Tecniche robuste per il filtraggio dei dati di qualità dell'aria acquisiti da sensori a basso costo

Dario Perregrini[1][0000-0001-8455-4892], Vittorio Casella[1][0000-0003-2086-7931]

[1] DICAR – Università degli Studi di Pavia, (dario.perregrini; vittorio.casella)@unipv.it

Abstract. L'obiettivo degli studi effettuati è quello di sviluppare una tecnica di filtraggio robusta per i dati provenienti dai sensori PurpleAir installati nel comune di Pavia per il monitoraggio del particolato presente nell'aria in varie zone della città. La massiccia raccolta di dati e la loro analisi rappresentano il punto di partenza, per esempio, per la stima della *personal exposure* ovvero la valutazione della quantità di inquinante assorbita da un soggetto. Attualmente la tecnica sviluppata restituisce ottimi risultati riuscendo a filtrare molto bene gli errori grossolani presenti nelle misurazioni dei sensori rispetto ad altri metodi già noti in letteratura, incentivandone future ottimizzazioni e migliorie.

Keywords: Inquinamento urbano, filtraggio robusto, PurpleAir, monitoraggio, qualità dell'aria, sensori a basso costo, metodi adattivi e robusti.

1 INTRODUZIONE: INQUINAMENTO NEL NORD ITALIA E RETE DI RILEVAMENTO

La qualità dell'aria influenza in modo determinante la qualità e la lunghezza della vita dei cittadini [1], infatti la causa di alcune morti premature è direttamente attribuibile all'esposizione a particolato sottile (PM 2,5). In base a quanto riportato dall'Agenzia europea dell'ambiente nel report annuale "Air quality in Europe", in Italia vengono stimate 59500 morti all'anno attribuibili al PM 2,5, facendoci purtroppo guadagnare la prima posizione a parimerito con la Germania rispetto agli altri paesi dell'Unione Europea [2]. I valori limite per la protezione della salute umana fissati per il PM 10 nella direttiva 2008/50/CE del Parlamento europeo, sono pari a 50 $\mu m/m^3$ da non superare per più di 35 volte per anno civile con una tolleranza prevista pari al 50% di tale concentrazione e un ulteriore limite di 40 $\mu m/m^3$ nell'arco dell'anno civile con una tolleranza del 20%, entrambi i limiti sono entrati in vigore dal 1° gennaio 2005.

Come riportato nel rapporto "Mal'aria di città" redatto annualmente da Legambiente sono 35, su 96, le città capoluogo nelle quali in almeno una centralina di monitoraggio viene registrato un valore oltre il limite giornaliero previsto per le polveri sottili citato precedentemente; undici le città nelle quali si sono avuti più del doppio dei giorni di superamento dei limiti. In particolare, questi limiti vengono superati per la maggior parte nelle città presenti nella zona della pianura padana che vede in prima posizione la città di Torino con ben 98 giorni di superamenti quasi il triplo del consentito [3]. Ciò è causato dalle naturali delimitazioni costituite da Alpi e Appennini che

fungendo da barriera naturale creano le condizioni ideali per l'accumulo delle polveri sottili e in generale di vari inquinanti in una delle zone più densamente popolate della penisola. Attraverso le immagini messe a disposizione da ESA (European Space Agency) è possibile avere un riscontro visivo immediate di tale fenomeno confrontando una ricostruzione dell'area padana libera da nuvole e smog con un'immagine acquisita senza particolari elaborazioni.

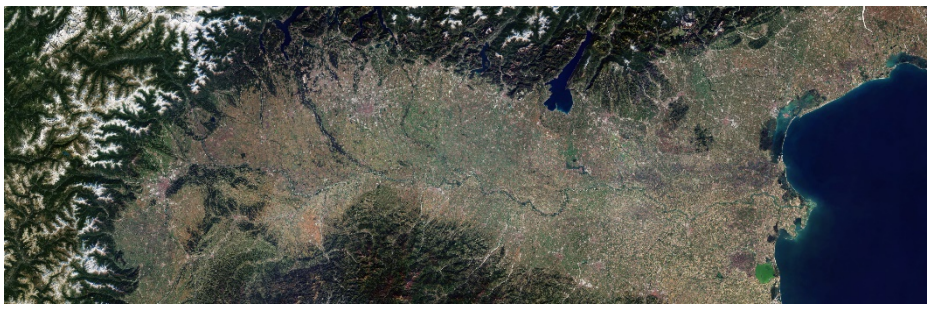

Fig. 1. Area della pianura padana libera da nuvole e smog ottenuta tramite la composizione di più immagini acquisite dal satellite Sentinel-2 del programma europeo Copernicus tra giugno 2018 e febbraio 2019. [4]

Fig. 2. Immagine catturata dal satellite Sentinel-3A il 16 febbraio 2019. [4]

La foschia che è possibile vedere nella seconda immagine molto probabilmente è dovuta ad un mix di nebbia e smog intrappolato alla base delle alpi non solo per i motivi legati alla topografia del territorio precedentemente descritti, ma anche a causa delle condizioni atmosferiche che in diversi momenti dell'anno ne agevolano l'accumulo.

Nell'ambito del progetto H2020 PULSE [5] che si è concluso il 30/4/2020 è stata creata una rete composta da sensori a basso costo nella città di pavia, lo scopo ultimo del progetto, infatti prevedeva l'acquisizione di una gran quantità di dati relativi alle concentrazioni atmosferiche del PM allo scopo di innovare l'ambito del monitoraggio dell'inquinamento e di sviluppare modelli numerici per la valutazione della qualità dell'aria. Attualmente a Pavia il PM ed altri inquinanti vengono monitorati tramite due centraline di proprietà di Arpa Lombardia che forniscono i valori delle concentrazioni su base giornaliera, la principale innovazione è avvenuta tramite la costituzione di una

rete di sensori a basso costo che è in grado di fornire una misurazione molto densa sia nello spazio che nel tempo. Vengono infatti monitorate le concentrazioni atmosferiche del particolato (PM10, PM2.5, PM1.0) ad intervalli di 2 minuti e ciò avviene per ciascuno dei 40 sensori installati uniformemente all'interno del tessuto urbano. In questo modo è possibile valutare le fluttuazioni del PM nelle varie zone della città in momenti diversi della giornata.

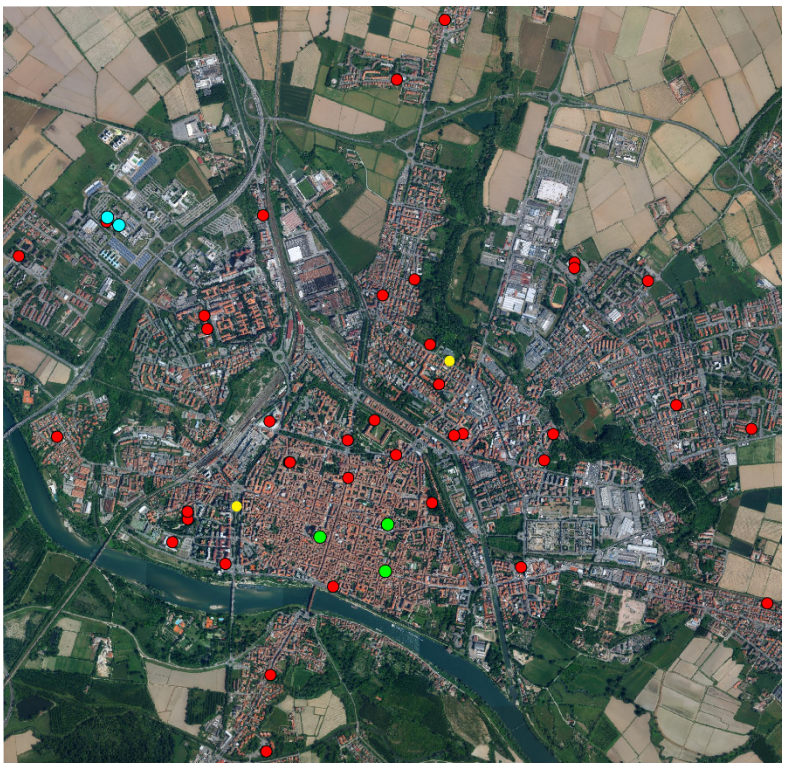

Fig. 3. Posizione dei sensori PurpleAir (Rosso) centraline di proprietà Arpa Lombardia (Giallo) presenti a Pavia, Sensori P@P-PA-4_1, P@P-PA-43_1, P@P-PA-34_1 (Verde), Sensori P@P-PA-2_1, P@P-PA-5_1 (Ciano).

Avere a disposizione un quadro così dettagliato di tali concentrazioni è un elemento fondamentale per poter implementare il calcolo della *personal exposure*, ovvero una tecnica avanzata attraverso la quale è possibile valutare l'esposizione di un individuo ad un determinato inquinante, ricavando la quantità di microgrammi assorbiti in funzione del tipo di attività svolta e degli spostamenti effettuati in un certo lasso di tempo. I sensori PurpleAir utilizzati per la costruzione della rete di monitoraggio a Pavia rappresentano un buon compromesso tra costo, semplicità d'installazione e qualità delle misurazioni effettuate, che però come tutti i sensori a basso costo presentano delle rilevazioni spesso influenzate da un elevato *noise* e un discreto numero di *outlier*. Si pone

quindi il problema di sviluppare delle procedure di filtraggio per limitare questi disturbi, ricavandone un segnale il più possibile di qualità che colga le reali concentrazioni del particolato atmosferico e le fluttuazioni che naturalmente si manifestano nelle stesse al variare delle condizioni atmosferiche nell'arco della giornata. Nell'ambito di un lavoro di tesi e di successivi studi è stata sviluppata una tecnica robusta e adattiva, il seguente articolo si pone l'obbiettivo di illustrarne il concetto e i risultati ottenuti applicandola alle misurazioni riportate dai sensori PurpleAir.

2 SENSORI UTILIZZATI

I sensori adottati sono i PurpleAir, grazie alla loro semplicità di installazione è stato possibile posizionarli non solo in varie aree pubbliche come il polo universitario ma anche presso le abitazioni di privati cittadini, un sensore per poter funzionare necessita solamente di una presa elettrica per l'alimentazione e una connessione WiFi per permettere la trasmissione delle misurazioni.

Fig. 4. Sensore PurpleAir PA-II-SD, [6].

Come è possibile notare dalle immagini si tratta di un sensore di dimensioni ridotte che tramite due contatori laser ricava il numero delle particelle che vengono convogliate all'interno del sensore da una ventola posta alla base di ogni misuratore. Per ogni sensore si hanno quindi due canali di misura, ciò costituisce per noi un elemento di supporto non indifferente nell'individuazione di errori grossolani nelle misurazioni, se solo uno dei due canali presenta un picco nelle concentrazioni molto probabilmente quel picco sarà un *outlier*. Inoltre, in questo modo è possibile continuare a raccogliere dati

nel caso in cui un misuratore si guastasse, garantendo una continuità nella raccolta dati nella zona in cui il sensore è installato.

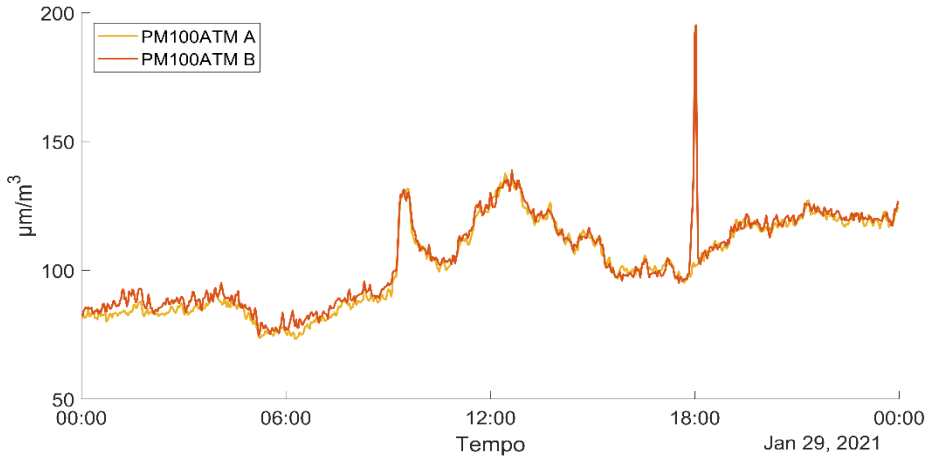

Fig. 5. Esempio di dati registrati dal sensore P@P-PA-4_1 il 29 gennaio 2021.

Osservando il risultato dei dati trasmessi da un sensore è possibile identificare molto prime ore della giornata si può notare molto bene la presenza di un disturbo di fondo detto *noise*. In corrispondenza delle 18 si verifica un *outlier* sul canale di misura B e possiamo identificarlo facilmente osservando di quanto il valore si discosta rispetto all'andamento generale e perché non si registra alcun aumento delle concentrazioni nel canale A. Le misurazioni riportate da questo tipo di sensori nonostante la presenza di svariarti errori mostrano una buona correlazione con i risultati ottenuti da delle centraline (MetOne BAM e GRIMM) molto più costose e di conseguenza più accurate, secondo quanto riportato nel report di AQ-Spec relativo alla Field Evaluation [7]. Viene evidenziato come il parametro R^2, da interpretare come 1 per misure in completo accordo e 0 per dati in completo disaccordo, sia molto elevato soprattutto per il PM 1.0 (R^2 > 0.96 vs GRIMM) e per il PM 2.5 (R^2 >0.93 vs GRIMM, R^2 >0.86 vs BAM) mentre per il PM 10 (R^2 >0.68 vs GRIMM, R^2 >0.60 vs BAM) valori considerevolmente più bassi mostrando una correlazione minore tra le misurazioni relative al PM 10 riportate dai tre sensori PurpleAir e quelle riportate dalle centraline BAM e GRIMM.

3 TIPOLOGIE DI ERRORE

Il *noise* è sempre presente all'interno dei segnali e la sua intensità in accordo con i dati tecnici relativi ai sensori forniti dal produttore, per le concentrazioni tra 0÷100 µm/m3 l'incertezza nella misurazione è pari a ± 10 µm/m3, mentre da 100÷500 µm/m3 l'incertezza nella misura è pari al ±10% del valore rilevato. Indentificare gli errori grossolani non è altrettanto semplice, se l'incremento non è elevato e puntuale non è da escludere che sia un evento reale anziché un errore, per poterlo determinare con certezza è necessario effettuare dei confronti con altri sensori posti nelle vicinanze.

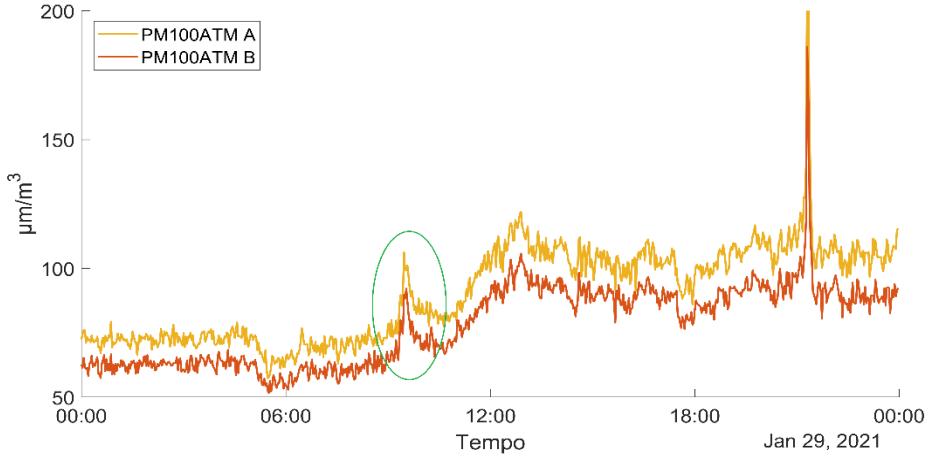

Fig. 6. Esempio di dati registrati dal sensore P@P-PA-43_1 il 29 gennaio 2021.

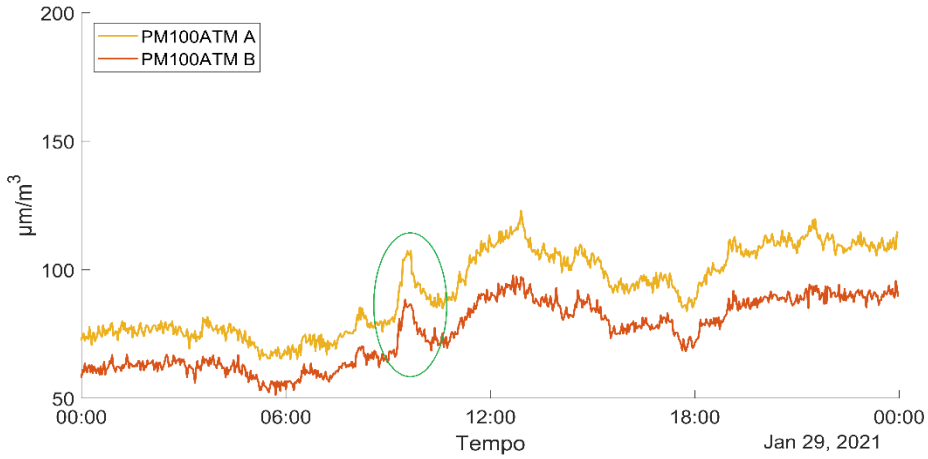

Fig. 7. Esempio di dati registrati dal sensore P@P-PA-34_1 il 29 gennaio 2021.

Siccome l'incremento evidenziato in verde si verifica indicativamente nello stesso orario nelle misurazioni di vari sensori presenti nella stessa zona ovvero quella del centro città, siamo portati a pensare che sia una misura reale, una brusca diminuzione delle concentrazioni del particolato probabilmente dovuto a fenomeni ventosi. L'unico modo per identificarli correttamente è appunto tramite confronti tra le misurazioni dei vari dispositivi, quindi sorge la necessità di sviluppare una tecnica di filtraggio che sia in grado non solo di eliminare gli errori ma allo stesso tempo di cogliere queste repentine fluttuazioni.

4 METODOLOGIA

La tecnica di filtraggio sviluppata permette di approssimare l'andamento delle concentrazioni riportate in modo adattivo, ovvero selezionando per ogni intervallo di tempo il polinomio che meglio interpola i dati grezzi dopo averli filtrati da errori grossolani. Ciò avviene separando una giornata in sotto intervalli, sei intervalli da quattro ore ciascuno, inoltre ogni intervallo prevede una sorta di fascia di tolleranza della durata di un'ora all'inizio e alla fine di ogni intervallo, questo per avere delle misurazioni in comune tra una fascia e l'altra e quindi risultati più concordi nei punti di contatto di ogni intervallo.

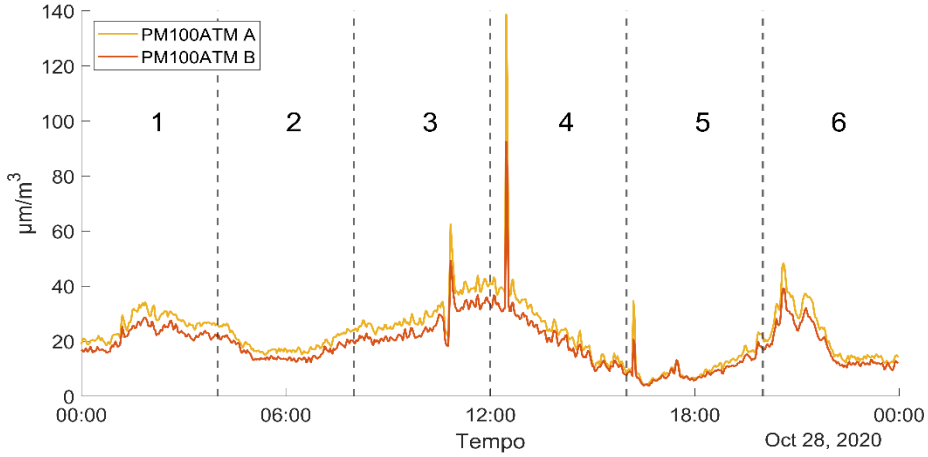

Fig. 8. Esempio di dati registrati dal sensore P@P-PA-2_1 il 28 ottobre 2020 con le fasce orarie di filtraggio.

Tabella 1. Intervalli di tempo adottati per il filtraggio all'interno di una giornata di misure

INTERVALLO	ORARIO	TOLLERANZA	INTERVALLO FILTRATO
1	00:00 ÷ 04:00	± 1 ora	23:00 ÷ 04:58
2	04:00 ÷ 08:00	± 1 ora	03:00 ÷ 08:58
3	08:00 ÷ 12:00	± 1 ora	07:00 ÷ 12:58
4	12:00 ÷ 16:00	± 1 ora	11:00 ÷ 16:58
5	16:00 ÷ 20:00	± 1 ora	15:00 ÷ 20:58
6	20:00 ÷ 00:00	± 1 ora	19:00 ÷ 00:58

Definiti gli intervalli di tempo dai quali prelevare i dati da filtrare, viene effettuata un'operazione iterativa a livello dell'interpolazione polinomiale, ovvero su ogni intervallo l'algoritmo effettua una prima interpolazione sui singoli canali in modo separato, allo scopo di eliminare tutti i valori che si trovano al di fuori della fascia di tolleranza che viene costruita intorno ad ogni interpolazione con ampiezza pari a 0.75 volte la deviazione standard dei dati grezzi osservati.

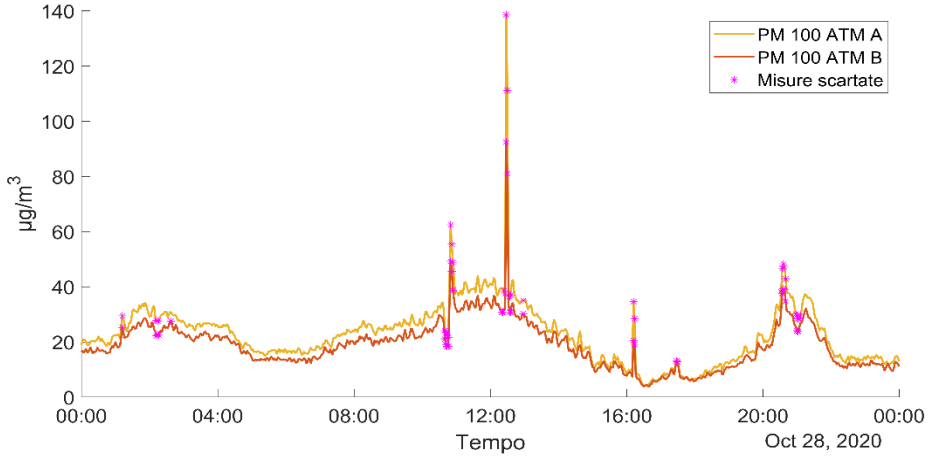

Fig. 9. Misure del sensore P@P-PA-2_1 il 28 ottobre 2020 scartate perché al di fuori della fascia di tolleranza.

Successivamente con i dati restanti di entrambi i canali si effettua un ulteriore interpolazione ottenendo un andamento che tiene quindi in considerazione tutte le misurazioni rimaste. Questa operazione viene effettuata per i gradi polinomiali da 1 a 8 in modo intelligente, l'algoritmo infatti confronta la bontà dell'interpolazione ottenuta con il polinomio di grado i con quello i + 1, se passando al grado successivo si ottiene un notevole incremento di qualità del polinomio interpolante si ripete la procedura fino al grado massimo 8. L'indicatore di qualità utilizzato deriva dalla combinazione di due parametri statistici relativi all'interpolazione effettuata, che sono R^2, descritto in precedenza e RMSE (errore quadratico medio) è una misura di errore assoluta dove elevando al quadrato le deviazioni si evita che i valori positivi e negativi si possano annullare a vicenda. Combinando questi due parametri con la seguente formula, otteniamo un parametro che ci dà una buona misura di quanto il grado d'interpolazione utilizzato sia buono:

$$Q = (1-R^2) * RMSE$$

Più tale valore è prossimo allo zero e maggiore sarà la qualità generale dell'interpolazione, dal confronto tra il valore di Q del grado polinomiale i con il grado polinomiale "i + 1" l'algoritmo valuta o meno se sia necessario procedere all'utilizzo di un grado polinomiale più elevato se riscontra un sostanzioso aumento di qualità. Ovvero nel caso in cui passando al grado successivo si verifichi una diminuzione del parametro Q maggiore del 10% del valore attuale, prosegue testando gradi più elevati altrimenti procede controllando che l'attuale grado adottato presenti un valore di $R^2 > 0.9$ e $Q < 1$, se tali condizioni si verificano l'algoritmo si arresta, adottando il grado polinomiale raggiunto e passa all'intervallo di tempo successivo.

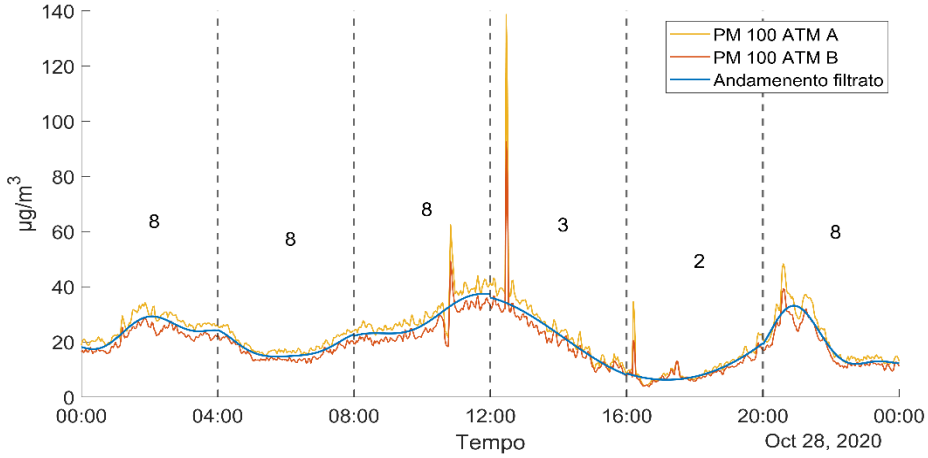

Fig. 10. Andamento risultante dopo l'interpolazione sui dati rimanenti del sensore P@P-PA-2_1 il 28 ottobre 2020, il numero contenuto in ogni intervallo rappresenta il grado polinomiale al quale l'algoritmo di è interrotto raggiunta la qualità minima imposta per l'interpolazione.

Osservando i risultati ottenuti anche su altri sensori risulta evidente come in alcuni casi per ottenere una buona approssimazione dei dati grezzi siano sufficienti polinomi di basso grado come rette e parabole, e l'andamento filtrato del sensore P@P-PA-5_1 ne è un ottimo esempio. Questo permette un notevole risparmio di tempo computazionale richiesto per le operazioni di filtraggio nel caso in cui fosse necessario filtrare una gran quantità di dati su base giornaliera.

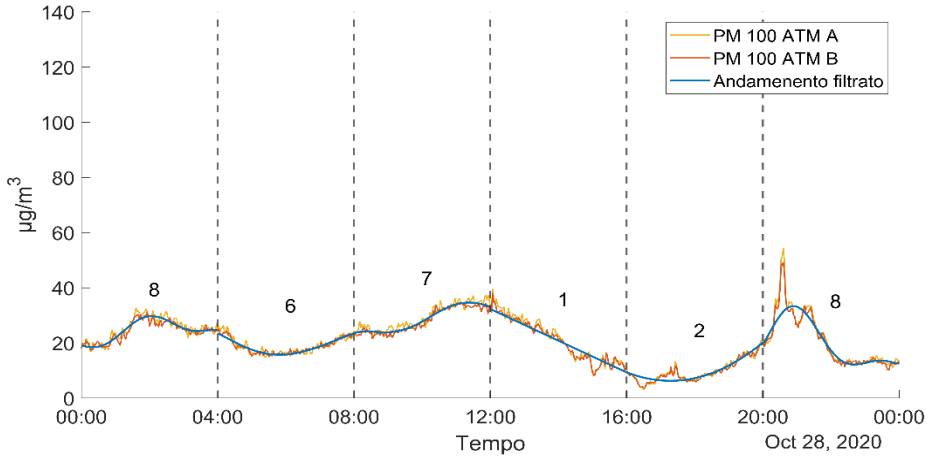

Fig. 11. Filtraggio applicato al sensore P@P-PA-5_1 il 28 ottobre 2020, con grado polinomiale adottato in ogni intervallo.

5 CONFRONTO CON ALTRI METODI

Per filtrare al meglio i dati riportati da questo tipo di sensori è stato necessario sviluppare una tecnica di individuazione e filtraggio apposita, in quanto metodi già noti in letteratura come il metodo di Hampel, non si sono dimostrati adatti a gestire la forte variabilità delle misurazioni.

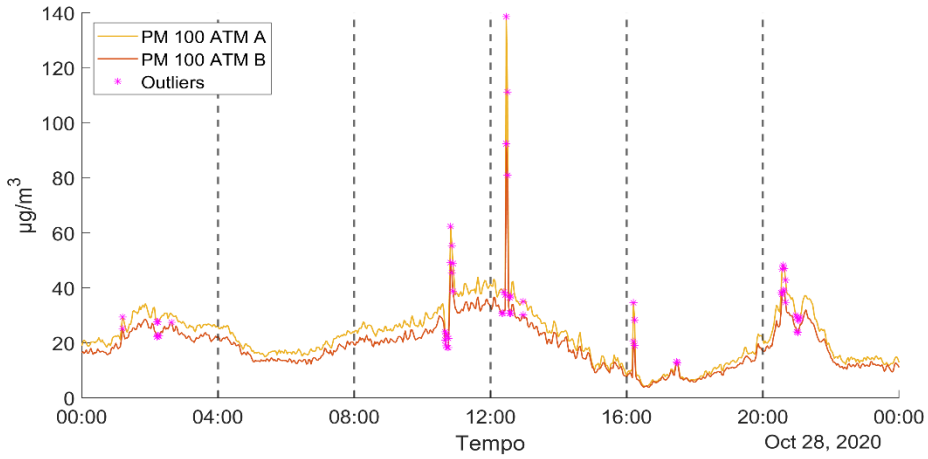

Fig. 12. Andamento risultante dal filtraggio effettuato con la nostra tecnica di rimozione degli *outlier* sulle misurazioni del sensore P@P-PA-2_1 il 28 ottobre 2020.

Il metodo di Hampel è una tecnica che sfrutta una finestra mobile di ampiezza pari a due volte il numero di elementi specificati più l'elemento considerato, su questa finestra viene calcolata la mediana e la deviazione standard degli elementi considerati rispetto alla mediana. Successivamente se un punto si discosta dalla mediana calcolata di tre volte il valore della deviazione standard il valore di quel punto viene sostituito con il valore della mediana. Nelle immagini successivamente riportate quindi vediamo i risultati del metodo di Hampel prima nella configurazione standard ovvero considerando un'ampiezza della finestra mobile pari a tre elementi, e successivamente estendendola a 90 elementi per simulare un intervallo di ampiezza pari a quella utilizzata dal nostro metodo. I risultati della prima configurazione mostrano come un *outlier* costituito non da un'unica misura che registra un valore elevatissimo, ma da un serie breve di tre o più misurazioni errate, venga ritenuto dall'algoritmo come un dato valido, quando in realtà anche dai confronti con i sensori nei dintorni è immediato attribuire l'aumento repentino delle concentrazioni atmosferiche del particolato ad un errore grossolano.

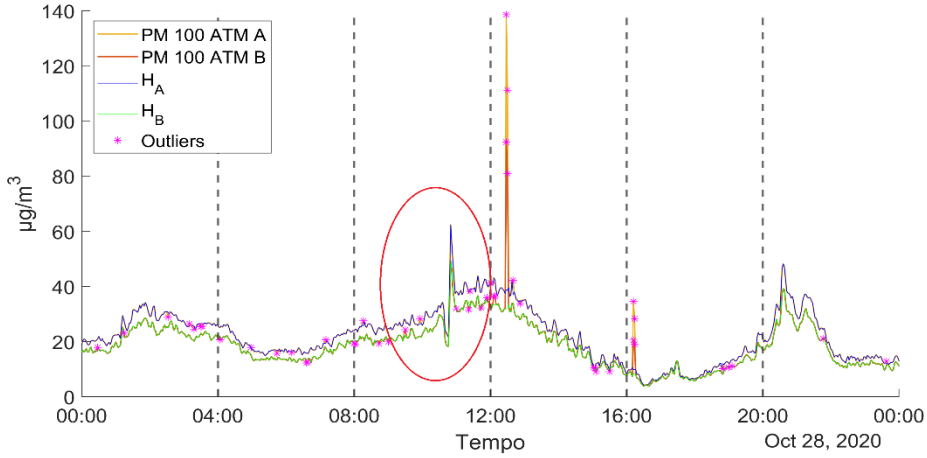

Fig. 13. Andamento risultante dal filtraggio utilizzando il metodo di Hampel standard per la rimozione degli *outlier* applicato alle misurazioni del sensore P@P-PA-2_1 il 28 ottobre 2020.

Nella seconda casistica, assumendo una finestra mobile con ampiezza pari agli intervalli utilizzati dalla nostra tecnica, non solo il picco precedentemente analizzato non viene filtrato completamente ma solo in parte, in più abbiamo un effetto di filtraggio esteso a tutto la zona evidenziata in verde. Dove si stanno quindi associando a tutte quelle misurazioni ritenute dal metodo di Hampel come errori grossolani i valori della mediana relativa a tutto l'ultimo intervallo, cambiando drasticamente i dati da utilizzare per il passaggio dell'interpolazione polinomiale.

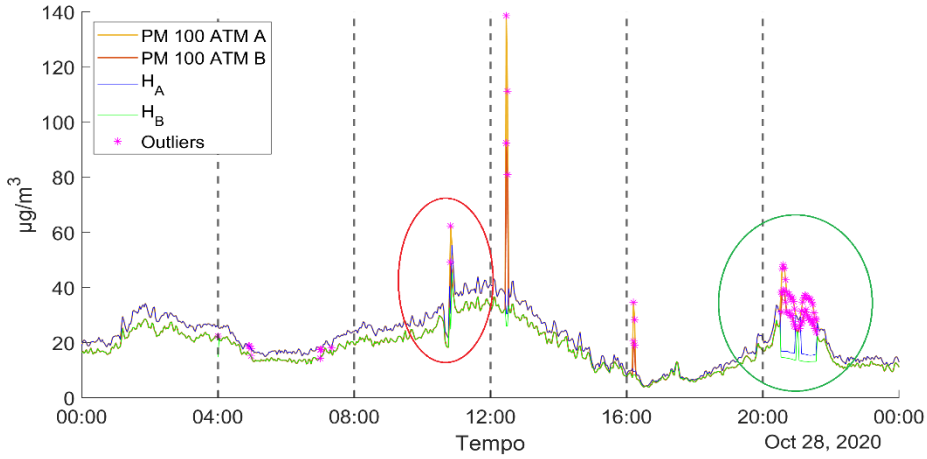

Fig. 14. Andamento risultatante dal filtraggio utilizzando il metodo di Hampel con 90 elementi per la rimozione degli *outlier* applicato alle misurazioni del sensore P@P-PA-2_1 il 28 ottobre 2020.

6 CONCLUSIONI

Per concludere, nell'ambito del progetto H2020 PULSE [4] sono stati acquisiti dati con centraline e sensori a basso costo, in questo articolo è stata descritta nel dettaglio una tecnica innovativa sviluppata, appositamente per il filtraggio di questa tipologia di dati. Come prevedibile, vista la relativa economicità dei sensori, i dati riportati contengono un *noise* significativo e una quantità piuttosto elevata di *outlier*; il lavoro che ci ha portato a sviluppare questa tecnica ha avuto inizio tramite l'utilizzo di metodi considerati standard nel mondo del *signal processing*. Nel dettaglio il metodo più preso in considerazione è stato quello di Hampel disponibile nell'ambiente matlab da noi utilizzato, pur variando i parametri di tale metodo si hanno delle difficoltà nella corretta individuazione degli outlier. Per cogliere al meglio le naturali fluttuazioni delle concentrazioni del particolato atmosferico è stata quindi sviluppata un'apposita tecnica di filtraggio che vanta diverse caratteristiche qualificanti. Principalmente di natura adattiva, dividendo una giornata di misurazioni in intervalli di tempo all'interno dei quali identifica il grado polinomiale che meglio approssima i dati, dopo averne filtrato gli *outlier,* inoltre siccome l'interpolazione viene effettuata con il metodo dei *bisquare weights* risulta essere un algoritmo robusto. L'articolo illustra attraverso diversi esempi il suo comportamento anche in comparazione col metodo di Hampel, mostrando come il metodo da noi sviluppato sia in grado di filtrare la maggior parte degli *outlier* ed efficacemente filtra il *noise* per altro conservando la gran parte delle brusche variazioni che vengono a volte registrate dai sensori. Possibili sviluppi futuri comprendono l'utilizzo metodi più complessi per la stima del grado polinomiale adeguato come il metodo "LASSO" e imporre la continuità del polinomio interpolante e della sua derivata prima tra un intervallo e l'altro, esistono inoltre delle casistiche caratterizzate da variazioni così brusche che attualmente l'algoritmo considera come degli *outlier* quando in realtà non lo sono, dal confronto visivo con le misure di sensori vicini.

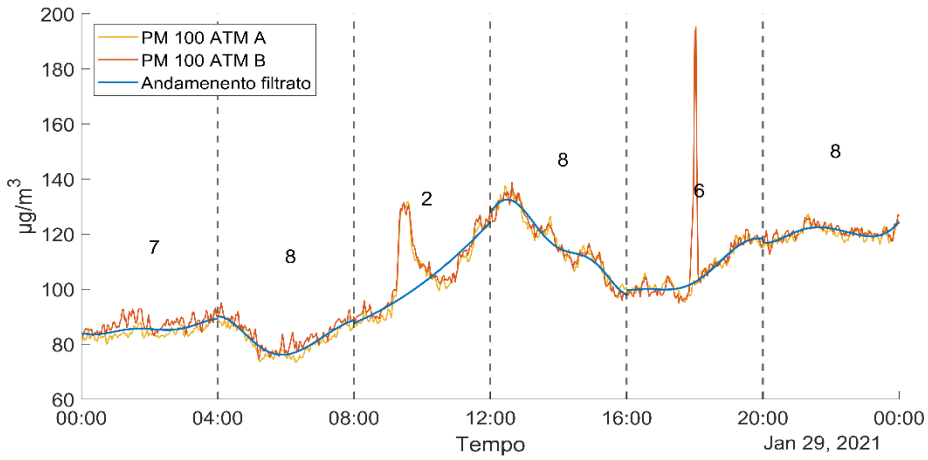

Fig. 15. Tecnica di filtraggio applicata alle misure riportate dal sensore P@P-PA-4_1 il 29 gennaio 2021.

Casi come questo non sono molto frequenti e spesso isolati, se si effettua il filtraggio sui sensori vicini, evidenziati in figura 3 lo stesso picco viene riconosciuto dall'algoritmo in quanto non si presenta così marcato. Trovare una soluzione a queste problematiche di fatto renderebbe l'algoritmo completo e affidabile per qualsiasi possibile misurazione effettuata dai sensori a basso costo, fornendo le fondamenta per futuri progetti come l'implementazione della *personal exposure* ovvero la valutazione della quantità di inquinante assorbita da un soggetto all'interno dell'area monitorata.

References

1. URL: https://www.arpalombardia.it
2. URL: https://www.eea.europa.eu/it/pressroom/newsreleases/molti-cittadini-europei-sono-ancora/morti-premature-attribuibili-allinquinamento-atmosferico
3. URL: https://www.legambiente.it
4. URL http://www.esa.int
5. URL: http://www.project-pulse.eu
6. URL https://www2.purpleair.com.
7. URL: http://www.aqmd.gov/aq-spec/sensordetail/purpleair-pa-ii

Acknowledgement

Geometra Paolo Marchese tecnico del Laboratorio di Geomatica dell'Università degli Studi di Pavia per le attività di installazione e manutenzione della rete di sensori PurpleAir.

PULSE (NO GA727816) è finanziato dal programma ricerca e innovazione Horizon 2020 dell'Unione Europea.

Il progetto CE4WE (ID 1139857), call di regione Lombardia nell'ambito del programma "progetti strategici di ricerca, sviluppo e innovazione volti al potenziamento degli ecosistemi lombardi della ricerca e dell'innovazione quali hub a valenza internazionale".