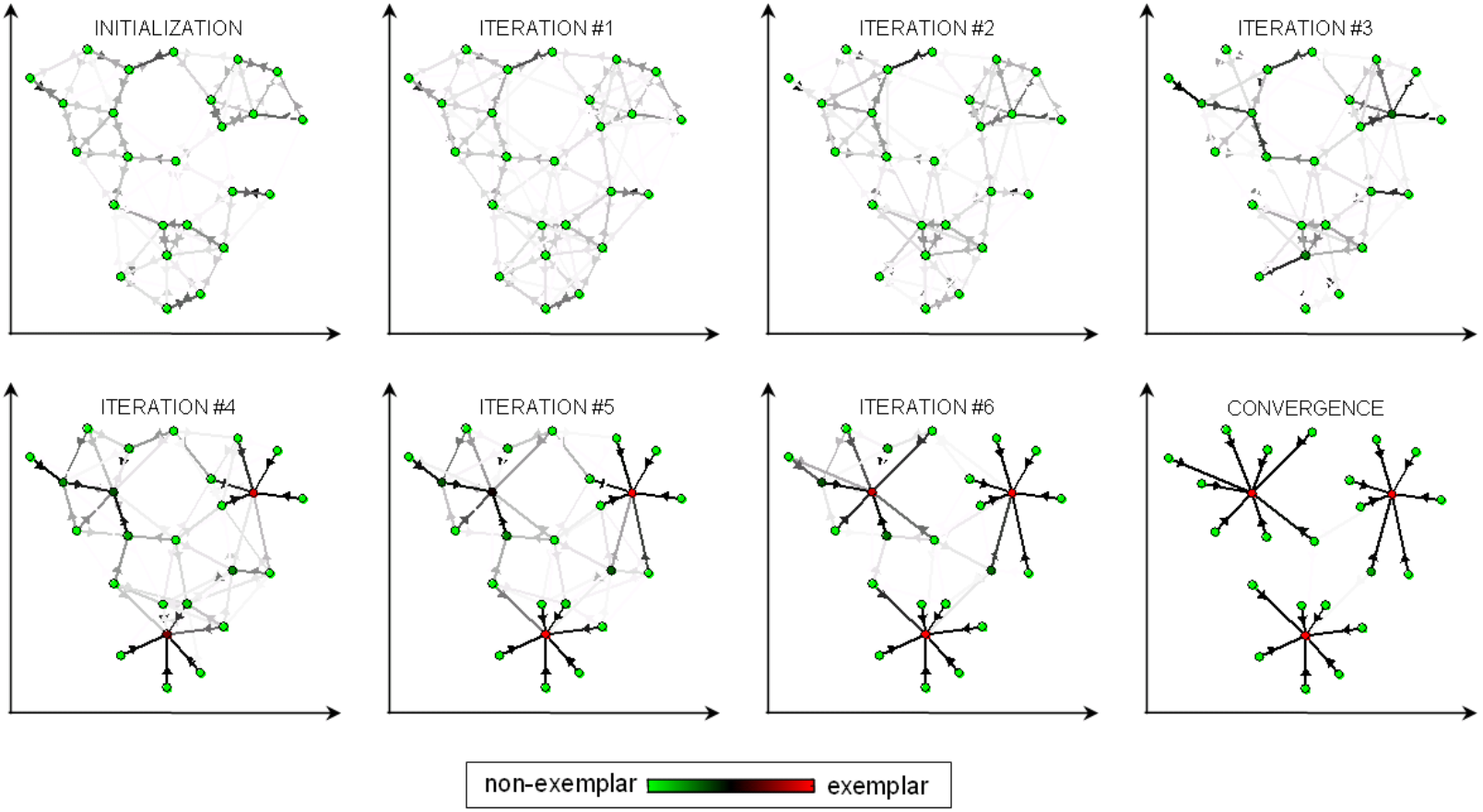
Steps:

- Updating all responsibilities r(i,k): $r(i,k) \leftarrow s(i,k) - \max_{j:j\neq k}(a(j,i) + s(i,j))$
- Updating all availabilities a(i,k): $a(k,i) \leftarrow \min\left(0, r(k,k) + \sum_{j:j\notin\{k,i\}} \max\{0, r(j,k)\}\right)$

  $a(k,k) \leftarrow \sum_{j:j\neq k} \max\{0, r(j,k)\}$
- Making decisions:

$$c_i^* \leftarrow \arg\max_k r(i,k) + a(k,i)$$

Example


INITIALIZATION
ITERATION #1
ITERATION #2
ITERATION #3
ITERATION #4
ITERATION #5
ITERATION #6
CONVERGENCE
non-exemplar
exemplar

Datasets & Methods

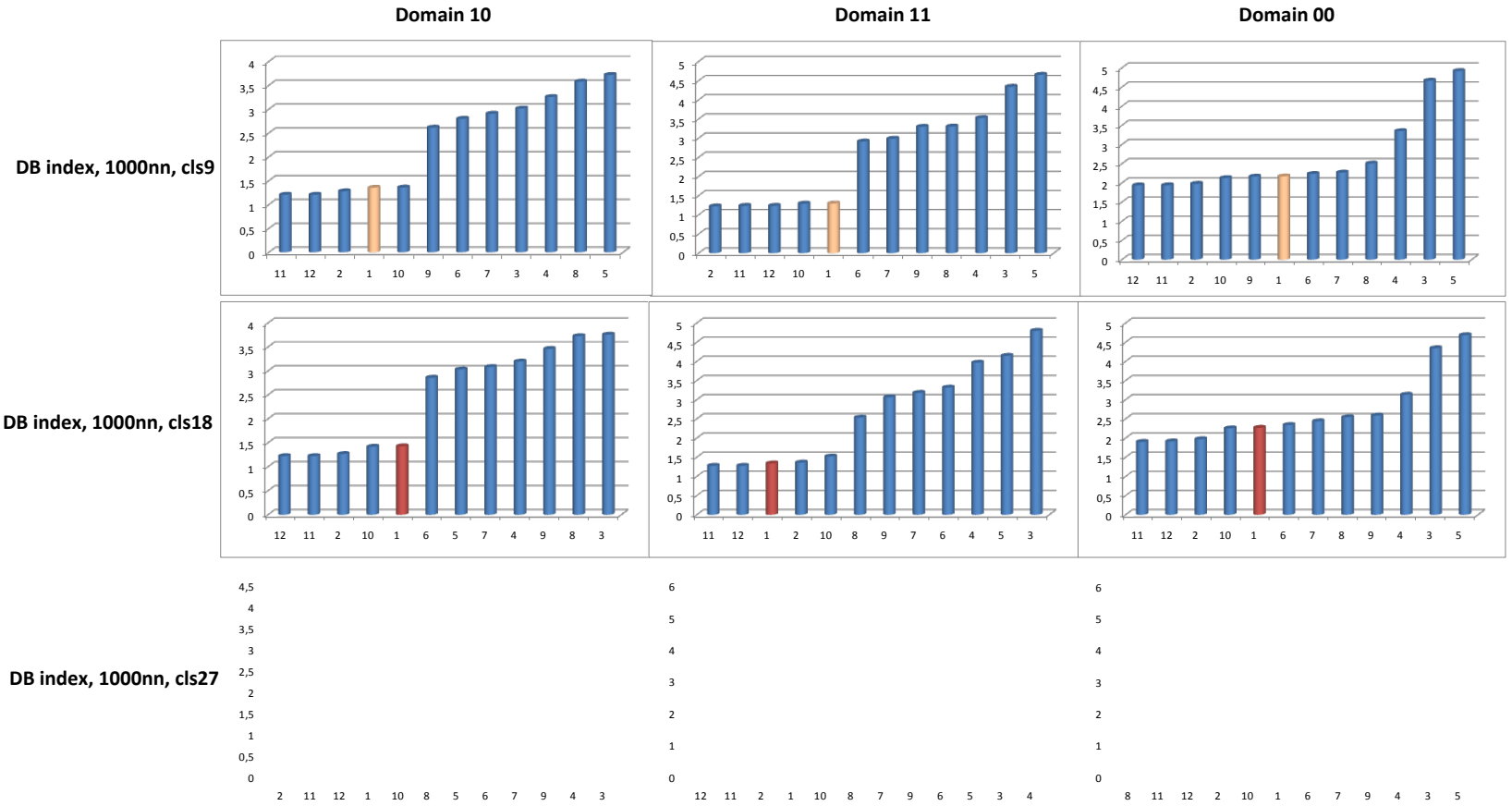
- Cost733 domain00,domain10,domain11
- Parameter: MSLP
- Period: 1-9-1957 to 31-8-2002
- Number of classes: 9,18,27
- Methods: AP,CKM,KIR,KRZ,LIT,LND,PCT,PTT,PXK,RAC, SAN,SOM
- Evaluation: 2mT, CP+LSP

Processing & indices

- Concatenate grid points to high dimensional vectors
- Step1: Vector quantization
  - sparse distance matrix of [500,1000,1500,3000] nearest neighbors
  - ~300-1000 classes
- Step2: user-specified number of clusters
  - heuristic version of AP
  - classification into 9,18,27 classes
- Validity indices
  - Davies-Bouldin index: ratio of the sum of within-cluster scatter to between-cluster separation
  - EV, PF, ECV, WSD, FSIL

$$DB = \frac{1}{n}\sum_{i=1, i\neq j}^{n} \max\left[\frac{d_i + d_j}{d(C_i + C_j)}\right]$$

DBi results



AP	1
CKM	2
KIR	3
KRZ	4
LIT	5
LND	6
PCT	7
PTT	8
PXK	9
RAC	10
SAN	11
SOM	12

AP performance tests

- Rand index: number of pairwise agreements between clusterings

	Rand Index	domain 00 - nn			
		500	1000	1500	3000
run1-2 / classes	9	0.781	0.795	0.803	0.803
	18	0.878	0.864	0.888	0.892
	27	0.910	0.907	0.919	0.929

Consecutive runs

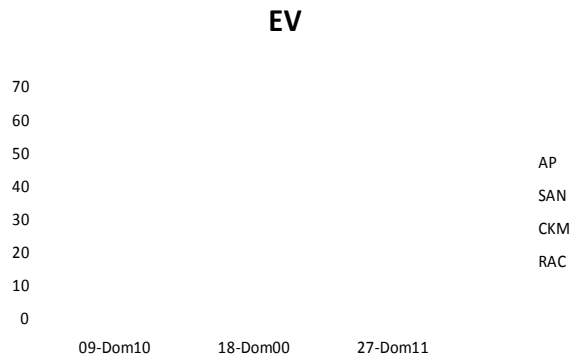
	Rand Index	domain 11 - nn			
		500	1000	1500	3000
run1-2 / classes	9	0.886	0.852	0.849	0.910
	18	0.917	0.931	0.914	0.930
	27	0.934	0.929	0.926	0.933

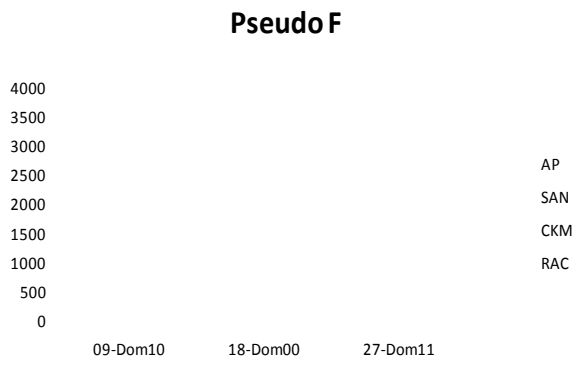
	Rand Index	domain 00 - nn			
		500	1000	1500	3000
1000nn / classes	9	0.780	1.000	0.817	0.817
	18	0.860	1.000	0.872	0.876
	27	0.908	1.000	0.913	0.919

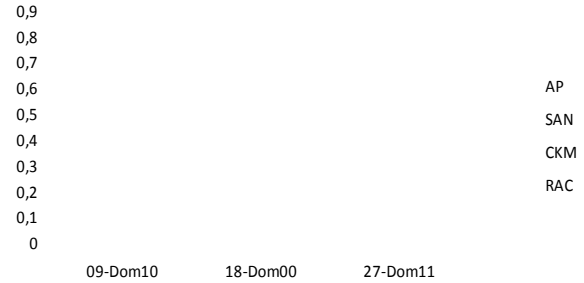
1000 nearest neighbors

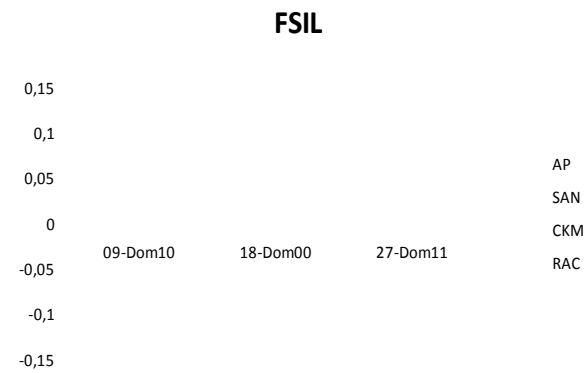
	Rand Index	domain 11 - nn			
		500	1000	1500	3000
1000nn / classes	9	0.862	1.000	0.856	0.859
	18	0.913	1.000	0.926	0.924
	27	0.942	1.000	0.937	0.931

AP vs SAN-CKM-RAC [MSLP]







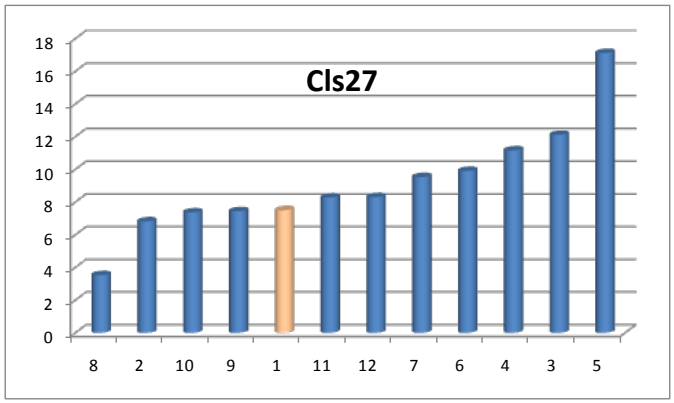


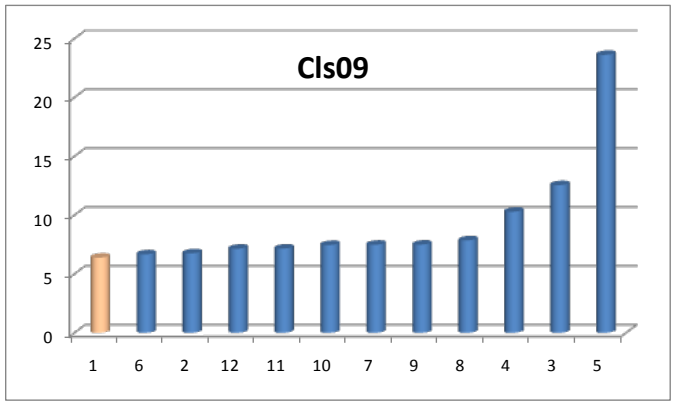
Evaluation tests: 2mT, CP+LSP

Domain 10

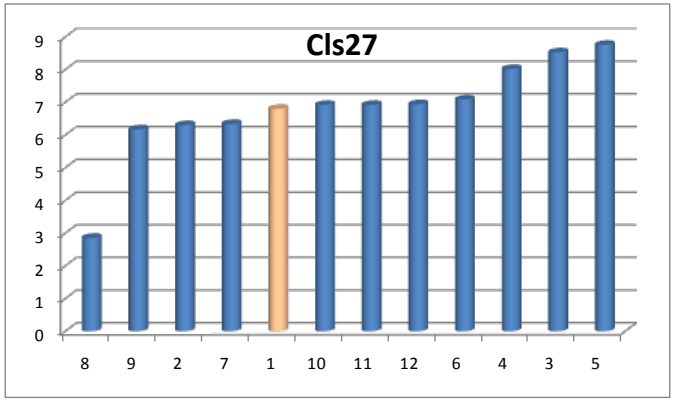
Domain 00

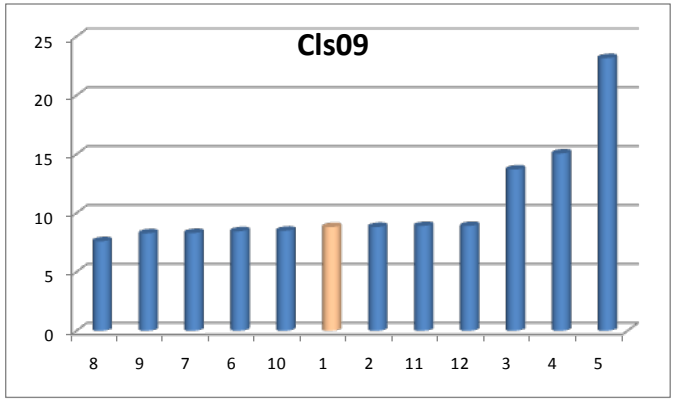
2mT, DB index, 1000nn





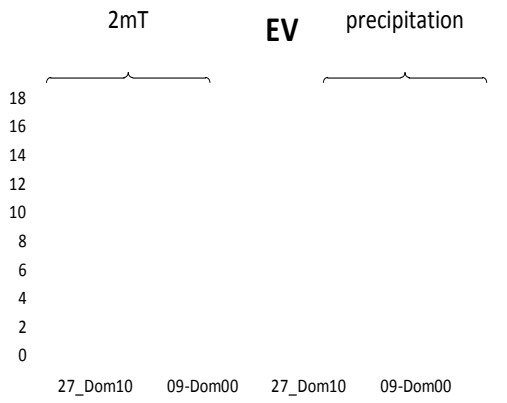
CP+LSP, DB index, 1000nn

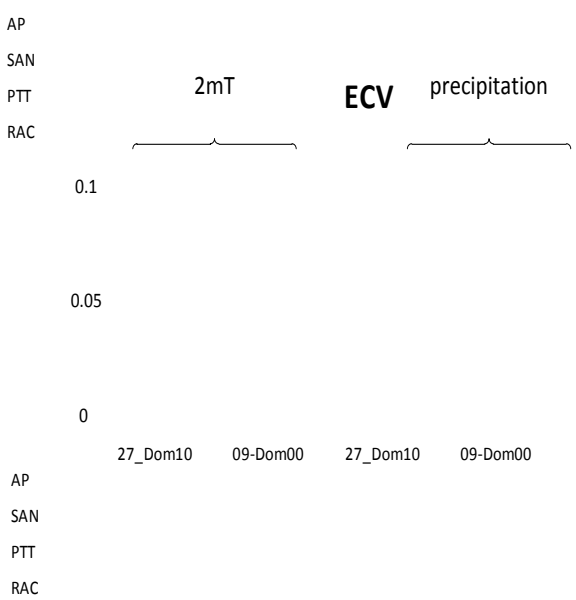


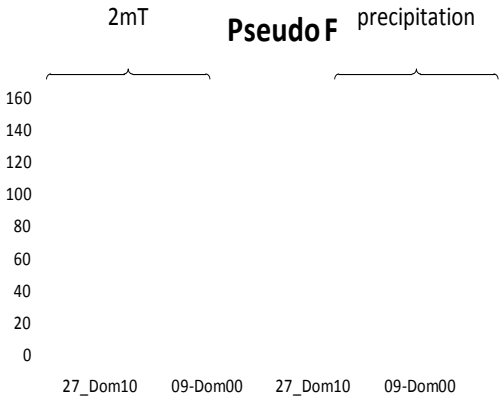


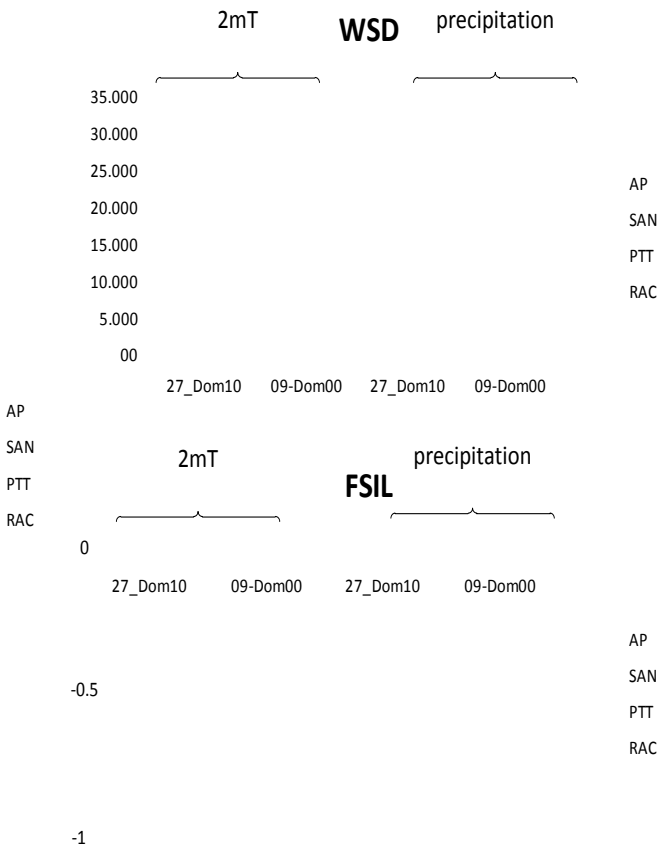
AP	1
CKM	2
KIR	3
KRZ	4
LIT	5
LND	6
PCT	7
PTT	8
PXK	9
RAC	10
SAN	11
SOM	12

Evaluation tests: 2mT, CP+LSP

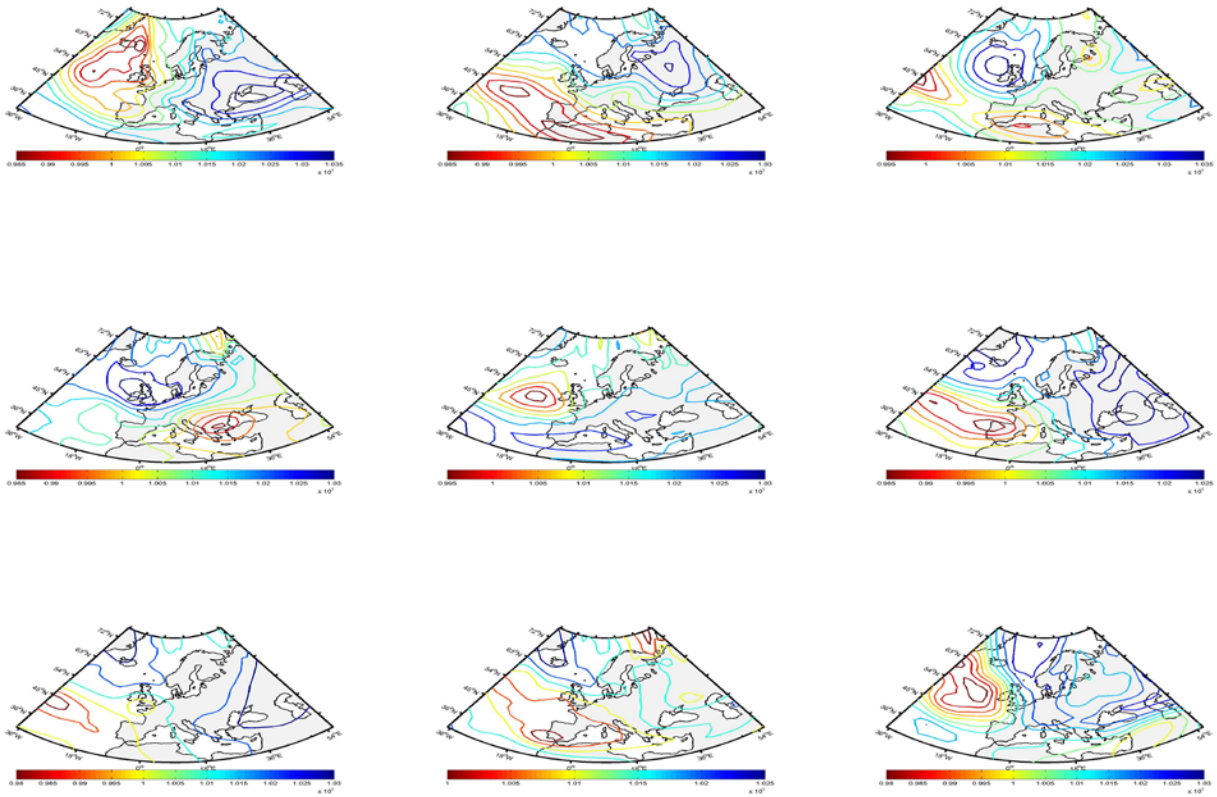




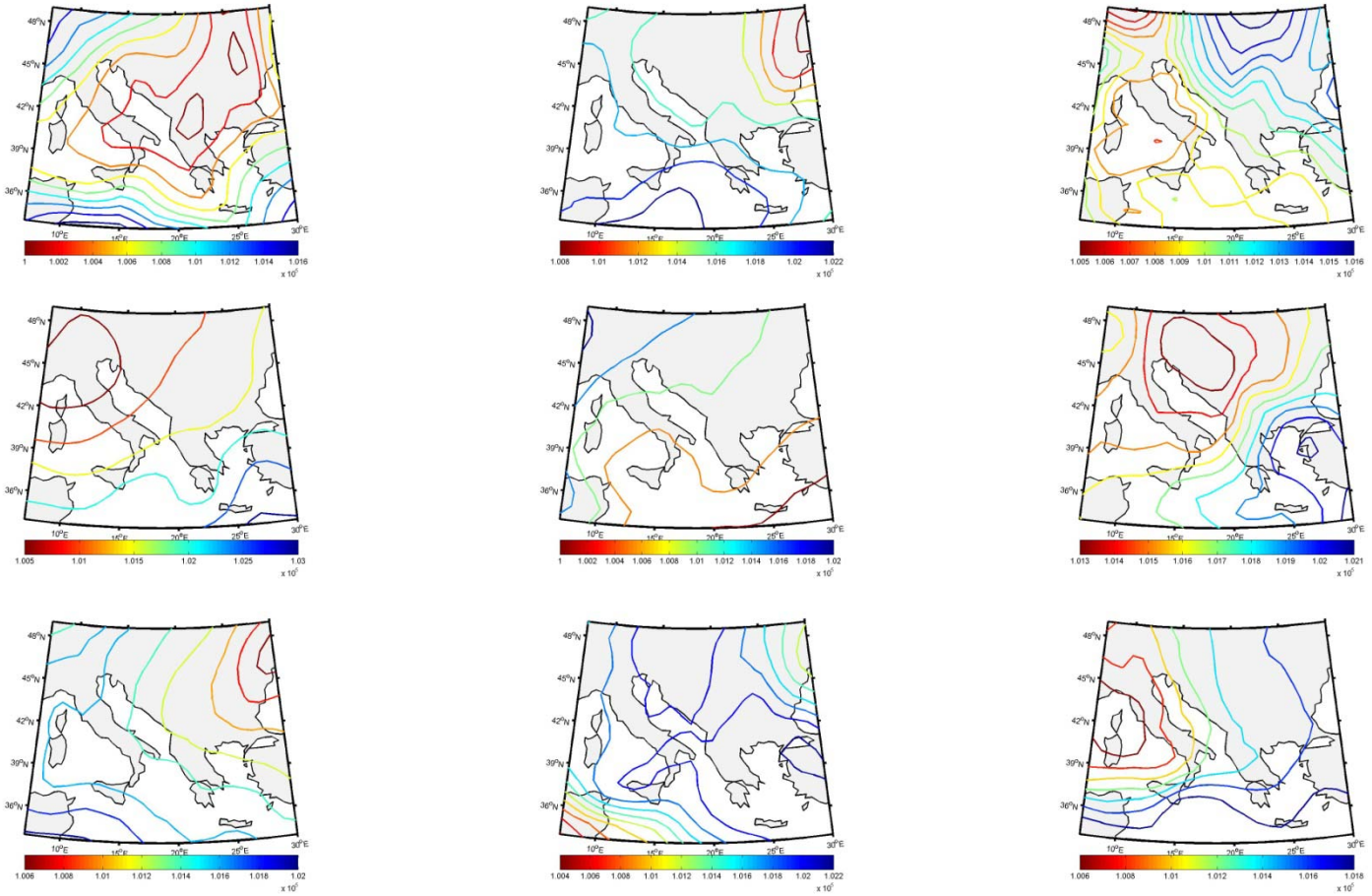




Cluster centroids (9cls-dom00)

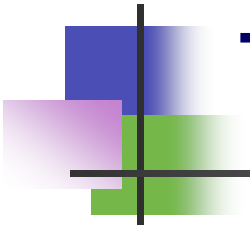


Cluster centroids (9cls-dom10)



Conclusions

- Good performance, close to best Cost733 methods
- Small domain + 27 classes: better results
- Evaluation (good perfomance)
- Preferences investigation
- 1000 runs
- Cluster stability investigation (derived from RI)
- Actual points (centroids)
- Computational time !



Thank you

- Thank you for your attention