Clustering algorithms applied to Air Quality data

Part II: stations
Pollutant: PM10

Maj 2nd, 2011

CRMA – Regional Center for Environmental Modeling
ARPA FVG
Palmanova - Italy

Francesco Montanari, francesco.montanari@arpa.fvg.it

Questions

Zonization (DLgs 155/2010) is based on **D**eterminants (**D**PSIR): orography, urbanization, micro-climatology…

Questions:

1. Do stations placed in the same Zone show the “same” data?

2. … or can we recognize **I**mpacts due to specific **P**ressures?

(so that stations should represent specific **Areas** inside **Zones**…)

3. Do stations placed in different Zones show “different” data?

Data availability

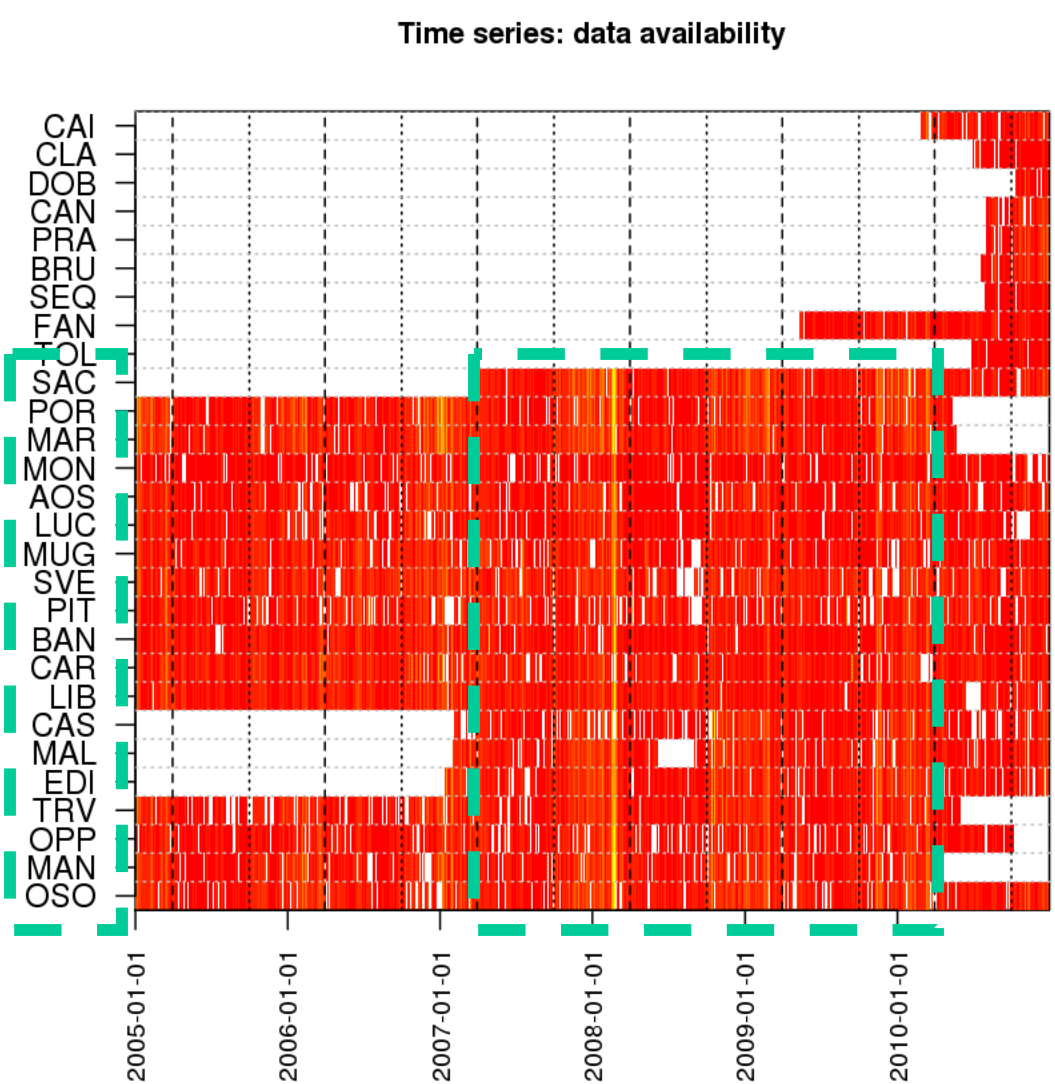
Query on `datiaria` server from LINUX cluster `nexus`

19 stations

3 years

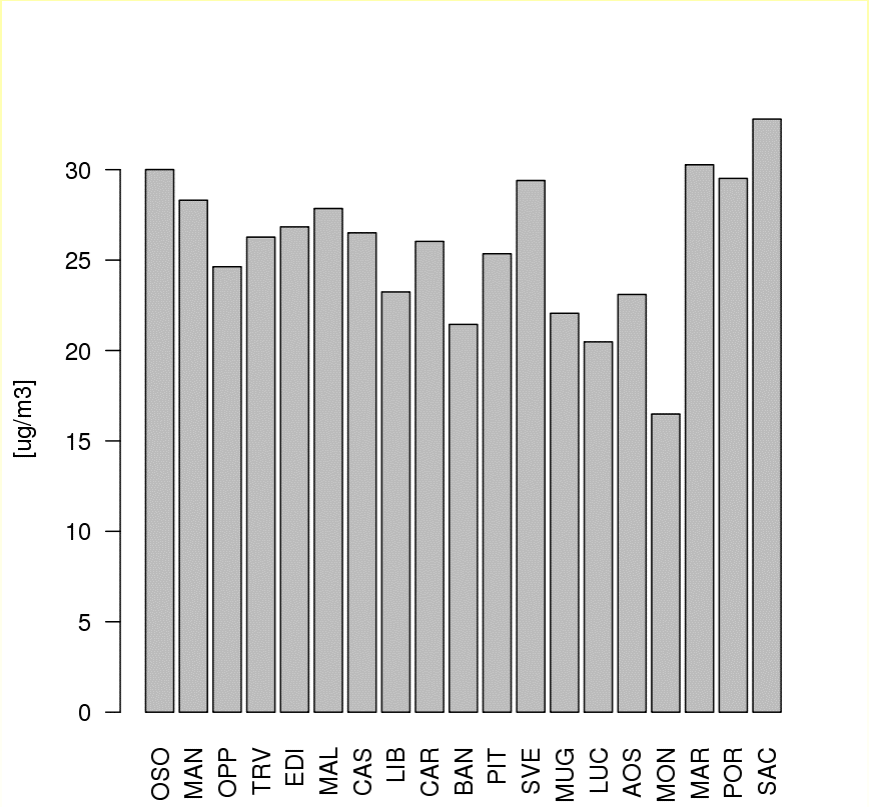
31.03.2007

31.03.2010

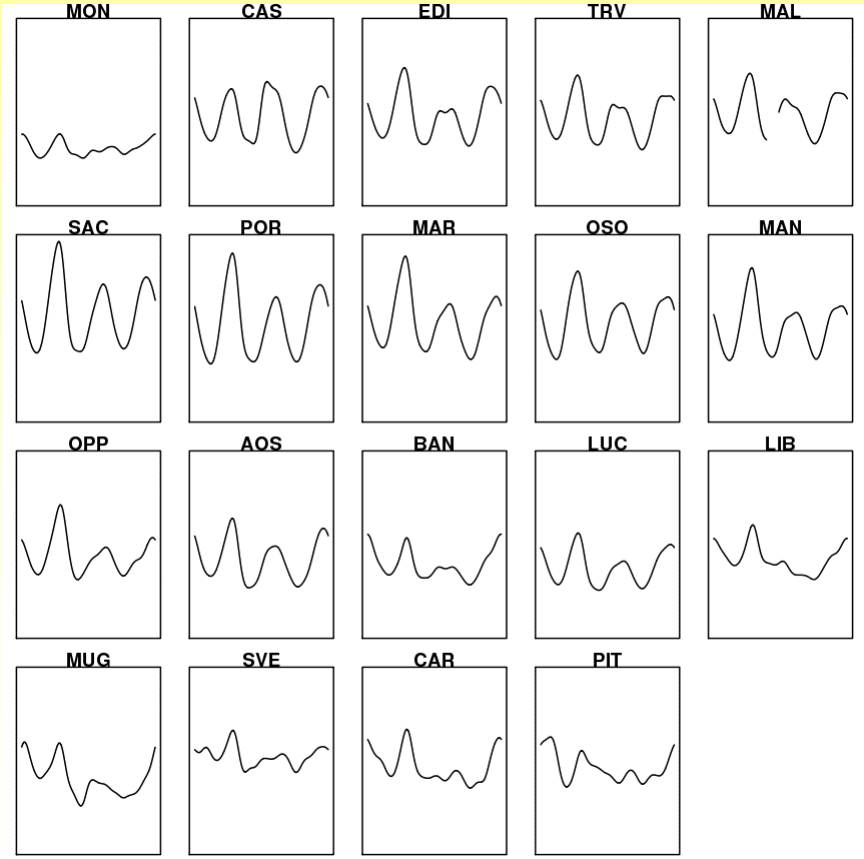


Temporal components

Baseline f_0 (mean value)



Inter-annual variation $f_{season}(d)$ (convolution... aka *weighted moving average*)



$$f_{season}(d) = (f * n)(d) = \int f(d-k)n(k)dk$$

$$n(d) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{d^2}{2\sigma^2}\right) \; ; \; \sigma \approx 38 \text{ days } (\text{FWHM} = 90 \text{ days})$$

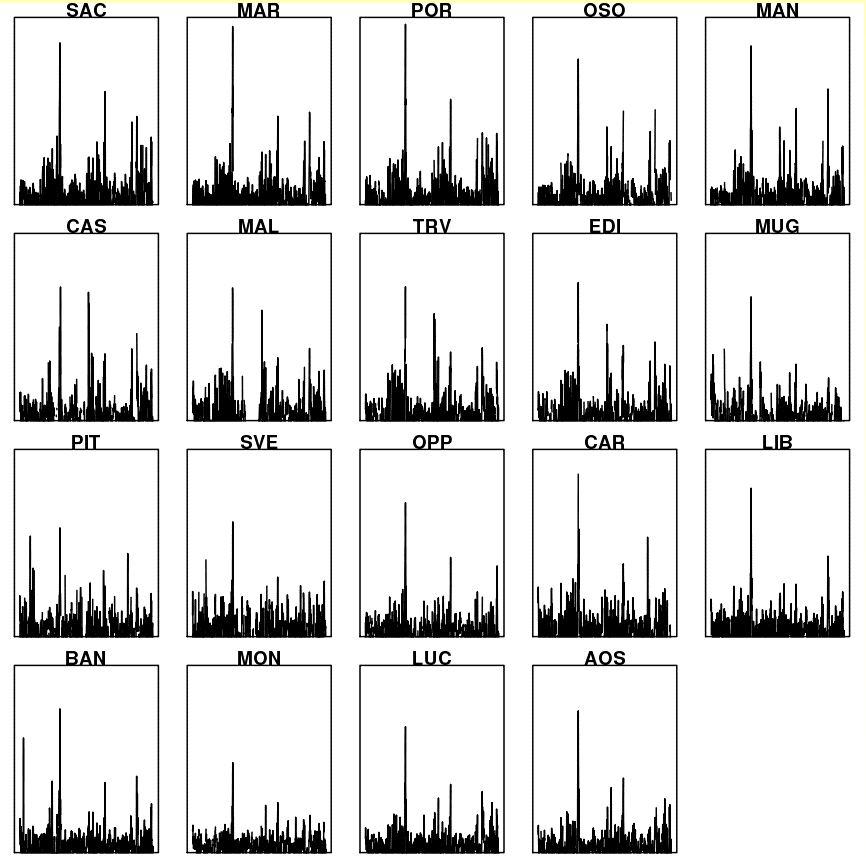
Temporal components (II)

Baseline f_0 and inter-annual f_{season}:

- Micro-climate and basin circulation
- Sources (heating, traffic, industries... maybe)

Rest series $r = f - f_{season}$

- Weather
- Sources (industries... maybe)



Hierarchical clustering

Grouping stations based on "similarity" in time series:

-19 rows (stations = *cases*) x 1096 columns (daily means = *fields*)

- Box Cox transforms (e.g. log), if desired…

- *distance* between cases (d_{ij} ; i =1,…,19 ; j = 1,…,19)… many choices!

$$\text{e.g. euclidean} : d_{ij}^{\,2} = \sum_{n=1}^{1096} (x_i - x_j)^2$$

- *clustering method* in R[1] function `hclust()` ("average", "complete", "single"…)

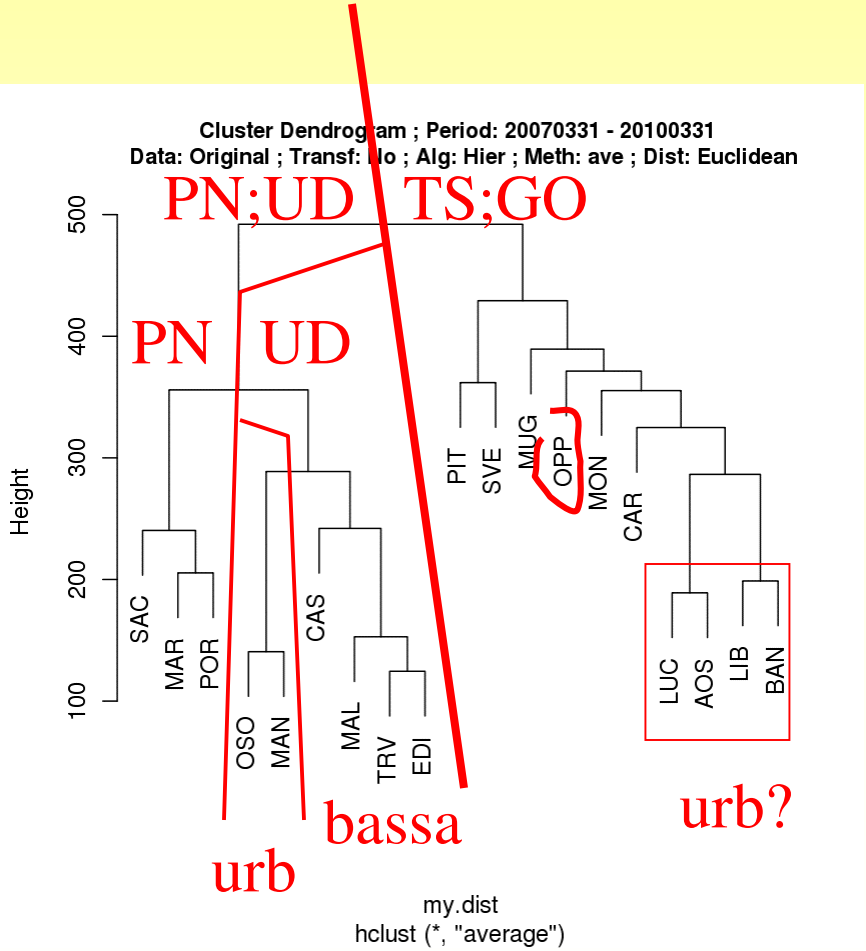
- problems due to *missing data*

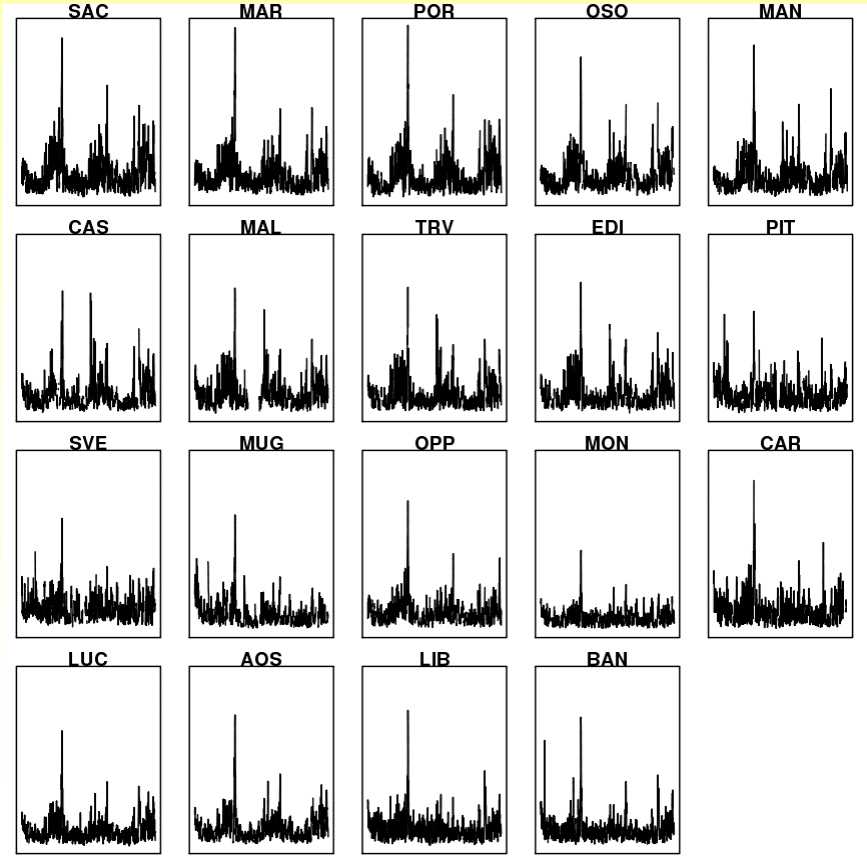
	Day 1	Day 2	…	Day 1096
Station 1	20.3	25.4	…	56.7
Station 2	22.2	23.6	…	89.5
…	21.5	26.2	…	78.2

[1] R Development Core Team (2007). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Hierarchical clustering: orginal data

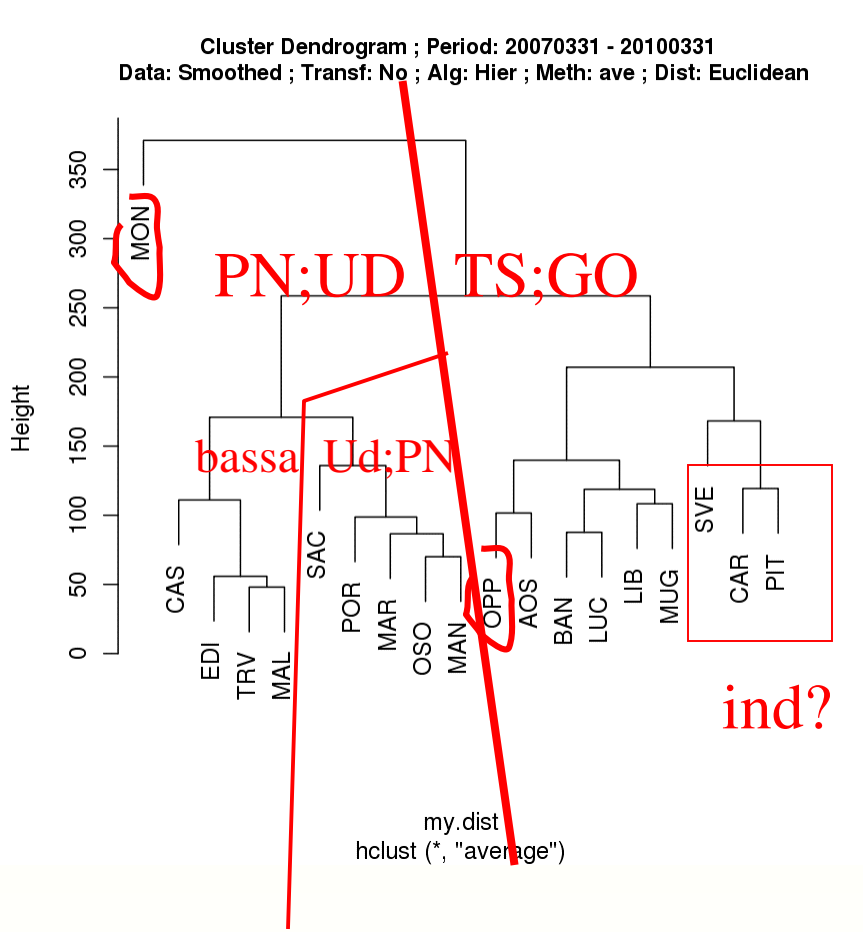
f(d):

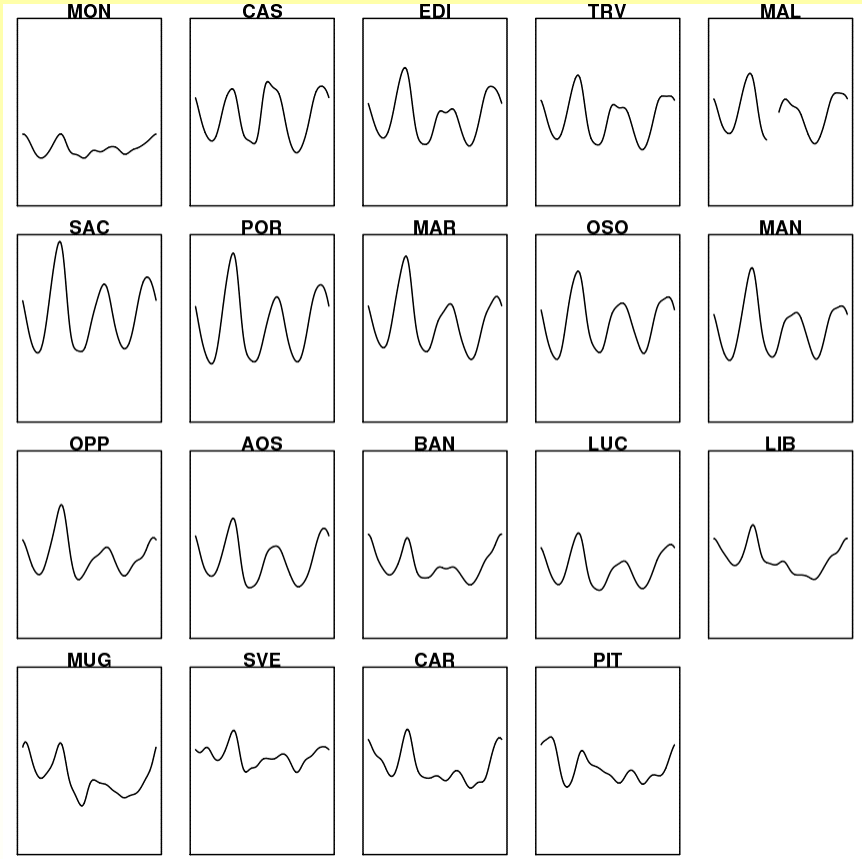




Hierarchical clustering: baseline + inter-annual variation

$f_{season}(d)$:

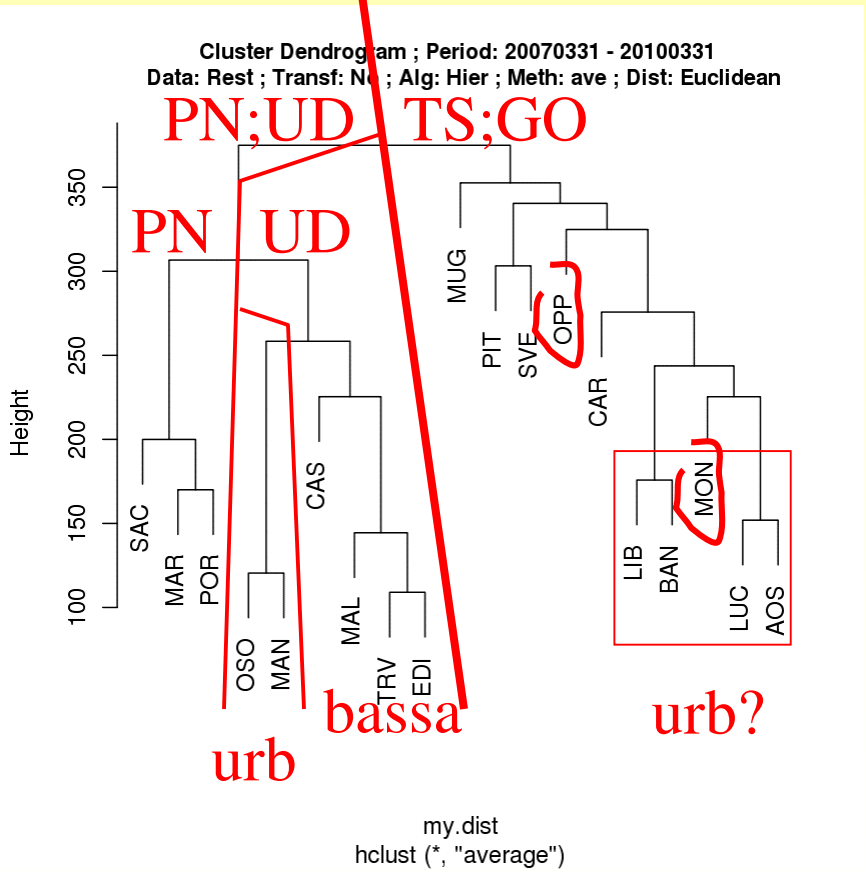


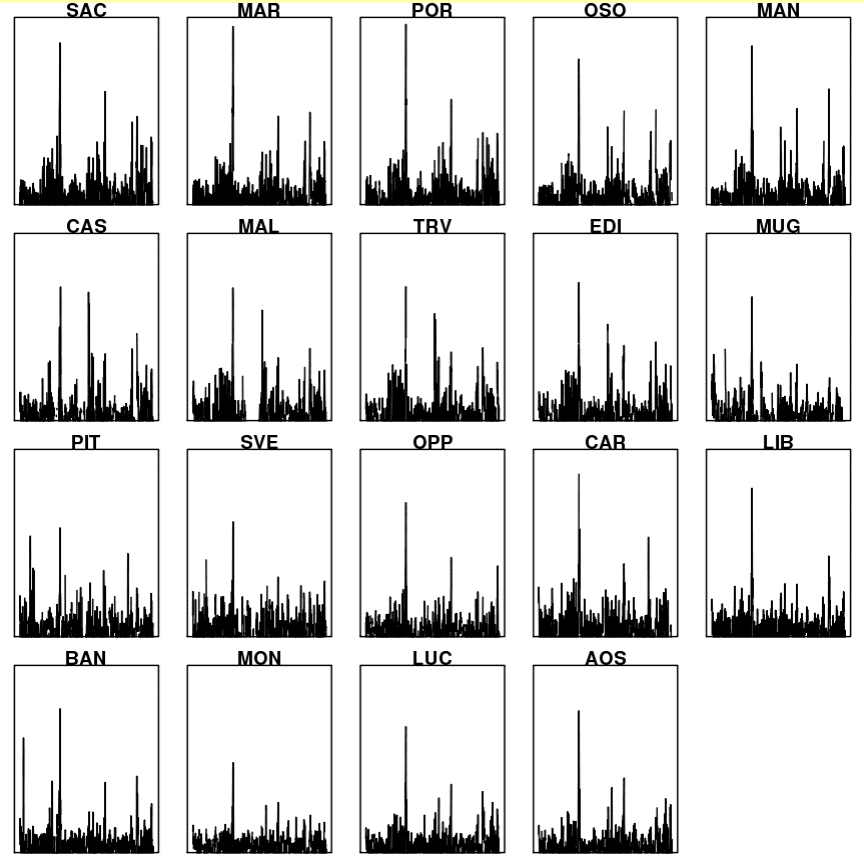


Hierarchical clustering: rest

$r = f - f_{season}$

- same dendrogram as original data
- MON is classified with GO and [LIB,BAN]

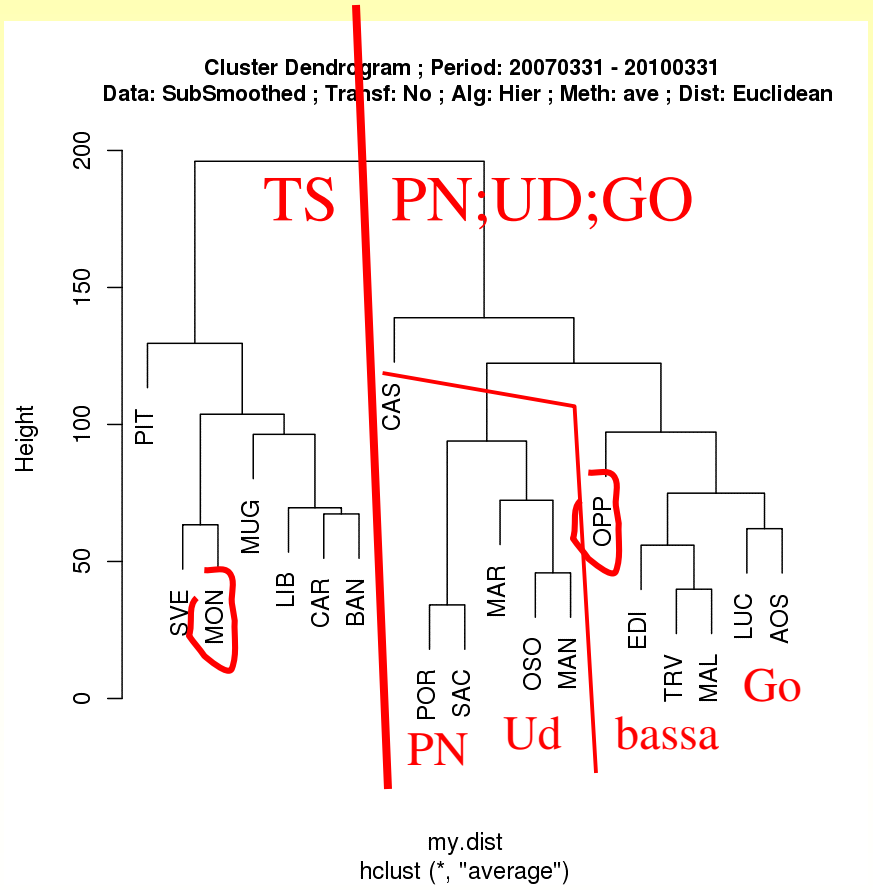


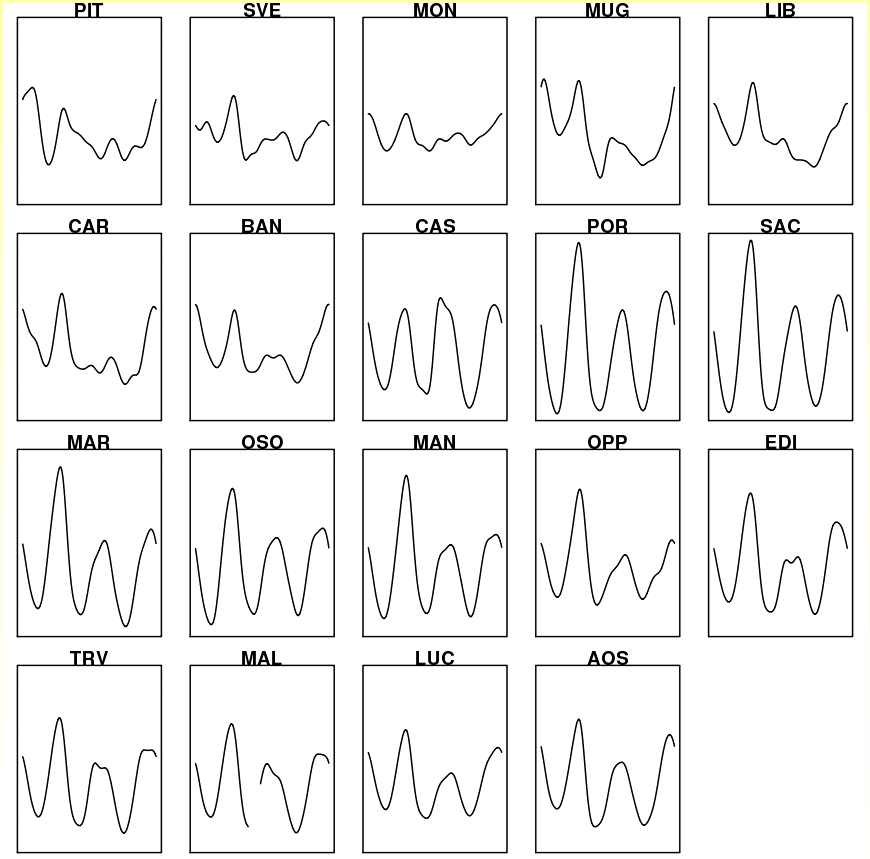


Hierarchical clustering: inter-annual variation

$f_{season} - f_0$

No specific sources, only micro-climate…(?):





Conclusions

- Original data (*f*) and rests (*r*) series: ~ same classification
  -> weight of the episodes
- PM10 decreases **from West to East**
- PM10 inter-annual variation decreases **from West to East**
- Inter-annual variation: LUC, AOS, EDI, TRV, MAL, OSO are grouped
- Monfalcone (**MON**):
  - very low level.
  - inter-annual variation similar to Trieste
- Osoppo (**OPP**): close to the South-Eastern area
- **r series**: LUC, AOS, MON, LIB, BAN are very similar
- (**SVE, PIT**) show a different behaviour than (**LIB, BAN**)