Supporto alla zonizzazione ed alla razionalizzazione della rete con algoritmi di clustering.
Parte II: le stazioni di rilevamento
Inquinante: PM10

CRMA

April 29, 2011

'If you torture your data long enough,
they will tell you whatever you want to hear'

James L. Mills[1], 1993
National Institute of Child Health
and Human Development
Bethesda, MD 20892

1 Abstract

Vengono prese in considerazione le serie temporali delle concentrazioni di PM_{10} misurate dalle Stazioni della Rete di Rilevamento Regionale nel periodo 2007-2010.

Le Stazioni vengono *raggruppate* (*clustering*) da un algoritmo automatico, sulla base della *similitudine* degli andamenti delle rispettive serie temporali.

La classificazione viene ripetuta sulle serie di dati *filtrati* (*smoothed*), che evidenziano l'andamento interannuale di ciascuna serie, e sui *residui* (risposta impulsiva alle varibili meteorologiche).

Se ne ricava che:

- la concentrazione di PM10 decresce sostanzialmente da Ovest verso Est (Gorizia). E' minima a Monfalcone, ma risale a Trieste: maggior effetto del ciglione carsico? delle sorgenti antropiche?
- l'intensità della variazione stagionale del PM10 decresce anch'essa da Ovest verso Est. Monfalcone si assimila a Trieste, più che a Gorizia.

[1]Mills J.L., *Data Torturing*, N Engl J Med, 1993; 329: 1196-1199

- in questi due effetti: stiamo vedendo il *bordo* del Bacino Padano? l'effetto del diverso consumo invernale della legna? l'effetto della bora o delle brezze?
- Monfalcone (MON) presenta concentrazioni medie bassissime e ridottissima variazione stagionale (in quest'ultimo aspetto, si assimila a Trieste). La risposta impulsiva è assimilabile a quella delle stazioni di Gorizia e delle sole [LIB,BAN] a Trieste.
- OPP (Osoppo) ha una *baseline* bassa, un andamento inter-annuale assimilabile a quello della Bassa Friulana e di Gorizia, una risposta impulsiva a se stante.
- la risposta impulsiva r (sottratte la media e la variazione interannuale) è molto simile a Gorizia, Monfalcone, LIB (Trieste p.zza Libertà) e BAN (Trieste, Tor Bandena). Meno nelle altre stazioni triestine: impatto dell'industria?
- da ogni punto di vista, SVE (Trieste, v. Svevo), CAR (Trieste, v. del Carpineto), PIT (Trieste, v.Pitacco), MUG (Muggia) sono classificate a parte rispetto a LIB e BAN.
- isolando la componente interannuale, tolta anche la media o *baseline* ($f_{smootehd} - \overline{f}$):
  - l'area triestina si ricompone (anche con Monfalcone)
  - Gorizia viene associata alla Bassa Friulana (e con OPP)
  - Udine Città è raggruppata a Pordenone

2 Introduzione

Si effettua una valutazione della Zonizzazione e della rappresentatività delle stazioni della Rete di Rilevamento Regionale per mezzo di opportuni algoritmi di ***clustering***.

Le domande cui si intende rispondere sono:

- le Zone individuate per il Piano Regionale di Miglioramento della Qualità dell'Aria, sulla base di orografia - microclimatologia - urbanizzazione (Determinanti), hanno un riscontro nei dati di Qualità dell'Aria? Vale a dire: stazioni inserite in una medesima zona, originano dati maggiormente correlati fra loro?
- viceversa: vi sono evidenze di stazioni collocate in Aree soggette a Pressioni specifiche, che le rendono non rappresentative per l'intera Zona?

Ne deve derivare un indirizzo per la razionalizzazione della Rete, complementare a quello ottenuto dalla Parte I del presente Studio, basato sulle simulazioni eseguite con il modello fotochimico.

L'inquinante qui considerato è il PM10, il parametro la media giornaliera.

L'idea di base è quella di considerare, per ciascuna delle n stazioni di rilevamento del PM10, gli m valori del parametro scelto, costruendo una matrice di n **righe** e m **colonne**.

Ciascuna riga è un *caso*; se si considera il periodo che va dal 2005-01-01 al 2010-12-31, esso sarà descritto da 365x6+1=2191 (il 2008 è bisestile) *qualificatori* o *campi*.

I casi saranno più o meno simili fra loro, sulla base di criteri che si vorranno assumere.

Sulla base di tali *similitudini* e criteri, i casi possono certamente essere raggruppati (*clustering*).

Come già realizzato per l'analisi di correlazione sulle centraline [5], è possibile ripetere l'analisi sulle diverse componenti del *segnale*: quella a bassa frequenza (andamento inter-annuale: climatica) o ad alta frequenza (short term: meteorologica).

Una valutazione definitiva sulle correlazioni si può avere da opportuni modelli di regressione: qualora i rilievi di una centralina siano *prevedibili*, con una determinata incertezza, dati i rilievi di una o più altre stazioni ed opportune grandezze meteorologiche.

3 Estrazione dati ed ambiente informatico

Il lavoro viene svolto sul cluster di calcolo del CRMA, `nexus`.

I dati sono estratti con opportune query sul server `aria` dell'Agenzia.

Le successive analisi sono gestite con script di R [6].

4 Analisi preliminare dati

Dalla Fig.1 si desumono i periodi di disponibilità dei dati per le stazioni che hanno misurato il PM10 nel sessennio 2005 - 2010.

In particolare, per 19 stazioni si dispone di serie storiche che coprono il triennio 2007 - 2009.

Un set diverso, di 23 stazioni, costituisce la rete PM10 attuale, per la quale i dati sono disponibili indicativamente da luglio 2010 a marzo 2011.

5 Componenti temporali

Si ritiene di individuare, nei segnali, una componente temporale *interannuale*, o a bassa frequenza, o microclimatica, ed una componente *impulsiva*, meteorologica. Nel caso delle stazioni industriali, tale componente può avere anche origine antropica.

In analogia a quanto fatto in [5], la componente a bassa frequenza viene isolata convolvendo le serie storiche originali con un *kernel* gaussiano di larghezza a metà altezza (Full Width Half Maximum) pari a 90 giorni:

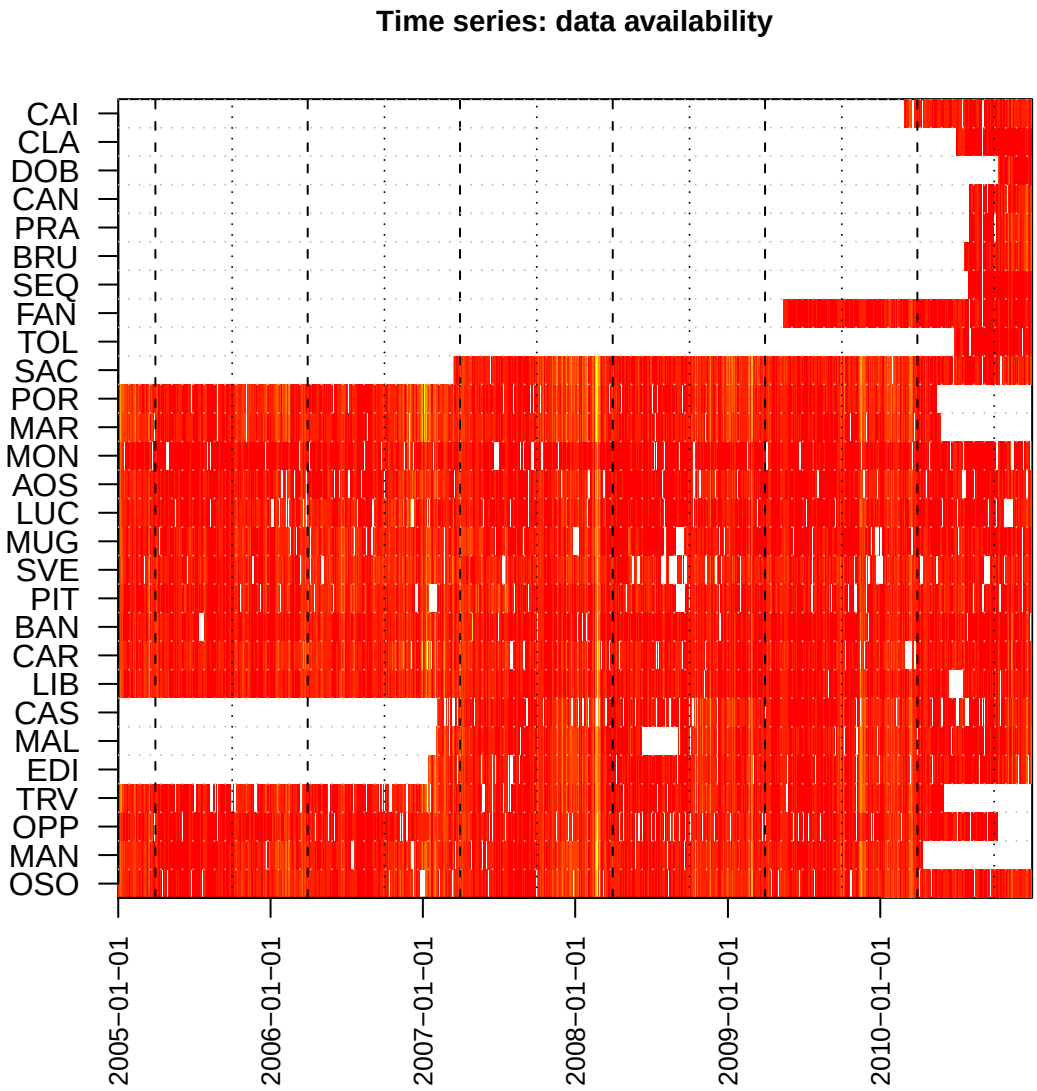

Figure 1: Available time series: length and quality. Red-yellow scale is related to the PM10 concentration; white spaces correspond to missing values. Dashed lines refer to the cold season (October 1st - March 31st)

$$
\begin{aligned}
FWHM &= 90[days] \\
\sigma = \frac{FWHM}{2\sqrt{2\log(2)}} &\approx \frac{90[days]}{2.35} \approx 38[days] \\
g(d) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{d^2}{2\sigma^2}\right) \\
f_{season}(d) = (f * g)(d) &= \sum_n f(n) g(d-n) \\
r(d) &= f(d) - f_{season}(d) \qquad (1)
\end{aligned}
$$

dove con $f(d)$ si è indicata la concentrazione di PM10 nel giorno d, con $f_season()$ la componente *interannuale* del segnale f e con r la componente *impulsiva* (r sta per *resto*).

Una particolare attenzione ha richiesto l'esecuzione della convoluzione:

- in presenza di dati mancanti
- nel trattamento degli estremi della serie temporale

La più semplice classificazione delle centraline può basarsi, ad esempio, sul *numero medio di superamenti annui* nel sessennio considerato: in tal caso, si sta confrontando la componente temporale a frequenza nulla fra i segnali (*baseline*).

Un passo successivo può farsi considerando il *numero di superamenti annui in ciascun anno* del sessennio considerato: in tal caso, si sta sostanzialmente confrontando il *trend* misurato nelle varie stazioni.

La scomposizione del segnale proposta mira a consentire l'individuazione di comportamenti simili dettati da fattori microclimatici (f_{season}) o meteorologici (r).

6 Classificazione

In Fig.3 e 4 sono riportati i dendrogrammi di classificazione gerarchica delle stazioni sulla base di f, f_{season} ed r, per il triennio che va dal 1 aprile 2007 al 31 marzo 2010.

In particolare, si è utilizzata la funzione `hclust(,method="average")` di `R` [6]; la distanza fra le serie storiche è quella euclidea, calcolata per mezzo della funzione `dist()` [6]; i dati non sono stati sottoposti a trasformazioni (es. logaritmica).

f ed r originano praticamente la medesima classificazione. L'algoritmo separa, al primo nodo, [Udine + Pordenone] da [Trieste + Gorizia].

Nei passaggi successivi, viene isolato il Pordenonese e quindi distinta la Bassa pianura friulana dal capoluogo.

Nella Venezia Giulia, SVE e PIT sono classificate a se stanti; AOS e LUC sono classificate simili a LIB e BAN.

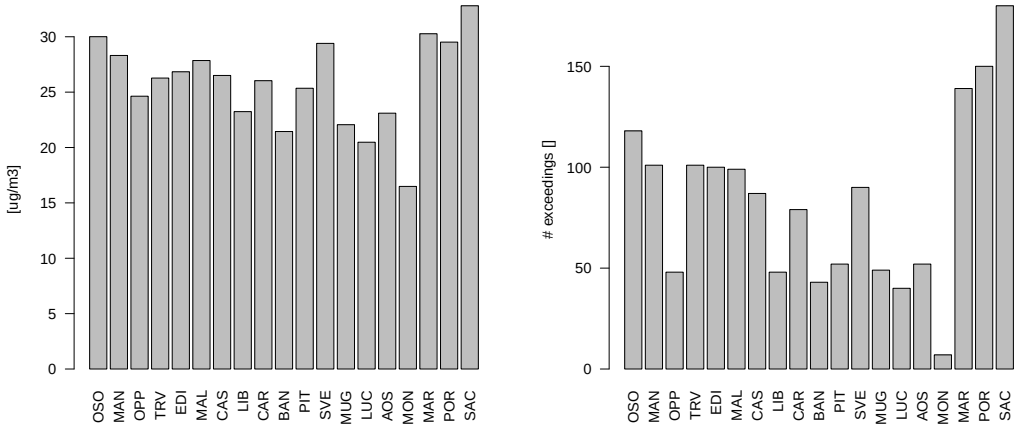

Figure 2: PM10 average and $50\mu g$ exceedings (2007-03-31 ; 2010-03-31)

Sfuggono a questo schema le stazioni di OPP, MON, CAR, MUG.
f_{season} origina una classificazione simile nelle divisioni principali.

```
 |-[UDINE,PORDENONE]
 |         |
 |         |-PORDENONE
 |         |
 |         |-UDINE
-|             |
 |             |-CITTA'
 |             |-BASSA FRIULANA
 |
 |-[TRIESTE,GORIZIA]
           |
           |-[LIB,BAN; AOS,LUC]
           |
           |-[SVE,PIT]
```

I *valori estremi* appaiono dunque avere un grosso peso nel calcolo della *distanza* fra le serie temporali: ciò è dovuto alla distribuzione tipica dei dati ed all'utilizzo della distanza Euclidea ai fini della classificazione.

Si esegue anche una classificazione delle serie temporali $f_{smoothed} - \overline{f}$, essendo $\overline{f}$ la media della serie temporale: si sottrae, cioè, la *baseline*.

Il passaggio al logaritmo riporta le distribuzioni dei dati a curve più vicine ad una normale, limitando il peso delle code (Fig.7)

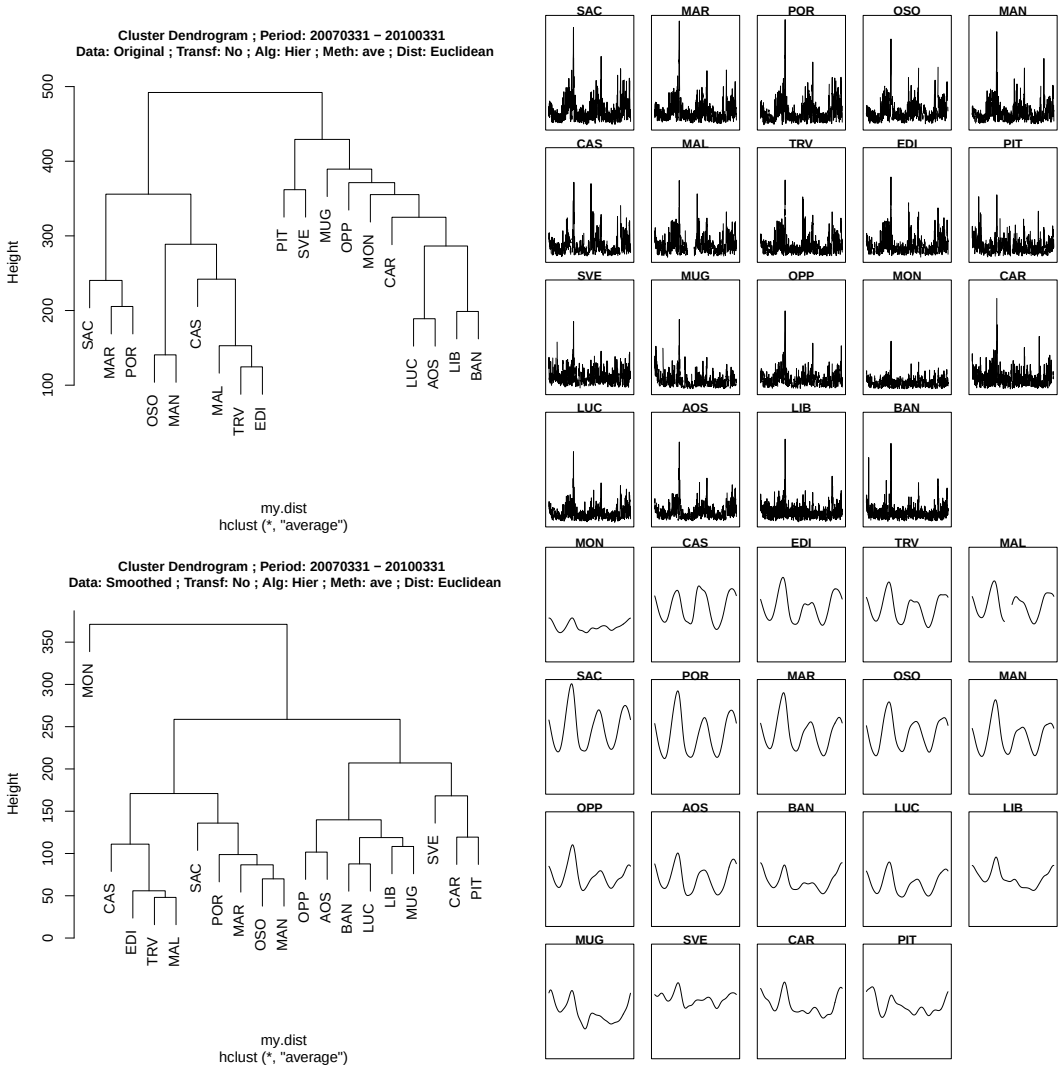

Figure 3: PM10 daily means (2007-03-31 ; 2010-03-31), no Box-Cox transform, Hierarchical Clustering, Euclidean Distance; Original and smoothed time series

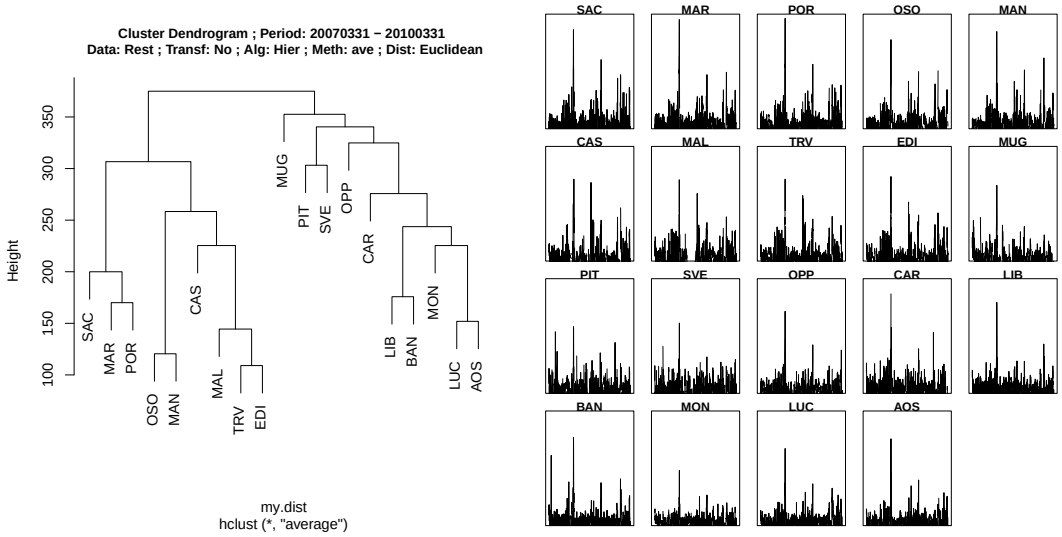

Figure 4: PM10 daily means (2007-03-31 ; 2010-03-31), no Box-Cox transform, Hierarchical Clustering, Euclidean Distance; Rest time series (Original data minus smoothed time series)

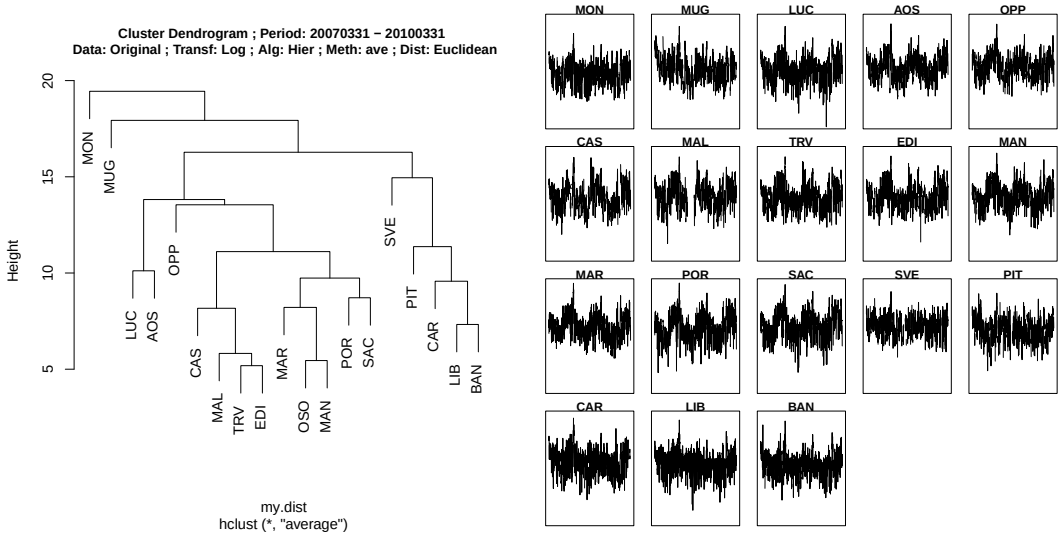

Figure 5: PM10 daily means (2007-03-31 ; 2010-03-31), Log transform, Hierarchical Clustering, Euclidean Distance; Original time series

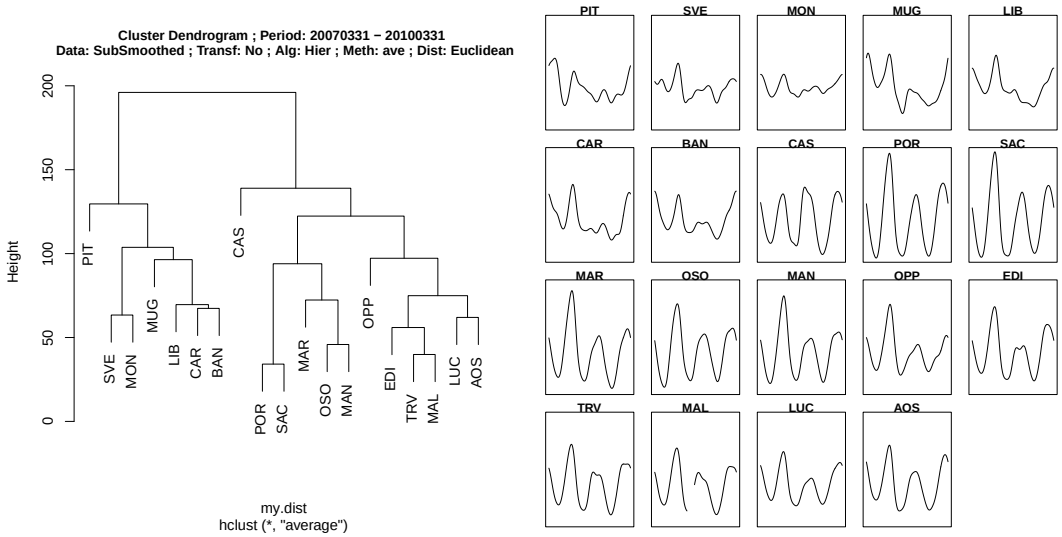

Figure 6: PM10 daily means (2007-03-31 ; 2010-03-31), No transform, Hierarchical Clustering, Euclidean Distance; Smoothed time series and baseline subtraction

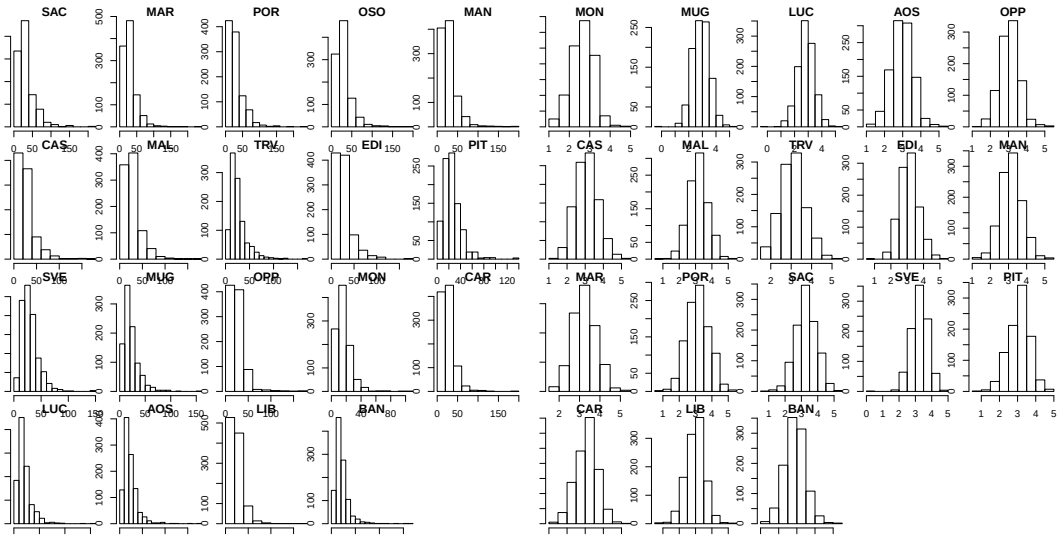

Figure 7: PM10 daily means (2007-03-31 ; 2010-03-31), Original and Log-transformed data

Dai dati trasformati per mezzo della funzione logaritmo, si ottiene invece una classificazione diversa:

```
 |-[UDINE,PORDENONE,GORIZIA]
 |         |
 |         |-GORIZIA
 |         |
-|         |-[UDINE,PORDENONE]
 |             |
 |             |-[UDINE CITTA',PORDENONE]
 |             |-BASSA FRIULANA
 |
 |-[TRIESTE]
          |
          |-[LIB,CAR]
```

Dalla serie temporale $f_{smoothed} - \overline{f}$, in cui si isola il comportamento interannuale delle stazioni, eliminando anche la *baseline*:

```
 |-[UDINE,PORDENONE,GORIZIA]
 |         |
 |         |-[GORIZIA, BASSA FRIULANA]
 |         |
-|         |-[UDINE CITTA',PORDENONE]
 |
 |-[TRIESTE,MONFALCONE]
```

7 Successivi sviluppi

Si individuano alcune direttrici per la prosecuzione dell'analisi dei dati:

- clustering degli episodi: date le serie storiche di PM10 in n stazioni ed opportune variabili meteo di m stazioni (vento-pioggia-temperatura), cluster analysis dei "'giorni"'. Si andrebbe a costruire un atlante regionale dell'inquinamento da polveri, con le situazioni tipiche...
- modelli stazione-stazione: date le serie storiche di 2 stazioni PM10 ed opportune variabili meteorologiche, sviluppare un modello (lineare, albero, rete neurale...) di previsione per una delle due stazioni. Una buona prestazione di un modello di questo tipo consente di estendere, al di sopra di ogni dubbio, la rappresentatività di una stazione fino all'area monitorata da una seconda stazione...
- metrica[9] basata sul *rango* dei dati (coefficienti di correlazione di Spearman[10] o Kendall): sarebbe estremamente interessante, ovviando i problemi di normalizzazione dei dati. A causa dei dati mancanti, la *distanza* fra le

serie temporali viene calcolata *a coppie* (problema già affrontato in [5]); da valutare l'introduzione di una metrica basata sul rango (*rank*) dei dati che aggiri questo problema.

- calcolo degli Spettri di Wiener. Motivo di interesse: individuare fattori antropici (picchi a 7gg = per lo più traffico). Problema: i dati mancanti, per la trasformata di Fourier.

8 PM10 e traffico veicolare

La prossimità alle strade è generalmente percepita come la principale causa di esposizione al PM_{10}, assieme alla prossimità di impianti industriali.

Nei grafici che seguono vengono proposti gli andamenti giornalieri estivi ed invernali della concentrazione di PM_{10}, CO, NO, NO_2 ed O_3, calcolati come *mediana* dei valori misurati in ciascuna ora del giorno nei periodi considerati (2007 - 2009).

La stagione estiva ed invernale sono definite sulla base dei giorni di passaggio fra ora legale e solare.

I grafici proposti vengono esaminati con maggior dettaglio nel documento dedicato agli NOx.

Si ritiene di identificare il *segnale* dovuto al traffico nella struttura giornaliera a 2 picchi della concentrazione di CO, NO ed NO_2, che si manifesta nelle stazioni da traffico e - in modo meno marcato - nelle stazioni di background urbano e suburbano.

Difficile dicriminare l'effetto del riscaldamento domestico rispetto al traffico; difficile riconoscerlo nelle differenze estate-inverno, a causa dell'effetto della maggiore altezza di rimescolamento.

L'effetto della crescita dell'altezza di rimescolamento durante le ore diurne è ben visibile negli andamenti del $PM10$: *il coperchio della pentola si alza e si abbassa*, concentrando o diluendo l'inquinante; l'effetto diretto delle emissioni dagli assi viari appare essere modesto, rispetto a tale dinamica principale.

Viceversa, per il CO e gli NOx sono le forti emissioni nelle ore di maggior traffico a determinare la principale dinamica del segnale.

Nel caso degli NO è evidente l'effetto delle interazioni con l'O_3.

Una stima indiretta dell'effetto diretto della prossimità di un'arteria stradale si può evere confrontando gli andamenti di OSO rispetto alle altre stazioni dell'udinese: MAN, in particolare, ed EDI, MAL, CAS, TRV, OPP.

9 Conclusioni

La chiave è: come cambia il *clustering* delle centraline a seconda delle diverse componenti temporali evidenziate.

- dati originali non trasformati (seire f) ed i "'resti"' (serie r) danno sostanzialmente la stessa classificazione: il nostro metodo (definizione di *distanza*

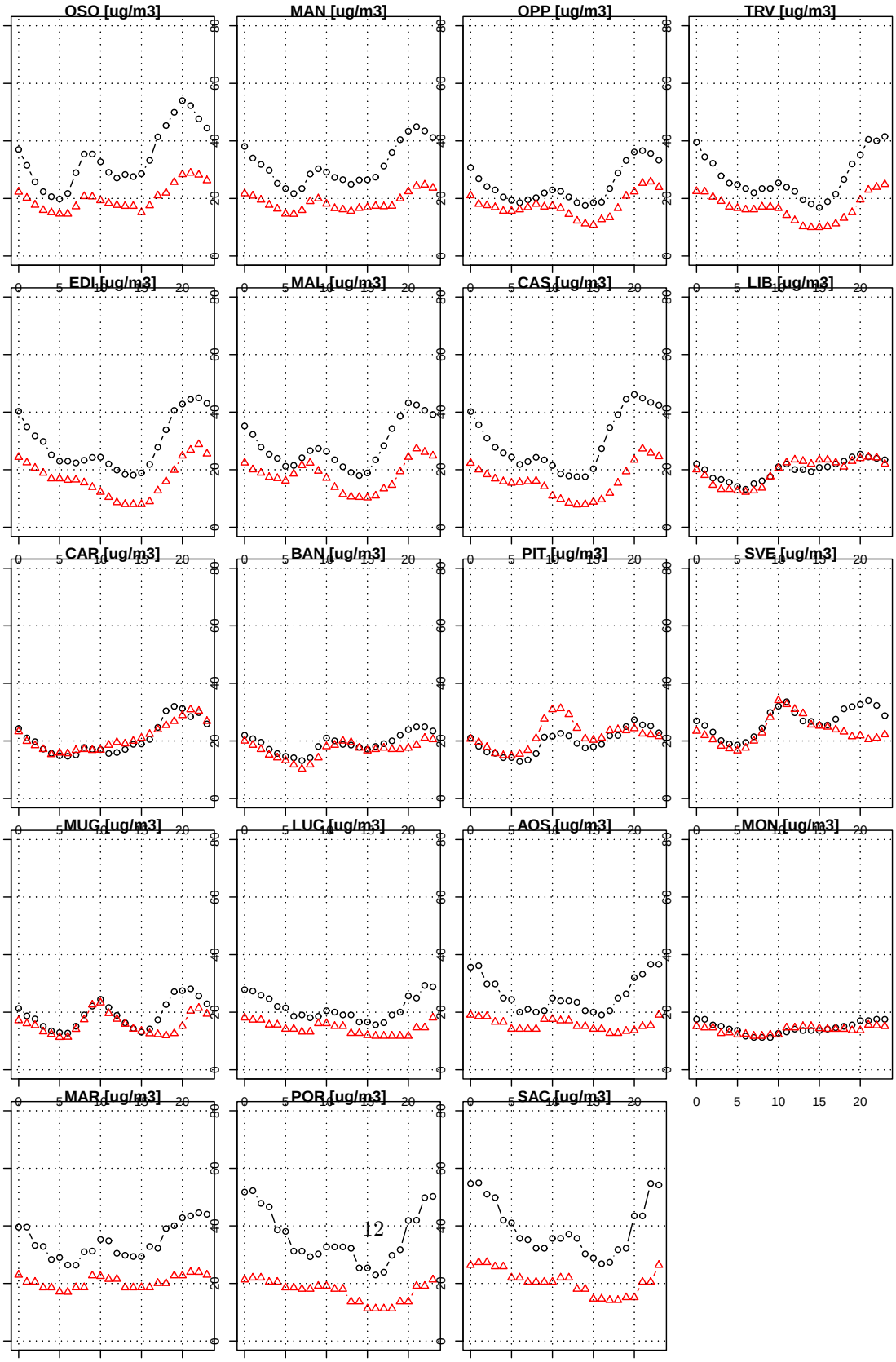

Figure 8: PM10 concentrations in $\mu g \cdot m^{-3}$, 2007 - 2009. Black circles represent winter median day; red triangles represent summer median day

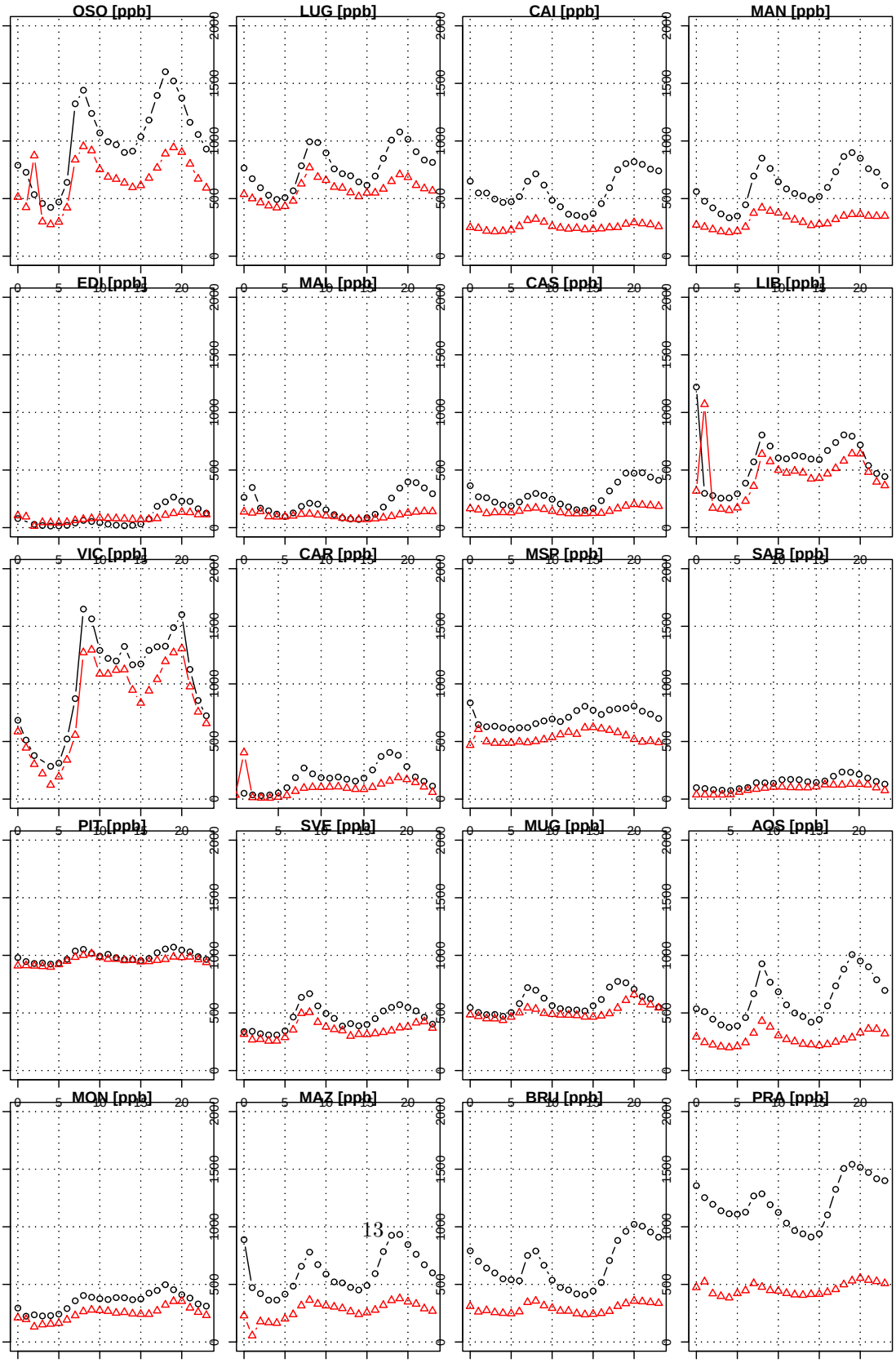

Figure 9: CO concentrations in [ppb], 2007-2009. Black circles represent winter median day; red triangles represent summer median day

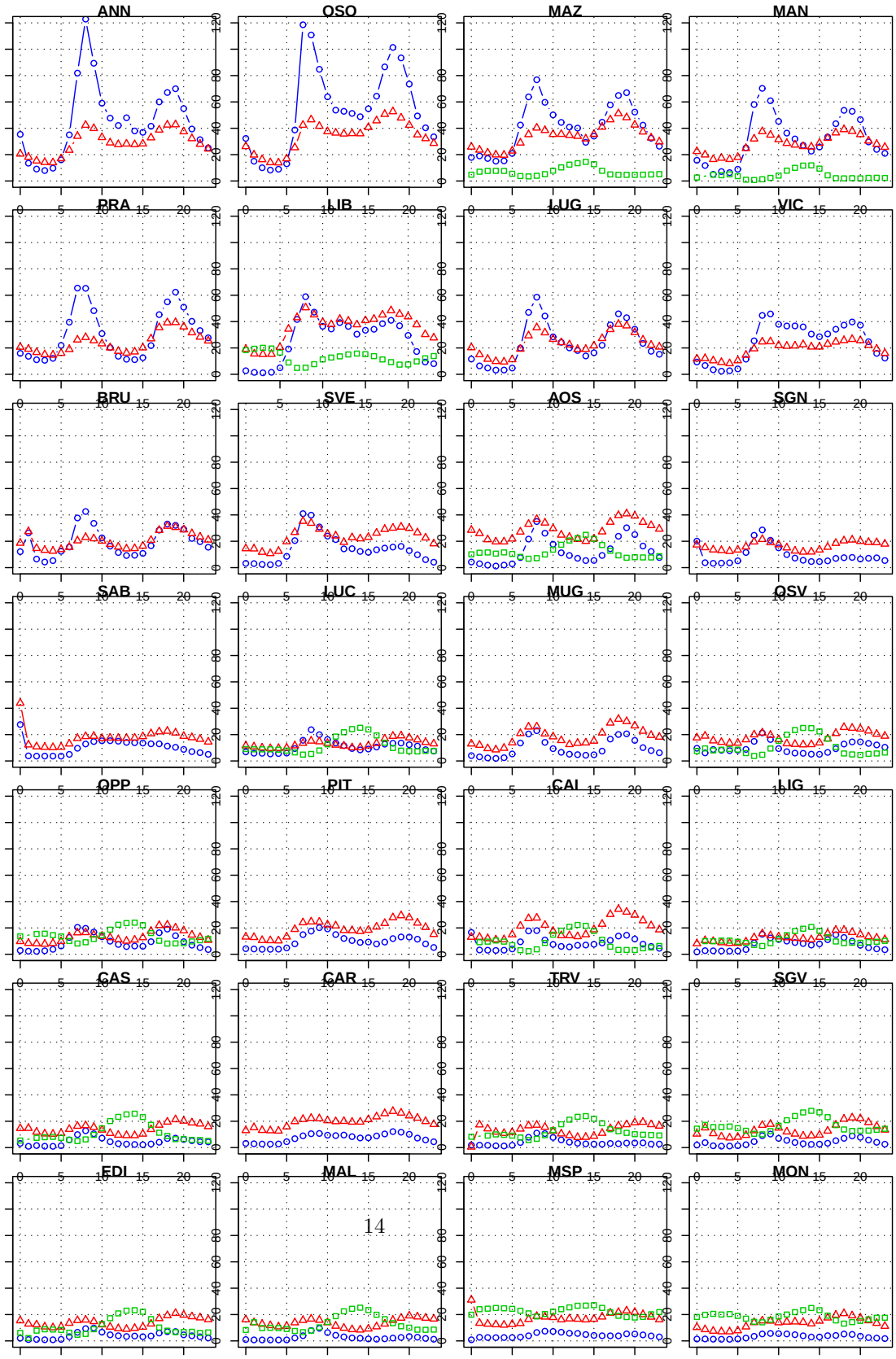

Figure 10: Daily patterns of NO (blue circles), NO_2 (red triangles) and O_3 (green squares), winter season. Concentration is expressed in ppb, end of scale is 120 ppb

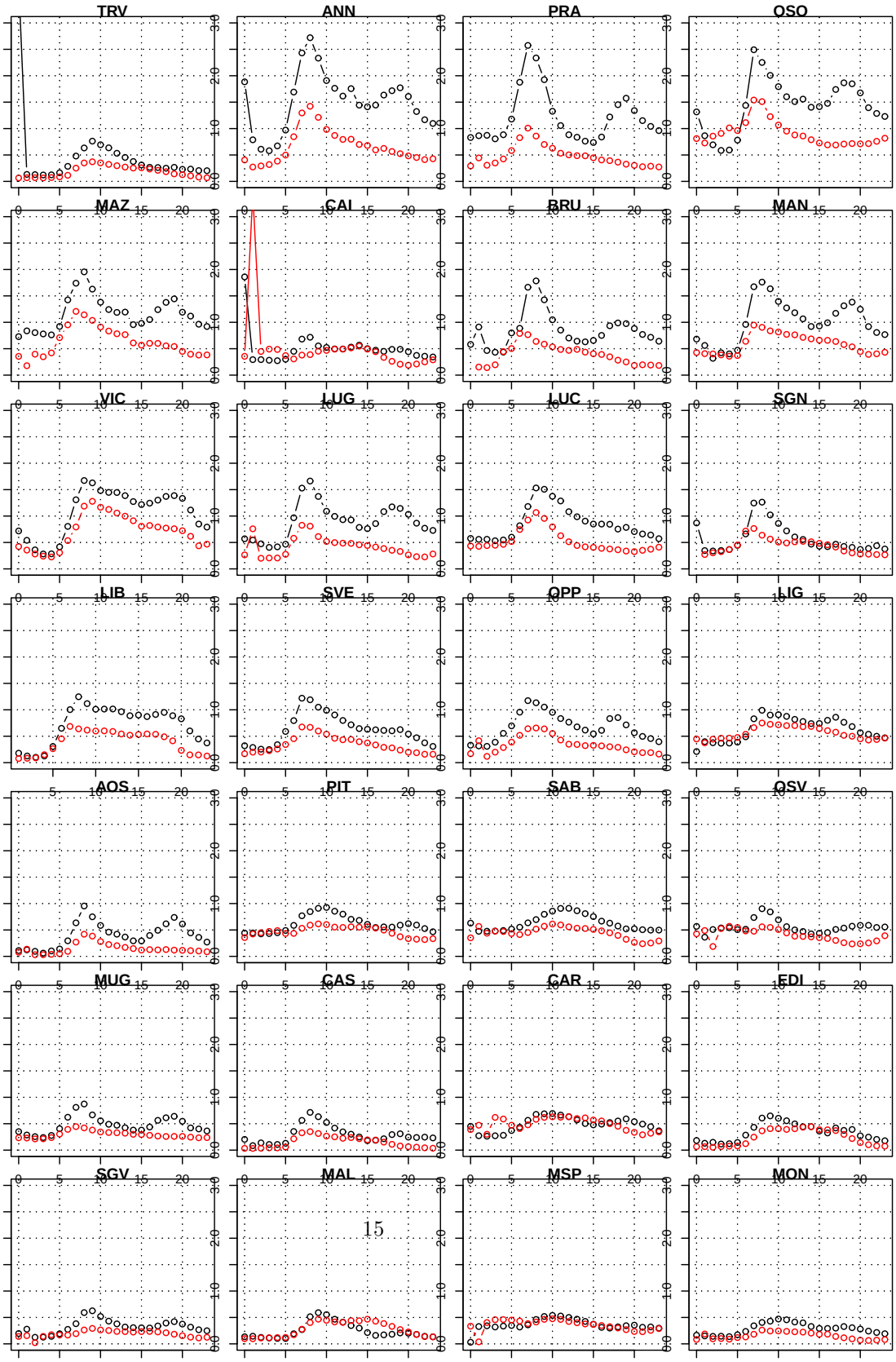

Figure 11: Winter (black) and summer (red) [NO]/[NO2] median value for each hour of the day. Summer values are 1h shifted to take into account for local Daylight Saving Time. [...] is $\mu g \cdot m^{-3}$; [NO] is actual [NO], not scaled to obtain *[NO2] equivalence* (i.e. [NO] is not multiplied by 46/30...)

fra le serie, primariamente, e metodo di clustering) sembra risentire fortemente degli extreme values, che - del resto - si vedono bene nei grafici delle time series e negli istogrammi. [2]

- la concentrazione di PM10 decresce sostanzialmente da Ovest verso Est (Gorizia). E' minima a Monfalcone, ma risale a Trieste: maggior effetto del ciglione carsico? delle sorgenti antropiche?
- l'intensità della variazione stagionale del PM10 decresce anch'essa da Ovest verso Est. Monfalcone si assimila a Trieste, più che a Gorizia.
- in questi due effetti: stiamo vedendo bordo del Bacino Padano? il consumo della legna? l'effetto della bora o delle brezze?
- Monfalcone (MON) presenta concentrazioni medie bassissime e ridottissima variazione stagionale (in quest'ultimo aspetto, si assimila a Trieste). La risposta impulsiva è assimilabile a quella delle stazioni di Gorizia e delle sole [LIB,BAN] a Trieste.
- OPP (Osoppo) ha una *baseline* bassa, un andamento inter-annuale assimilabile a quello della Bassa Friulana e di Gorizia, una risposta impulsiva a se stante: da approfondire con una classificazione basata sul rango.
- la risposta impulsiva r (sottratta la baseline + variazione interannuale) è molto simile a Gorizia, Monfalcone e (LIB,BAN). Meno nelle altre stazioni triestine: impatto dell'industria?
- da ogni punto di vista, SVE, CAR, PIT, MUG vanno per conto loro rispetto a LIB e BAN.
- isolando la componente interannuale, tolta anche la baseline ($f_{smootehd} - \overline{f}$):
  - l'area triestina si ricompone (anche con Monfalcone)
  - Gorizia va con la Bassa Friulana (e con OPP)
  - Udine Città va con Pordenone

References

[1] Mills J.L., *Data Torturing*, New England Journal of Medicine, 1993; 329: 1196-1199

[2] Arianet, *ARIA suites tools - Reference guide Release 1.2*, R2007.22, November 2007

[2] Può essere opportuno approfondire, andando a vedere cosa succede passando al logaritmo dei dati, o usando altre trasformazioni (normalizzazione (-1,1) della serie dei residui), o usando definizioni diverse di distanza (rank based - ma c'è il problema dei dati mancanti -, metrica 0/1 basata sui superamenti,...).

[3] Tarlao I., *Modelli di previsione dell'inquinamento atmosferico da ozono mediante alberi di classificazione e Random Forest: area urbana di Udine*, Università di Padova - Tesi di Laurea in Scienze Statistiche (Relatore: prof. Guido Masarotto, correlatore: dott. Francesco Montanari), A.A. 2006 - 2007 (dicembre 2006)

[4] Feresin T., *Concentrazioni di metalli nel particolato atmosferico presso un'acciaieria: un'analisi statistica*, Università di Padova - Tesi di Laurea in Scienze Statistiche (Relatore: prof. Silvano Bordignon, correlatori: dott. Francesco Pauli, dott. Francesco Montanari), A.A. 2006 - 2007 (febbraio 2007)

[5] Pillon A., *Analisi delle correlazioni fra le serie storiche dei dati di qualità dell'aria rilevati dalla Rete*, report interno (2008)
`http://172.19.216.87/wiki/index.php/VQA#Analisi_delle_correlazioni`

[6] 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.

[7] Oksanen J., *Cluster Analysis: Tutorial with R*, January 20, 2010
`http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio3.pdf`

[8] Oksanen J., mail in *R-sig-ecology – R SIG for the use of R in ecological data analysis*, Mon Nov 17 21:12:27 CET 2008
`https://stat.ethz.ch/pipermail/r-sig-ecology/2008-November/000431.html`

[9] Wikipedia, *Distance*, as of Mar. 28, 2011, 09:41 GMT
`http://en.wikipedia.org/wiki/Distance`

[10] Wikipedia, *Spearman's rank correlation coefficient*, as of Mar. 28, 2011, 09:45 GMT
`http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient`