

GIORNATE di STUDIO
LA CARATTERIZZAZIONE CHIMICA
DEL PARTICOLATO ATMOSFERICO
V EDIZIONE
Terni, 21-22 Novembre 2022

**La validazione del dato per la
caratterizzazione chimica del PM**

Arianna Trentini, ARPAE
Fabiana Scotto, Dimitri Bacco, ARPAE



Pre-trattamento/validazione del dato: che cosa significano?

La **validazione dei dati** è «il processo attraverso il quale si valuta se l'informazione può essere considerata consona alle **finalità** per le quali è stata prodotta» (ISTAT, 2001). E come si effettua? Tramite «l'insieme delle operazioni attraverso le quali si giudica lo scarto esistente fra gli obiettivi di qualità programmati in sede di progettazione dell'indagine [...] e i risultati effettivamente conseguiti»

Pre-trattamento/validazione del dato: che cosa significano?

I risultati delle analisi chimiche sono utilizzati per comprendere le complesse dinamiche che portano alla formazione di nuovi composti in atmosfera e sono l'input per metodi statistici (PMF, PCA, modellistica QA,..) necessari per l'interpretazione di un determinato fenomeno o l'identificazione di una sorgente inquinante.

Gli output derivanti da tali modelli saranno successivamente utilizzati come base per decisioni in merito a problematiche ambientali e valutazione del rischio nonché legate alla salute.

Pre-trattamento/validazione del dato: che cosa significano?

È quindi necessario validare i dati prima di utilizzarli e dimostrare la loro attendibilità. La “pulizia” di un gruppo di dati è finalizzata a creare un pacchetto dati “pulito”, **di qualità** adatta al tuo obiettivo di studio. Il dato non considerato idoneo si può quindi eliminare, rimpiazzare o correggere, ma anche **mettere da parte***.

Gli errori e i dati non validi sono diversi a seconda della tipologia del dataset che si sta analizzando e dello scopo dell’analisi. Per studiare il dataset possiamo analizzare i dati manualmente, laddove il dataset sia un numero di valori limitati, o in maniera automatica, sia graficamente sia utilizzando alcuni procedimenti statistici.

Che cos'è la qualità?

Rispetto della normativa

La “chimica” nelle agenzie

-->

IPA: Benzo(a)Pirene

Metalli: Nichel, Cadmio,

Arsenico e Piombo

Alcune ARPA hanno queste
 analisi accreditate

Obiettivi di qualità dei dati

I seguenti obiettivi di qualità dei dati vengono indicati come guida per la garanzia della qualità.

Direttiva 2004/107/CE	Benzo(a)pirene	Arsenico, cadmio e nickel	Idrocarburi policiclici aromatici diversi dal benzo(a)pirene, mer- curio gassoso totale
— Incertezza			
Misure fisse ed indicative	50 %	40 %	50 %
Modelli	60 %	60 %	60 %

Obiettivi di qualità dei dati per la valutazione della qualità dell'aria ambiente

DIRETTIVA 2008/50/CE D. Lgs. 155/2010	Biossido di zolfo, biossido di azoto e ossidi di azoto, monossido di carbonio	Benzene	Particolato (PM ₁₀ /PM _{2,5}) e piombo	Ozono e NO e NO ₂ connessi
Misurazioni in siti fissi (!)				
Incertezza	15 %	25 %	25 %	15 %
Raccolta minima dei dati	90 %	90 %	90 %	90 % in estate 75 % in inverno

I dati devono rispettare
 il periodo minimo di
 copertura ed
 essere equamente
 distribuiti.

Che cos'è la qualità? Rispetto della normativa

e per le altre analisi? EC/OC, componente ionica?

per la frazione carboniosa (UNI EN 16909) ci sono diversi protocolli (CEN/Technical Report 16243:2011), ad es.: **EUSAAR2** e **NIOSH-like**

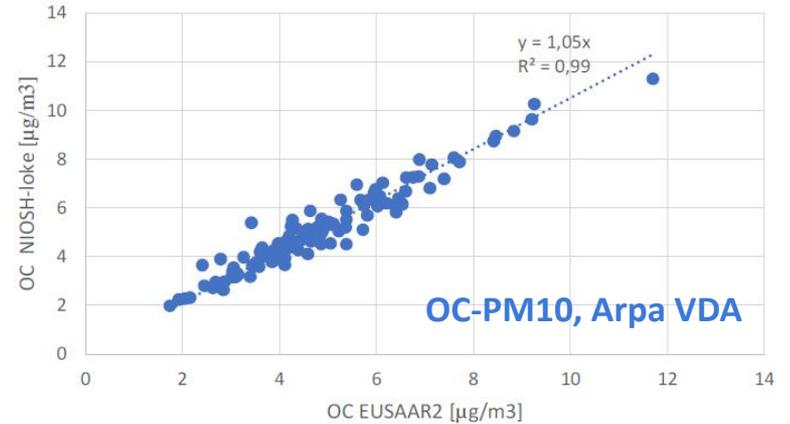
per la determinazione degli ioni si segue la UNI EN 16913

Che cos'è la qualità?

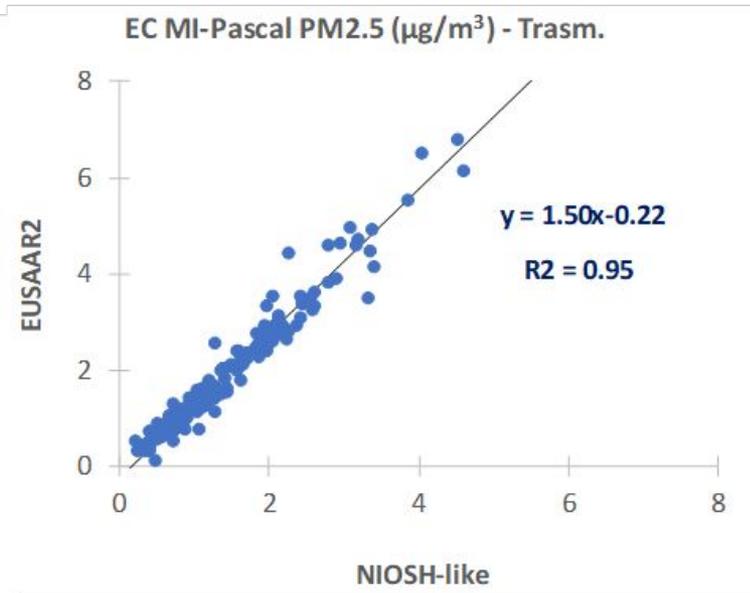
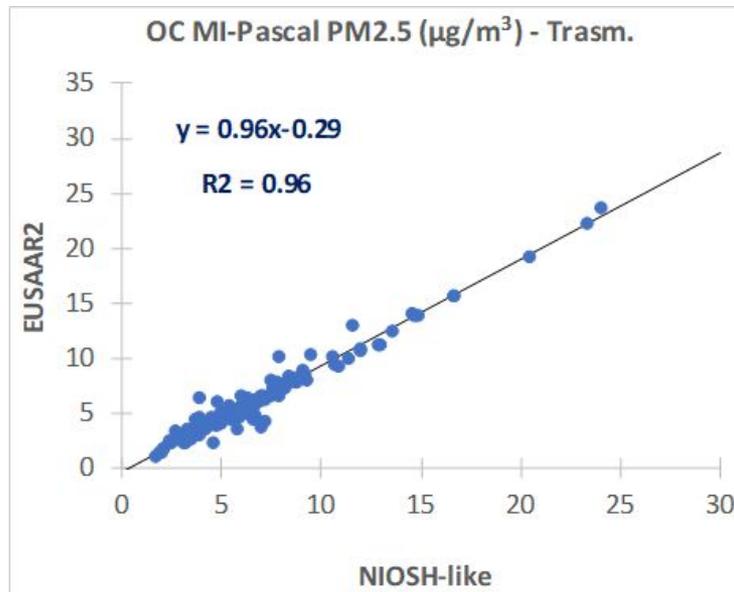
Rispetto della normativa

Confronti tra due metodi o più sono molto utili perchè permettono di rapportare il dato, ad es. tramite coefficienti moltiplicativi ricavati dalle rette di correlazione

<https://www.lifeprepare.eu/>

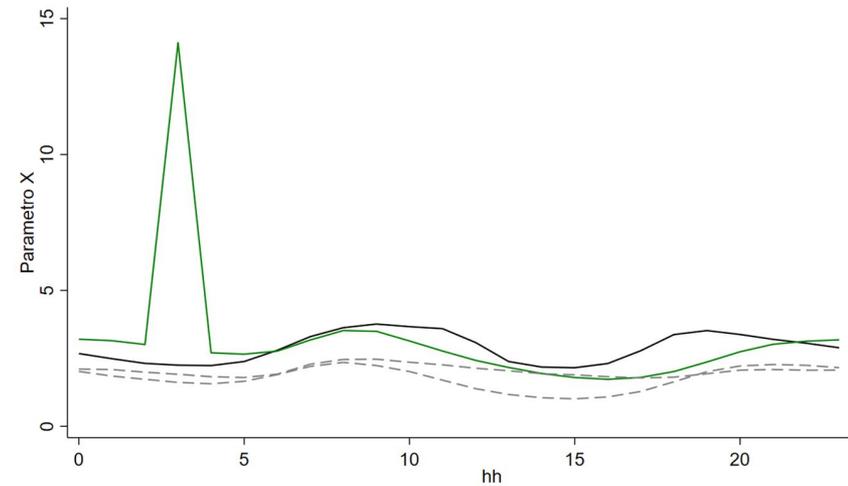
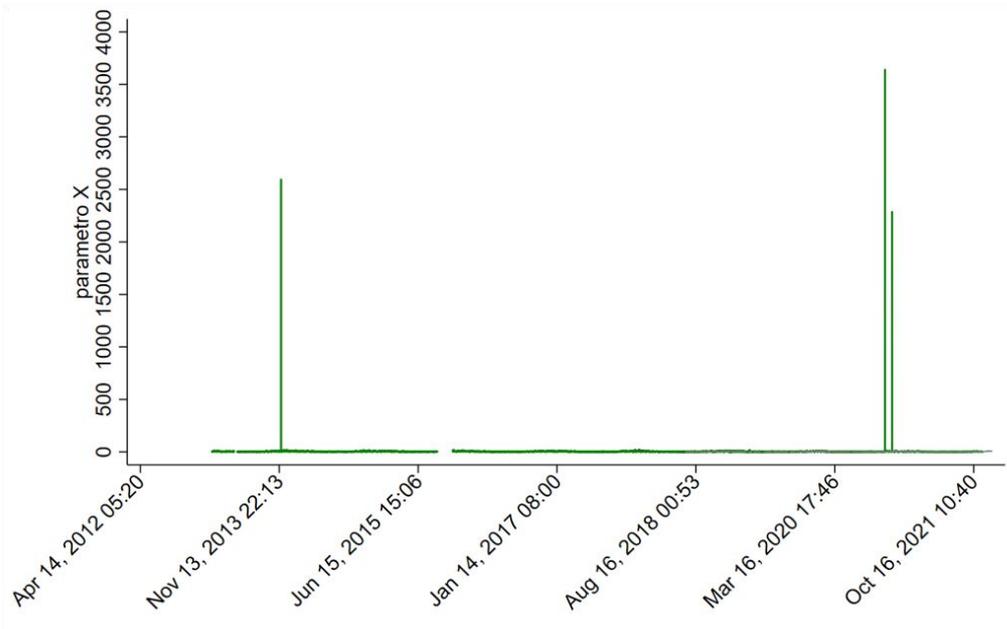


EC/OC-PM2.5, Arpa Lombardia



Pre-trattamento del dato

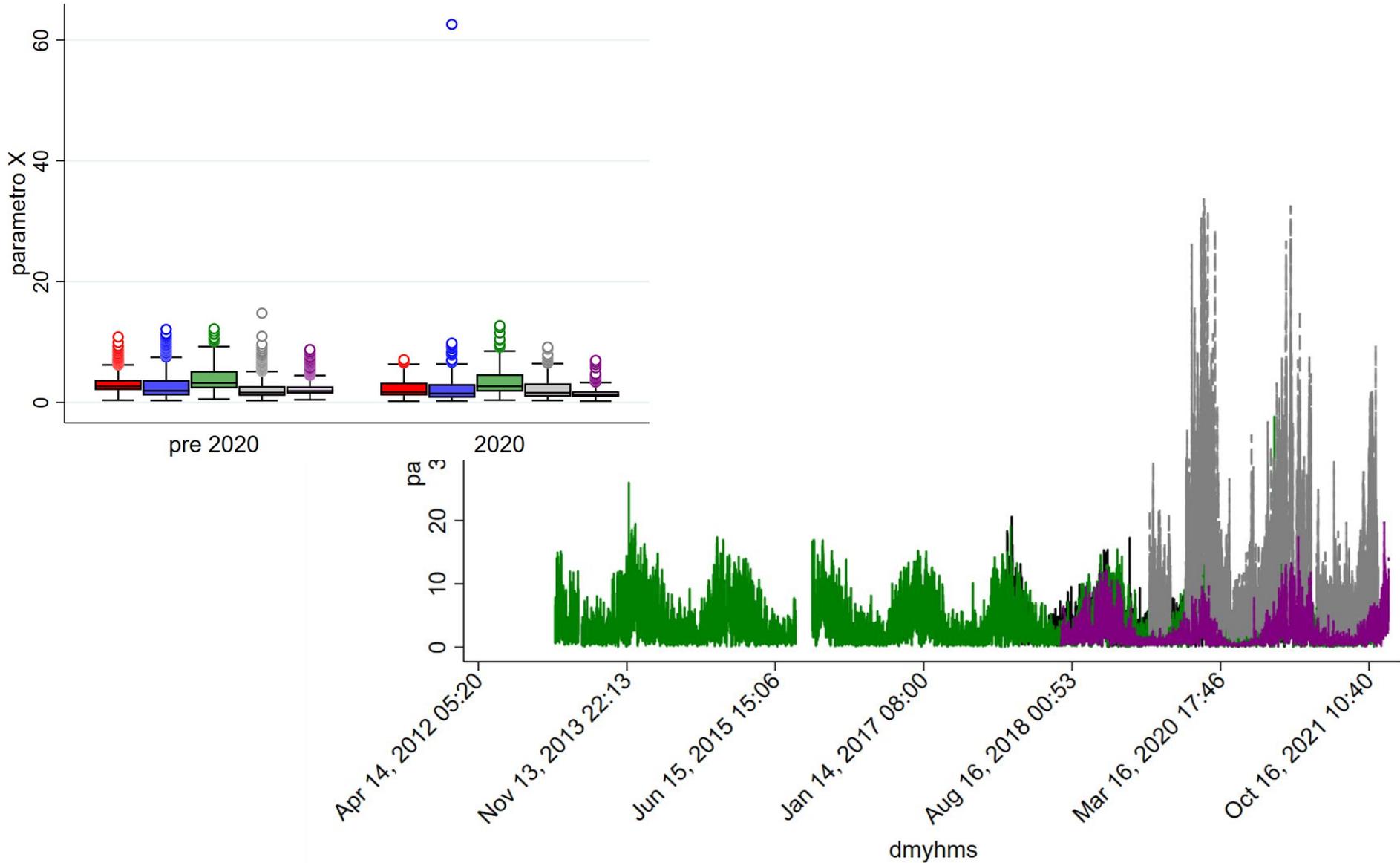
La prima cosa da fare è osservare i dati da diversi punti di vista



- Andamento temporale, gg, hh, al minuto..
- Giorno medio
- Anno medio

- Boxplot
- Barplot
- ecc.

Pre-trattamento del dato



Pre-trattamento del dato

- limite di rilevabilità
- outlier o valori anomali
- verifica componente PM
- dati mancanti

Limite di rilevabilità: esempio di un singolo valore

- ✗ 1. escluderli dall'analisi
- ✓ 2. sostituirli con $\frac{1}{2}$ del limite o il limite stesso
- ✓ 3. sostituirli con il valore strumentale

IPR dà indicazioni in tal senso:

i dati con valori = \pm LOD possono essere sostituiti con LOD o LOD $\frac{1}{2}$. Se il dato è di **poco** inferiore si può approssimarlo per difetto, es. LOD=2, valore -2.1 \rightarrow lo approssimo a -2)

https://www.eionet.europa.eu/aqportal/doc/IPR%20guidance_2.0.1_final.pdf

Limite di rilevabilità: esempio di diversi valori

Le scelte possono essere diverse a seconda dello scopo delle analisi e devono basarsi su diversi fattori: numero di dati disponibili, numero di dati inferiori, loro distribuzione, contesto/finalità dello studio (ambientale, epidemiologico?..), ecc.

	siti				
	1	2	3	4	5
n° gg	1126	1126	1126	1103	1123
<LOD	476	429	424	109	285
n° dati (gg-MD)	1026	923	1038	653	888
% (<LOD/n°)	46%	46%	41%	17%	32%

In una campagna di misura di 2 settimane con 8 dati con valori < LOD
 ..cosa fareste?

Limite di rilevabilità: esempio di diversi valori

Un dato inferiore non è un dato mancante, possiamo però decidere di eliminare una variabile con un numero di dati inferiore al limite maggiore di una certa percentuale

4.94	2.07	7.01	367.9	<9.1	<0.17	1.8	0.16	<0.04	<0.32	<0.11	<0.04	4.09	1.19	<0.16	<361	0.56	<2.0	139	3	1.5	0.39	17.
2.73	0.72	3.45	143.9	<9.1	0.2	0.72	0.07	<0.04	0.36	<0.11	<0.04	0.96	0.8	<0.16	<361	0.3	<2.0	86	3.4	<0.5	0.17	<11
6.60	1.78	8.38	400.7	<9.1	<0.17	3	0.26	<0.04	<0.32	0.19	<0.04	9.08	1.1	<0.16	<361	0.59	<2.0	135	4.5	0.8	0.64	27.
2.78	0.66	3.44	122	<9.1	<0.17	4.43	0.06	<0.04	<0.32	<0.11	<0.04	12.25	1.56	<0.16	<361	0.55	<2.0	95	6.7	3.2	0.82	45.
5.47	1.14	6.61	326	<9.1	<0.17	5.99	0.22	<0.04	<0.32	0.16	<0.04	15.87	2.84	<0.16	<361	0.32	<2.0	37	2.4	<0.5	0.19	<11
3.55	1.14	4.69	194.9	<9.1	<0.17	2.86	0.1	<0.04	<0.32	<0.11	<0.04	8.42	1.23	<0.16	<361	0.22	<2.0	31	<2.0	2.9	0.2	<11
2.18	0.64	2.82	216.8	<9.1	<0.17	0.54	0.06	<0.04	<0.32	<0.11	<0.04	0.6	0.54	<0.16	<361	0.11	<2.0	44	<2.0	<0.5	0.08	<11
1.82	0.51	2.33	120.2	<9.1	<0.17	0.6	<0.05	<0.04	<0.32	<0.11	<0.04	0.73	0.52	<0.16	<361	0.31	<2.0	48	3	0.7	0.11	13.
3.23	1.01	4.24	194.9	<9.1	<0.17	2.05	0.08	<0.04	<0.32	<0.11	<0.04	4.88	1.04	<0.16	<361	0.21	<2.0	84	<2.0	1.5	0.45	20.
3.23	0.93	4.16	187.6	<9.1	<0.17	0.56	0.08	<0.04	<0.32	<0.11	<0.04	0.61	0.78	<0.16	<361	0.21	<2.0	84	<2.0	1.5	0.45	20.
3.94	1.00	4.94	244.1	<9.1	<0.17	0.72	0.13	<0.04	<0.32	<0.11	<0.04	0.81	0.81	<0.16	<361	0.21	<2.0	76	<2.0	<0.5	0.63	<11
7.27	2.10	9.37	373.4	<9.1	<0.17	3.11	0.24	<0.04	<0.32	0.16	<0.04	9.3	1.62	<0.16	<361	0.63	<2.0	135	3.1	1.6	1.71	38.
4.70	1.29	5.99	187.6	<9.1	<0.17	3.19	0.17	<0.04	<0.32	<0.11	<0.04	8.92	2.22	<0.16	<361	0.49	<2.0	141	4.2	2.9	3.09	14.
3.63	1.22	4.85	262.3	<9.1	<0.17	1.71	0.12	<0.04	<0.32	<0.11	<0.04	3.68	1.37	<0.16	<361	0.51	<2.0	118	3	1	1.21	<11
3.30	0.95	4.25	240.4	<9.1	<0.17	1.64	0.11	<0.04	<0.32	<0.11	<0.04	3.5	1.07	<0.16	<361	0.23	<2.0	66	3.3	1.2	0.76	20.
2.81	0.77	3.58	153	<9.1	<0.17	1.56	0.1	<0.04	<0.32	<0.11	<0.04	2.64	1.41	<0.16	<361	0.22	<2.0	43	<2.0	<0.5	0.86	<11
4.00	0.81	4.81	185.8	<9.1	<0.17	1.46	0.2	<0.04	<0.32	0.14	<0.04	2.43	1.53	<0.16	<361	0.39	<2.0	60	<2.0	1.3	1.84	<11
1.97	0.94	2.91	82	<9.1	<0.17	0.49	0.06	<0.04	<0.32	<0.11	<0.04	0.44	0.74	<0.16	<361	0.1	<2.0	58	<2.0	0.8	0.51	<11
3.65	1.29	4.94	145.7	<9.1	<0.17	2.63	0.11	<0.04	<0.32	<0.11	<0.04	6.99	1.67	<0.16	<361	0.61	<2.0	53	2.8	0.8	0.16	18.
3.35	1.34	4.69	147.5	38.4	<0.17	0.63	0.09	<0.04	<0.32	<0.11	<0.04	0.54	1.04	<0.16	<361	0.4	<2.0	111	2.9	0.7	0.28	<11
0.80	0.33	1.13	60.1	<9.1	<0.17	0.2	<0.05	<0.04	<0.32	<0.11	<0.04	0.19	0.25	<0.16	<361	<0.05	<2.0	20	<2.0	<0.5	<0.04	<11
2.34	0.76	3.10	176.7	<9.1	<0.17	0.37	0.06	<0.04	<0.32	<0.11	<0.04	0.43	0.45	<0.16	<361	0.14	<2.0	86	2.7	<0.5	0.19	<11
2.24	0.43	2.67	89.3	<9.1	<0.17	0.56	<0.05	<0.04	<0.32	<0.11	<0.04	0.93	0.64	<0.16	<361	0.12	<2.0	36	<2.0	1.3	0.43	<11
3.45	0.77	4.22	162.1	<9.1	<0.17	3.63	0.09	<0.04	<0.32	<0.11	<0.04	11.54	1.86	<0.16	<361	0.84	<2.0	29	<2.0	<0.5	0.1	13.
2.66	0.69	3.35	116.6	<9.1	<0.17	2.68	0.07	<0.04	<0.32	<0.11	<0.04	6.97	1.93	<0.16	<361	0.5	<2.0	40	<2.0	<0.5	0.17	14.

Outlier: definizione

Un valore anomalo (grande o piccolo che sia) o un insieme di valori anomali che differiscono notevolmente dal contesto del restante dataset (non riconducibile alla maggior parte dei valori presenti nel dataset)

Possono essere dei veri e propri errori di misura, sia strumentali che legati all'operatore oppure dei dati "veri" ma che per qualche motivo sono rappresentativi di condizioni eccezionali.

Knocke, B. and P. Mee (2002) "Statistics for social data analysis"

Outlier: che cosa sono?

Qualche esempio di possibili outlier “veri”:

- Fuochi d’artificio
- Falò (feste di paese, nazionale, ..)
- Trasporti di sabbia
- Incendi limitrofi alla stazione di misura..
- (Covid19?)

--> valori "veri" ma a volte non rappresentativi..

Outlier: perchè ci interessano?

Gli outlier non devono essere trascurati perchè spesso ci indicano un errore nella misura/analisi.

Possono alterare le statistiche descrittive di base (media, coeff. di corr, ecc.) e portare a considerazioni/interpretazioni sbagliate.

→

- 1 possono essere utili
- 2 possono essere dannosi
- 3 entrambe

Outlier: perchè ci interessano?

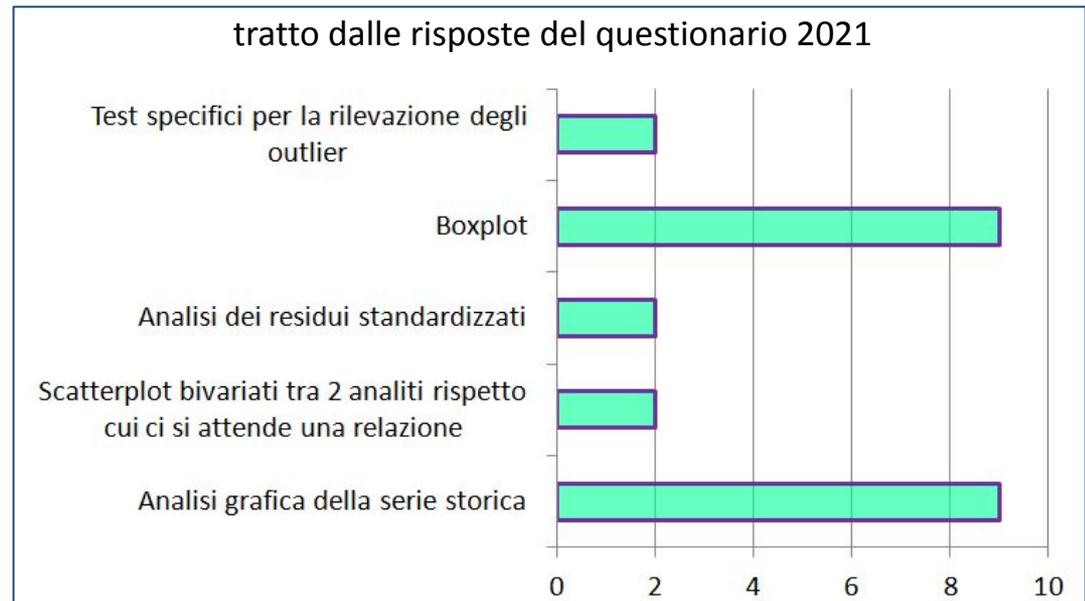
Esempio 1: ad es. una festa di paese con molti falò rappresenta un'occasione preziosa per verificare la presenza di metalli pesanti nel particolato emesso dalla combustione di biomassa.

Esempio 2: al tempo stesso la sorgente falò può produrre uno o più outlier che possono andare a pesare sulla regressione lineare



Outlier: come trovarli?

- Approccio grafico
 - serie storica
 - boxplot (IQR)
 - scatterplot bivariati tra due analiti rispetto ai quali ci si aspetta una relazione
- Approccio statistico
 - Test specifici per la rilevazione degli outlier
 - Analisi dei residui standardizzati
 - Media $\pm 2 \cdot \text{dev. std.}$



Alcuni approcci sono più invasivi di altri

Outlier: come trovarli?

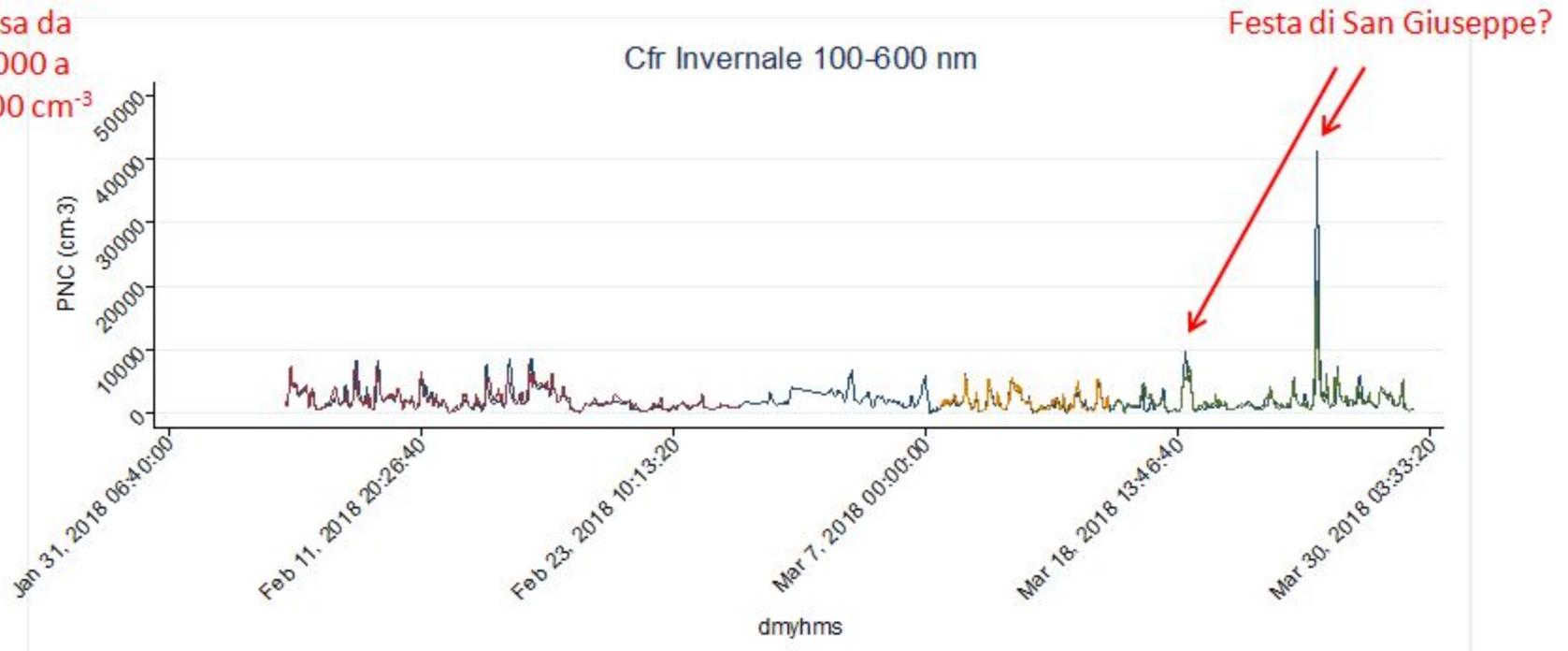
Dipende anche che cosa si sta indagando:

- . un singolo dato (es. validazione giornaliera rete QA sistema agenzie ARPA) → confronto il dato con altre stazioni di misura simili, controllo l'andamento nei giorni precedenti, verifico un'altra frazione nello stesso sito di misura, ecc.
- . un serie di dati (es. validazione semestrale o annuale) o una serie storica → ho bisogno di qualche metodo che mi aiuti velocemente a identificare valori anomali

Outlier: come trovarli?

Andamento orario: particelle non ultrafini
(DATI 24/25 MARZO NON CENSURATI)

Il fondo scala
passa da
10000 a
50000 cm⁻³

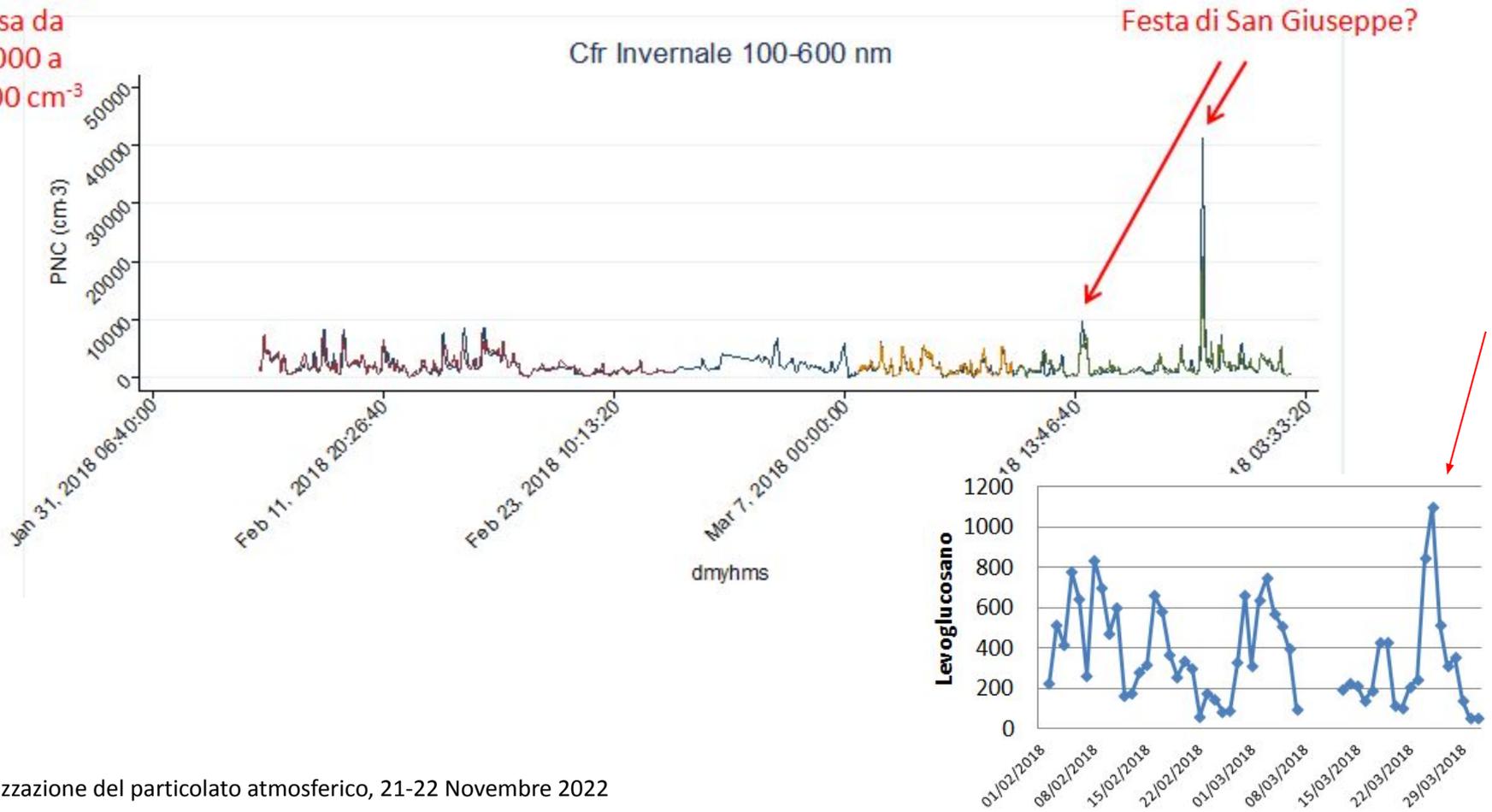


Festa di San Giuseppe: festa nel riminese dove vengono accesi molti falò

Outlier: come trovarli?

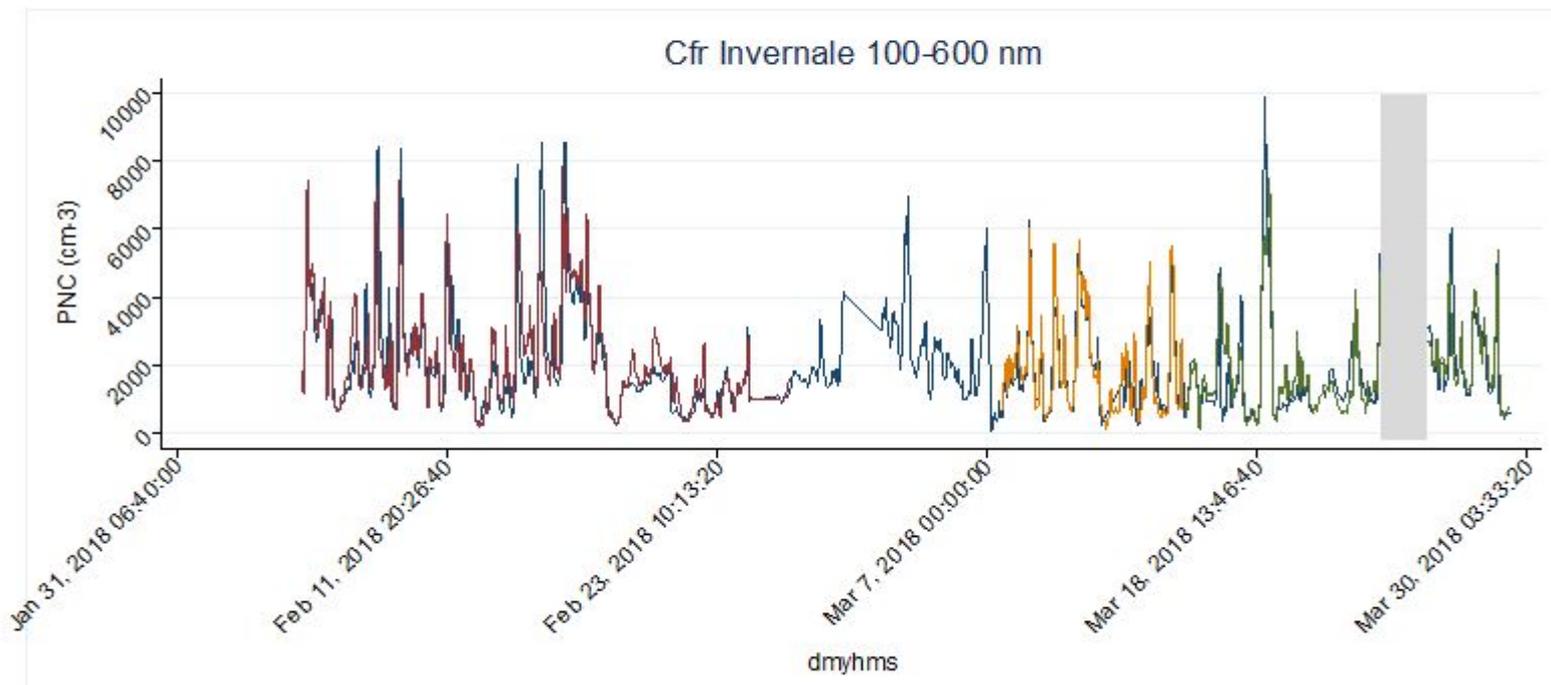
Andamento orario: particelle non ultrafini
 (DATI 24/25 MARZO NON CENSURATI)

Il fondo scala
 passa da
 10000 a
 50000 cm⁻³



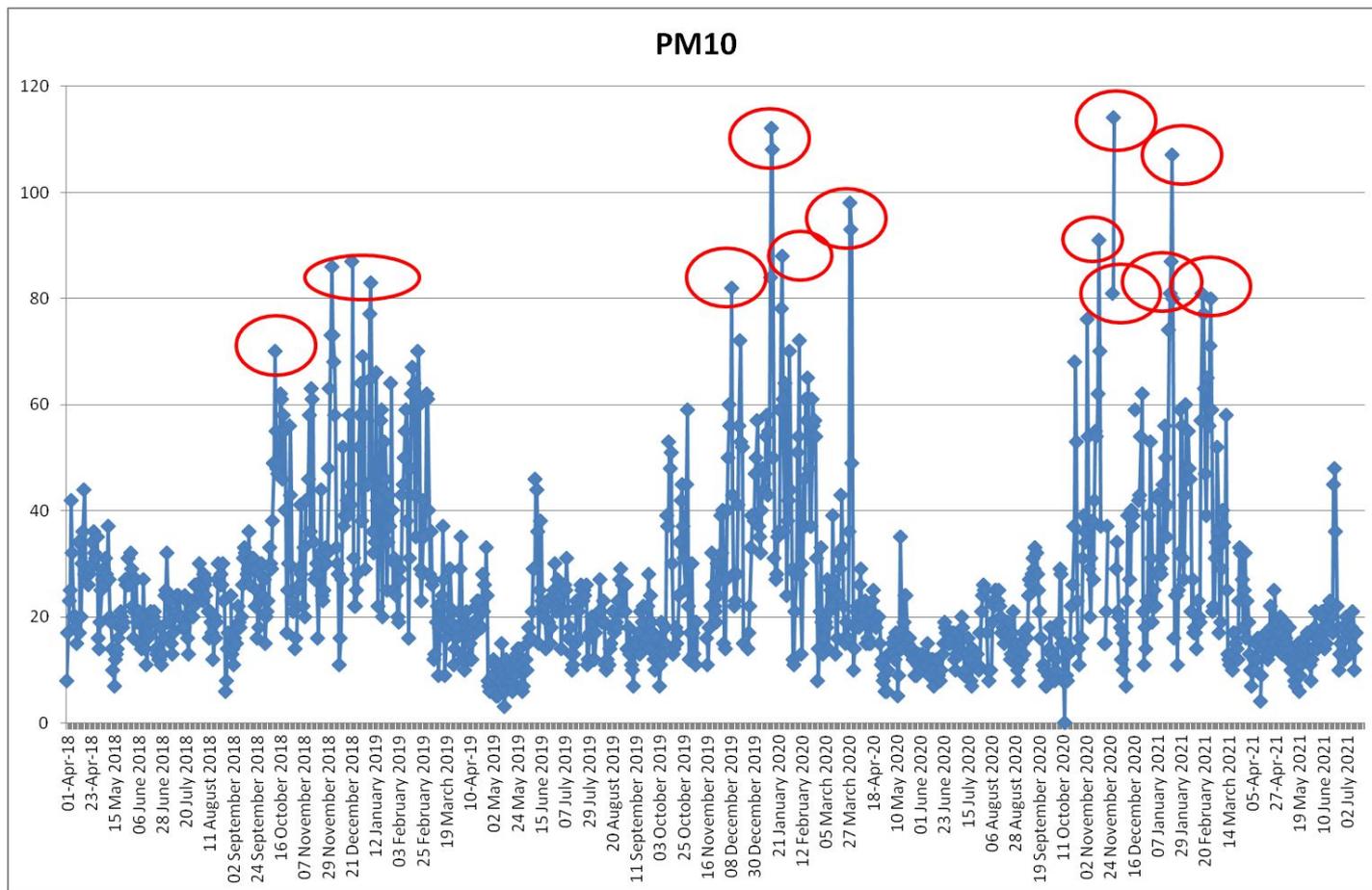
Outlier: come trovarli?

Andamento orario: particelle non ultrafini
(DATI 24/25 MARZO CENSURATI)



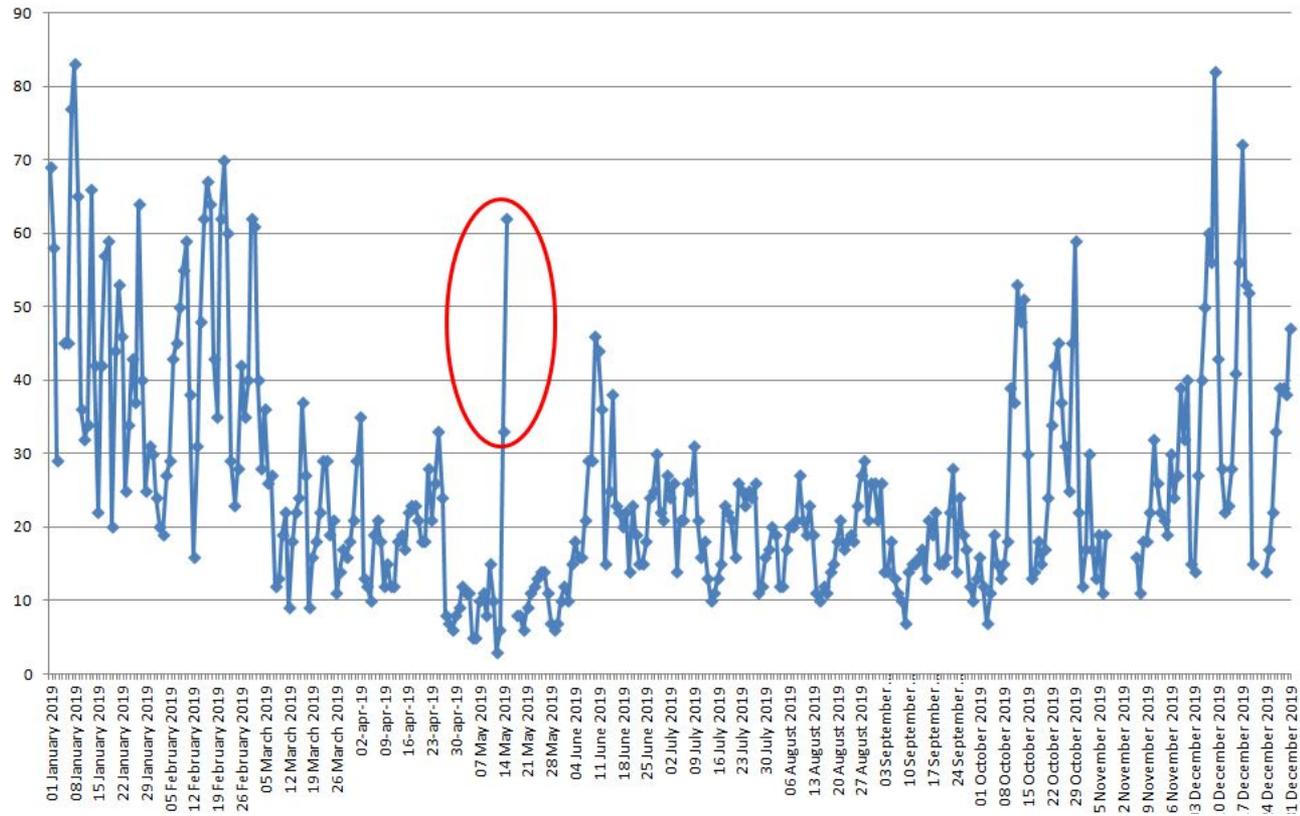
Outlier: come trovarli?

Serie storica PM10

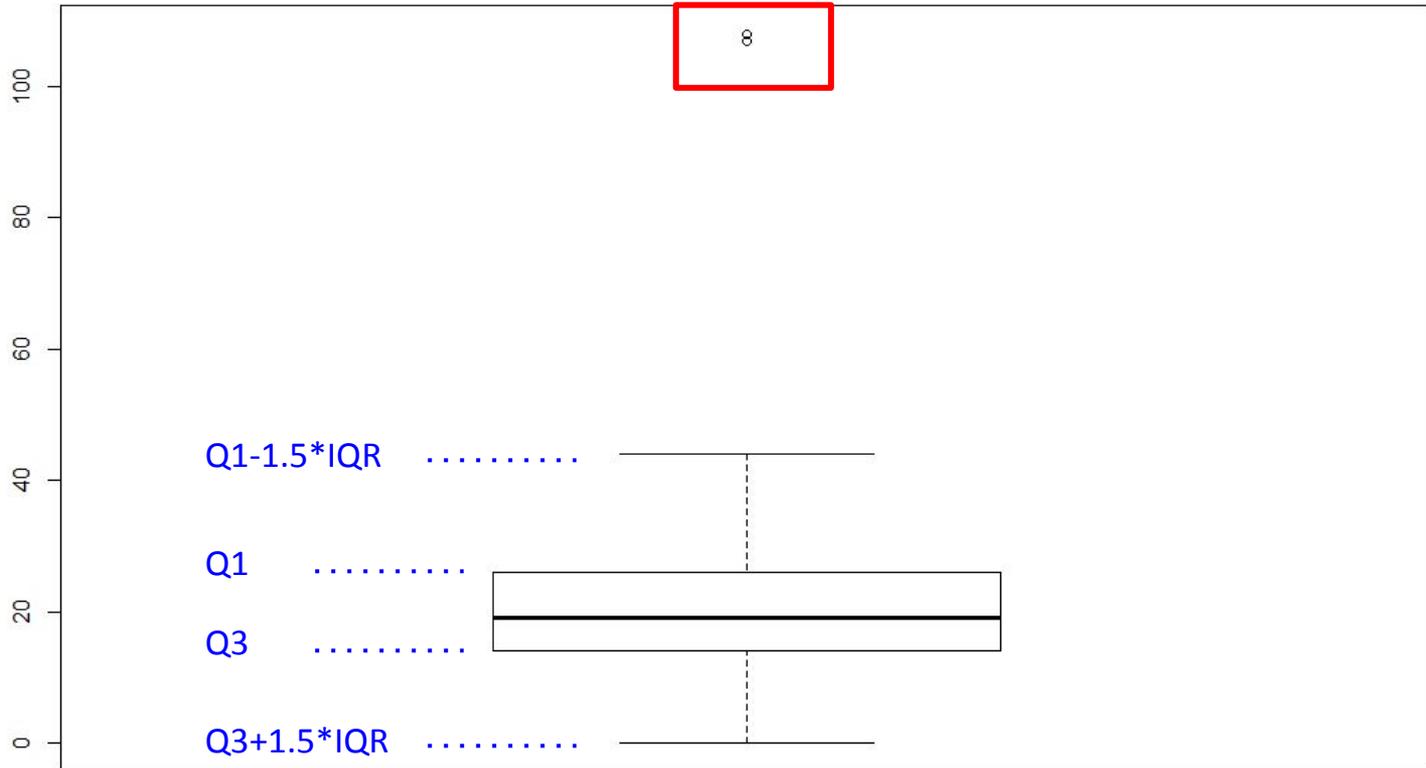


Outlier: come trovarli?

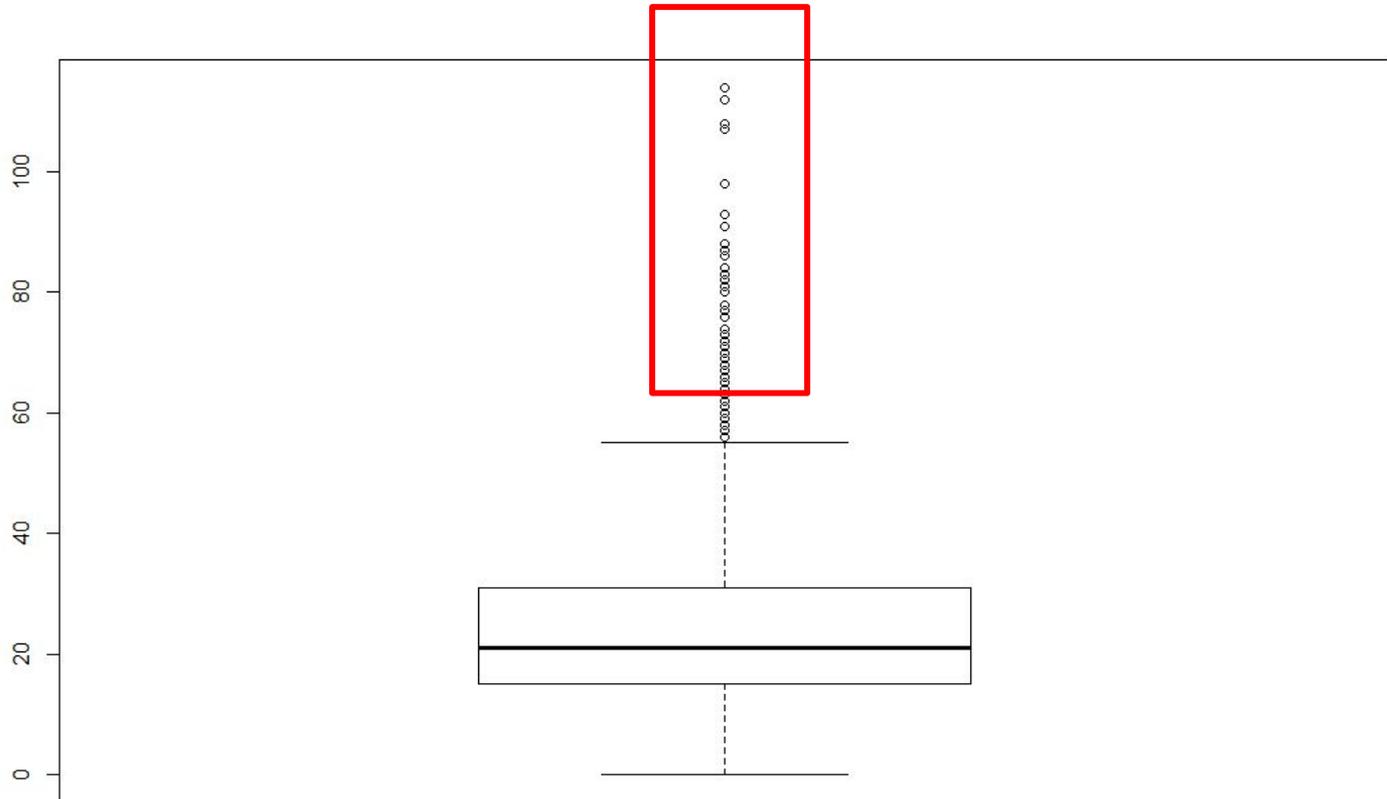
Serie storica PM10



Outlier: come trovarli?

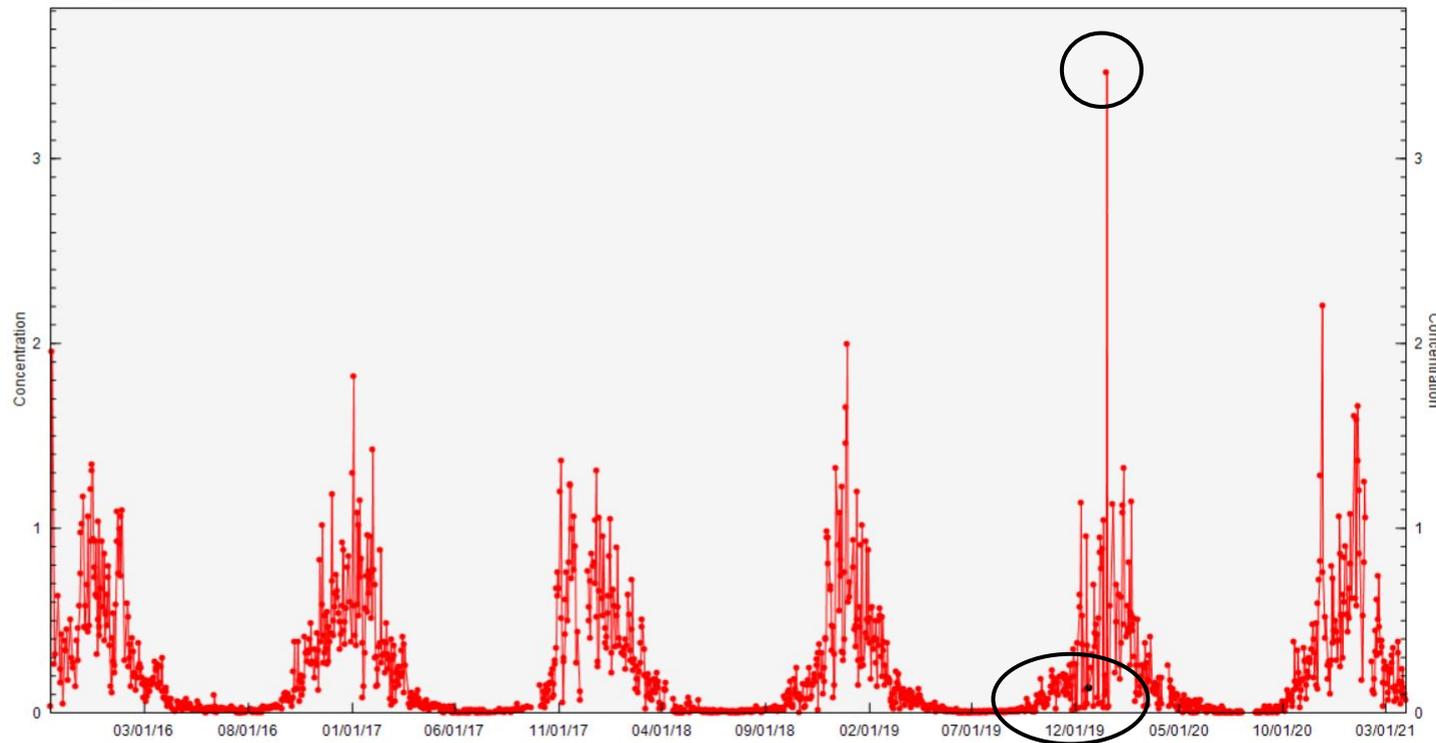


Outlier: come trovarli?

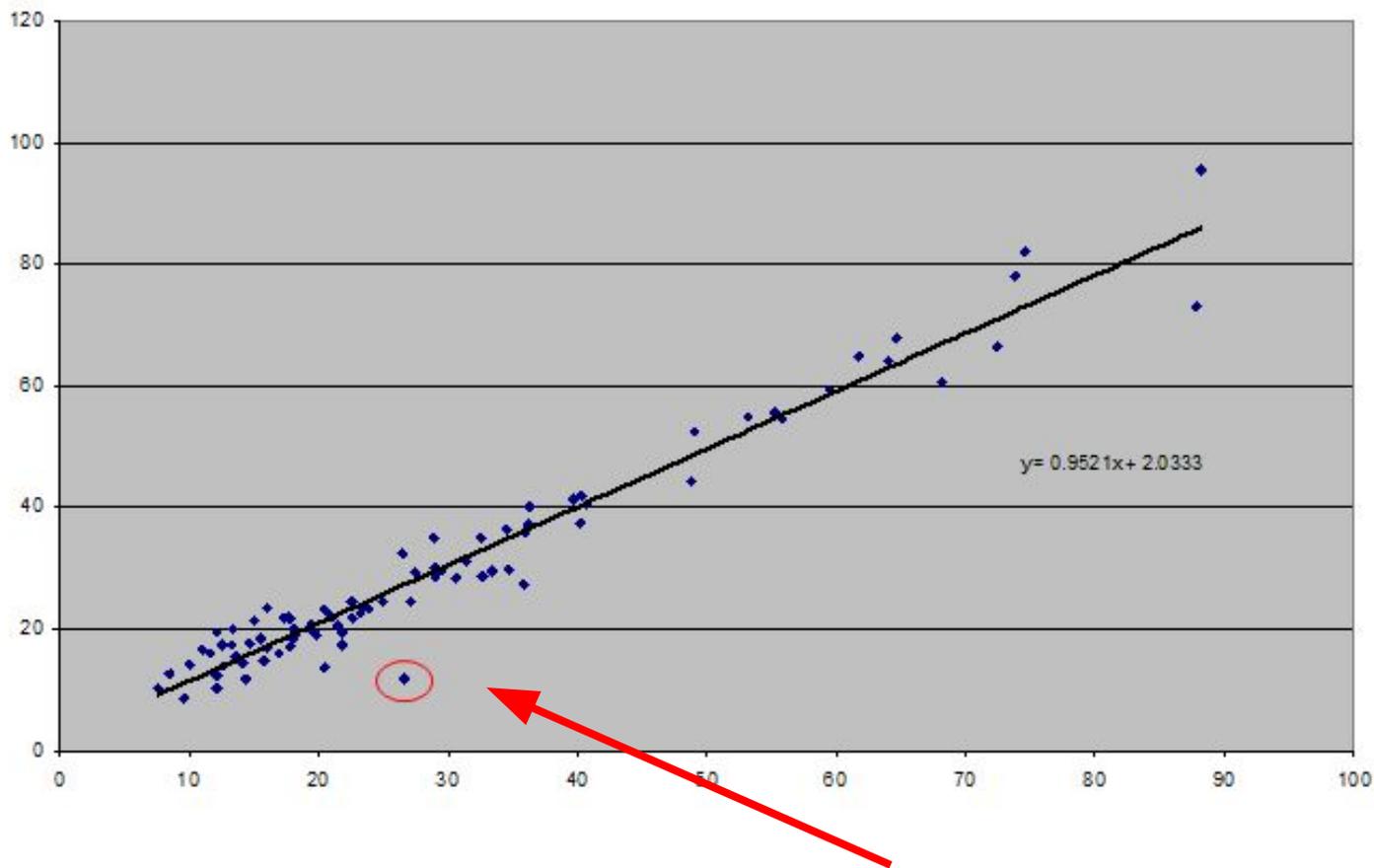


Outlier: come trovarli?

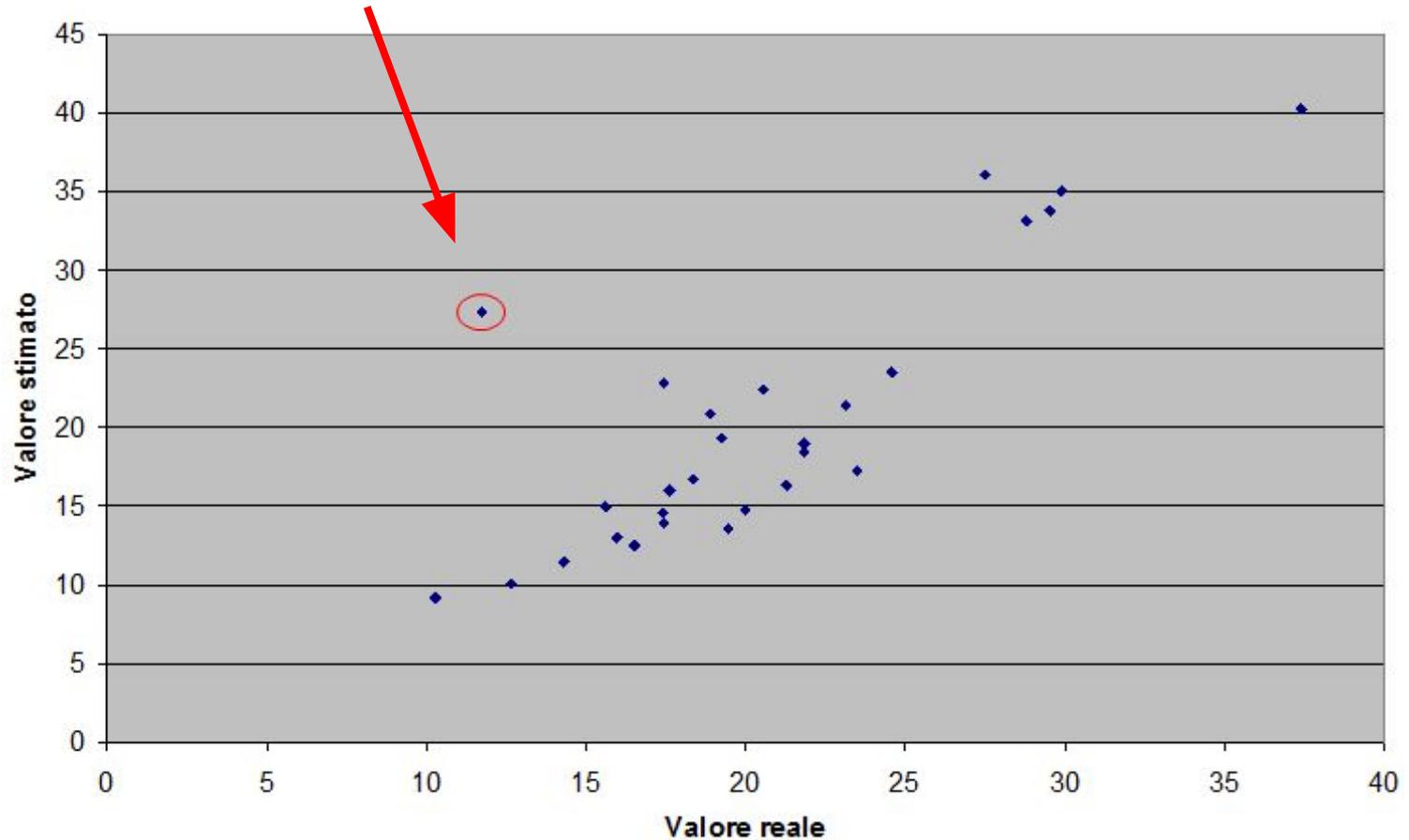
Anche valori **“troppo bassi”** possono essere degli outlier



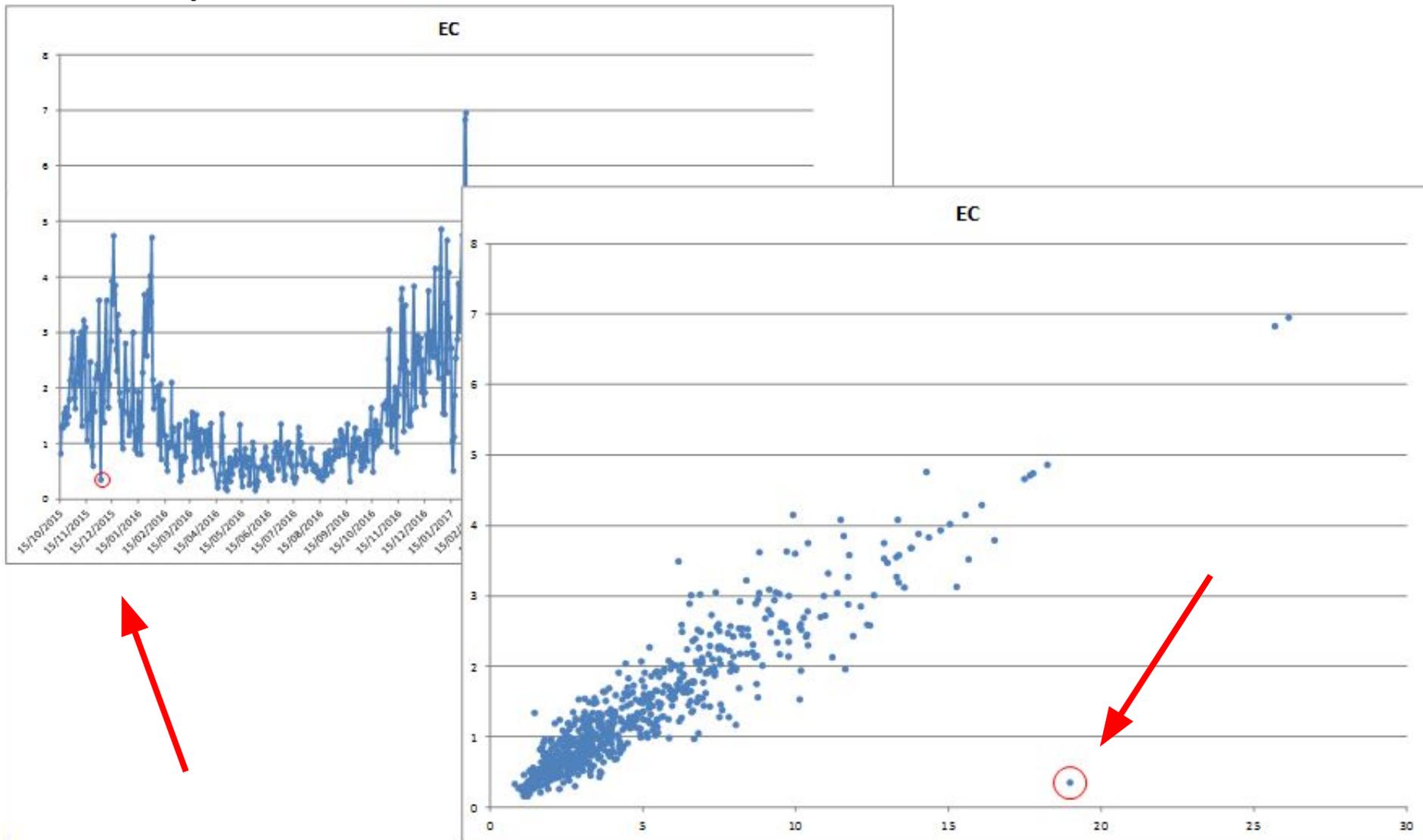
Outlier: come trovarli?



Outlier: come trovarli?



Outlier: come trovarli?



Outlier: come trovarli?

Test statistici

Test di Dixon

$$Q_n = \frac{|x_n - x_{n-1}|}{|R|},$$

dove $R = \max_j(x_j) - \min_j(x_j)$ è l'ampiezza dell'intervallo contenente tutti i valori osservati.

Test di Grubbs

$$G_{max} := \frac{\max(|y - \text{mean}(y)|)}{\text{stdev}(y)} = 3.631$$

Questi test hanno gli stessi problemi dei boxplot: non individuano valori fuori posto in una serie storica, ma si esprimono solo **a livello probabilistico** sui **valori estremi** (valori che possiamo identificare senza l'aiuto di un test e che valuteremo se sono reali o no sulla base di altre considerazioni --> andremo a vedere eventuali altre specie correlate o eventuali centraline correlate)

Altri test: Rosner test, test della discordanza ecc...

Outlier: come trovarli?

Data	CAS	MXS	Scarto Star
02/06/2008	13.38	19.99	-1.035238
03/06/2008	13.2	17.42	-0.542415
04/06/2008	11.55	15.96	-0.565214
05/06/2008	9.9	14.31	-0.548975
06/06/2008	14.67	17.62	-0.295947
07/06/2008	20.35	23.12	-0.314864
08/06/2008	12.1	19.45	-1.174682
09/06/2008	15.03	21.28	-0.977511
10/06/2008	15.95	23.49	-1.251609
11/06/2008	17.78	21.83	-0.552561
12/06/2008	15.43	18.35	-0.297263
13/06/2008	13.57	15.61	-0.098152
14/06/2008	7.52	10.28	-0.186542
15/06/2008	8.44	12.66	-0.495569
16/06/2008	12.47	17.43	-0.687272
17/06/2008	19.8	18.89	0.446645
18/06/2008	11	16.51	-0.785808
19/06/2008	21.42	20.54	0.424537
20/06/2008	21.81	17.43	1.139812
21/06/2008	26.58	11.74	3.241988
22/06/2008	18.15	19.25	0.049907
23/06/2008	17.23	21.82	-0.658097
24/06/2008	22.54	24.57	-0.184376
25/06/2008	32.66	28.79	0.928246
26/06/2008	34.64	29.88	1.09162

Rispetto all'analisi grafica, **l'analisi dei residui standardizzati** ha il vantaggio che può essere fatta in **automatico** e non da un operatore, **evidenziando tutti i residui maggiori di una soglia prestabilita**. Questa procedura viene ad es. utilizzata in alcune Arpa in fase di validazione del dato giornaliero: se si presenta un residuo maggiore di un valore soglia viene dato un warning e gli operatori addetti alla validazione controllano.

Utile per i dati che necessitano di essere validati in tempo reale e quando i dati da controllare sono tanti!

Outlier: come trattarli?

Gli outlier possono essere inclusi o eliminati a seconda della loro tipologia e a seconda delle finalità dello studio dall'analisi.

- ✓ • Li escludiamo dalle elaborazioni se sono valori che riteniamo sbagliati
- ✓ • Anche se sappiamo essere veri possiamo decidere di escluderli da talune elaborazioni

Per alcune elaborazioni può essere lecito, anzi consigliabile eliminare outliers, anche se rappresentano situazioni reali.

- ✓ • Anche se sappiamo essere veri non li escludiamo dalle elaborazioni, ma cerchiamo di effettuare elaborazioni che non siano troppo influenzate da tali valori

x es., una mediana anziché una media, un 95° percentile anziché un valore massimo. In generale molta statistica non parametrica aiuta in questo senso.

Outlier: come trattarli?

- ✗ ✓ • Non li escludiamo dei dati dalle elaborazioni perchè non possiamo essere sicuri che siano sbagliati → verificare la possibilità di eseguire ulteriori indagini
- ✗ ✓ • Se sono valori che sappiamo essere veri li consideriamo in tutte le elaborazioni → Legittimo, ma outlier veri possono essere esclusi da alcune elaborazioni anche qualora non siano sbagliati, dipende dalla finalità del lavoro

Outlier: come trattarli?

Posso decidere di togliere i giorni con i fuochi d'artificio dalla PMF perché il modello non riesce ad individuare/isolare questa sorgente e andrebbe a spalmarla sulle altre "sporcandole".

Posso decidere di eliminare il giorno dei fuochi della festa da un modello che prova a prevedere la massa del PM dalle variabili meteo.

MA, NON posso decidere di escludere né i fuochi d'artificio né i falò della festa se devo calcolare la media o il numero di superamenti in un sito* (limiti normativi)

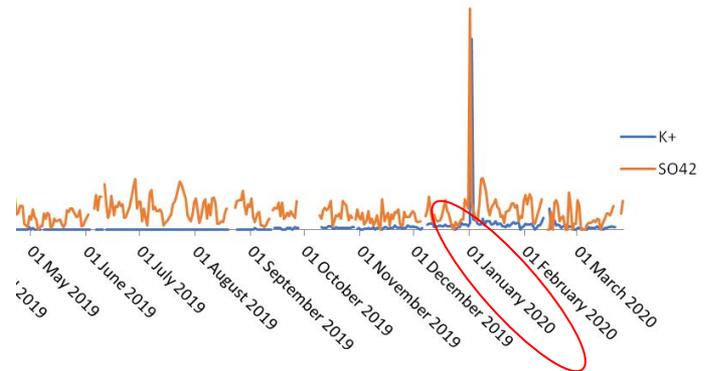
* ma i trasporti di sabbia sì, perché sono eventi naturali

Outlier: considerazioni per input PMF

Ogni sito ha le sue peculiarità, quindi nel trattamento di un dataset che diventerà l'input per es della PMF bisogna tenere a conto della realtà di quel sito.

L'eliminazione dei **fuochi d'artificio a Capodanno** non è sempre la soluzione corretta, può andare bene per un sito ma non per altro, o magari ci sono feste di paese che su quel sito incidono particolarmente, ecc.), ecc.

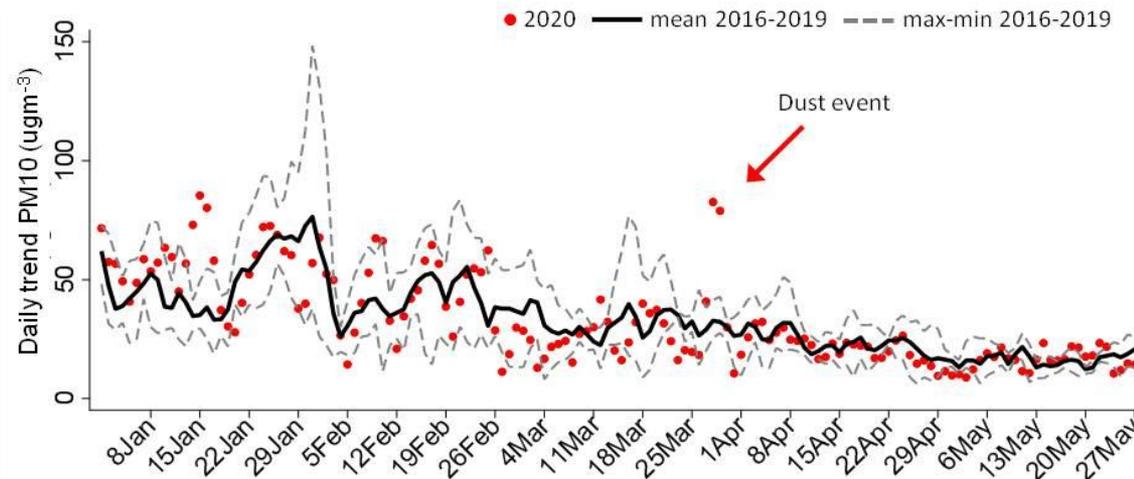
Di nuovo analisi del vostro dataset!



Outlier: considerazioni per input PMF

Gli eventi di trasporto di polvere dal deserto, che sappiamo essere sempre più frequenti e più lunghi, li togliamo come i fuochi d'artificio se vanno a "sporcare" i fattori: tutto dipende dall'intensità dell'evento, se è arrivato al suolo, se abbiamo gli elementi per caratterizzarlo (Al, Si, ..) e se la PMF riesce a vedere la sorgente/fattore "Saharan dust"!

Non tutti gli eventi sono ben definiti .. oppure un evento combacia a livello temporale con un altro che ha alcuni marker in comune.



Outlier: considerazioni per input PMF

Anche il fattore navale non può “esistere” senza marcatori caratteristici (o rapporti tra marcatori, come Ni/V)

Nel caso dell'es. della festa coi falò di San Giuseppe la sorgente fa ovviamente del Biomass Burning (BB), un fattore che la PMF identifica (in diversi siti della pianura padana)

Posso quindi decidere di lasciarlo anzi, può essere utile perché gli eventi estremi spesso aiutano a definire i risultati statistici. Magari però i due profili chimici non sono proprio uguali ...(materiale bruciato differente, metodo di combustione, ..)

Outlier: considerazioni per input PMF

Cosa tolgo, cosa tengo?

Dipende!

Valutare di volta in volta:

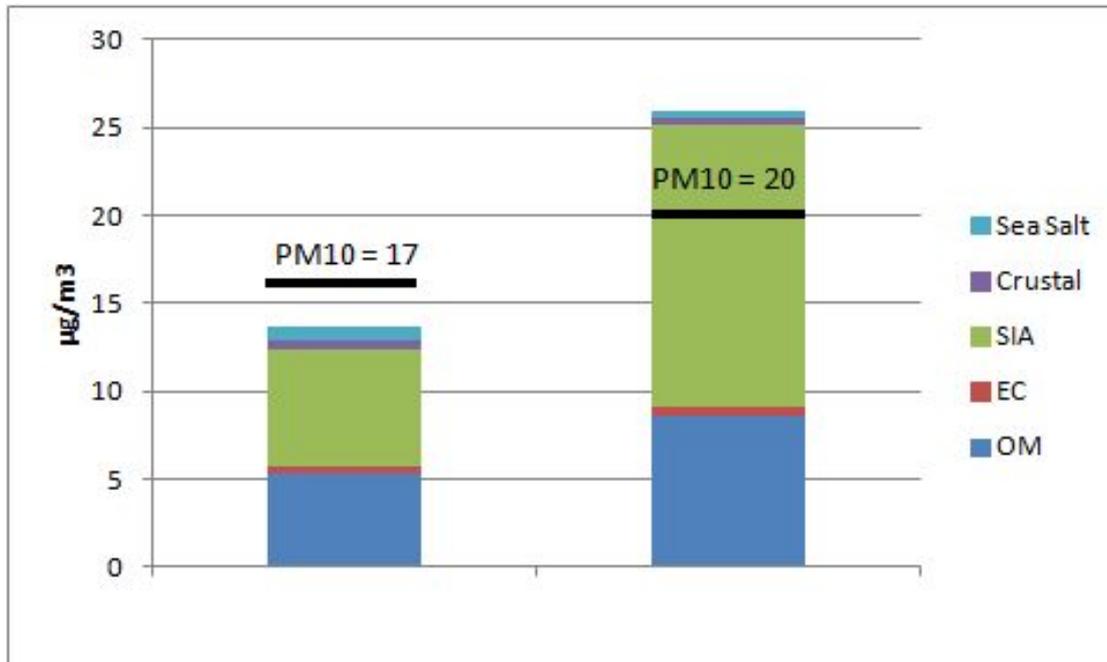
- tipo di dataset
- analisi chimica disponibile
- frazione granulometrica
- conoscenza del territorio
- numero di siti disponibili
- lavoro ad hoc o a posteriori
- bibliografia
- ...

Come “verificare” i risultati della componente chimica del PM?

- Chiusura del bilancio di massa
- Bilanciamento ionico
- Stima della componente crostale
- Scatterplot bivariati tra due analiti rispetto ai quali ci si aspetta una relazione lineare
- Rapporti diagnostici tra analiti di riferimento

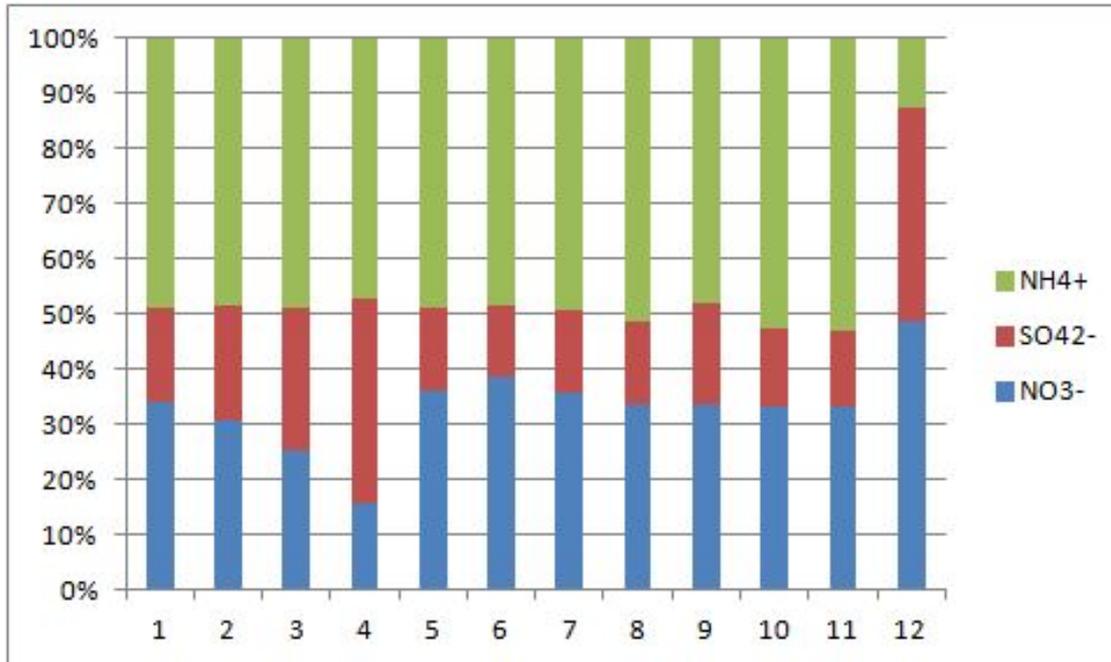
Come “verificare” i risultati della componente chimica del PM?

Esempio di chiusura di bilancio di massa sul PM10



chiusura
maggiore della
massa totale del
PM= errore
nella
concentrazione
della massa?

Come “verificare” i risultati della componente chimica del PM?



l'ultimo giorno c'è probabilmente un errore nel calcolo di 1 dei tre ioni (sovrastima di un anione o sottostima del catione)

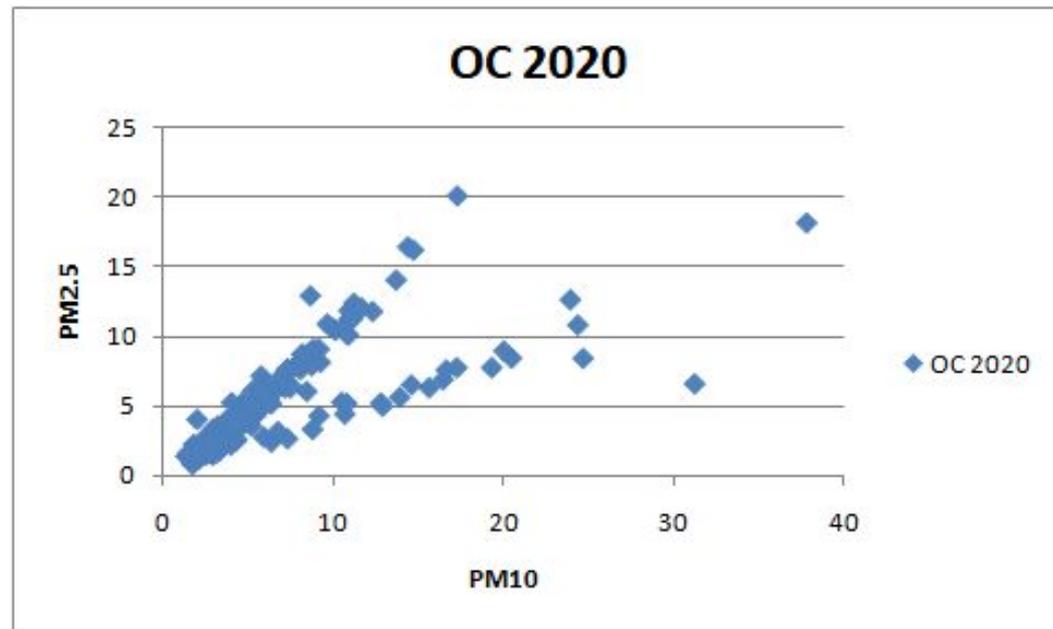
Come “verificare” i risultati della componente chimica del PM?

Rapporti diagnostici tra analiti di riferimento

Ratio	Value (Mean ± error)
SO_4^{2-}/S	3.59 ± 0.35
Molar NH_4^+/SO_4^{2-}	2.06 ± 0.04
OC/EC	5.88 ± 0.35
K^+/K	0.72 ± 0.06
Al/Si	0.48 ± 0.03
K/Fe	1.06 ± 0.06
Al/Ca	1.76 ± 0.07

I rapporti diagnostici sono di solito usati per valutare le possibili sorgenti (es. Ni/V per le navi, K+/EC per distinguere fra BB o fossil fuel combustion, oppure i rapporti tra gli IPA per valutare il tipo di combustione, ecc...) ma sono utili anche per valutare eventuali stranezze nei dati (Becagli et al., 2012; Viana et al., 2014; Fabbri et al. 2009)

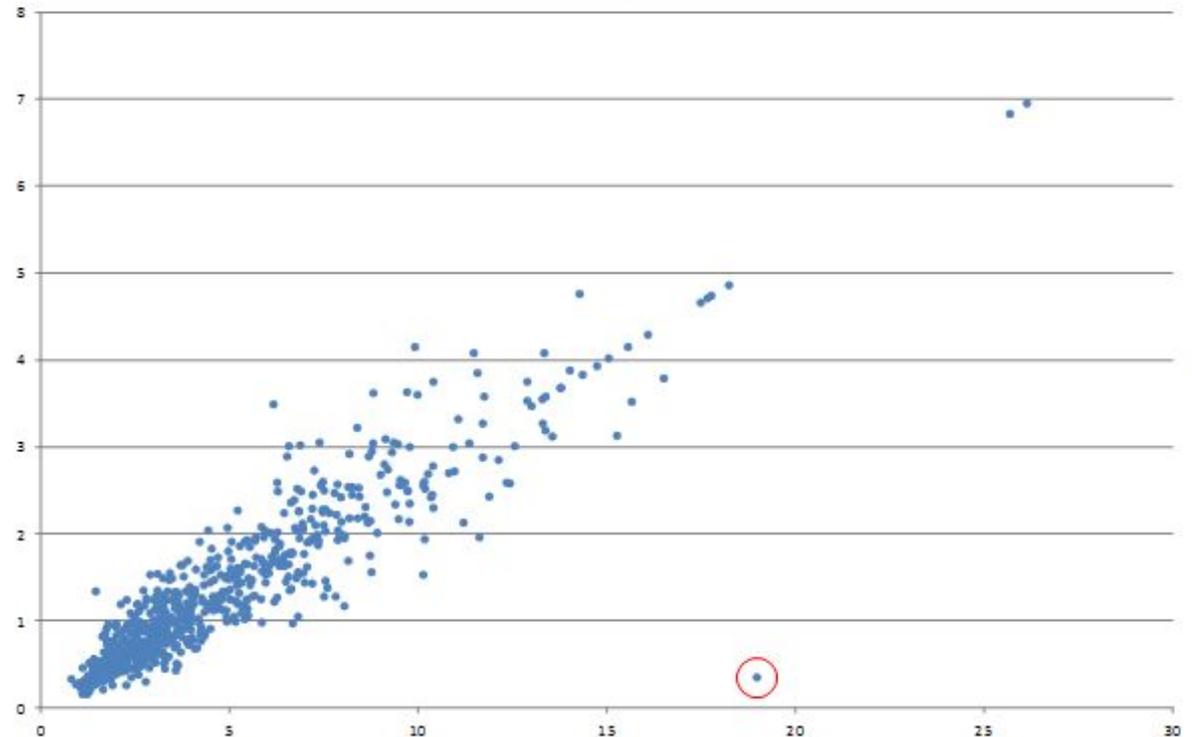
Come “verificare” i risultati della componente chimica del PM?



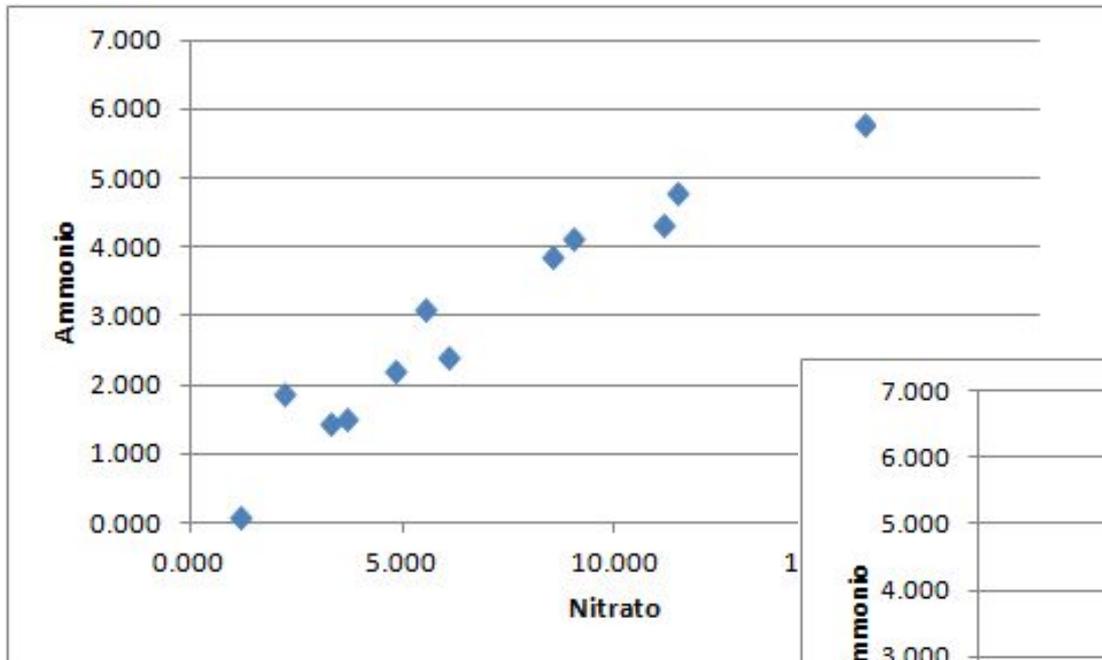
relazione lineare tra due tagli dimensionali a due code
→ necessario indagare

Come “verificare” i risultati della componente chimica del PM?

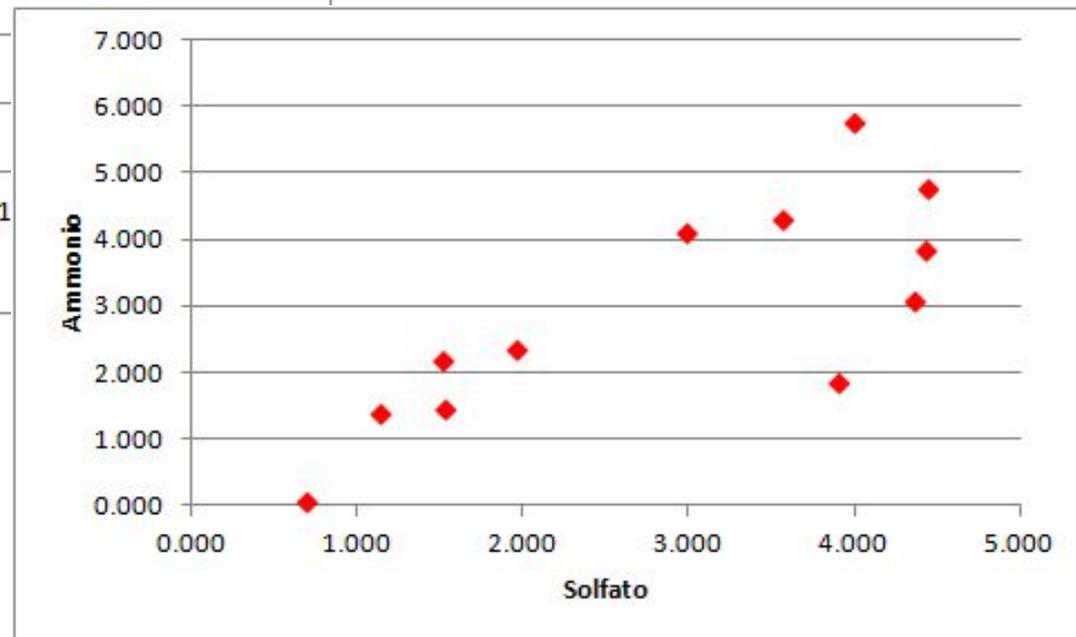
Scatterplot bivariati tra due analiti (EC e OC) rispetto ai quali ci si aspetta una relazione lineare



Come “verificare” i risultati della componente chimica del PM?



risposta lineare
tra due analiti



sorgente esterna non
secondaria del solfato (come
un'eruzione vulcanica o
trasporto dust)?

Come “verificare” i risultati della componente chimica del PM?

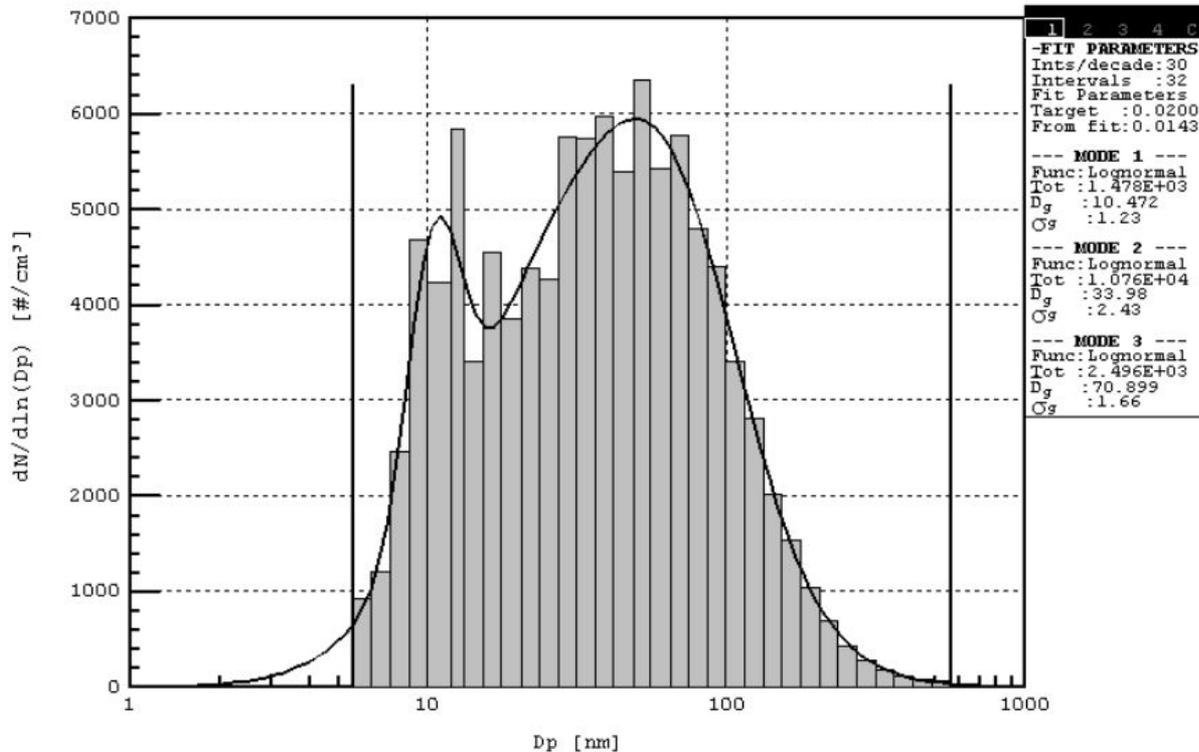
Stima della componente crostale, possiamo andare a valutare:

- la percentuale di crostale sul PM totale (se ricade fuori dall'intervallo atteso può indicare la presenza di contributi di dust, o altri eventi anomali o anche dati sbagliati)
- il rapporto tra le maggiori componenti

Può essere utile, per chi ne dispone, analizzare il dato di distribuzione dimensionale delle particelle non ultrafine (OPC)

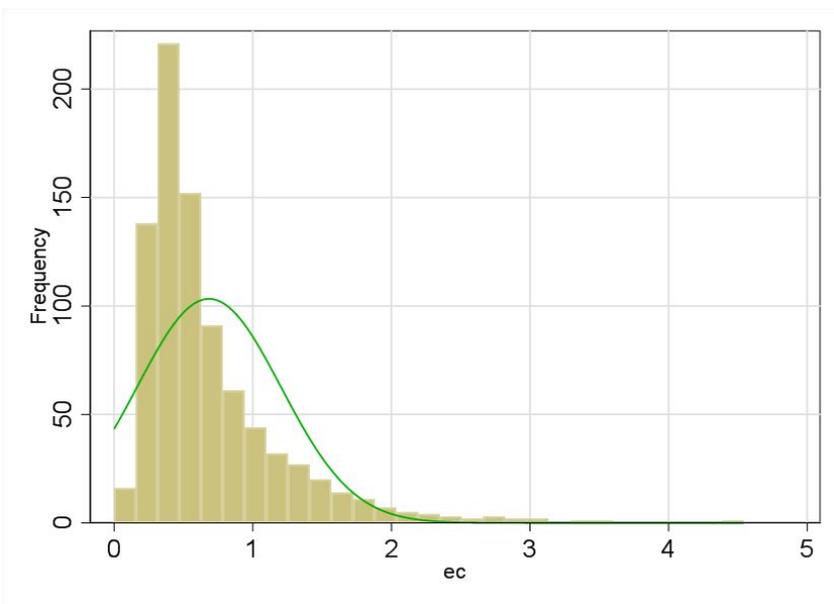
Verificare (anche solo graficamente) la forma distributiva degli analiti

→ dimensione delle particelle o del PM

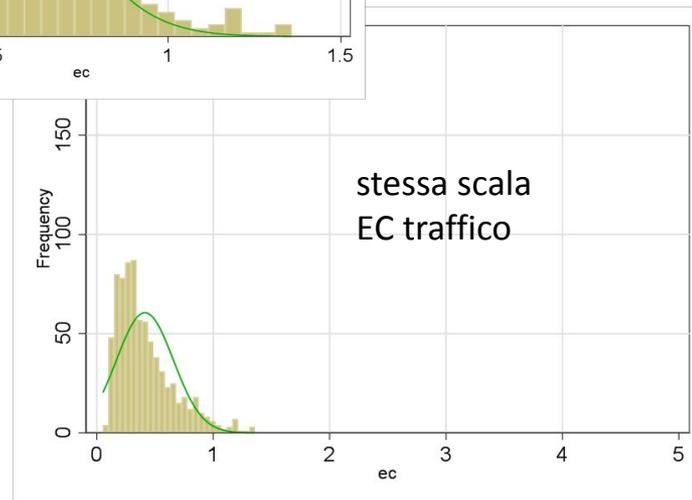
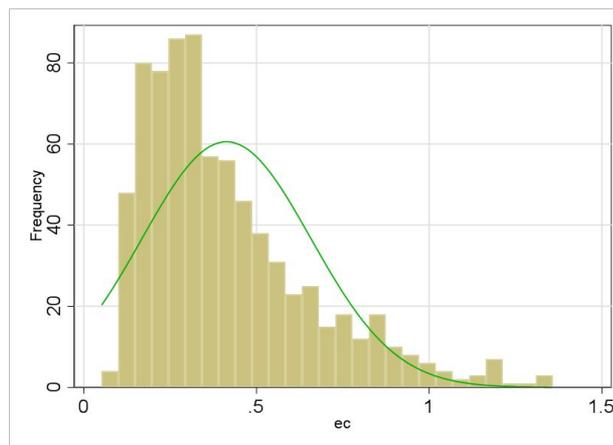


Verificare (anche solo graficamente) la forma distributiva degli analiti

→ EC sito da traffico



→ EC sito rurale

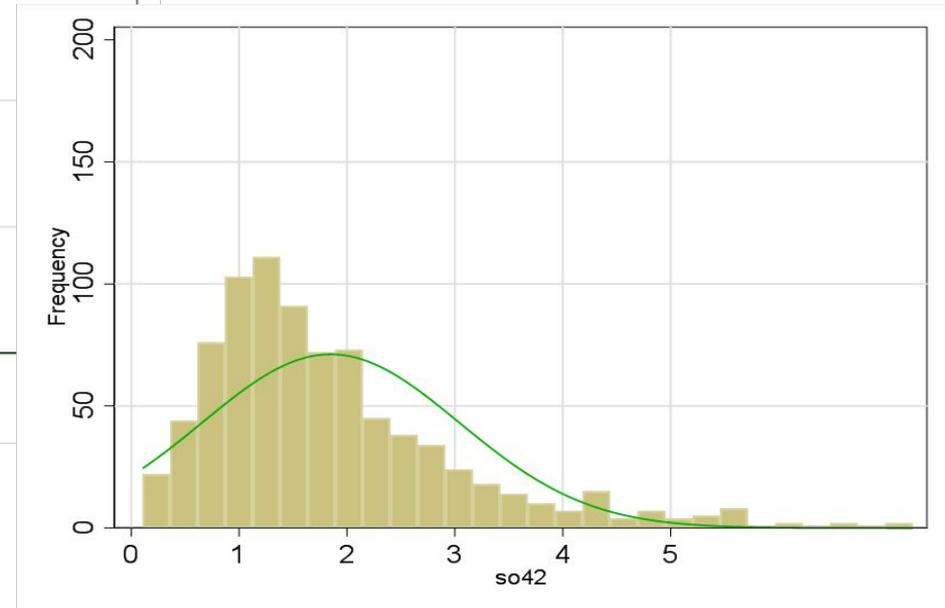
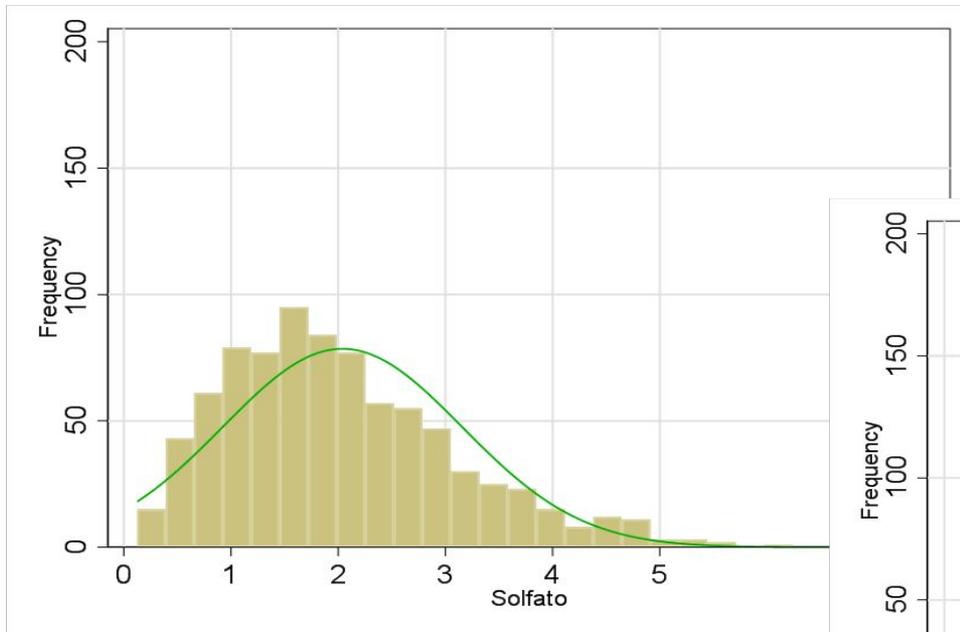


Verificare (anche solo graficamente) la forma distributiva degli analiti

→ solfato sito da traffico

→ solfato sito urbano

stessa scala



Dati mancanti (MD): perchè ci interessano?

La presenza di dati mancanti può influire

- sulle proprietà degli stimatori (prima tra tutti la media)
- può ridurre la rappresentatività del campione

→ possono quindi condurre a interpretazioni e considerazioni non corrette.

Dati mancanti (MD): perchè ci interessano?

Un dato derivante dalle analisi chimiche può essere “missing” per diversi motivi: può mancare il campione, può esserci stato un errore umano o strumentale (macchina rotta, blocco corrente..) nell’eseguire l’analisi o nel campionamento (filtro caduto, es. di casualità) oppure, per es. per la componente carboniosa, il filtro può aver campionato ma causa “impaccamento” (errore non casuale), es. non si riesce a differenziare EC da OC..

Infine un dato presente ma invalidato diventa un dato mancante

→ cercare di capire se il dato è mancante per motivi casuali o non casuali

Dati mancanti (MD): come trattarli?

SOSTITUZIONE / STIMA

Valore medio o mediano oppure valore di un dato “simile” di un'altra centralina che sia ben correlata oppure media di altre centraline ben correlate

Stimo il dato da una specie correlata o da una centralina correlata. Per es. se manca EC/OC per filtro “impaccato” lo stimo da TC poichè altamente correlato. Ovviamente non possibile per tutti gli analiti e servono delle info di background.

ESCLUSIONE DALLA ANALISI

La scelta deve basarsi su diversi fattori, numero di dati disponibili, contesto di analisi, numero di dati mancanti, loro distribuzione..

Dati mancanti (MD)

SOSTITUZIONE / STIMA

Valore medio o mediano oppure valore di un dato “simile” di un'altra centralina che sia ben correlata oppure media di altre centraline ben correlate

Stimo il dato da una specie correlata o da una centralina correlata. Per es. se manca EC/OC per filtro “impaccato” lo stimo da TC poichè altamente correlato. Ovviamente non possibile per tutti gli analiti e servono delle info di background.

ESCLUSIONE DALLA ANALISI

La scelta deve basarsi su diversi fattori, numero di dati disponibili, contesto di analisi, numero di dati mancanti, loro distribuzione..

ATTENZIONE!

OGNI TRASFORMAZIONE CAMBIA IL DATO ORIGINALE!

NON BUTTARE MAI IL DATO ORIGINALE!

TENERE TRACCIA DELLE MODIFICHE/SCELTE!



Focus dati esterni: specifiche per chi riceve e per chi invia

Quando si inviano file a colleghi esterni è necessario creare file leggibili, chiari, con **legenda** e specifiche
MA cosa si intende per file chiari e leggibili?

- evitare **colori**
- evitare **SiGLe** o codici numer1c1 (o in legenda specificare cosa significano)
 - controllare formato celle (data, numero, ..)
 - non raggruppare più celle
 - se possibile fare il minor n° di fogli
 - lasciare se possibile un riferimento email



Focus dati esterni: specifiche per chi riceve e per chi invia

→ implementazione in script

utilità: gestisco moltissimi dati, ho lo storico delle scelte che faccio

```

1 #basics dashboardPage
2
3 #install.packages("shinydashboard")
4
5 ## ui.R ## creo la pagina
6
7 ## app.R ##
8
9 library(shiny)
10 library(shinydashboard)
11
12 #server <- function(input, output) {
13
14 #shinyapp(ui, server)
15
16
17 ## app.R ## it's better if you put something in!
18
19 # ci metto qualcosa dentro
20
21 ui <- dashboardPage(
22   dashboardHeader(title = "You can Do It!"),
23   dashboardSidebar(),
24   dashboardBody(
25     # boxes need to be put in a row (or column)
26     "fluidrow",
27     box(plotOutput("plot1", height = 250)),
28     box(
29       title = "controls",
30       sliderInput("slider", "number of observations:", 1, 100, 50)
31     )
32   )
33 )
34
35 #}
36
37 #server <- function(input, output) {
38   set.seed(122)
39   histdata <- rnorm(500)
40
41   output$plot1 <- renderPlot({
42     data <- histdata[seq_len(input$slider)]
43     hist(data)
44   })
45 }
46
47 #topLevel
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
  
```

```

1 *****
2 *****
3 *****
4 *** FILE bologna_gg
5 *** DISTRIBUZIONI MULTIPLE
6 *****
7 *****
8 *****
9
10 clear
11 cd "M:\Aria\CTR_Aree Urbane\Supersito\da\lp1\spc\dmps\dati controllati"
12 use dmps_spc_2014_controllati_hh.dta, clear
13 *gen date =mdy(month,day,year)
14 order date
15 gen ss=0
16 gen mm=0
17 gen double dayhms=Cmdyhm3(month, day, year, hh, mm, ss)
18 *** unisco FMS MS e PD
19 mmerge year month day hh using "D:\Elaborazioni\smps_fm3\fm3_poair_hh_corr_da_long.dta"
20 keep if _merge=3
21 drop _merge
22
23 drop DateOfYear-dmyhms
24 drop pd_c6_04-pd_c523_3
25 *label var hh "hour (UTC+1)"
26
27 egen m_day=count(hh), by (year month day)
28 *drop if m_day<18
29 collapse (mean) ms_c6_04-ms_c523_3,by (year month day)
30
31 ren ms_c6_04 msc6
32 ren ms_c6_98 msc7
33 ren ms_c8_0 msc8
34 ren ms_c10_8 msc10
35 ren ms_c12_4 msc12
36 ren ms_c14_3 msc14
37 ren ms_c16_5 msc16
38 ren ms_c19_1 msc19
39 ren ms_c22_1 msc22
40 ren ms_c25_5 msc25
41 ren ms_c29_4 msc29
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
  
```

Focus dati esterni: specifiche per chi riceve e per chi invia

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X								
1	year	month	day	hh	tr	a	b	c	d	e	f	g	h	i	l	m	n	o	p	q	r	s	t									
2	2019	3	1	0	21609	8.6	10.86	10.86	6.427	98.75	15.72	5.871	1.385	0.291	0.02	0.005	4E-04	7.471	7.044	5.982	3.301	3301	4991	3105								
3	2019	3	1	1	21609	7.24	10.86	10.86	6.427	96.24	13.54	4.467	0.868	0.194	0.02	0.006	5E-04	6.873	6.42	5.632	3.422	3422	5497	3222								
4	2019	3	1	2	21609	6.24	10.86	10.86	6.427	94.36	13.03	4.18	0.783	0.178	0.018	0.006	7E-04	6.683	6.213	5.498	3.391	3391	5376	3204								
5	2019	3	1	3	21609	4.67	10.86	10.86	6.427	84.61	11.54	3.693	0.702	0.176	0.02	0.006	7E-04	6.115	5.627	4.912	3.24	3240	4883	3045								
6	2019	3	1	4	21609	4.28	10.86	10.86	6.427	85.29	11.79	3.736	0.703	0.171	0.018	0.005	8E-04	6.117	5.64	4.963	3.207	3207	4820	3021								
7	2019	3	1	5	21609	4.53	10.86	10.86	6.427	82.48	11.24	3.519	0.655	0.153	0.016	0.004	3E-04	5.738	5.407	4.785	3.252	3252	4919	3069								
8	2019	3	1	6	21609	5.35	10.86	10.86	6.427	82.64	11.27	3.519	0.657	0.164	0.019	0.005	6E-04	5.893	5.46	4.79	3.241	3241	4957	3076								
9	2019	3	1	7	21609	6.98	10.86	10.86	6.427	81.56	11.14	3.473	0.674	0.17	0.019	0.005	3E-04	5.81	5.42	4.737	3.205	3205	4871	3005								
10	2019	3	1	8	21609	8.86	10.86	10.86	6.427	79.43	10.84	3.438	0.705	0.195	0.024	0.007	8E-04	5.968	5.4	4.61	3.068	3068	4593	2887								
11	2019	3	1	9	21609	8.88	10.86	10.86	6.427	81.85	11.54	3.803	0.834	0.23	0.028	0.007	9E-04	6.342	5.703	4.788	3.026	3026	4554	2845								
12	2019	3	1	10	21609	7.57	10.86	10.86	6.427	85.22	11.79	3.813	0.808	0.215	0.023	0.006	0.001	6.382	5.797	4.968	3.065	3065	4648	2888								
13	2019	3	1	11	21609	5.99	10.86	10.86	6.427	68	9.695	3.157	0.683	0.172	0.016	0.004	6E-04	5.018	4.637	3.99	3.075	3075	4674	2895								
14	2019	3	1	12	21	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
15	2019	3	1	13	21	year	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	
16	2019	3	1	14	21	month	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
17	2019	3	1	15	21	day	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
18	2019	3	1	16	21	hh	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
19	2019	3	1	17	21	tr	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	21610	2
20	2019	3	1	18	21	a	8.6	7.24	6.24	4.67	4.28	4.53	5.35	6.98	8.86	8.88	7.57	5.99	5.31	3.6	2.92	3.12	3.28	5.74	6.69	6.99	7.91	8.99	10.86	9.64	6.86	
21	2019	3	1	19	21	b	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	7.55	
22	2019	3	1	20	21	c	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	10.86	7.55	
23	2019	3	1	21	21	d	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	6.427	
24	2019	3	1	22	21	e	98.75	96.24	94.36	84.61	85.29	82.48	82.64	81.56	79.43	81.85	85.22	68	52.6	46.91	46.73	44.32	53.48	76.23	62.16	66.82	71.77	69.7	68.93	64.32	98.75	
25	2019	3	1	23	21	f	15.72	13.54	13.03	11.54	11.79	11.24	11.27	11.14	10.84	11.54	11.79	9.695	7.691	6.77	6.939	6.619	8.49	11.1	9.78	10.47	10.67	10.05	9.918	9.875	15.72	
26	2019	3	2	0	21	g	5.871	4.467	4.18	3.693	3.736	3.519	3.519	3.473	3.438	3.803	3.813	3.157	2.578	2.285	2.4	2.326	3.378	4.402	3.8	3.983	3.874	3.534	3.483	3.38	5.871	
27	2019	3	2	0	21	h	1.385	0.868	0.783	0.702	0.703	0.655	0.657	0.674	0.705	0.834	0.808	0.683	0.603	0.534	0.567	0.575	0.959	1.257	1.082	1.05	0.957	0.855	0.827	0.781	1.385	
28	2019	3	2	1	21	i	0.291	0.194	0.178	0.176	0.171	0.153	0.164	0.17	0.195	0.23	0.215	0.172	0.17	0.15	0.157	0.163	0.293	0.417	0.366	0.328	0.278	0.244	0.227	0.235	0.291	
29	2019	3	2	1	21	j	0.02	0.02	0.018	0.02	0.018	0.016	0.019	0.019	0.024	0.028	0.023	0.016	0.019	0.018	0.018	0.022	0.042	0.068	0.061	0.047	0.033	0.025	0.022	0.025	0.02	
30	2019	3	2	2	21	k	0.005	0.006	0.006	0.006	0.005	0.004	0.005	0.005	0.007	0.007	0.006	0.004	0.006	0.006	0.004	0.002	0.012	0.02	0.017	0.013	0.009	0.007	0.005	0.007	0.005	
31	2019	3	2	3	21	l	4E-04	5E-04	7E-04	7E-04	8E-04	3E-04	6E-04	3E-04	8E-04	9E-04	0.001	6E-04	8E-04	4E-04	6E-04	5E-04	0.001	0.002	0.002	9E-04	0.001	7E-04	6E-04	5E-04	4E-04	
32	2019	3	2	4	21	m	7.471	6.873	6.683	6.115	6.117	5.738	5.893	5.81	5.968	6.342	6.382	5.018	4.263	3.8	3.778	3.788	5.382	7.977	6.71	6.347	6.123	5.628	5.383	5.243	7.471	
33	2019	3	2	5	21	n	7.044	6.42	6.213	5.627	5.64	5.407	5.46	5.42	5.4	5.703	5.797	4.637	3.76	3.375	3.37	3.308	4.44	6.298	5.302	5.352	5.343	5.042	4.91	4.715	7.044	
34	2019	3	2	6	21	o	5.982	5.632	5.498	4.912	4.963	4.785	4.79	4.737	4.61	4.788	4.968	3.99	3.098	2.762	2.763	2.63	3.237	4.53	3.738	4.018	4.262	4.107	4.067	3.818	5.982	
35	2019	3	2	7	21	p	3.301	3.422	3.391	3.24	3.207	3.252	3.241	3.205	3.068	3.026	3.065	3.075	2.884	3.013	3.065	3.044	2.867	2.799	2.778	2.767	2.747	2.79	2.799	2.806	2.832	
36	2019	3	2	8	21	q	3301	3422	3391	3240	3207	3252	3241	3205	3068	3026	3065	3075	2884	3013	3065	3044	2867	2799	2778	2767	2747	2790	2799	2806	2832	
37	2019	3	2	9	21	r	4991	5497	5376	4883	4820	4919	4957	4871	4593	4554	4648	4674	4511	4632	4727	4722	4536	4488	4467	4469	4445	4483	4470	4452	4462	
38	2019	3	2	10	21	s	3105	3222	3204	3045	3071	3069	3076	3005	2887	2845	2888	2895	2687	2829	2868	2857	2674	2596	2592	2586	2559	2592	2610	2632	2627	

“un file noioso è un file felice”

Focus dati esterni: specifiche per chi riceve e per chi invia

quando il numero dei dati è elevato si rende necessario/utile l'utilizzo di linguaggi di programmazione

→ **implementazione in script**

dati al minuto o mediate su 5-10 minuti, dati orari con un numero elevato di informazioni (> 100 canali per strumenti come i MPSS), ecc.

l'analisi della PMF è utilizzata anche per dati orari (misure di chimica oraria (AMS), di particelle (OPC, MPSS), black carbon, ecc.)

importanza anche qui di guardare il dato, grafici descrittivi, ecc..

Focus dati orari

Linee Guida WHO (update Sep 2021, **Good practice statements about other PM types**): parametri non normati ma di cui si consiglia la misura come UFP e BC.

Nuova proposta della Commissione europea nell'ambito della Direttiva della qualità dell'aria ambiente (2008/50/CE) e di quella concernente l'arsenico, il cadmio, il mercurio, il nickel e gli IPA nell'aria ambiente (2004/107/CE)

Proposal: DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on ambient air quality and cleaner air for Europe (recast)

Article 10 introduces monitoring supersites and regulates their number and location. [...] Introducing additional sampling points for **unregulated air pollutants** of emerging concern, such as UFP, BC, NH₃ or the oxidative potential of particulate matter, will support scientific understanding of their effects on health and the environment.



Grazie