Clustering algorithms applied to Air Quality data

Part II: stations
Pollutant: PM10

May 2nd, 2011

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

Francesco Montanari, francesco.montanari@arpa.fvg.it
Zonization (DLgs 155/2010) is based on Determinants (DPSIR): orography, urbanization, micro-climatology…

Questions:

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

2. … or can we recognize Impacts due to specific Pressures?
   (so that stations should represent specific Areas inside Zones…)

3. Do stations placed in different Zones show “different” data?
Query on datiaria server from LINUX cluster nexus

19 stations
3 years
31.03.2007
31.03.2010

Data availability
Clustering algorithms applied to Air Quality data - PM10

Temporal components

Baseline $f_0$
(mean value)

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

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

$$n(d) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{d^2}{2\sigma^2}\right) \quad ; \quad \sigma = 38 \text{ days (FWHM = 90 days)}$$
Baseline $f_0$ and inter-annual $f_{\text{season}}$:
- Micro-climate and basin circulation
- Sources (heating, traffic, industries... maybe)

Rest series $r = f - f_{\text{season}}$
- Weather
- Sources (industries... maybe)
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 \(d_{ij}; i = 1, \ldots, 19; j = 1, \ldots, 19\)… many choices!

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

- clustering method in \(R^1\) function hclust() (“average”, “complete”, “single”…)
- problems due to missing data

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 2</th>
<th>…</th>
<th>Day 1096</th>
</tr>
</thead>
<tbody>
<tr>
<td>Station 1</td>
<td>20.3</td>
<td>25.4</td>
<td>…</td>
<td>56.7</td>
</tr>
<tr>
<td>Station 2</td>
<td>22.2</td>
<td>23.6</td>
<td>…</td>
<td>89.5</td>
</tr>
<tr>
<td>…</td>
<td>21.5</td>
<td>26.2</td>
<td>…</td>
<td>78.2</td>
</tr>
</tbody>
</table>

f(d):

Cluster Dendrogram; Period: 20070331 - 20100331
Data: Original; Transf: Jo; Alg: Hier; Meth: ave; Dist: Euclidean

PN; UD  TS; GO
PN  UD
urb bassa
urb?
Hierarchical clustering: baseline + inter-annual variation

\[ f_{season}(d) : \]
\[ r = f - f_{\text{season}} \]
- same dendrogram as original data
- MON is classified with GO and [LIB,BAN]
Clustering algorithms applied to Air Quality data - PM10

Hierarchical clustering: inter-annual variation

\[ f_{season} - f_0 \]

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

Cluster Dendrogram; Period: 20070331 - 20100331
Data: SubSmoothed; Transf: No; Alg: Hier; Meth: ave; Dist: Euclidean

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