

Fabio D'Andrea

LMD – 4^e étage

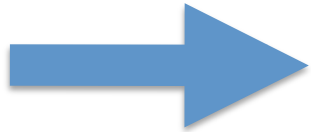
01 44 32 22 31

dandrea@lmd.ens.fr

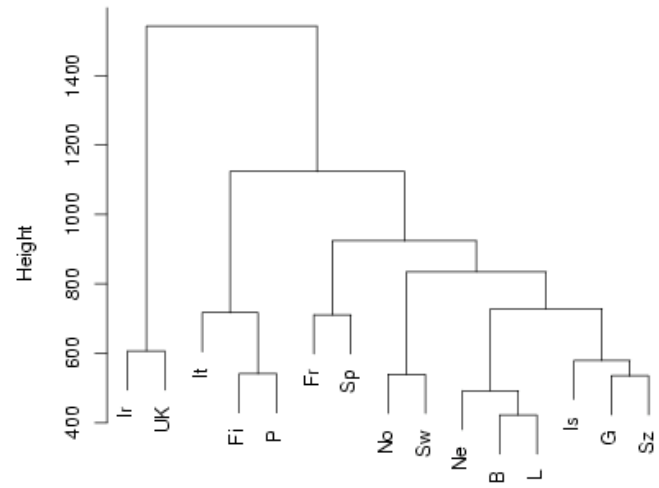
<http://www.lmd.ens.fr/dandrea/TEACH>

Program

13/1	Elementary statistics - 1
20/1	Elementary statistics - 2
27/1	Exercises – Computer room
10/2	Fourier Analysis - 1
17/2	Fourier Analysis - 2, stochastic processes
24/2	Exercises – Computer room
2/3	Exercises – Computer room
9/3	Principal component analysis - 1
16/3	Principal component analysis - 2
23/3	Exercises – Computer room
13/4	Exercises – Computer room
27/4	Cluster analysis, Neural networks
4/5	Exercises – Computer room
25/5	Principal component analysis: Complements
6/6	Exam.



. Lesson 3. Cluster Analysis



Cluster analysis or clustering is the assignment of a set of observations into subsets (the *clusters*) so that observations in the same cluster are similar in some sense.



Robert Tryon

Found online:

Pang-Ning Tan, Michael Steinbach, Vipin Kumar Addison-Wesley Introduction to Data Mining, 2005. ISBN : 0321321367.

<http://www-users.cs.umn.edu/~kumar/dmbook/index.php> - FREE!!

Data Mining. - What is it?

It is the computational process of analysing large amount of data in order to discover patterns, and summarizing it into useful information for further use.

It uses methods at the intersection of statistics, computer science, artificial intelligence.

Applications: Business and marketing, Science and engineering (genetics, electric network, traffic, , Medical sciences and policy, GIS, Visual data analysis, security/surveillance,

With the deveopment of data gathering and storing techniques, the limiting factor is using information, not obtaining it.

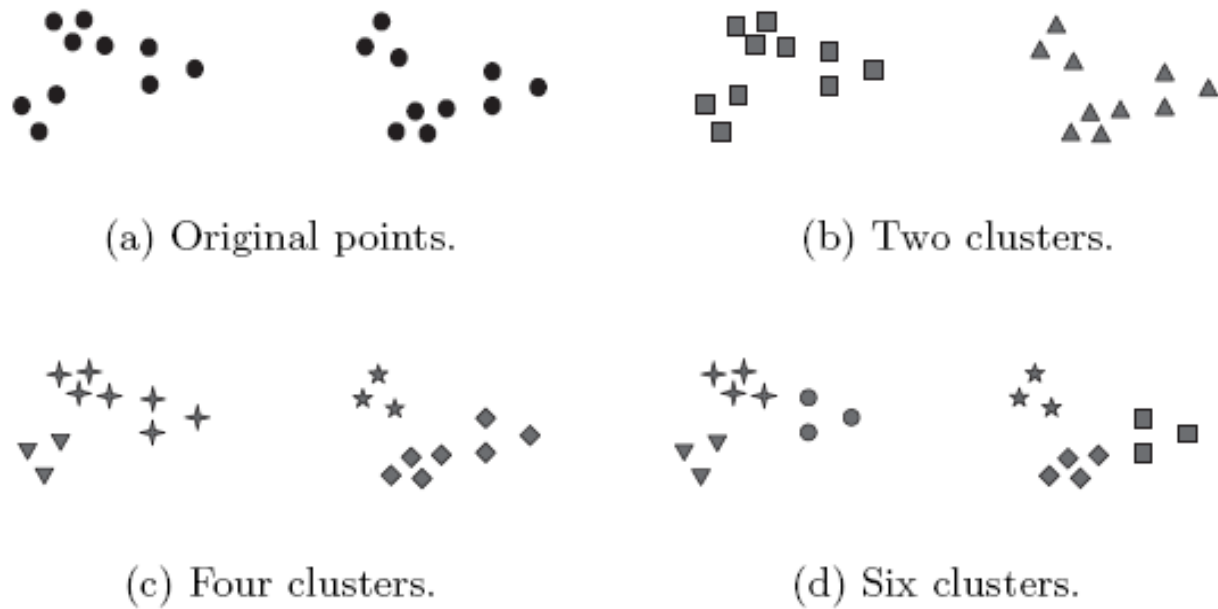
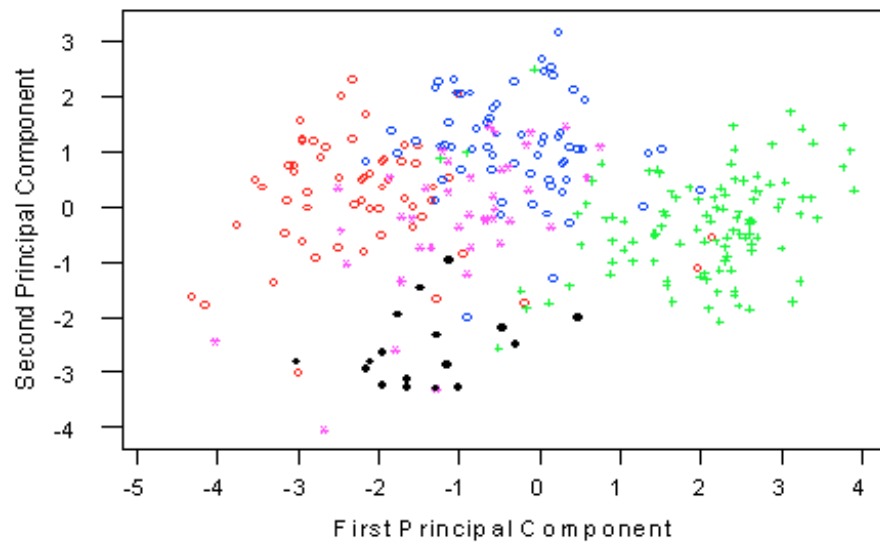
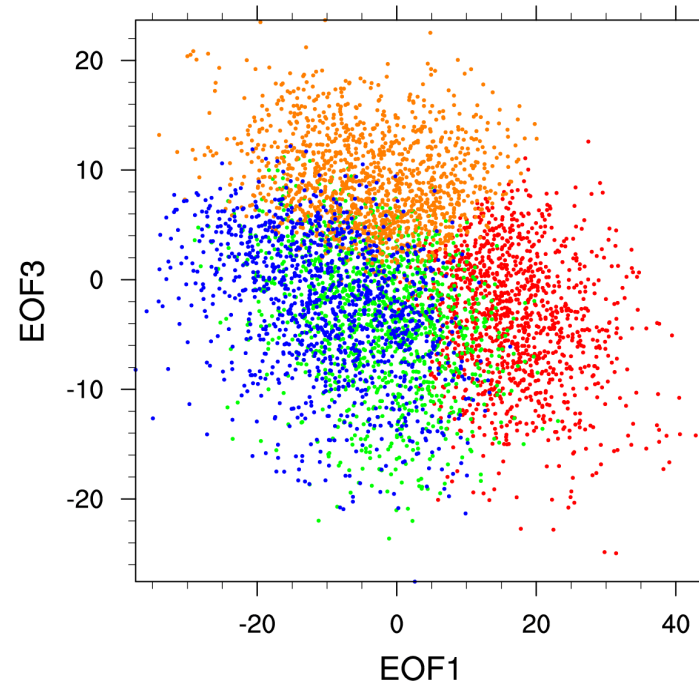
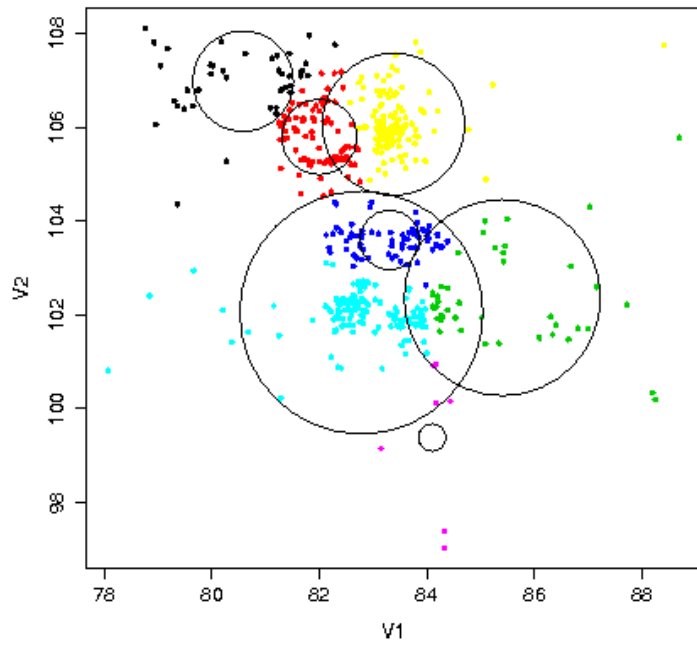


Figure 8.1. Different ways of clustering the same set of points.



- Bede
- + Alfred
- GD
- * OR
- PP

Examples

First step is the choice of a distance function in the space of the observations

Euclidean distance $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}$

Other distances:

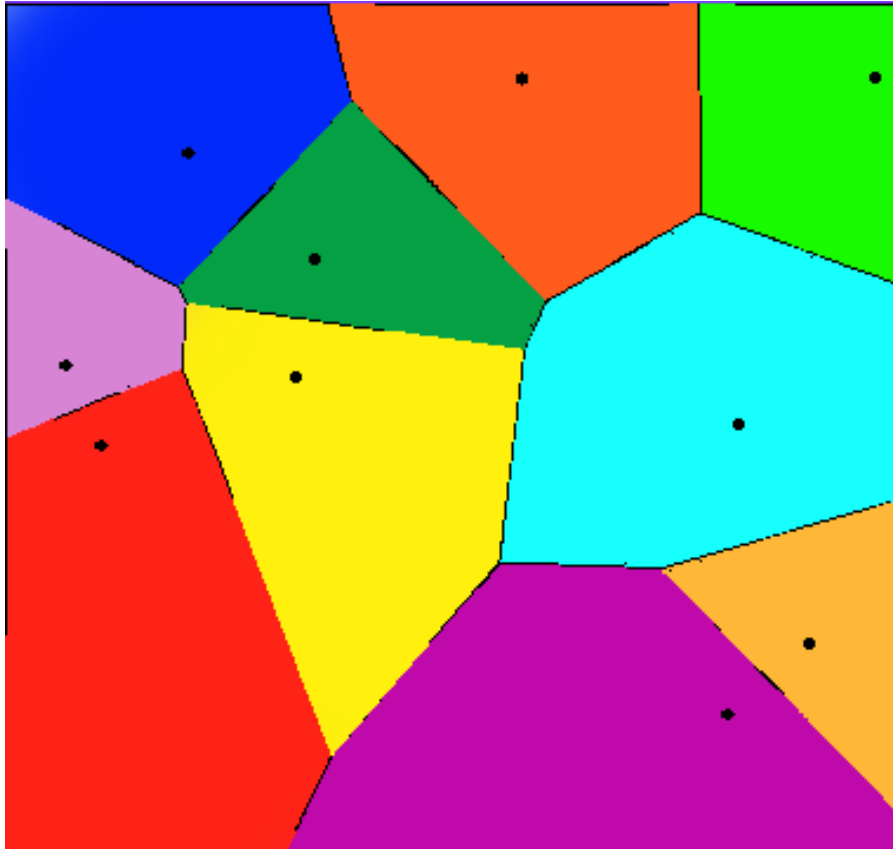
Taxi distance $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^N |x_i - y_i|}$

Cosine similarity $d(\mathbf{x}, \mathbf{y}) = 1 - \cos(\mathbf{x}, \mathbf{y})$

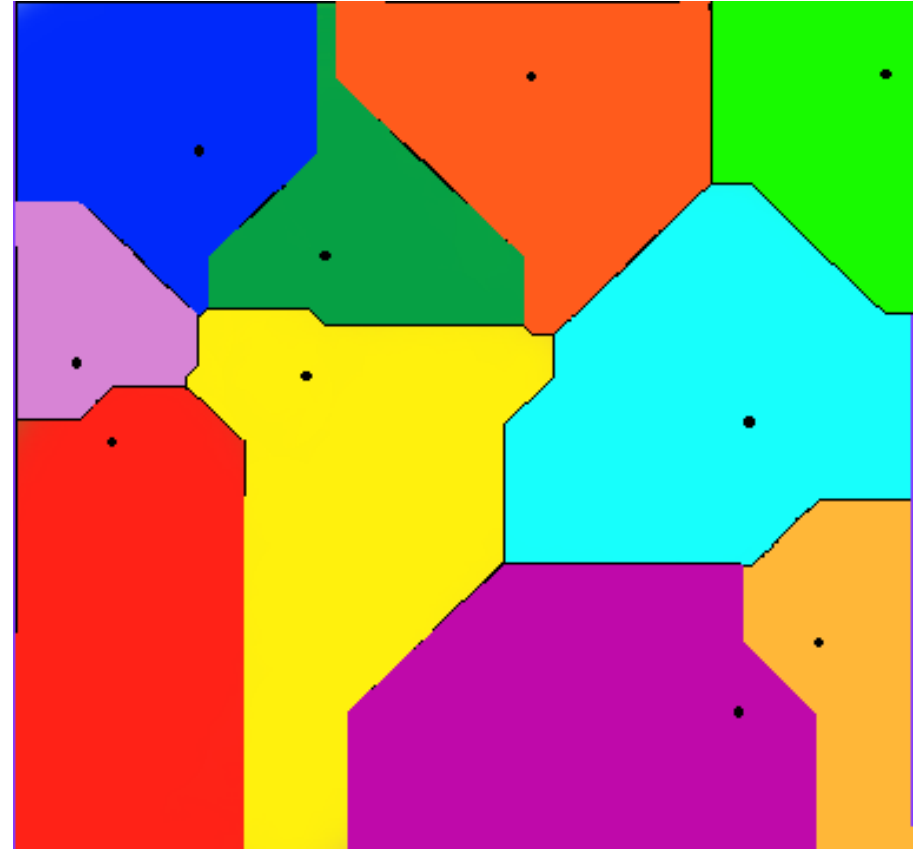
Reminder: the distance follows the choice of a norm in the vector space of observations:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

10 Bakeries in a city map. Which one is the nearest?



10 shops in a flat city and their Voronoi cell (Euclidean Distance)



The same 10 shops, now under the Manhattan Distance. This shows that the Voronoi cells depend significantly on the metric used.

A “statistically interesting” norm (Mahalanobis norm):

$$\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{C}^{-1} \mathbf{x} \rangle$$

If \mathbf{C} is diagonal, the distance becomes:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^N \frac{(x_i - y_i)^2}{\sigma_i^2}}$$

Which means that each component is normalized by its own variance. It is useful in case of vectors of heterogeneous observations.

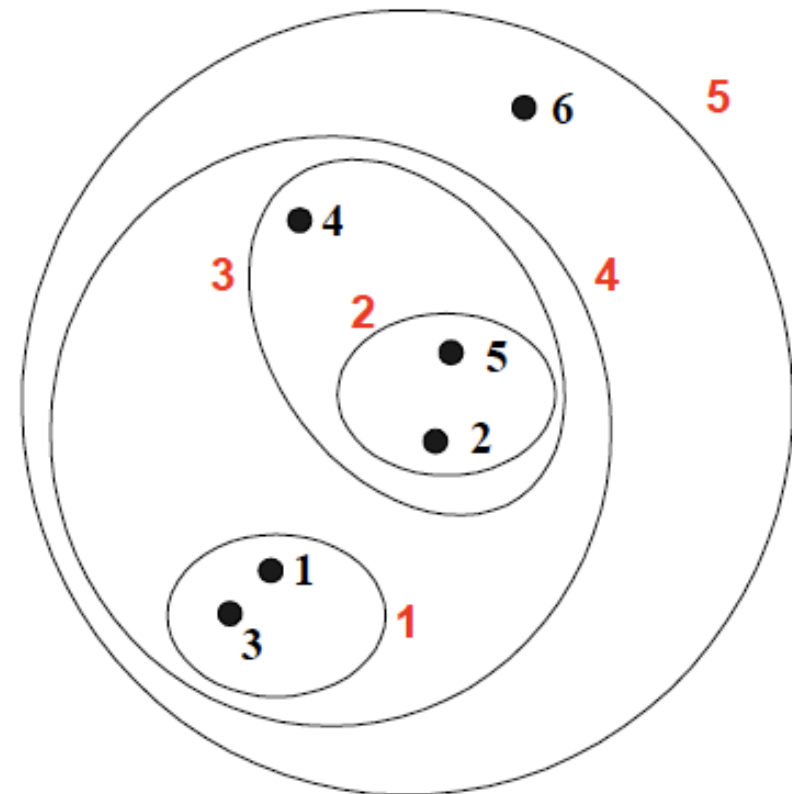
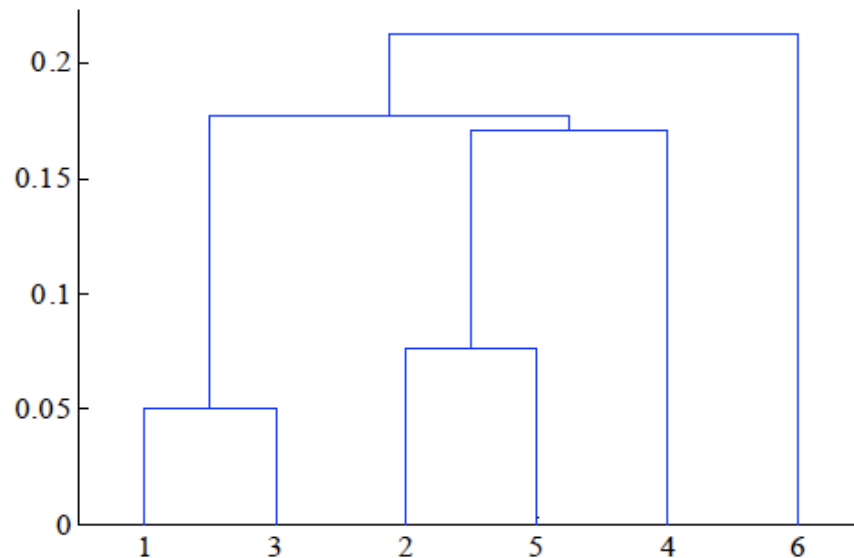
Cluster Analysis

There are two main kind of clustering algorithms:
Hierarchical and Non-Hierarchical.

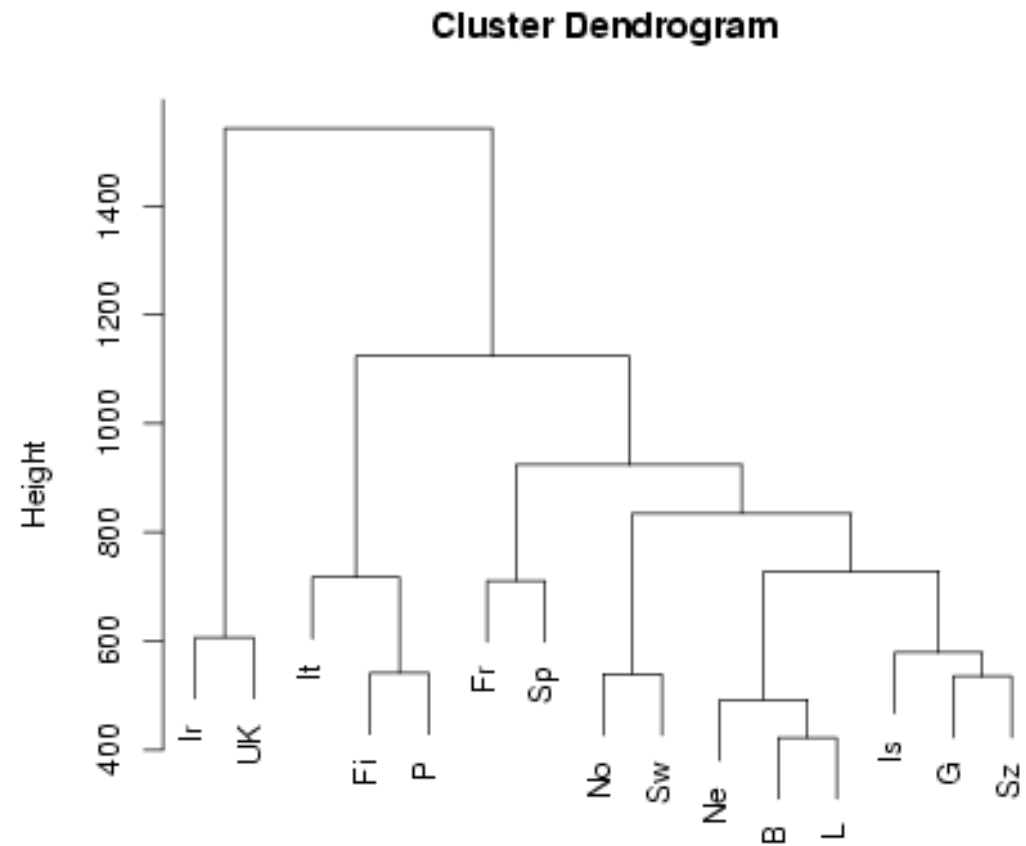
We will see one algorithm per type, with examples.

Hierarchical clustering

1. Compute the distance of each point to each point (proximity matrix)
2. Let each data point be a cluster
- 3. Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
- 6. Until** only a single cluster remains



Example 1
Judges voting at the "Eurovision"



<http://www.admin.ox.ac.uk/po/Eurovision.shtml>

```
as.dist(equiv.dist)  
hclust (*, "complete")
```

The Eurovision Song Contest Is Voting Political or Cultural?¹

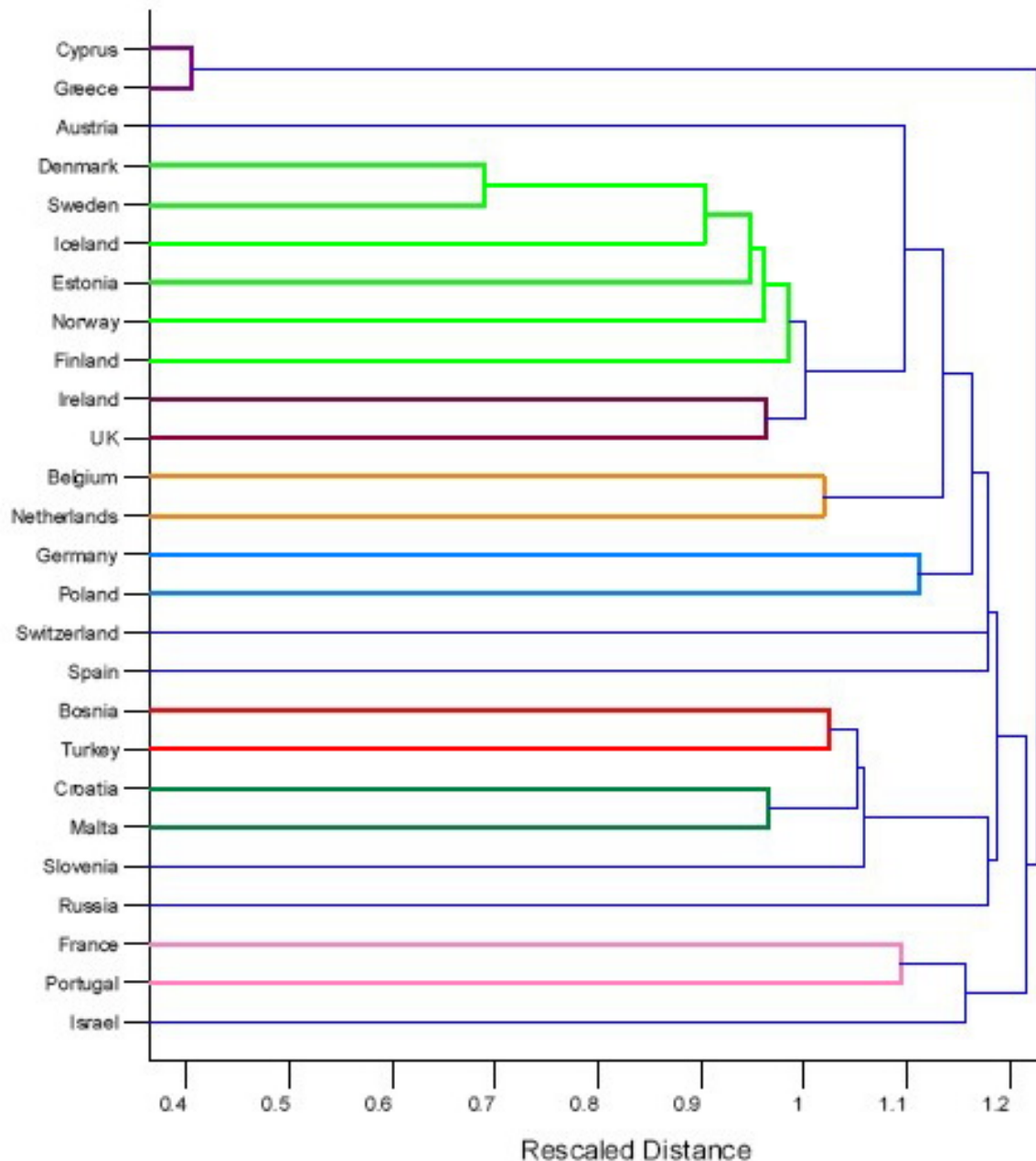
Victor Ginsburgh
ECARES, Université Libre de Bruxelles and
Center for Operations Research and Econometrics, Louvain-la-Neuve

Abdul Noury
ECARES, Université Libre de Bruxelles

November 2004
Revised October 2006

Abstract

We analyze the voting behavior and ratings of judges in a popular song contest held every year in Europe since 1956. The dataset makes it possible to analyze the determinants of success, and gives a rare opportunity to run a direct test of vote trading. Though the votes cast may appear as resulting from such trading, we show that they are rather driven by quality of the participants as well as by linguistic and cultural proximities between singers and voting countries. Therefore, and contrary to what was recently suggested, there seems to be no reason to take the result of the Contest as mimicking the political conflicts (and friendships).



Ginsburgh and Noury,: The Eurovision Song Contest. Is voting political or cultural? European Journal of Political Economy. 2008, p.41-52

Example: Remote sensing of vegetal species by airborne lidar:

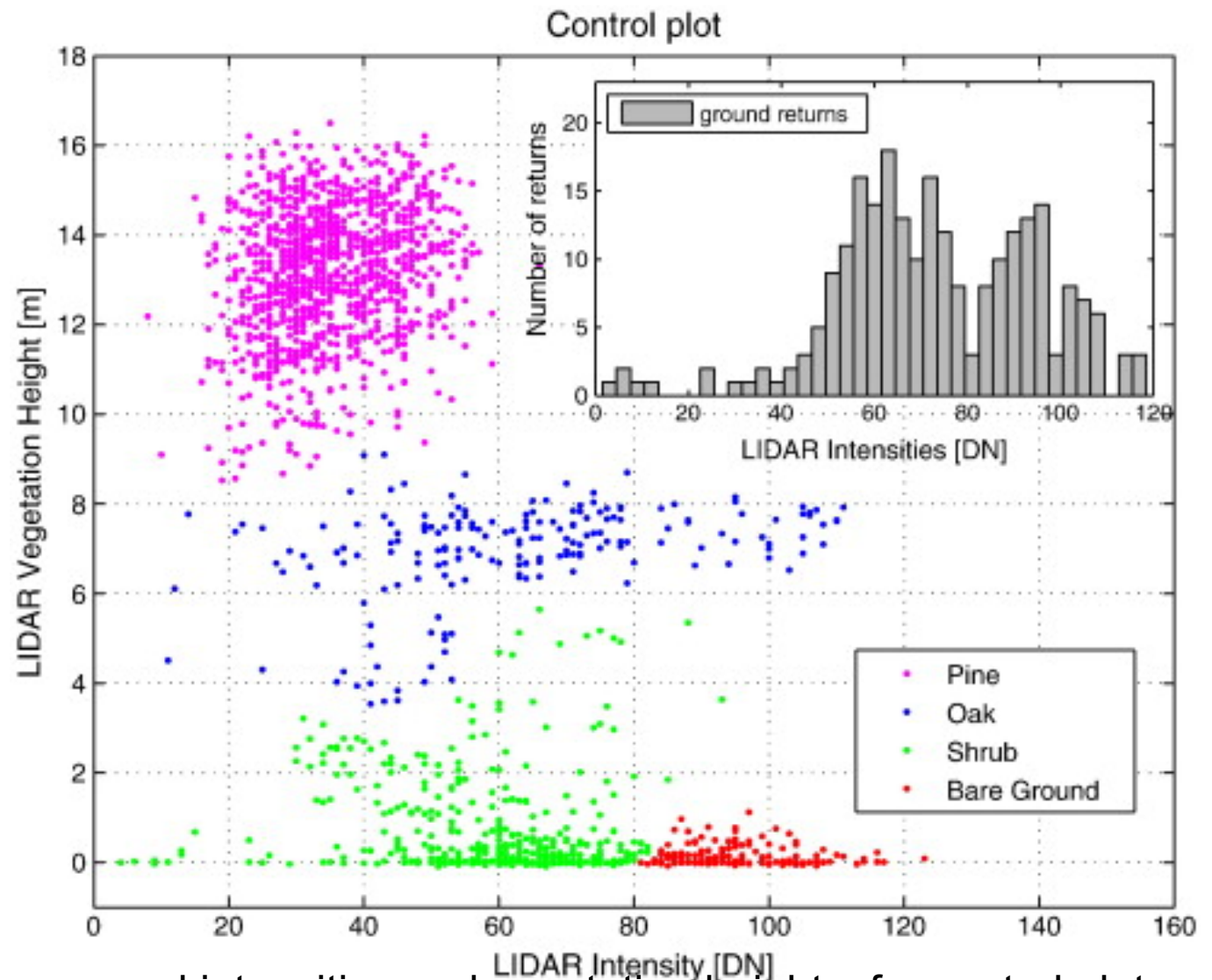
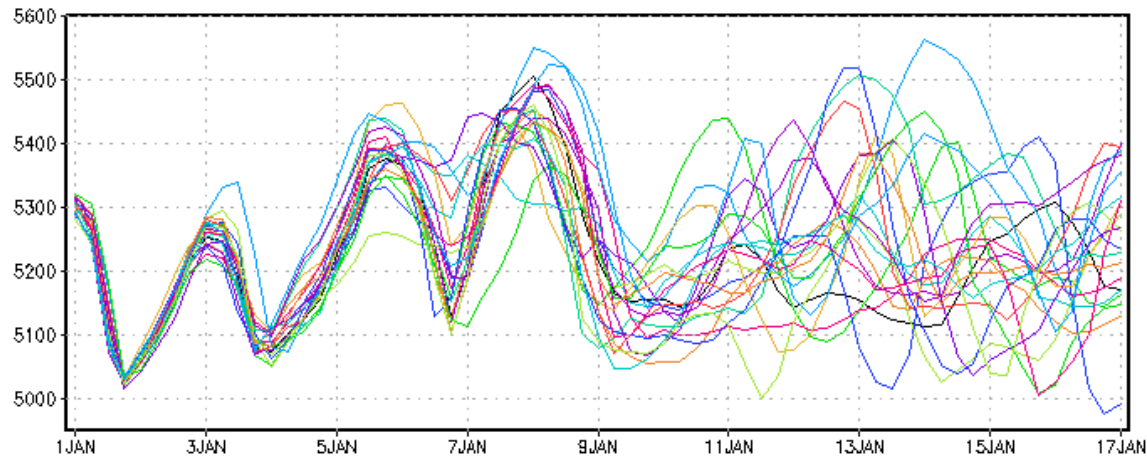
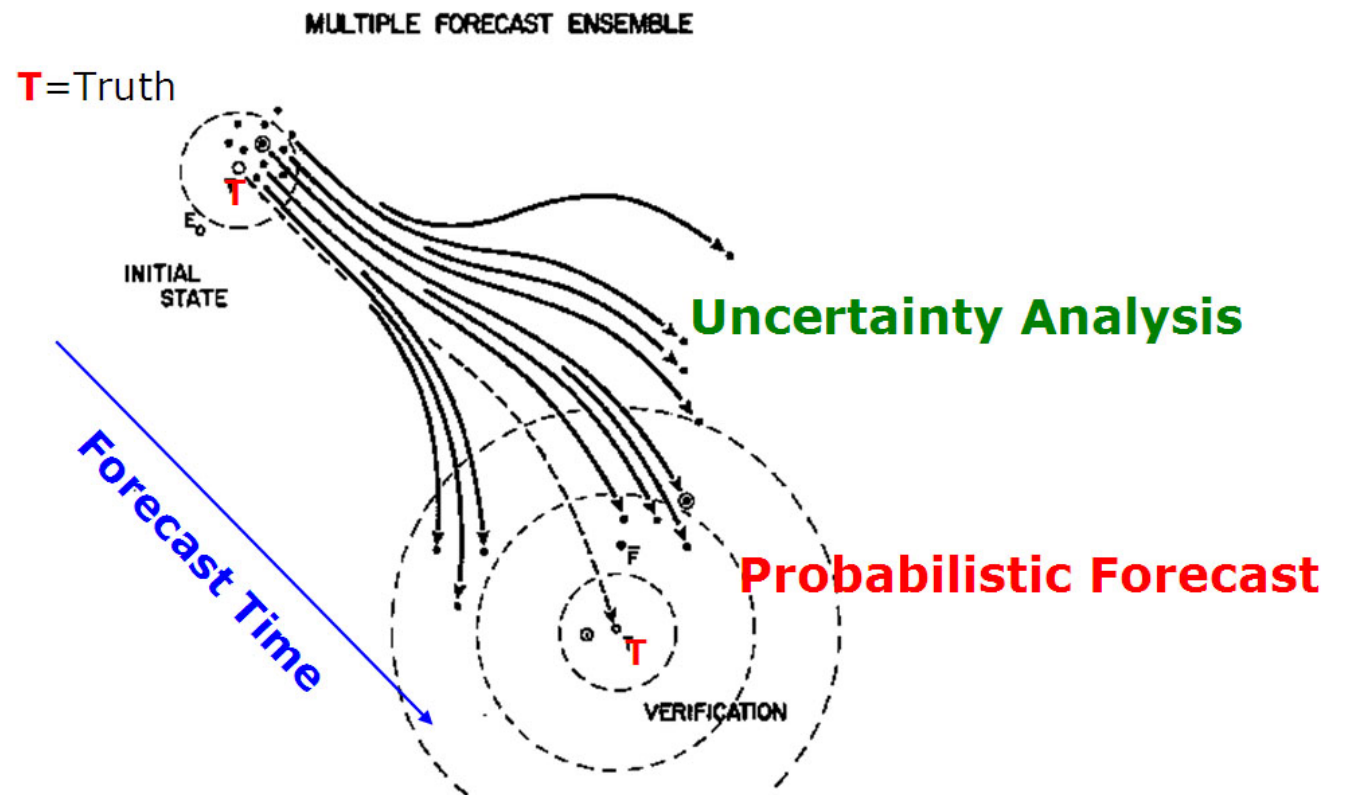


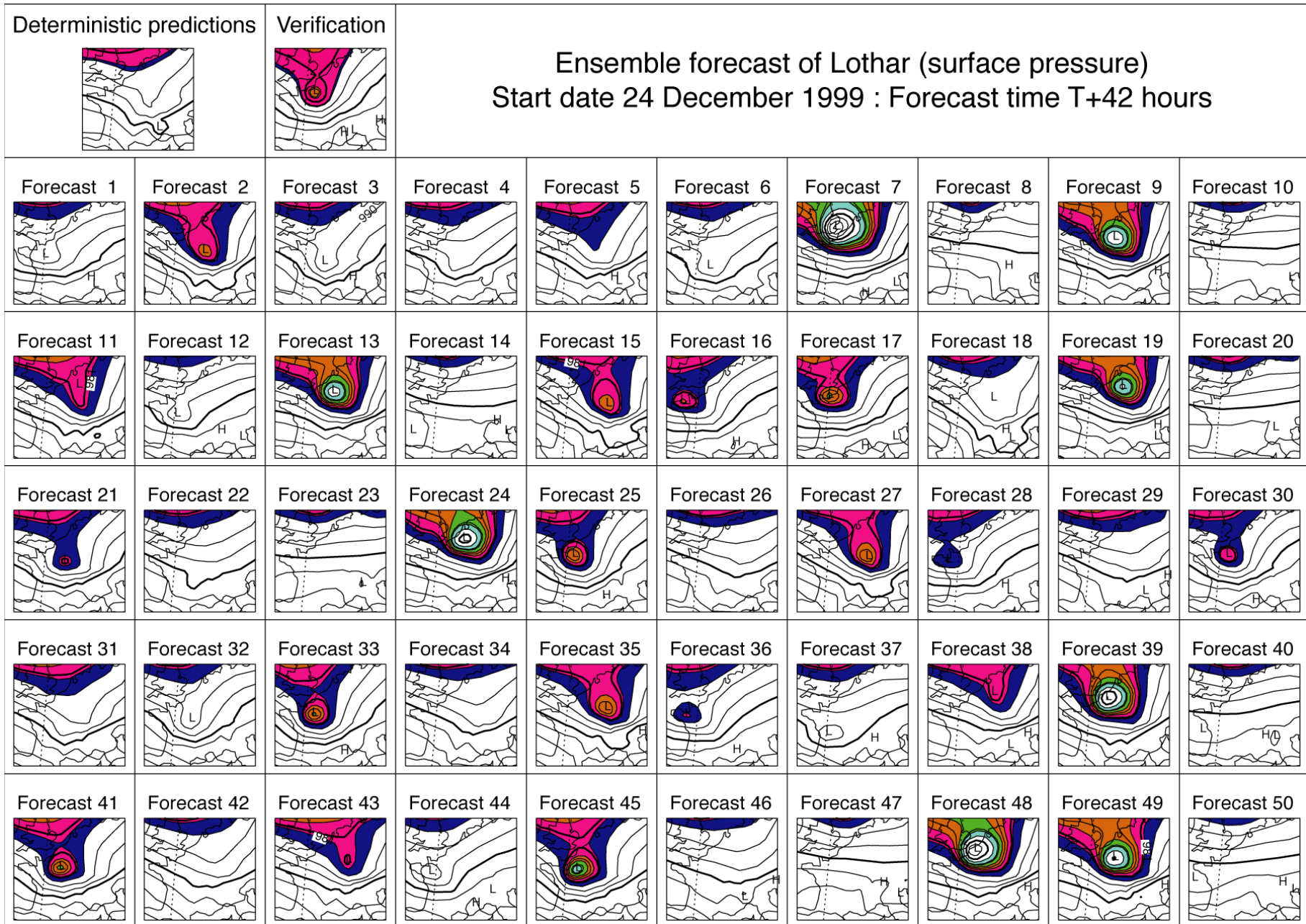
Fig. 4. Scatter plot of ALS measured intensities and vegetation heights, for control plot including shrub and ground layers. Colors represent assignment to different clusters based on a supervised classification of the point cloud using four seedpoints. All laser echos, including those classified as ground return, have been used for this analysis.

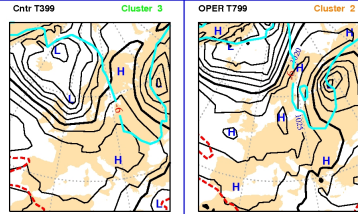
Morsdorf et al. "Discrimination of vegetation strata in a multi-layered Mediterranean forest ecosystem using height and intensity information derived from airborne laser scanning". Remote Sensing of Environment, 2010

Another example:

clustering of the ensemble prediction members forecasts.







ECMWF ENSEMBLE FORECASTS
 Wednesday 7 October 2009 12UTC ECMWF Forecast t+168 VT: Wednesday 14 October 2009 12UTC Surface: Mean sea level pressure
 MSLP (contour every 5hPa) and Temperature at 850hPa (only -6 and 16 isolines are plotted)



ECMWF Ensemble Forecast Clusters

Wednesday 7 October 2009 12UTC
1000hPa Geopotential

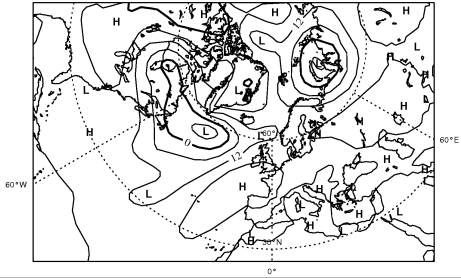
Exp 0001

Operational Forecast in cluster 2

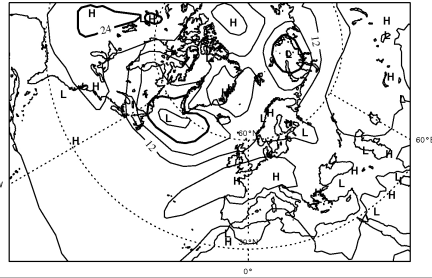
Control Forecast in cluster 3

- Cluster 1 : 23 Forecast(s)
- Cluster 2 : 16 Forecast(s)
- Cluster 3 : 12 Forecast(s)
- Cluster 4 : 0 Forecast(s)
- Cluster 5 : 0 Forecast(s)
- Cluster 6 : 0 Forecast(s)

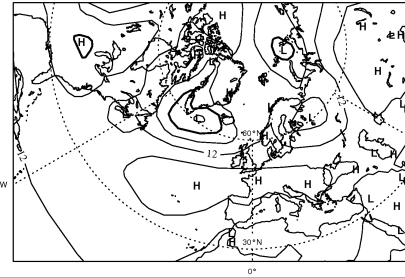
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +96 VT: Sunday 11 October 2009 12UTC
1000hPa Geopotential - Cluster Number 1 (23 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



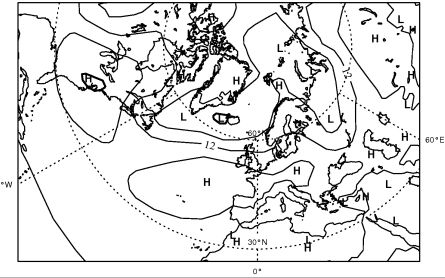
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +120 VT: Monday 12 October 2009 12UTC
1000hPa Geopotential - Cluster Number 1 (23 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



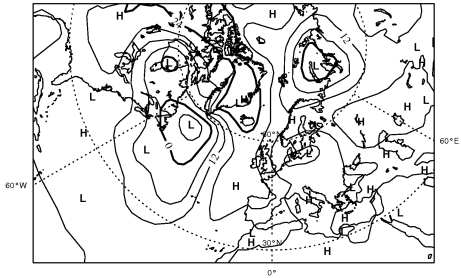
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +144 VT: Tuesday 13 October 2009 12UTC
1000hPa Geopotential - Cluster Number 1 (23 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



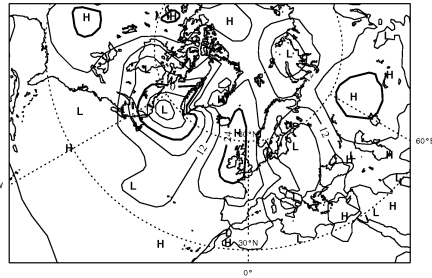
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +168 VT: Wednesday 14 October 2009 12UTC
1000hPa Geopotential - Cluster Number 1 (23 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



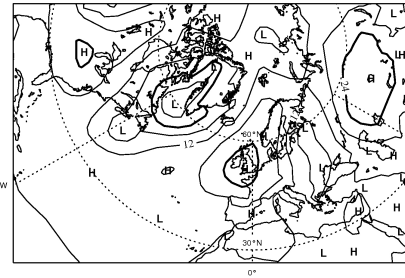
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +96 VT: Sunday 11 October 2009 12UTC
1000hPa Geopotential - Cluster Number 2 (16 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



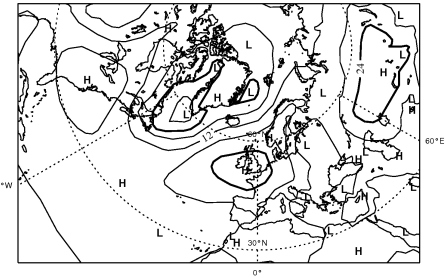
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +120 VT: Monday 12 October 2009 12UTC
1000hPa Geopotential - Cluster Number 2 (16 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



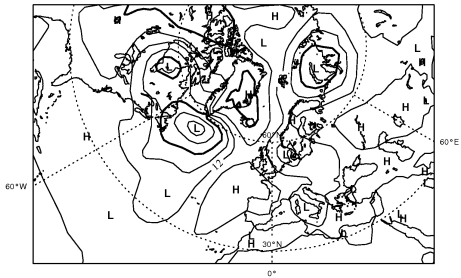
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +144 VT: Tuesday 13 October 2009 12UTC
1000hPa Geopotential - Cluster Number 2 (16 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



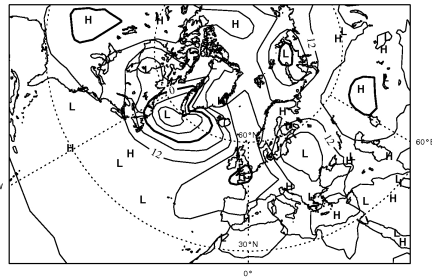
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +168 VT: Wednesday 14 October 2009 12UTC
1000hPa Geopotential - Cluster Number 2 (16 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



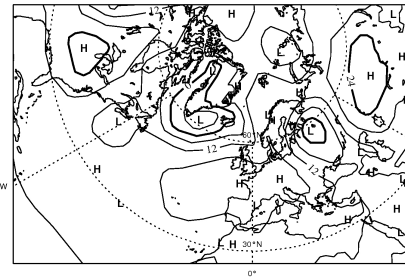
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +96 VT: Sunday 11 October 2009 12UTC
1000hPa Geopotential - Cluster Number 3 (12 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



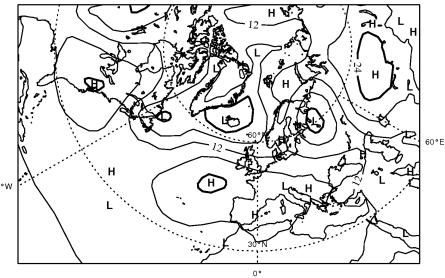
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +120 VT: Monday 12 October 2009 12UTC
1000hPa Geopotential - Cluster Number 3 (12 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +144 VT: Tuesday 13 October 2009 12UTC
1000hPa Geopotential - Cluster Number 3 (12 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3

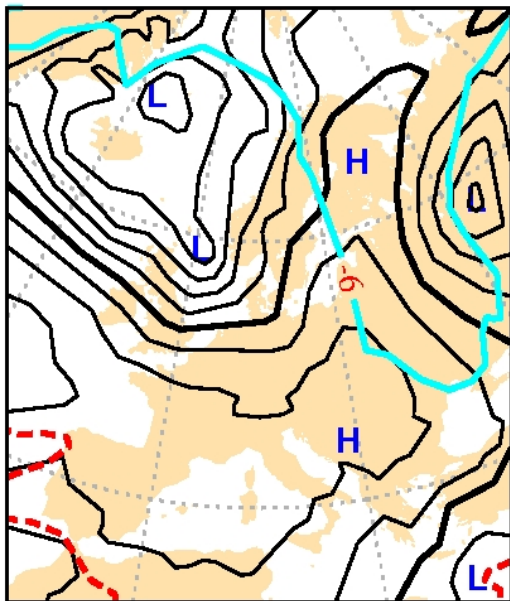


Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast +168 VT: Wednesday 14 October 2009 12UTC
1000hPa Geopotential - Cluster Number 3 (12 members) of 3
Reference step +120-168, Domain 75-340/20-65, Op FC in clus 2, Ctrl FC in clus 3



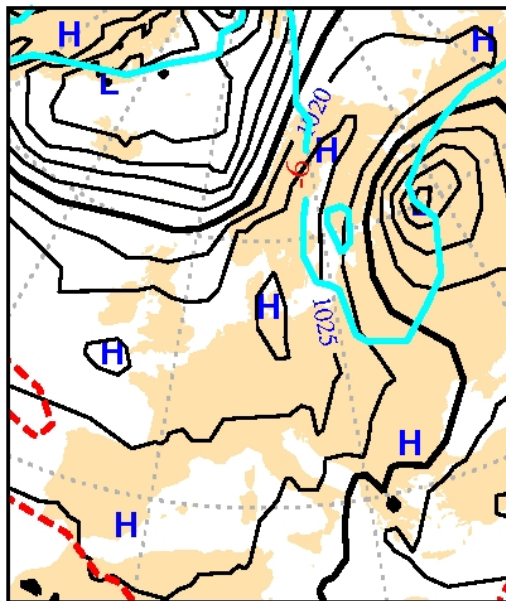
Cntr T399

Cluster 3

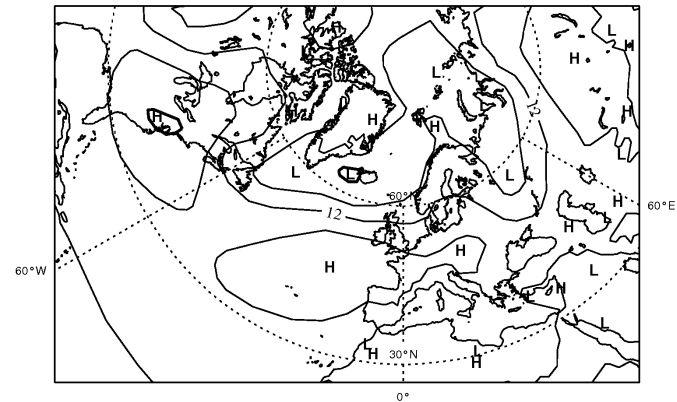


OPER T799

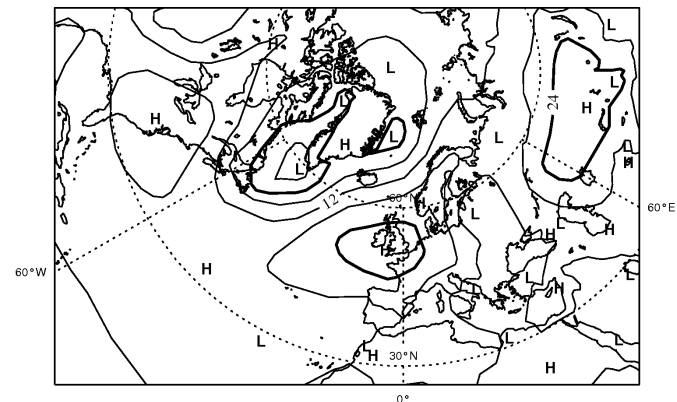
Cluster 2



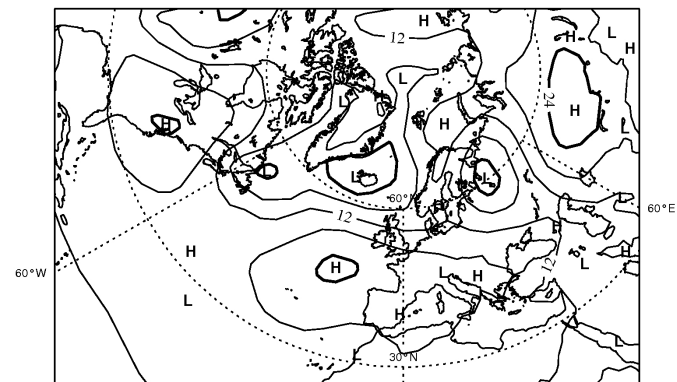
Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast t+168 VT: Wednesday 14 October 2009 12UTC
1000hPa Geopotential - Cluster Number 1 (23 members) of 3
Reference step t+120-168, Domain 75/340/30/45, Op FC in clus 2, Ctrl FC in clus 3



Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast t+168 VT: Wednesday 14 October 2009 12UTC
1000hPa Geopotential - Cluster Number 2 (16 members) of 3
Reference step t+120-168, Domain 75/340/30/45, Op FC in clus 2, Ctrl FC in clus 3

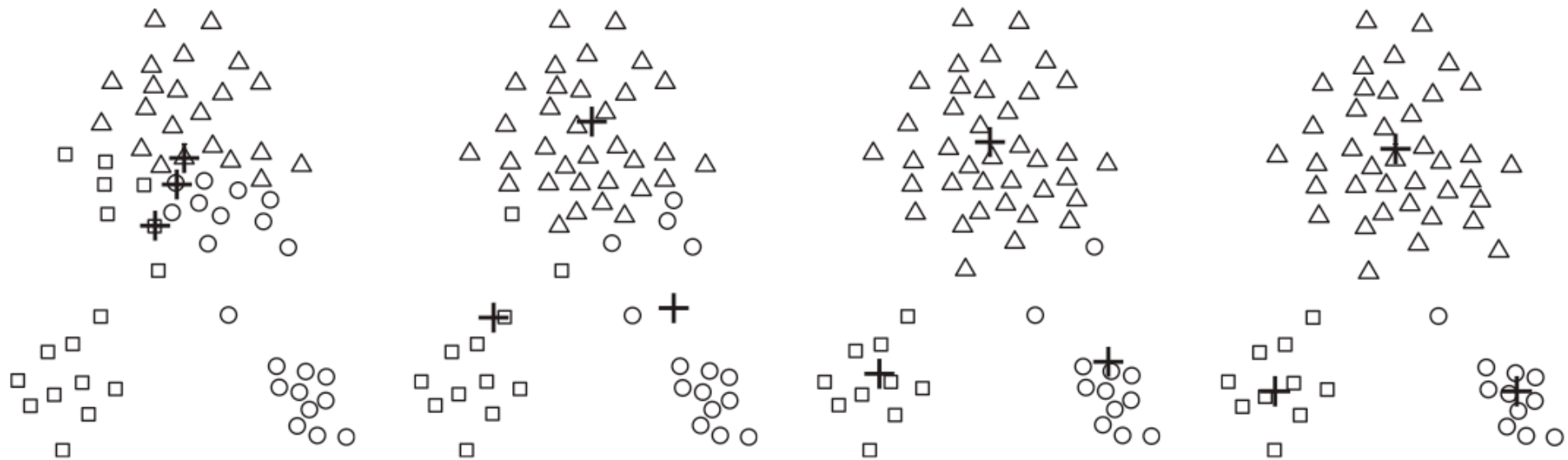


Wednesday 7 October 2009 12UTC ECMWF EPS Cluster Mean Forecast t+168 VT: Wednesday 14 October 2009 12UTC
1000hPa Geopotential - Cluster Number 3 (12 members) of 3
Reference step t+120-168, Domain 75/340/30/45, Op FC in clus 2, Ctrl FC in clus 3



Non-hierarchical clustering: The K-Means algorithm

- 1: Select K points as initial centroids.
- 2: **repeat**
- 3: Form K clusters by assigning each point to its closest centroid.
- 4: Recompute the centroid of each cluster.
- 5: **until** Centroids do not change.



(a) Iteration 1.

(b) Iteration 2.

(c) Iteration 3.

(d) Iteration 4.

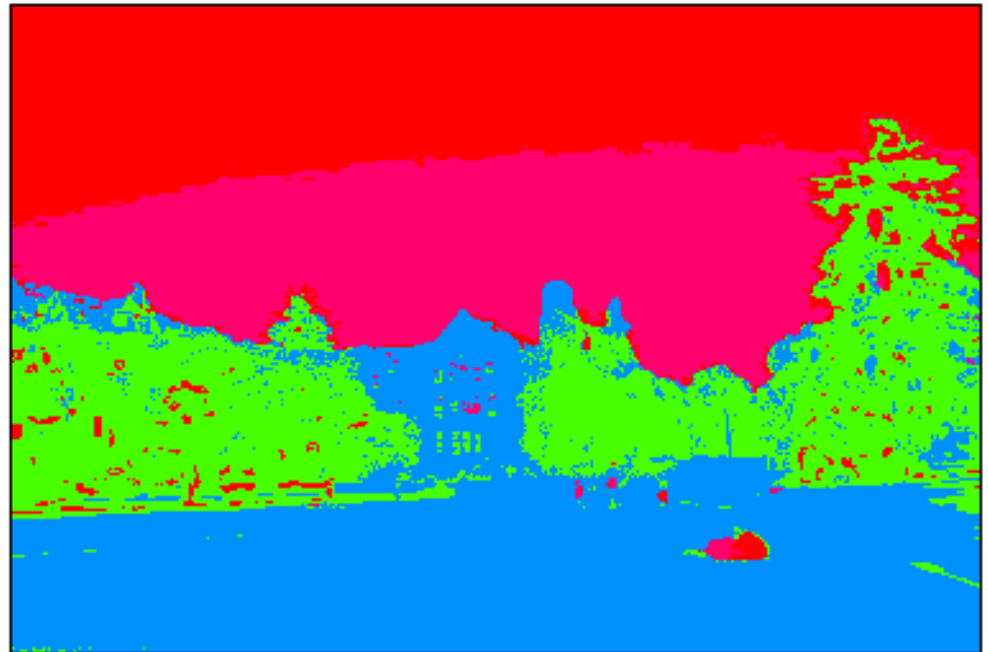
Warning: the results of the algorithm depend on the initial random choice of centroids.

Image elaboration examples:
K-means analysis on the color vectors [r,g,b]

Original image

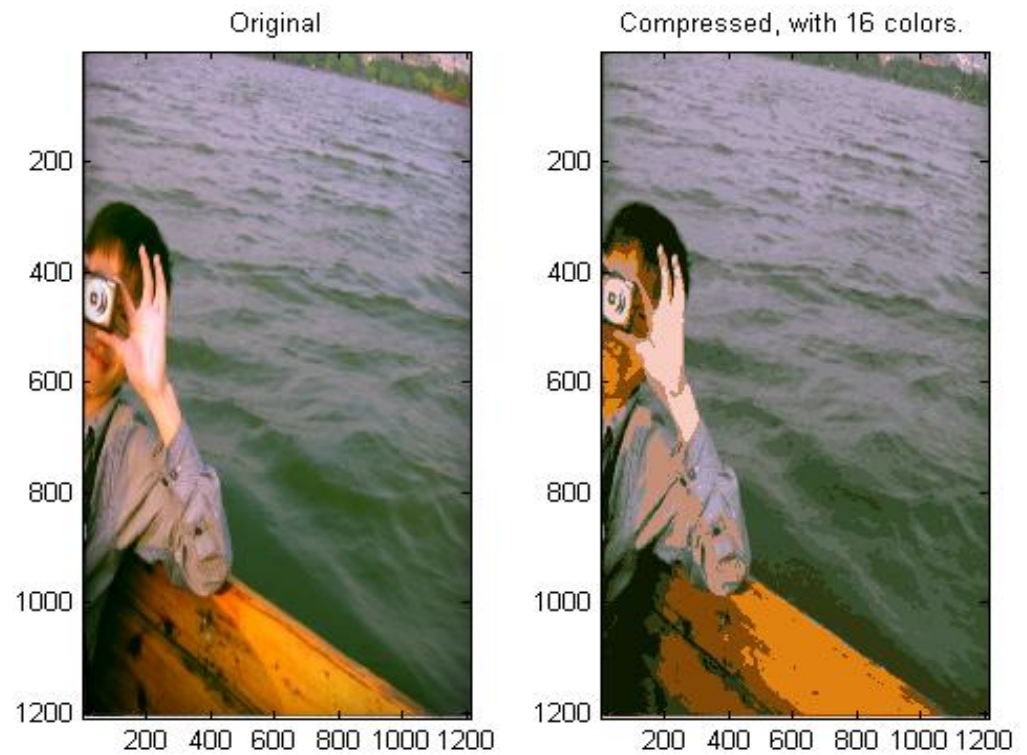


Four cluster membership



Attributing the mean color of the cluster to each member

16 clusters



3 clusters

Example: Euro-Atlantic weather regimes

Michelangeli, P.-A., R. Vautard and B. Legras, 1995: Weather Regimes: recurrence and quasi-stationarity. J. Atmos. Sci., 52, 8, 1237-1256.

Looking for persistent and recurrent states in the large scale atmospheric circulation.

1) Truncate the data on the first few EOFs

$$\mathbf{z}(t) = \sum_{i=1}^T c_i(t) \mathbf{e}_i, \quad \text{where } c_i = \langle \mathbf{z}, \mathbf{e}_i \rangle$$

2) Apply K-Means algorithm to truncated data c_i

3) Devise a statistical test to i) find the optimal number of cluster and ii) test the significance of the results

4) Filter back the cluster centroids onto the physical space and plot.

4) Filter back the cluster centroids onto the physical space and plot.

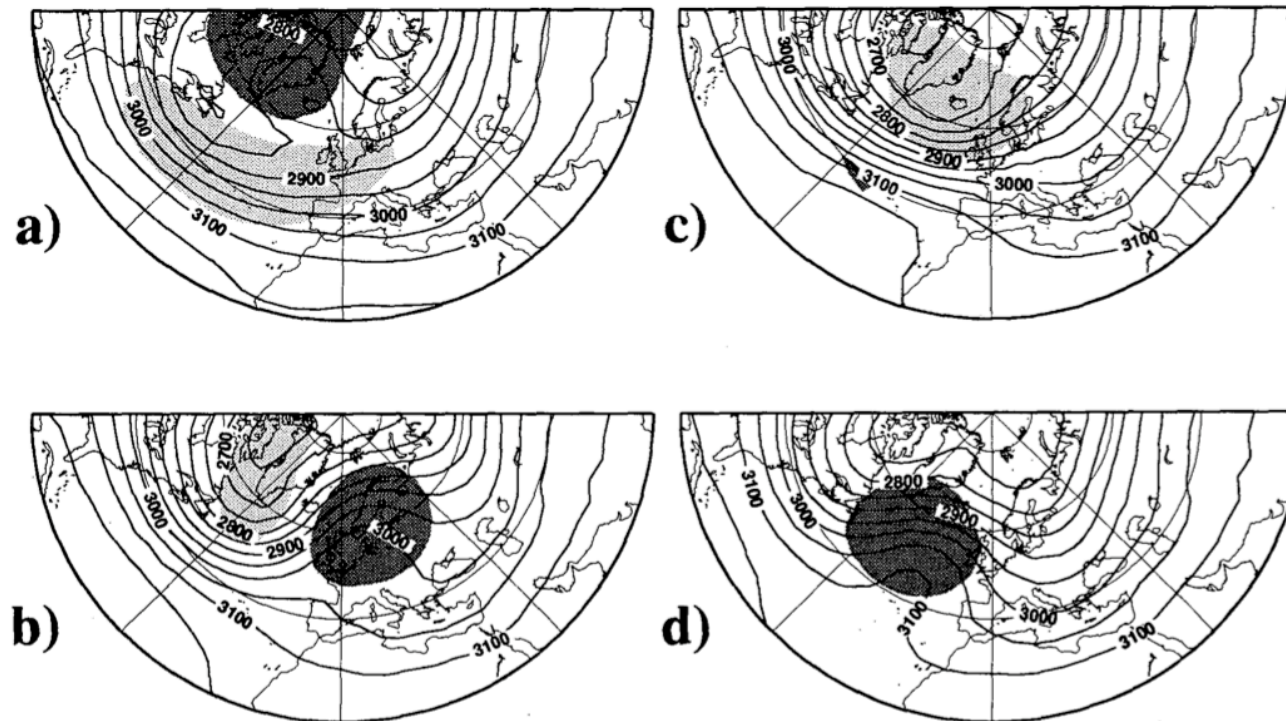


FIG. 4. Composites of the 700-hPa geopotential heights for the four clusters found over the ATL sector. Contour interval is 50 m. Dark shaded areas show areas where the anomaly of the composite with respect to the wintertime average is larger than 50 m. Light shaded areas correspond to anomalies lower than -50 m. Clusters are sorted by their consistency: (a) cluster 1; (b) cluster 2; (c) cluster 3; (d) cluster 4.

Weather Regimes

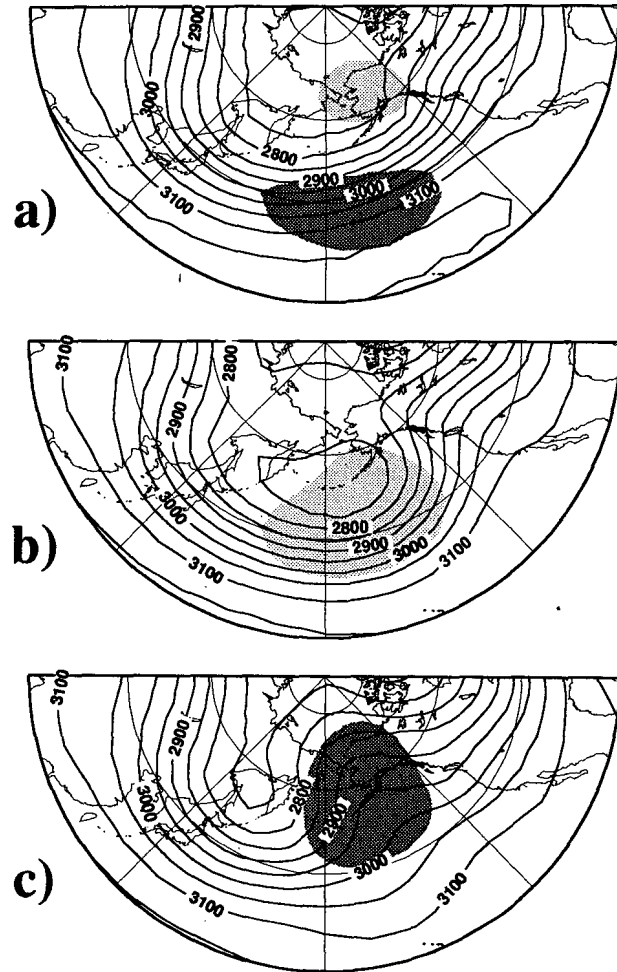


FIG. 7. Same as Fig. 4 for the three clusters of the PAC domain: (a) cluster 1; (b) cluster 2; (c) cluster 3.

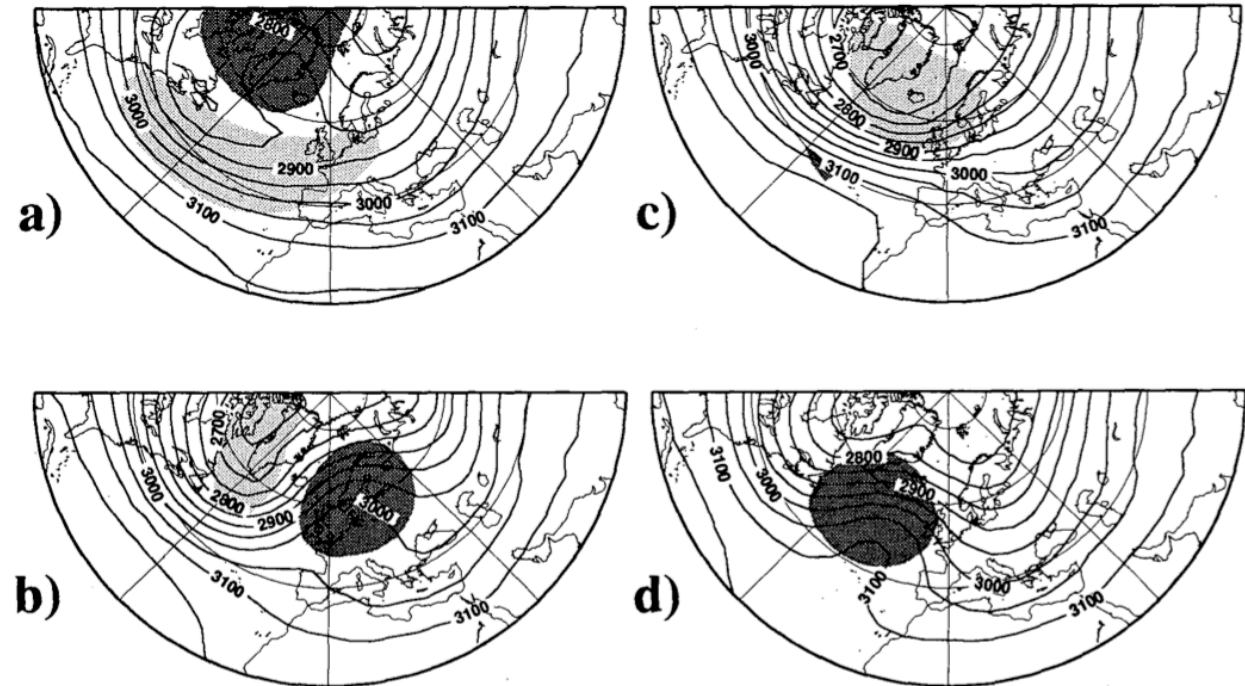
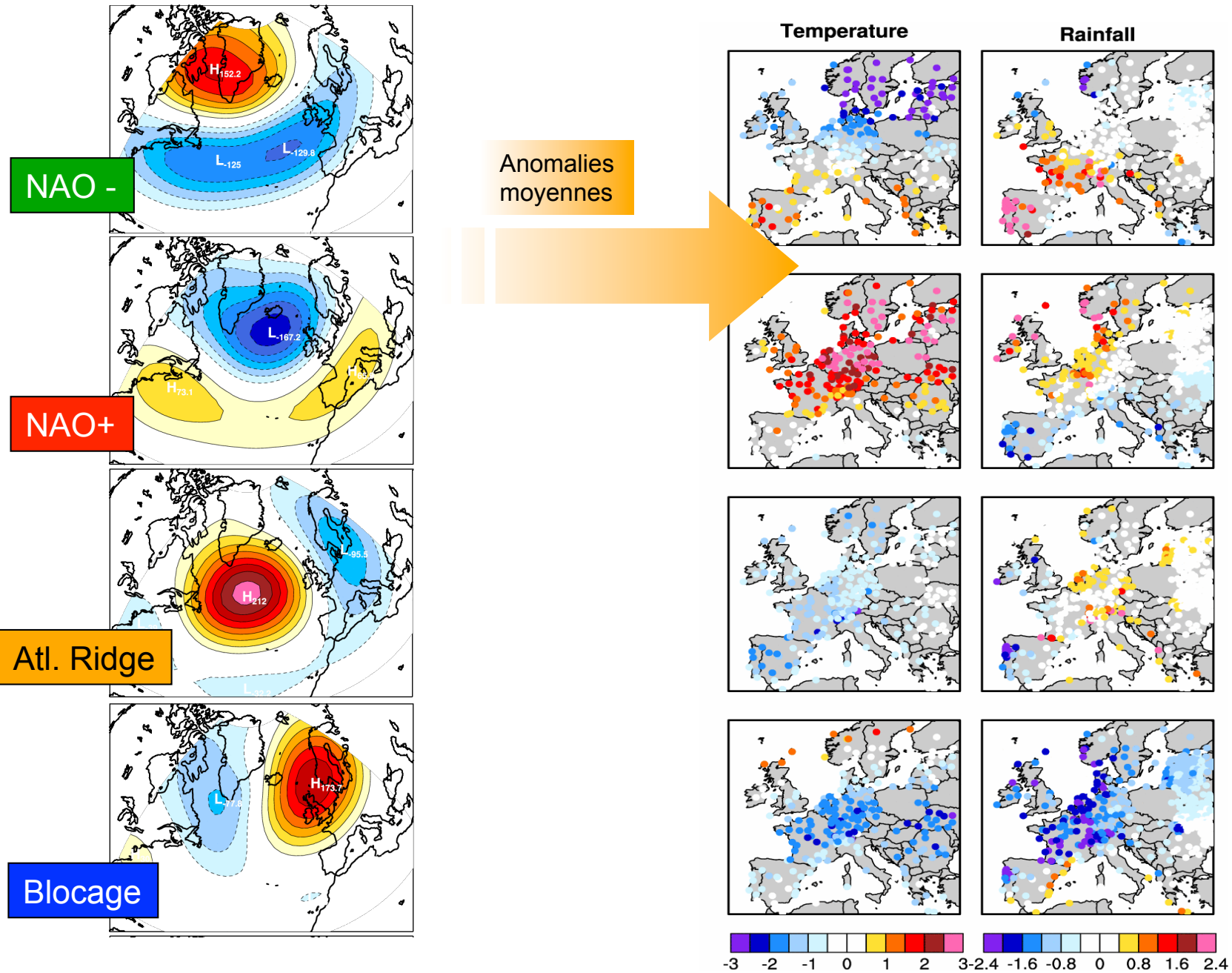
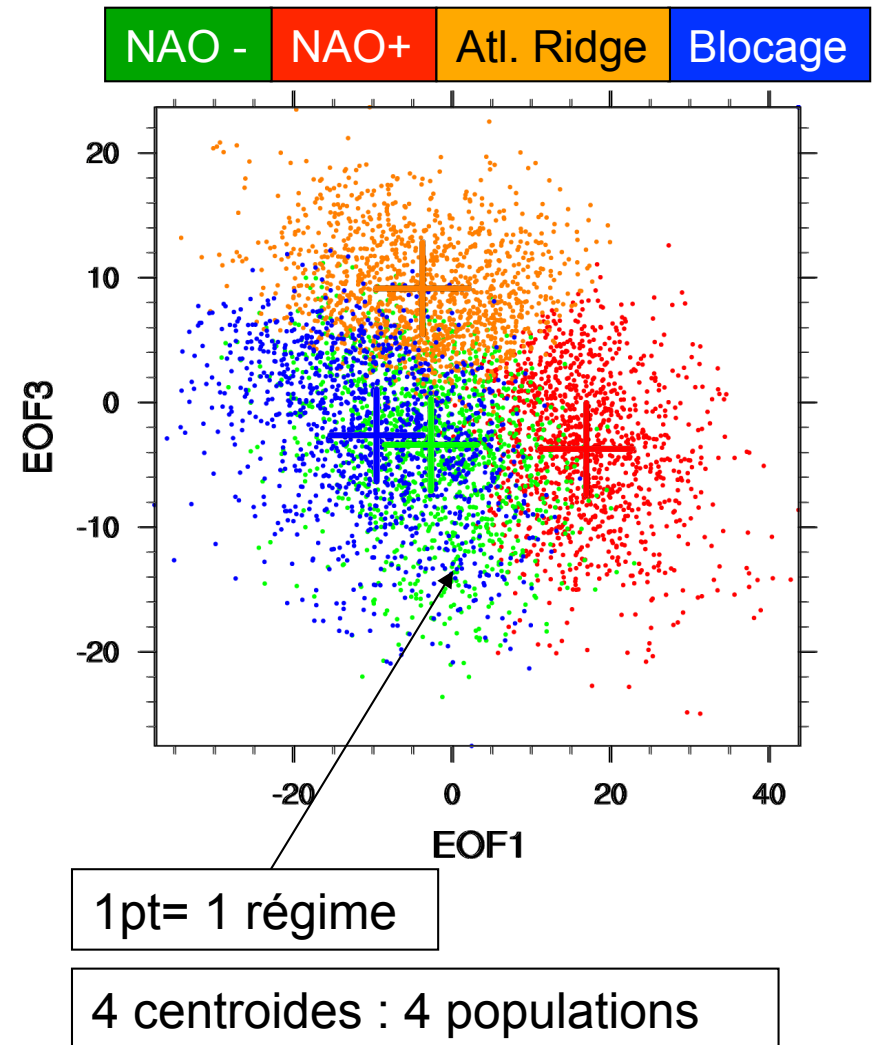
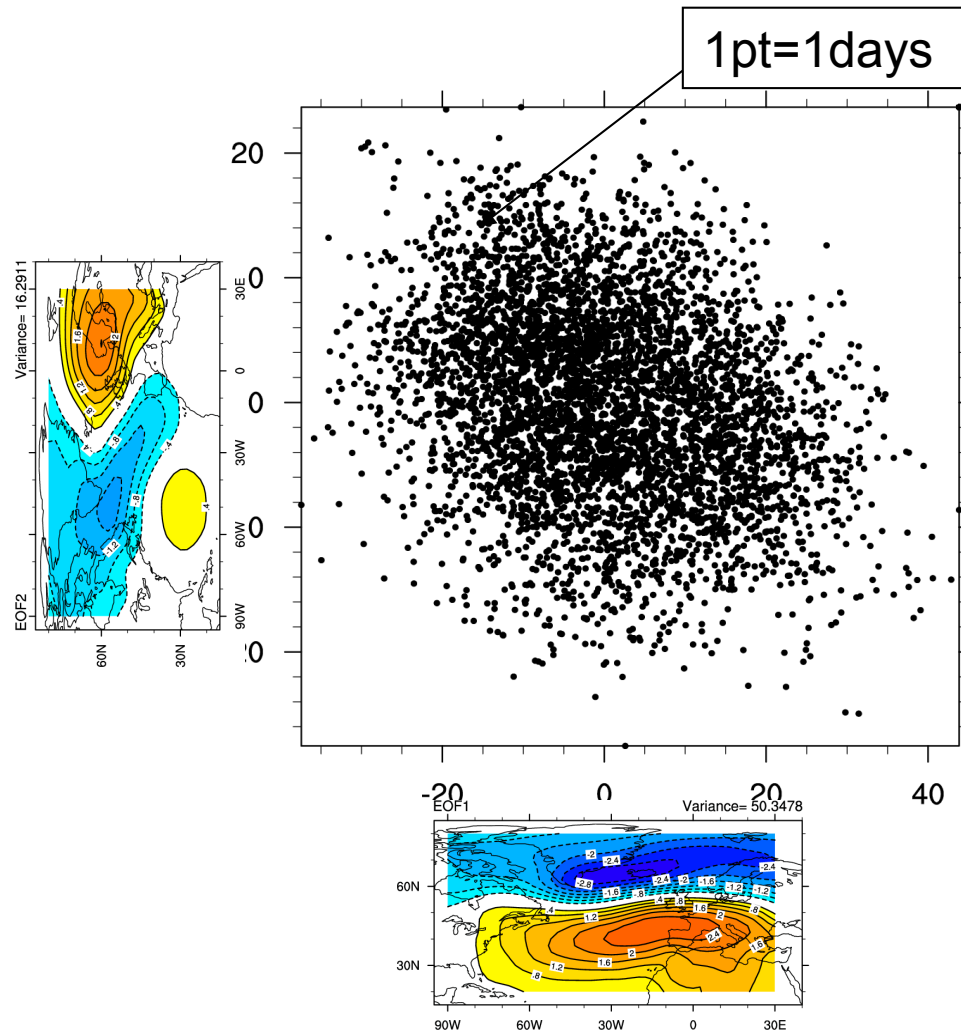


FIG. 4. Composites of the 700-hPa geopotential heights for the four clusters found over the ATL sector. Contour interval is 50 m. Dark shaded areas show areas where the anomaly of the composite with respect to the wintertime average is larger than 50 m. Light shaded areas correspond to anomalies lower than -50 m. Clusters are sorted by their consistency: (a) cluster 1; (b) cluster 2; (c) cluster 3; (d) cluster 4.

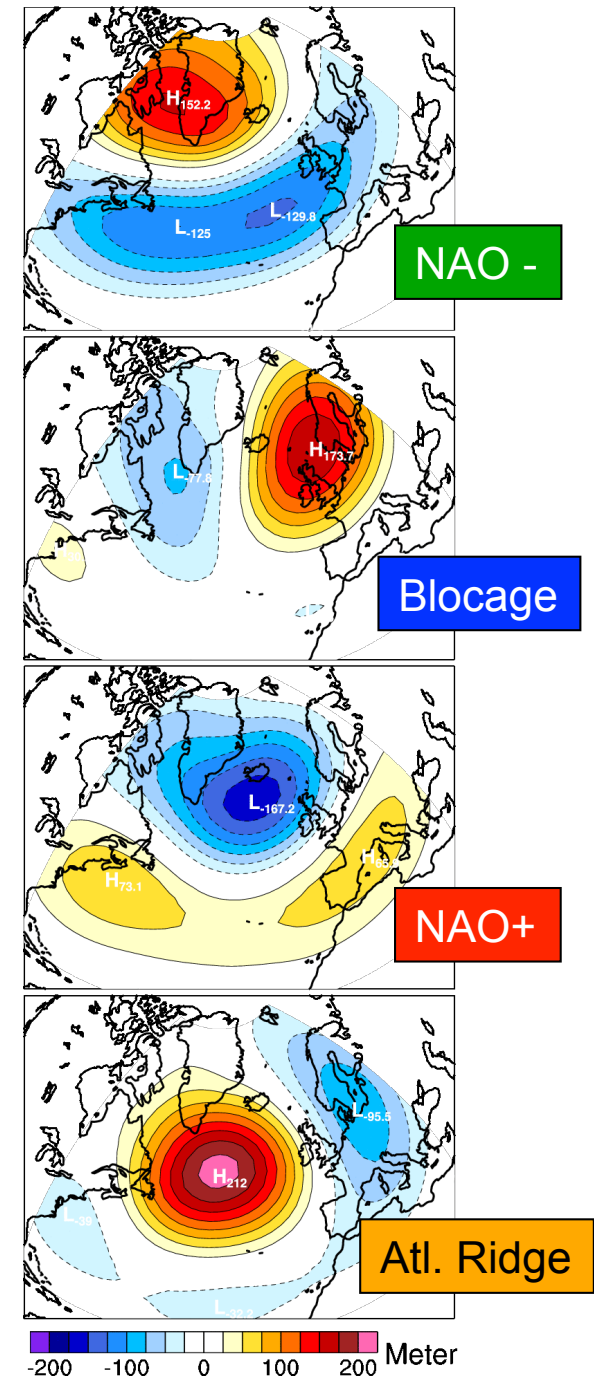
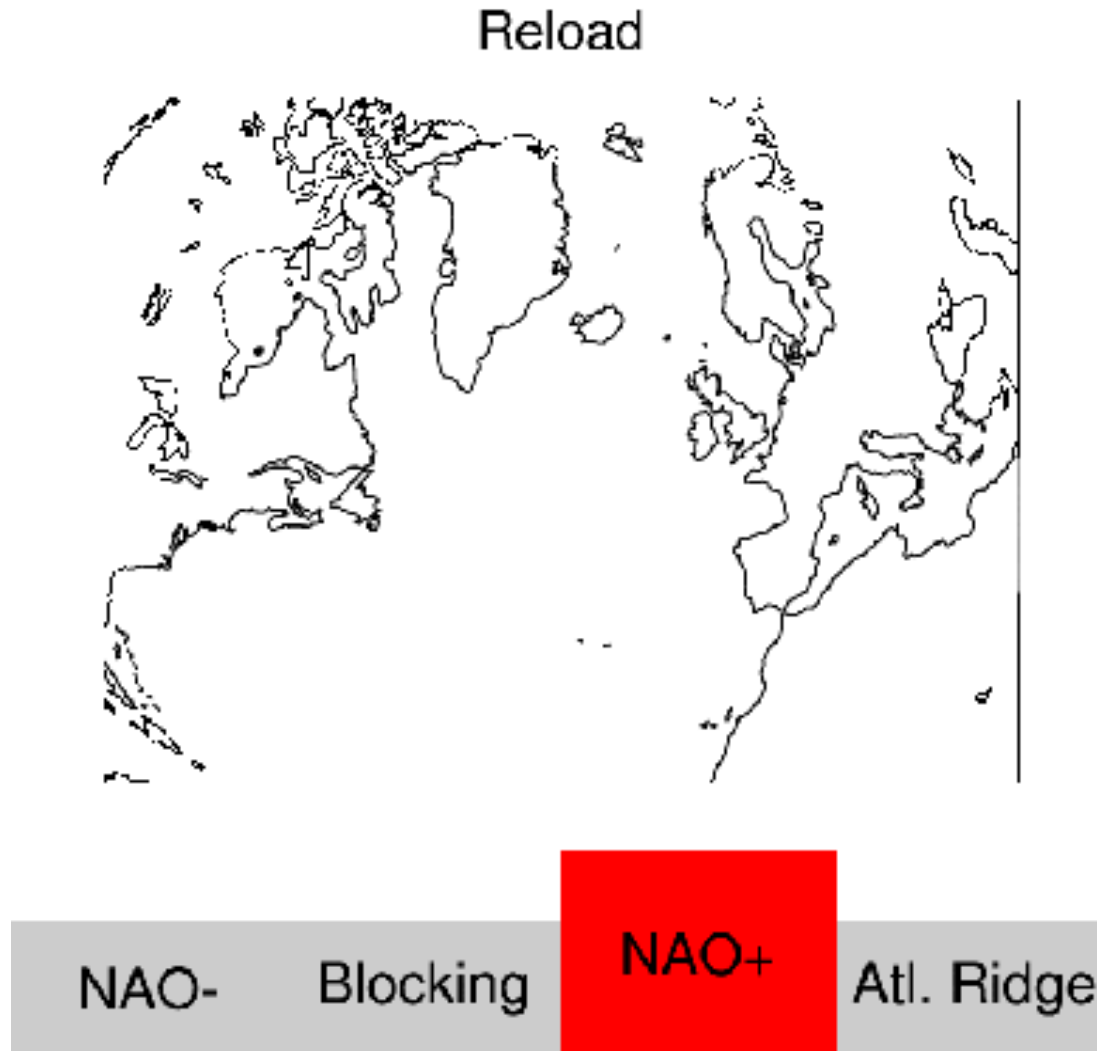
Lien Régimes / Températures / Précipitation : la moyenne



Les régimes dans l'espace des EOFs



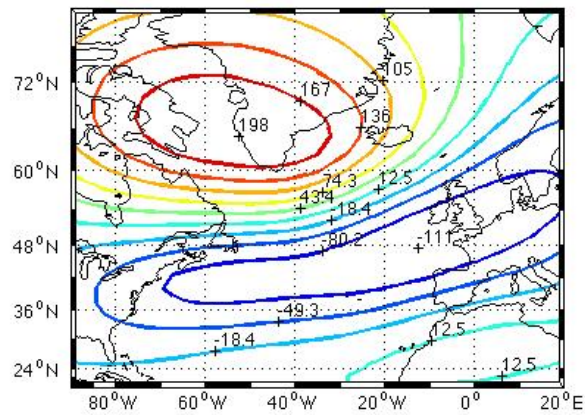
December-january 2012. Circulation and classification into regimes



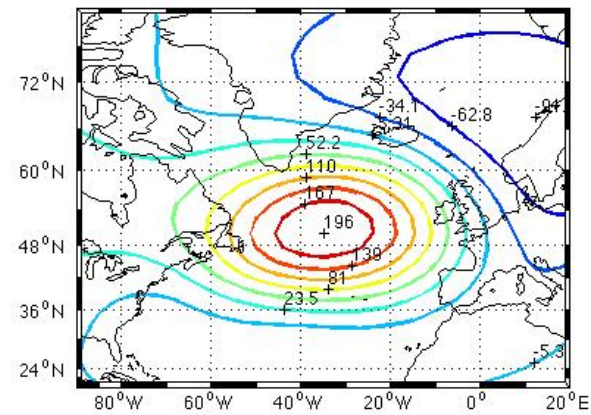
Courtesy of Christophe Cassou (CERFACS, Toulouse)

Exercise : compute weather regimes by k-means clustering on the same data as last week

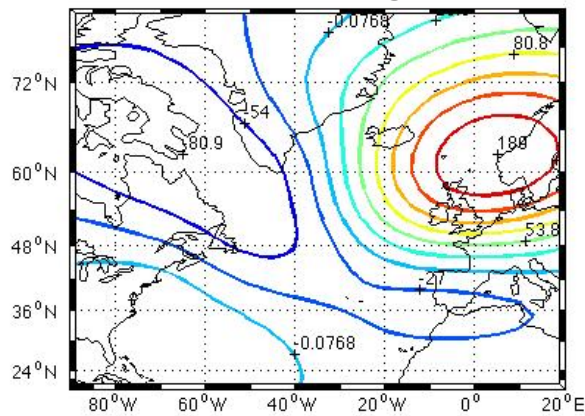
Greenland ridge



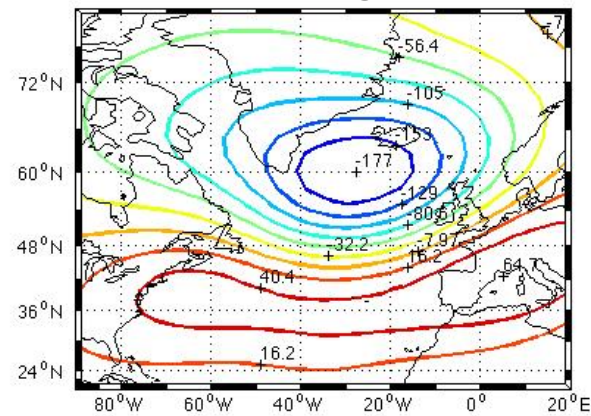
Atlantic ridge



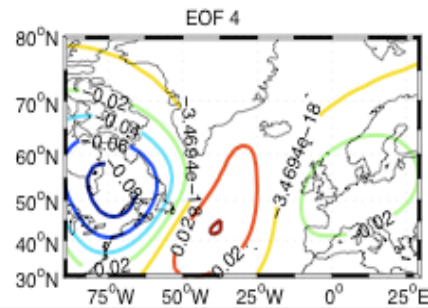
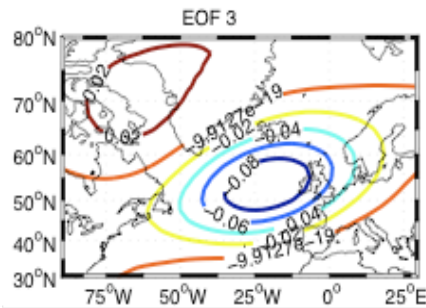
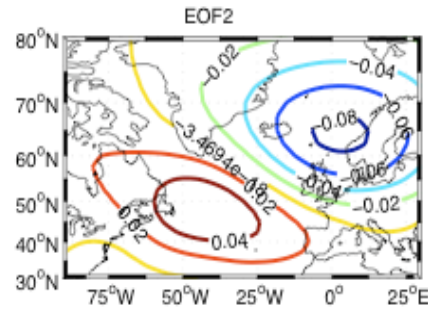
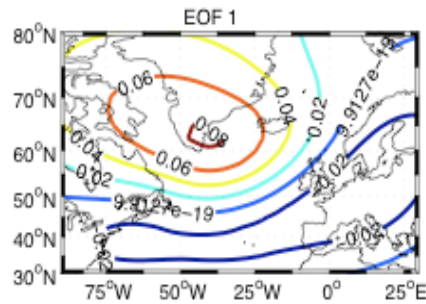
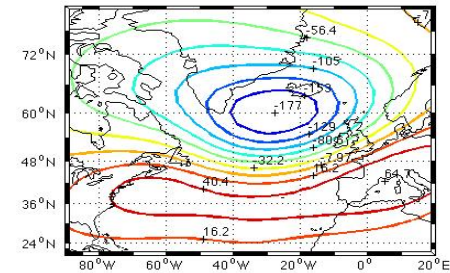
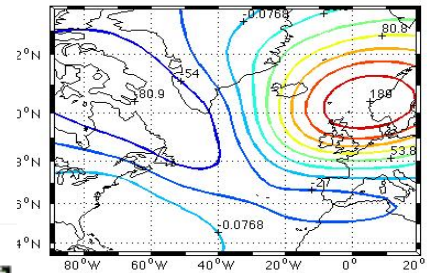
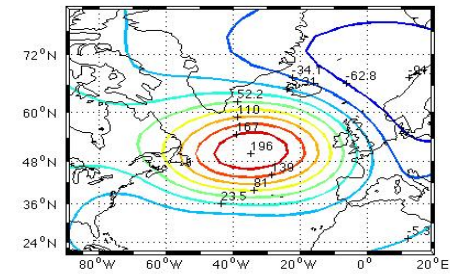
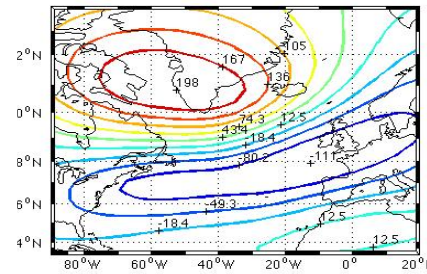
Blocking



Zonal regime



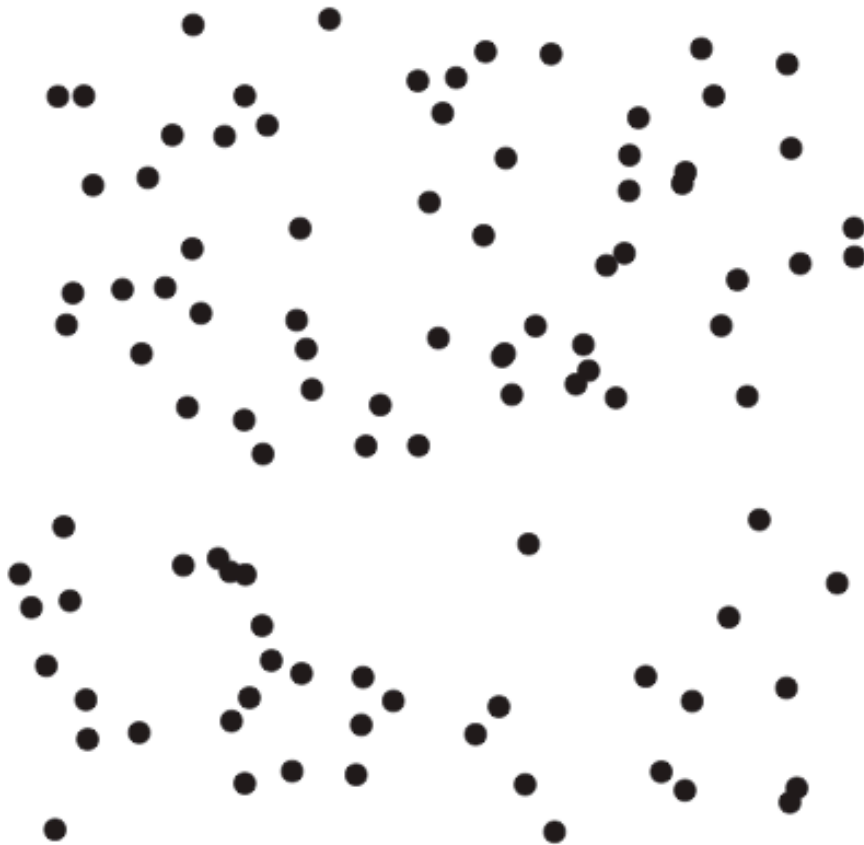
Compare the clusters with the EOFs:



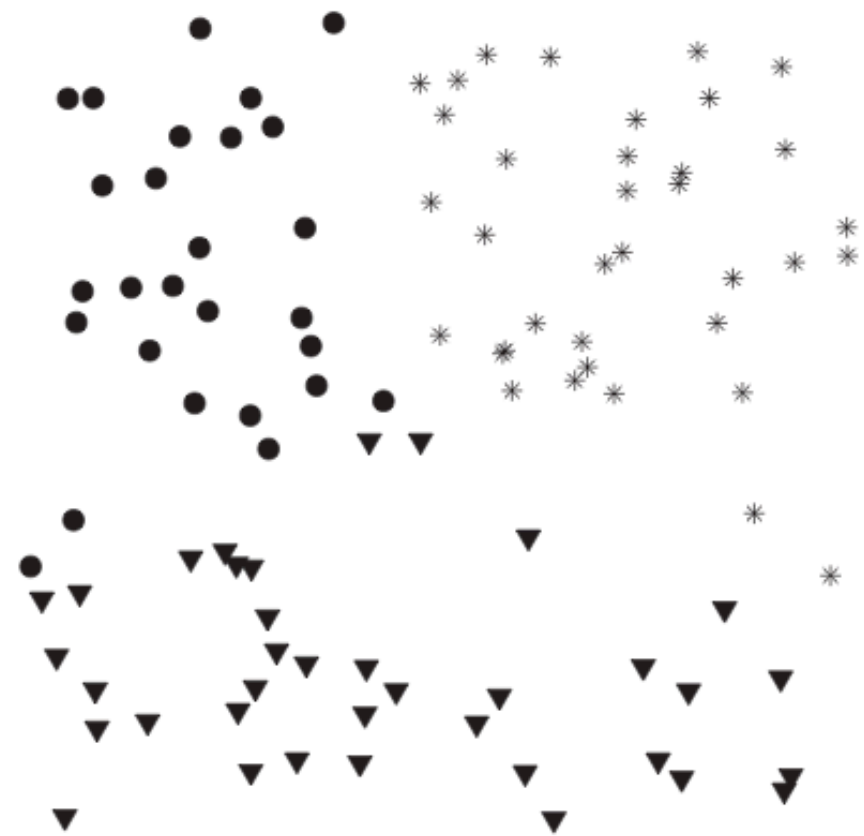
The data were projected on the 8 first EOFs.

Validity of Clustering.

How do we estimate how good is the clustering, or rather how well the data can be described as a group of clusters? What is the optimal number of clusters K , or, where to cut a dendrogram? Are the clusters statistically significant?



(a) Original points.



(c) Three clusters found by K-means.

Measures of cohesion and separation:

$$SSE = \sum_i \sum_{x \in C_i} (x - c_i)^2$$

$$SSB = \sum_i |C_i| (c_i - c)^2$$

It can be shown that SSE+SSB is constant for a given sample, so optimizing one is like optimizing the other.

The Silhouette coefficient

For an individual point, x_i

– Compute $a_i = \text{average distance of } x_i \text{ to the points in its cluster}$

$$a_i = \frac{1}{|C_i|} \sum_{y \in C_i} d(x_i, y)$$

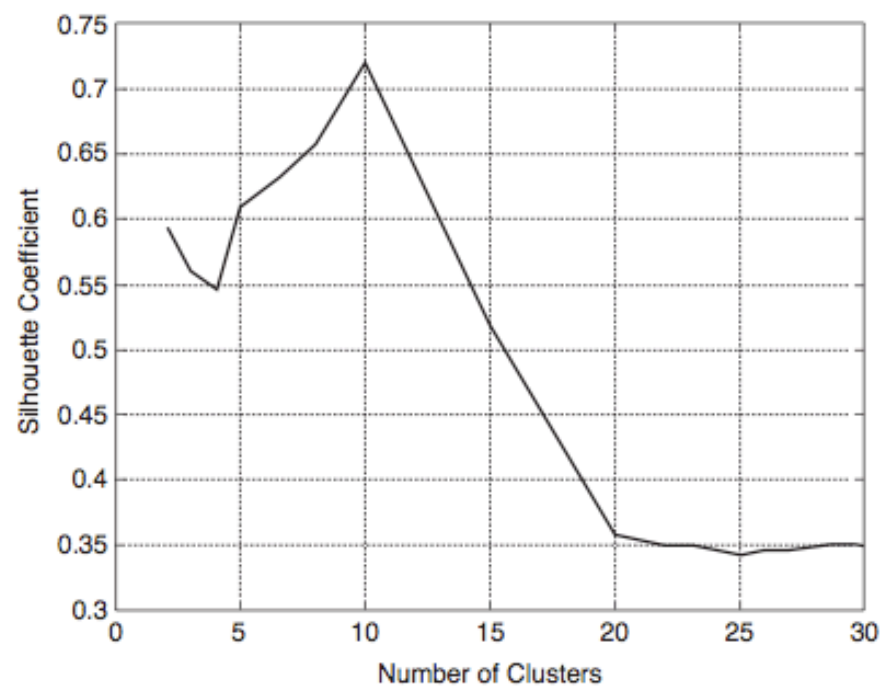
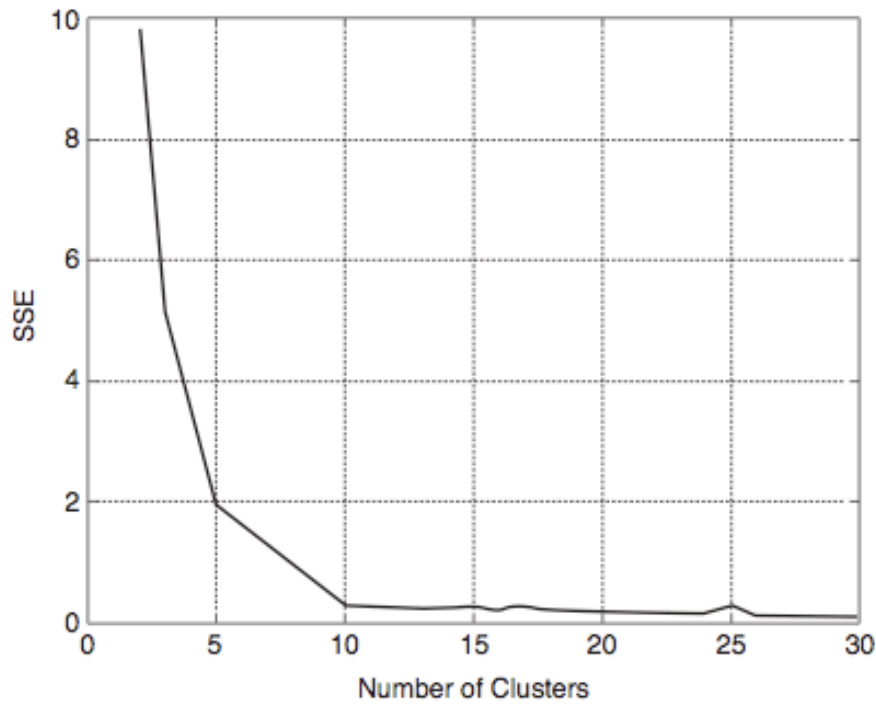
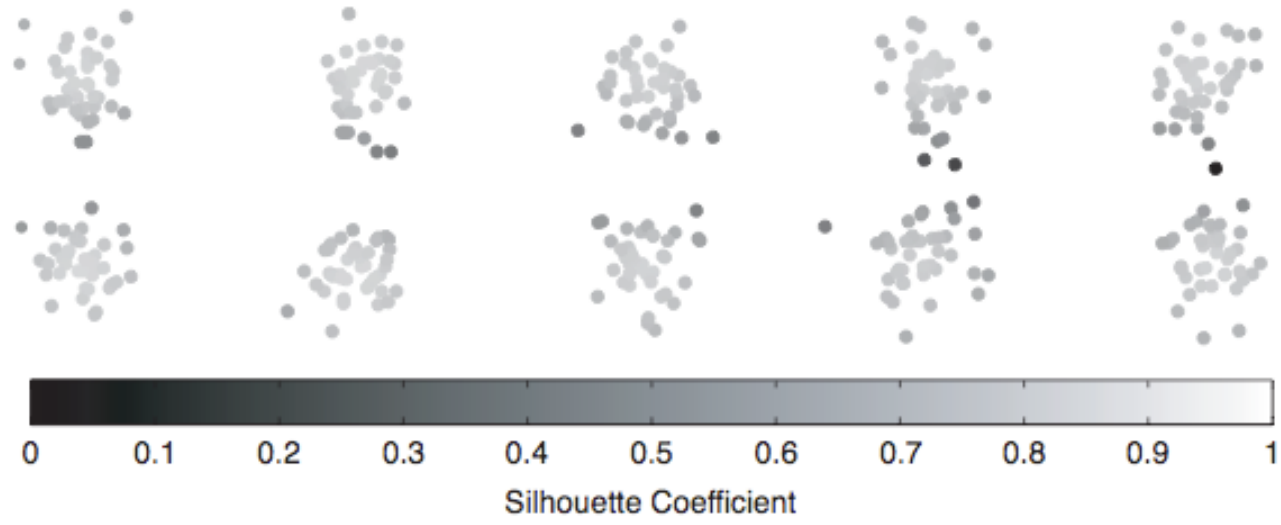
– Compute $b_i = \min (\text{average distance of } x_i \text{ to points in another cluster})$

$$b_i = \min_j \frac{1}{|C_j|} \sum_{y \in C_j} d(x_i, y)$$

– The silhouette coefficient for a point is then given by

$$1 - \frac{a_i}{b_i} \quad (\text{or } 1 - \frac{b_i}{a_i} \text{ if } a_i \geq b_i)$$

Example



The example of Michelangeli et al 1995.

- 1) Define the classifiability of data.
- 2) Test the null hypothesis that the data are “as classifiable as a random process”

1) Classifiability

The idea is that data can be classifiable into clusters or not. If they are, the K-Means algorithm should give always the same partition into the same clusters, independently of the initial random choice of centroids.

They defined an index of classifiability:

$$I(K) = \frac{1}{N(N-1)} \sum_{m \neq m'} c(P_m(K), P_{m'}(K))$$

That states that any two partitions are very similar if the data are very classifiable.

What is the similarity of two partitions. c ?

$$c(P_m(K), P_n(K)) = \sum_{i=1, K} \langle c_i^n, c_i^m \rangle$$

It is the sum of the scalar products of corresponding cluster centroids.

2) Test the null hypothesis.

What is the probability distribution of the null hypothesis?

To estimate the probability distribution of the null hypothesis they used a “Montecarlo” method. They computed the classifiability of 100 random samples and they compared the value of the classifiability index obtained from them to the one of the data.

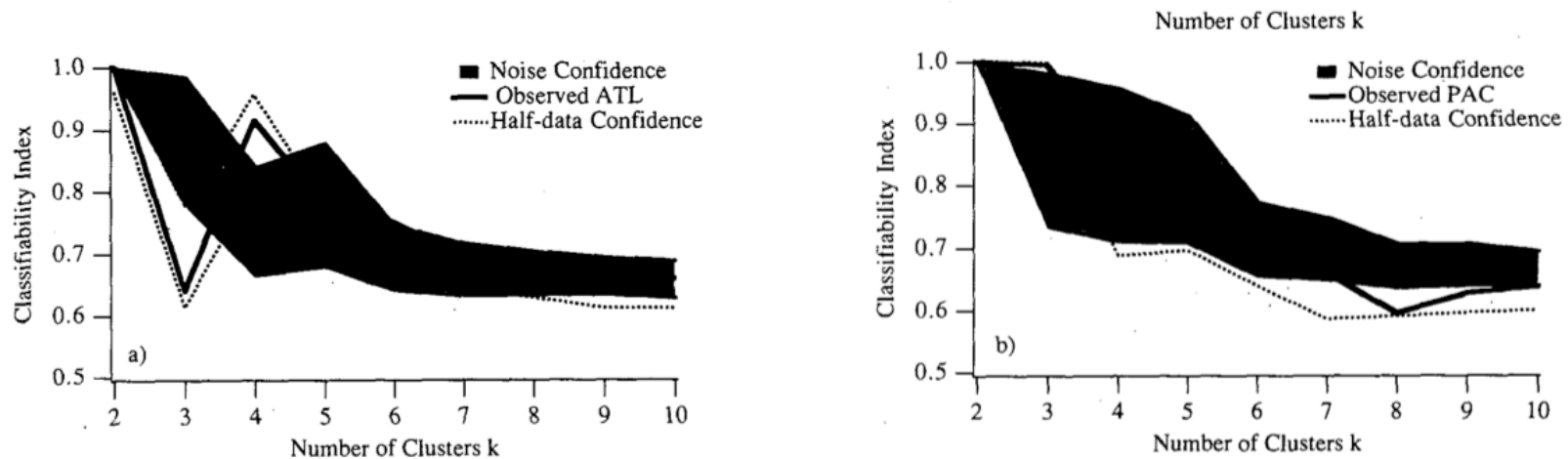


FIG. 2. The classifiability index as a function of the number of clusters k (heavy solid). Shaded area represents the 10%–90% bounds of the classifiability index distribution calculated from the first-order Markov process (noise model). Dotted lines represent the same bounds for random halves of the atmospheric data: (a) Atlantic sector; (b) Pacific sector.

NEURAL NETWORKS



Warren McCulloch
1898- 1969



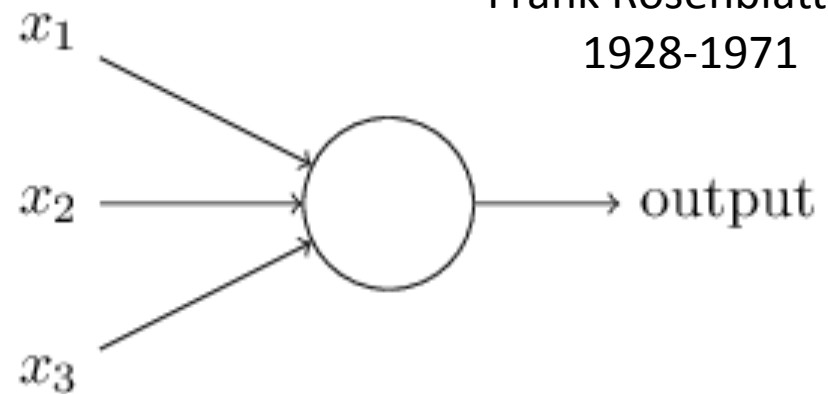
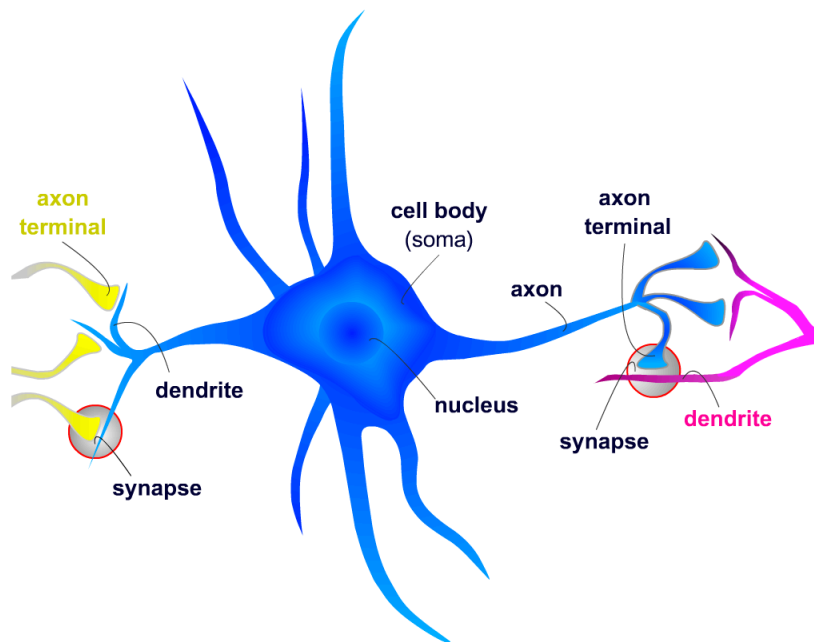
Walter Pitts
1923 –1969

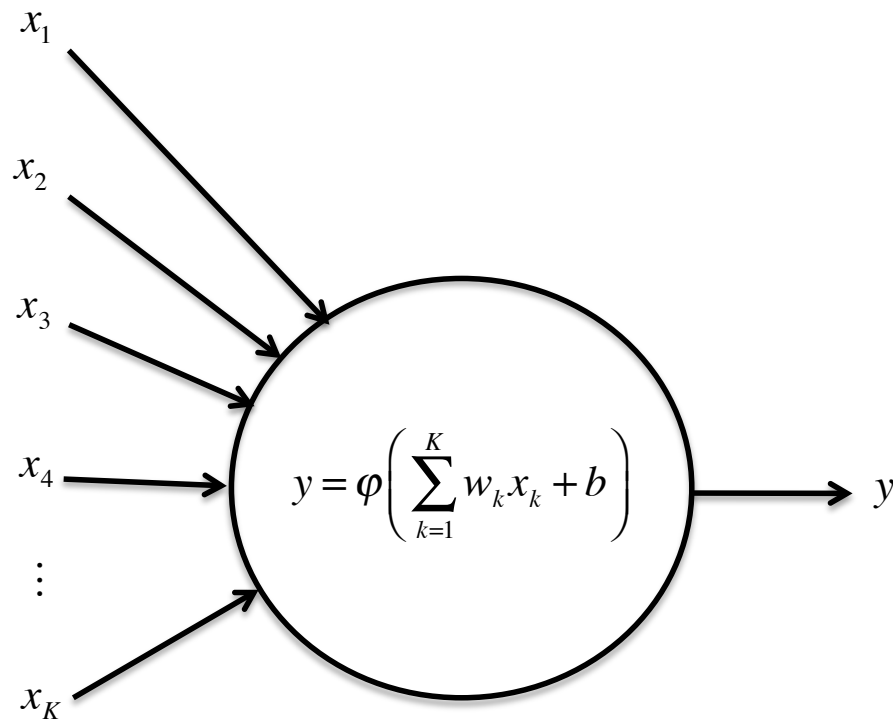


Jerome Lettvin
1920-2011



Frank Rosenblatt
1928-1971





This thing can compute any Boolean operation.

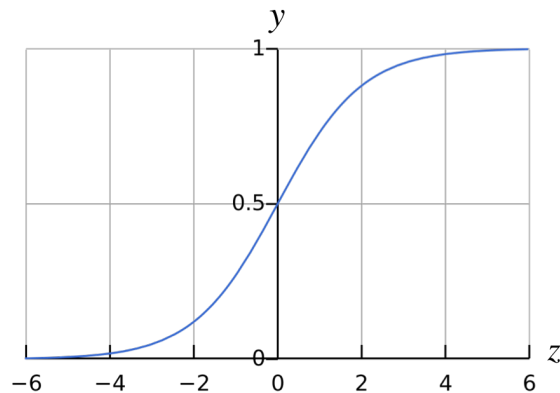
Proof/example, the NAND operator.

$$w_1 = -2, w_2 = -2, b = 3$$

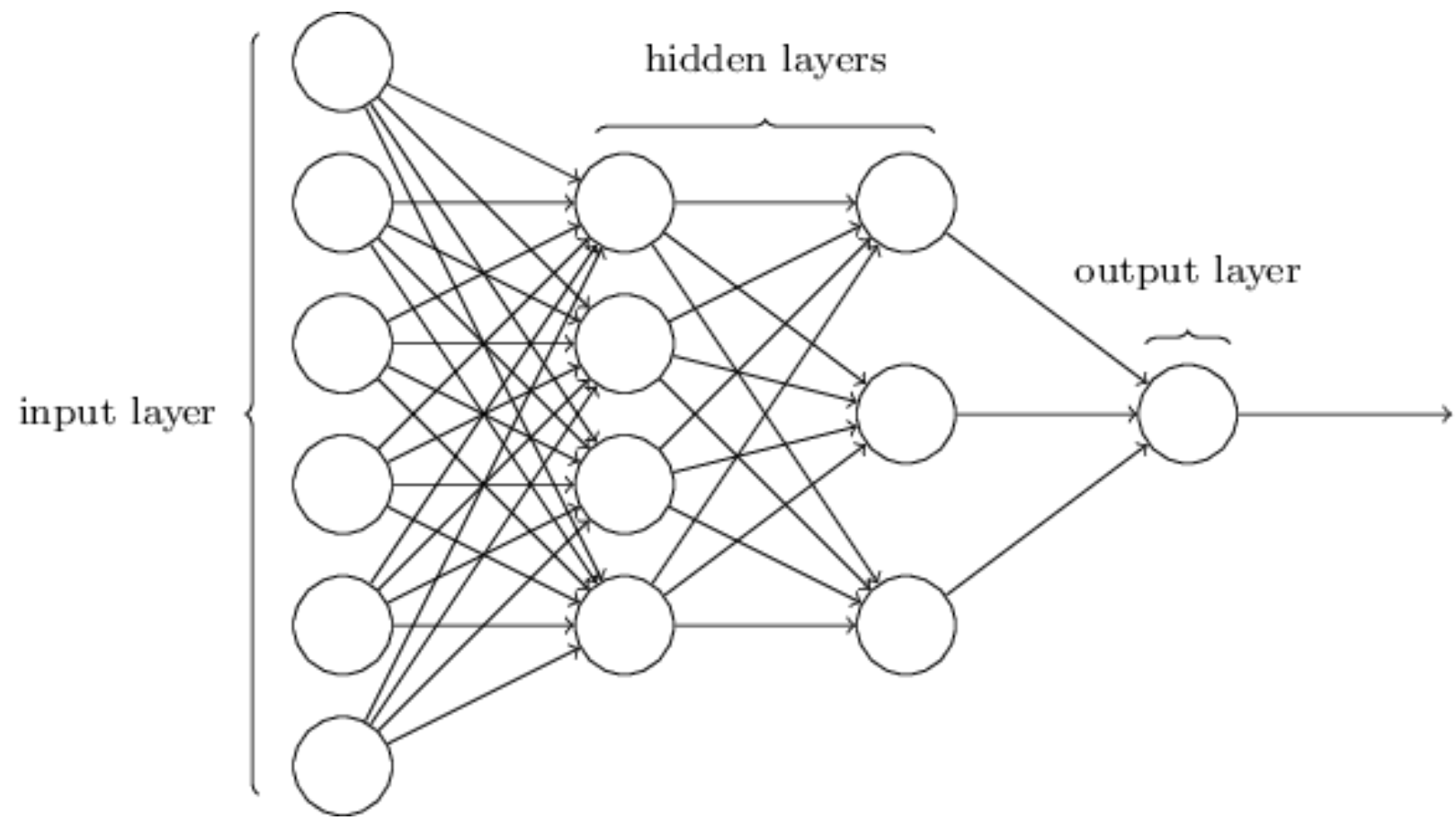
$$x_1 \quad x_2 \quad y = \varphi(w_1 x_1 + w_2 x_2 + b) \quad \text{int}(y)$$

1	1	0.2689	0
1	0	0.7311	1
0	1	0.7311	1
0	0	0.9526	1

$$\varphi(z) = \frac{1}{1 + e^{-z}}$$



Ok, it can compute anything, then !




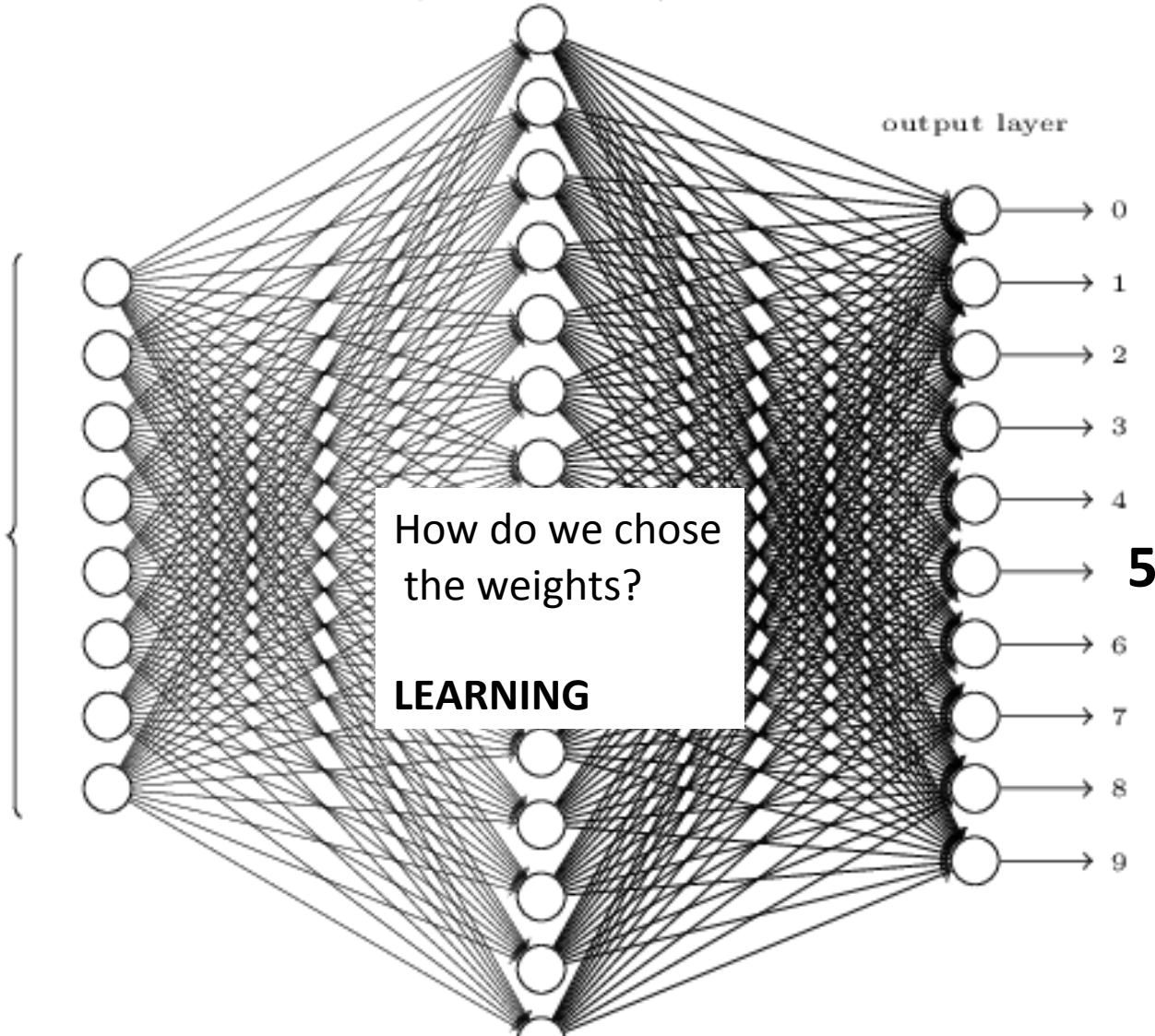
Example: handwriting recognition

504192

hidden layer
(n = 15 neurons)

output layer


input layer
(784 neurons)



Use a training dataset to define a cost function

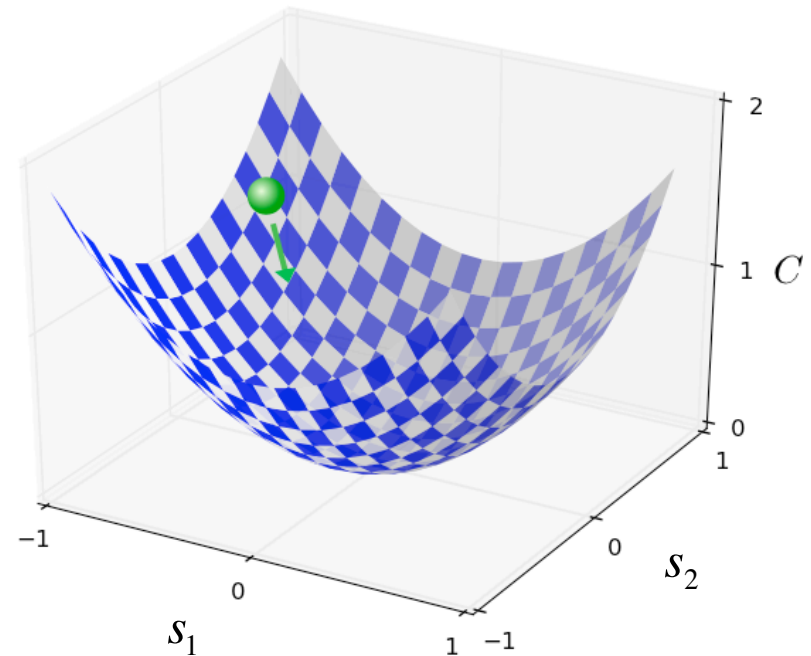


$$C(\mathbf{w}, \mathbf{b}) \equiv \frac{1}{2n} \sum_x \|\mathbf{y}(\mathbf{x}) - \mathbf{a}\|^2$$

A variational approach. Use a gradient descent method to find the minimum of $C(\mathbf{w}, \mathbf{b})$

$$\delta C(\mathbf{w}, \mathbf{b}) = \nabla_{\mathbf{w}} C \cdot \delta \mathbf{w} + \nabla_{\mathbf{b}} C \cdot \delta \mathbf{b} = \nabla C \cdot \delta \mathbf{s}$$

$$\mathbf{s} \rightarrow \mathbf{s}' = \mathbf{s} - \eta \nabla C$$

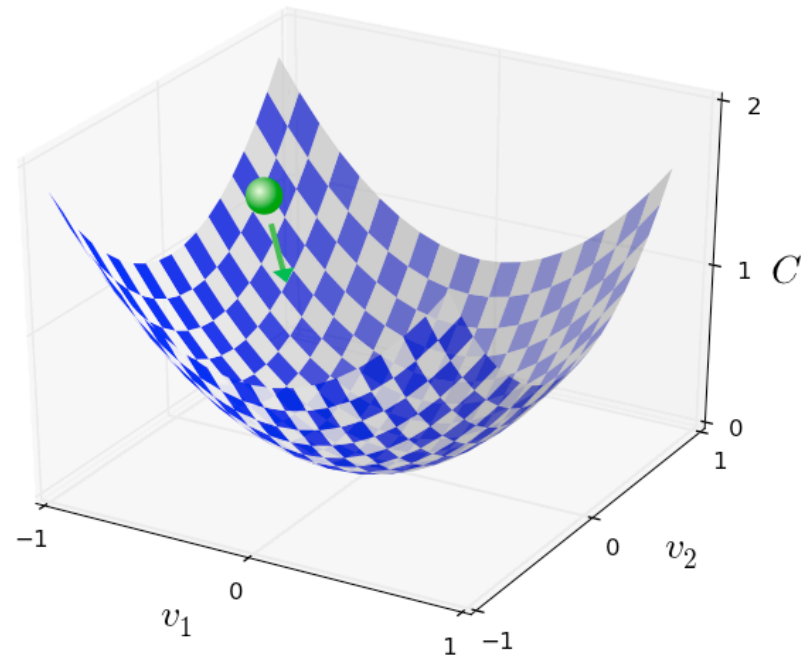


Then the main problem becomes finding the gradient of the cost function.

Active area of research.

Back-propagation algorithms.

We are at the eve of a technological
revolution: deep machine learning.

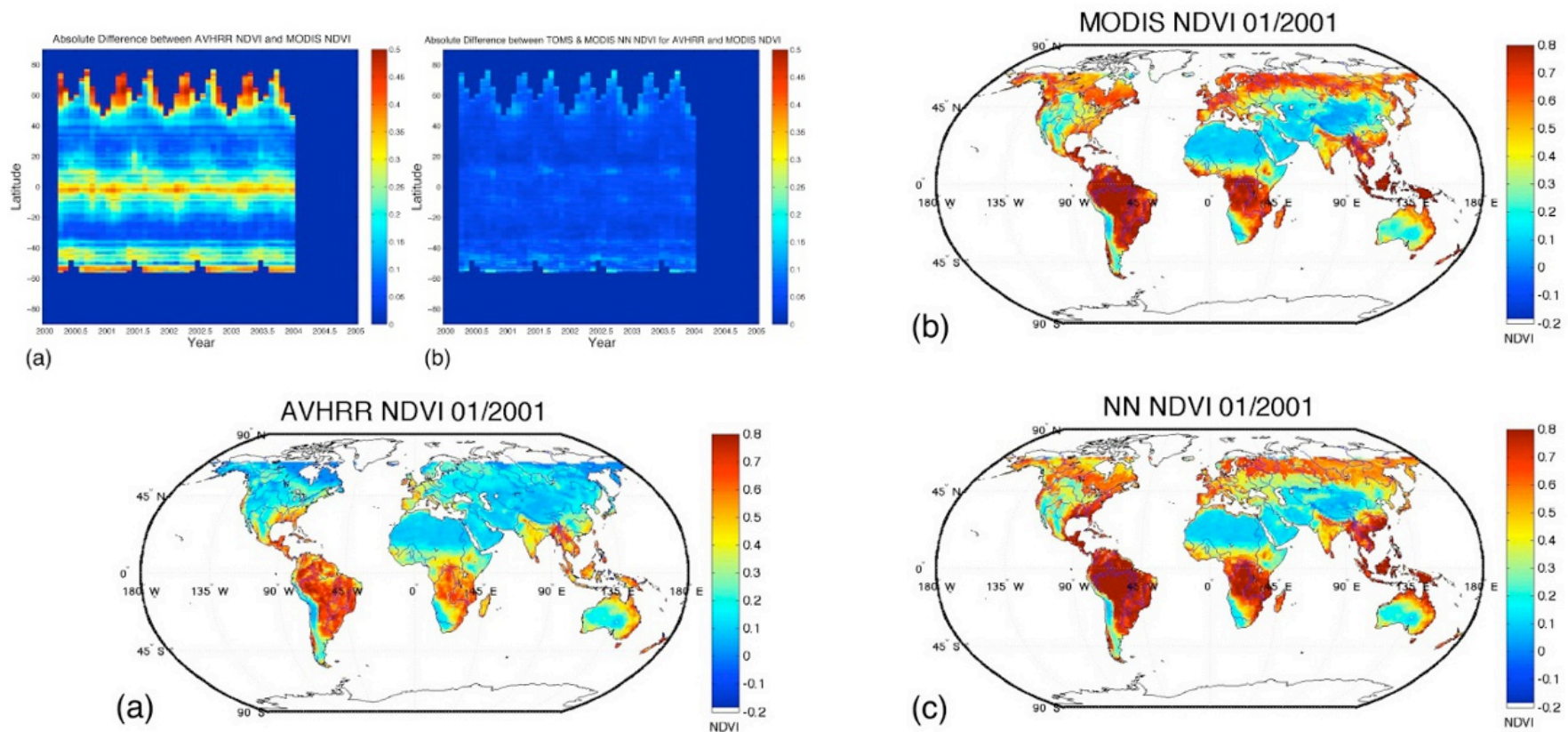


Geophysical applications

- 1) Bias correction: comparison between products of different instruments
- 2) Proxy evaluation: inferring the value of an observable from another with a complex relation.
- 3) Codes acceleration. Substitute physically based schemes (Ex. The radiative scheme)

Geophysical applications

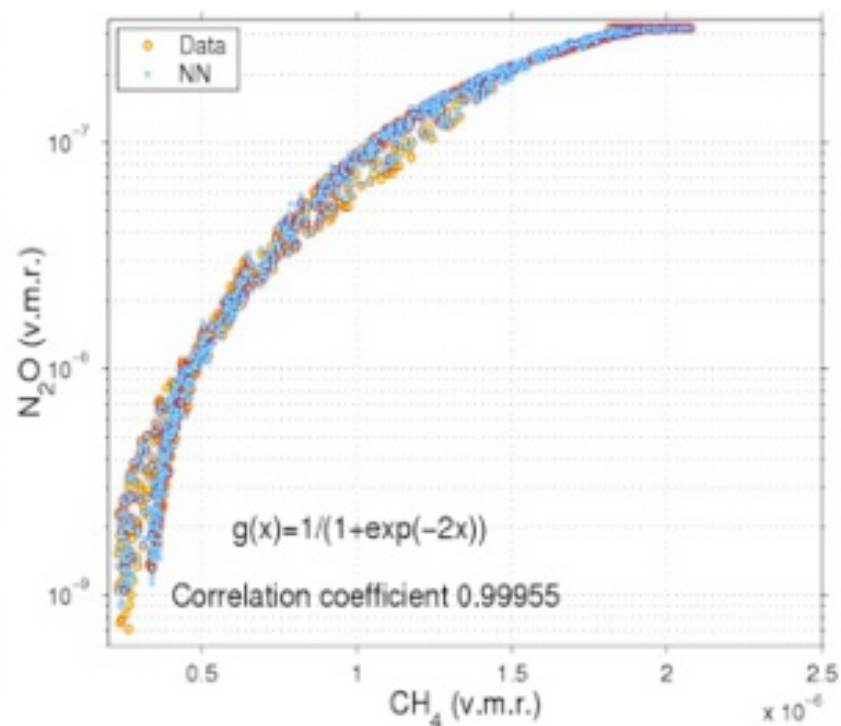
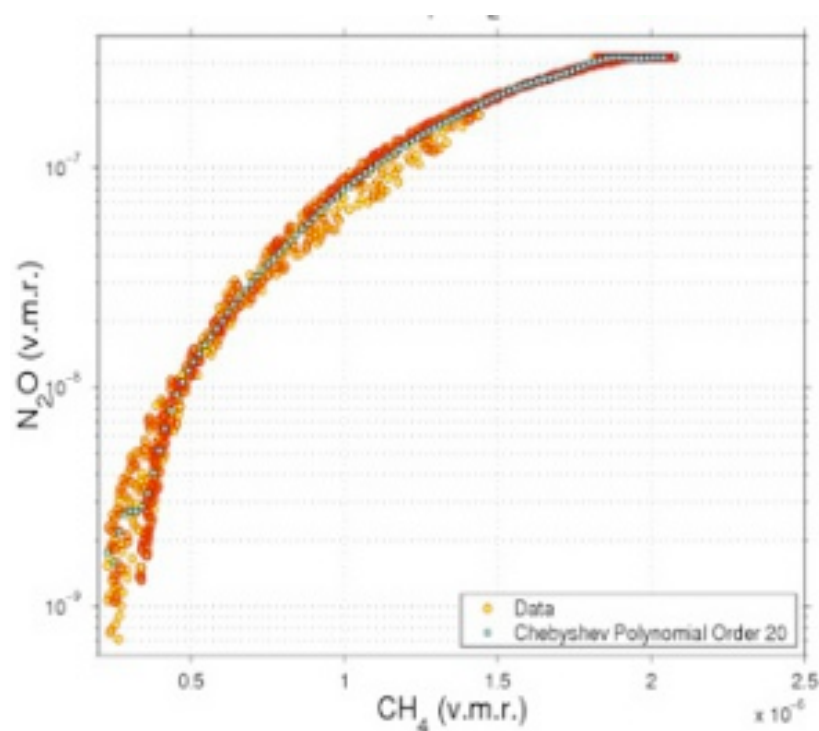
- 1) Bias correction: comparison between products of different instruments
- 2) Proxy evaluation: inferring the value of an observable from another with a complex relation.
- 3) Codes acceleration. Substitute physically based schemes (Ex. The radiative scheme)



NDVI from AVHR (panel a), MODIS (panel b)

Geophysical applications

- 1) Bias correction: comparison between products of different instruments
- 2) Proxy evaluation: inferring the value of an observable from another with a complex relation.
- 3) Codes acceleration. Substitute physically based schemes (Ex. The radiative scheme)



global N_2O - CH_4 relation