Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le Tecnologie e le Scienze

Relazione Finale

Identificazione di valori anomali nelle serie storiche per dati ambientali

Relatrice: Prof.ssa Luisa Bisaglia
Dipartimento di Scienze Statistiche

Laureando: Giovanni Capasso
Matricola N. 1169119

Anno Accademico 2021/2022

Indice

Introduzione

I valori anomali o outlier sono uno degli argomenti maggiormente affrontati nel campo della statistica. È inusuale trovare un insieme di dati che non contenga osservazioni anomale.

Nella maggior parte dei casi, gli outlier sono visti come una fonte di disturbo che possono portare a stime distorte dei parametri di un modello, cattive previsioni e statistiche fuorvianti. In molte applicazioni, i ricercatori si limitano a identificarli ed eliminarli dall'insieme dei dati, senza indagare il meccanismo che porta un'osservazione a mostrare un comportamento differente dal resto dei dati.

Lo scopo di questa relazione è differente, in quanto vuole sottolineare che il comportamento anomalo di una o più osservazioni può dipendere da differenti fattori specifici legati al contesto di applicazione. Questi possono essere fonti di informazioni preziose che ci permettono di comprendere più approfonditamente il comportamento generale dei dati.

Attualmente non esiste una metodologia universale per identificare gli outlier, ma possono essere adoperate differenti tecniche, che portano a prediligere un metodo rispetto ad un altro. La scelta del metodo si basa sulla natura del dato e il tipo di valore anomalo che si desidera individuare.

Nel capitolo 1, dopo aver definito cosa si intende per osservazione anomala nel contesto delle serie storiche, viene fornita una panoramica generale sui principali tipi di outlier che possono caratterizzare le serie storiche univariate e multivariate. Nel capitolo 2 viene sviluppato un metodo di rilevamento au-

tomatico degli outlier appositamente sviluppato per il contesto dei dati ambientali. La forza del metodo proposto risiede nel non dover fare assunzioni parametriche. Nel Capitolo 3 viene applicato ad una serie storica univariata un algoritmo al fine di rilevare le concentrazioni anomale di particolato. Oltre all'analisi dei dati, vengono identificati i fattori che maggiormente contribuiscono all'incremento dei livelli di particolato nella Pianura Padana.

Capitolo 1

I valori anomali nelle serie storiche

Lo scopo di questo capitolo è di fornire una definizione di outlier e sottolineare l'importanza della loro rilevazione nelle serie temporali. Attualmente non esiste una classificazione *"standard"* nel contesto delle serie storiche e i possibili metodi e algoritmi che possono essere applicati sono numerosi e in molti casi elaborati *"ad hoc"* per un determinato contesto applicativo. Viene qui proposta una classificazione sulle principali tecniche di rilevazione in base al tipo di outlier che si desidera identificare e al tipo di dato in ingresso.

1.1 Il concetto di outlier

1.1.1 Definizione di outlier

La prima domanda a cui è di interesse dare una risposta è:
"Che cos'è un outlier?".
Il primo a rispondere a tale quesito è stato Grubbs (1969) il quale afferma:
"Un'osservazione anomala, o outlier, è un'osservazione che sembra discostarsi notevolmente dagli altri membri del campione in cui si trova";
Barnett and Lewis (1994), invece, lo definiscono come:

"Un'osservazione (o un sottoinsieme di osservazioni) che sembra essere incoerente con il resto dell'insieme di dati".
Infine, Hawkins (1980) descrive un outlier come:
"Un'osservazione che si discosta così tanto da altre osservazioni da suscitare il sospetto che sia stata generata da un meccanismo differente."
Queste definizioni sottolineano due aspetti principali:

- La distribuzione degli outlier si discosta notevolmente da quella generale dei dati;
- La maggior parte del dataset è formata da osservazioni "normali". Gli outlier costituiscono solo una piccola parte del dataset.

1.1.2 Cause e contesti di applicazione

In questo contesto, l'interpretazione di un valore anomalo assume lo stesso peso del rilevarlo.
Un'osservazione può mostrare un comportamento anomalo per uno dei seguenti motivi: *"errore umano, errore dello strumento, deviazioni naturali nelle popolazioni, comportamenti fraudolenti, cambiamenti nel comportamento dei sistemi o difetti nei sistemi"* (Hodge, 2004, p. 87).
Dopo l'identificazione di un valore anomalo è necessario valutare se mantenerlo o eliminarlo dall'insieme dei dati. Se l'outlier è dovuto ad errori umani o tecnici, solitamente l'osservazione dovrebbe essere rimossa per salvaguardare l'integrità e la coerenza dei dati. Tuttavia, se l'outlier è imputabile a deviazioni naturali o cambiamenti che sono di interesse al contesto specifico dei dati, è importante che esso venga mantenuto.
L'applicazione di tecniche di identificazione degli outlier è fondamentale in diversi contesti applicativi, per esempio (Hodge, 2004):

- Monitoraggio delle attività: rilevamento di frodi telefoniche o di scambi anomali nel mercato azionario;

- Monitoraggio delle condizioni mediche: rilevamento di valori anomali in esami strumentali come ECG o Elettroencefalogramma;

- Diagnosi di guasti: monitoraggio di linee di produzione al fine di rilevare pezzi non conformi e/o guasti;

- Rilevamento delle intrusioni: rilevamento di accessi non autorizzati in reti informatiche.

1.1.3 Gli outlier possono segnalare l'insorgere di epidemie

Si desidera introdurre la rilevazione degli outlier tramite un esempio e sottolineare due aspetti chiave da tenere sempre in considerazione a prescindere dal metodo utilizzato. Nella sorveglianza sanitaria è vitale rilevare nel minor tempo possibile il cambiamento da una situazione "normale" ad una situazione potenzialmente "critica".
Chen (2014) propone un metodo per rilevare gli outlier nelle serie storiche univariate attraverso l'utilizzo di una carta di controllo Cumulative Sum Control Chart (CUSUM) applicata ai residui. Lo scopo dello studio è rilevare la presenza di conteggi anomali negli accessi ospedalieri giornalieri per sindromi respiratorie a Taiwan dal 2005 al 2008. Le carte di controllo con memoria possono essere un ottimo strumento per rilevare sia singoli outlier nonchè pattern anomali. Tramite l'adattamento di un modello parametrico si è in grado di spiegare i molteplici fattori (stagionalità, giorno della settimana, festività, eventi naturali, *etc.*) che influenzano la serie dei conteggi. Successivamente, è stata applicata una statistica di controllo CUSUM sui residui stimati dal modello.
La statistica identifica vari giorni con conteggi elevati classificati come anomali. Tuttavia, uno studio approfondito delle cause non ha rilevato un'associazione con l'inizio di una nuova epidemia.
È importante sottolineare due aspetti: i) classificare correttamente gli outlier

rilevati in veri o falsi outlier (un'osservazione rilevata come outlier quando non è anomala); ii) il tempo che intercorre dal momento in cui si verifica un'anomalia all'istante di tempo in cui il metodo la identifica deve essere il minore possibile.

1.1.4 Differenti tipi di outlier

Data una serie temporale $X = (x_1, \ldots, x_{\tau-k}, \ldots, x_\tau, \ldots, x_{\tau+k}, \ldots, x_t)$ di lunghezza t, è di interesse esplicitare tre tipi di outlier che si possono rilevare nelle serie storiche univariate e multivariate (Lai, 2021):

1. **Punto anomalo :** Se al tempo τ l'osservazione x_τ presenta un valore particolarmente elevato o particolarmente piccolo (massimo o minimo assoluto) rispetto a tutti gli altri punti (osservazioni) della serie X, allora è un punto anomalo. In questo caso, il punto individuato sarà classificato come outlier *globale*.

   Per esempio, nel contesto bancario, una transazione di importo particolarmente elevato rispetto a tutte le altre transazioni effettuate dal cliente può essere considerato, come un punto anomalo.

2. **Sottosequenza anomala**: Questo tipo di outlier è spesso presente nelle serie storiche a causa della forte dipendenza tra le osservazioni.

   Se x_τ non è un punto anomalo, ma appartiene alla sottosequenza di punti nella finestra $[x_{\tau-k}, x_{\tau+k}]$ che congiuntamente mostra un comportamento anomalo rispetto al resto della serie X, allora fa parte di una sottosequenza anomala.

   Per esempio, nel contesto bancario, un cliente che preleva una grossa somma di denaro per diversi giorni consecutivi può rappresentare una sottosequenza anomala.

3. **Outlier contestuali** (*locali*): Gli outlier contestuali, solitamente, hanno valori relativamente più grandi/piccoli (massimi/minimi relativi)

nel loro contesto (finestra).

Infatti, la differenza tra gli outlier specificati nei punti 1, 2 e gli outlier contestuali dipende se guardiamo l'intera serie X di lunghezza t o un segmento della serie di lunghezza inferiore (ad esempio, $[x_{\tau-k}, x_\tau]$). Gli outlier contestuali vengono anche chiamati *locali* e possono riguardare singoli punti o sottosequenze.

Una temperatura di $25°C$ in estate a Padova è normale, ma la stessa temperatura in inverno per uno o più giorni potrebbe essere considerata come anomala.

Gli outlier *globali* sono anche outlier *locali*, ma non tutti gli outlier *locali* sono *globali*. Conoscere a priori il tipo di outlier che è di interesse rilevare permette di selezionare il metodo di rilevamento più appropriato.
Nei paragrafi 1.2 e 1.3 viene proposta una classificazione delle principali tecniche di rilevazione che si basano sul tipo di dati in ingresso, il tipo di outlier che si desidera rilevare e la natura del metodo. Questa tassonomia è stata proposta da Blazquez-Garcìa (2021).

1.2 Serie storiche univariate

In questo paragrafo analizzeremo il caso in cui la serie storica di interesse è formata da una singola variabile tempo-dipendente.
Definiamo una serie storica univariata come:

Definizione 1.1 (serie storica univariata)**.** Una serie storica univariata $X = \{x_t\}_{t\in T}$ è un insieme ordinato di osservazioni a valore reale, dove ogni osservazione è registrata ad uno specifico tempo $t \in T \subseteq \mathbb{Z}^+$.

x_t è definito come un *punto* o un'osservazione raccolta al tempo t. Si assume che ogni osservazione x_t sia un valore realizzato di una certa variabile casuale X_t.

In alcuni contesti applicativi, i dati sono raccolti in modo continuo (in streaming) ed è di interesse marcare al tempo t un'osservazione come anomala oppure no prima dell'arrivo dell'osservazione successiva al tempo $t+1$.

1.2.1 Metodi di identificazione di punti outlier

In letteratura sono stati proposti numerosi metodi di identificazione di punti anomali nell'ambito delle serie storiche. In questo paragrafo vengono presentati i metodi *basati sul modello*.
Possiamo definirli come approcci strettamente legati alle definizioni di outlier discusse nel sottoparagrafo 1.1.1, in quanto si fondano sul misurare la distanza per ogni punto temporale della serie tra il valore osservato e il valore previsto, calcolato dall'adattamento di un opportuno modello ai dati.

Definizione 1.2 (Metodo basato sul modello)**.** Data una serie storica univariata X, indentifichiamo un punto come outlier al tempo t se la distanza tra il valore osservato e il valore stimato è maggiore di una soglia prefissata δ:

$$|x_t - \widehat{x}_t| > \delta \tag{1.1}$$

dove x_t e $\widehat{x}_t$ sono rispettivamente il valore osservato e stimato al tempo t.

La scelta del parametro di soglia rappresenta una parte critica del suddetto metodo. Se da un lato permette allo sperimentatore di scegliere δ in base al contesto di applicazione, dall'altro, un parametro di soglia eccessivamente piccolo o grande può portare a identificare un numero elevato o basso di outlier.
In aggiunta, il tipo di outlier (*globale* o *locale*) che si desidera rilevare influenza la scelta della soglia δ (Lai, 2021).
Per identificare outlier globali, possiamo porre δ:

$$\delta = \lambda\sigma\left(X\right) \tag{1.2}$$

dove $\sigma(X)$ è la deviazione standard della serie temporale X e λ è una costante che controlla il range.

Se, invece, è di interesse rilevare punti outlier *locali* in una finestra della serie temporale X, si può porre:

$$\delta = \lambda\sigma\left(X_{t-k,t+k}\right) \tag{1.3}$$

dove $X_{t-k,t+k} = (x_{t-k}, x_{t-k+1}, \ldots, x_{t+k})$ è un segmento della serie X di lunghezza k.

Nell'esempio dei dati raccolti in streaming, alcuni algoritmi di rilevamento utilizzano modelli che si aggiornano all'arrivo dei nuovi dati. Questi modelli incrementali riescono a cogliere con maggior flessibilità il comportamento della serie, specialmente in presenza di un cambiamento importante in un certo istante temporale (ad esempio, uno shift della media).

Esistono due strategie differenti per il calcolo di $\widehat{x}_t$: il primo metodo si basa sulla costruzione di un modello di previsione che considera solamente i valori precedenti al tempo t, ovvero $(x_1, \ldots, x_{t-2}, x_{t-1})$; il secondo metodo si basa sull'identificazione di un modello di stima che considera le osservazioni passate, correnti e future al tempo t $(x_1, \ldots, x_{t-2}, x_{t-1}, x_t, x_{t+1}, \ldots)$.

Un approccio interessante è stato fornito da Chen (2010). Egli propone un metodo di rilevamento degli outlier basato sulla costruzione di un intervallo di confidenza puntuale. Il primo passo consiste nel supporre che il processo di generazione dei dati sottostante possa essere modellato come una funzione continua:

$$x_i = m\left(t_i\right) + \varepsilon_i \tag{1.4}$$

dove x_i, $i = 1, \ldots, n$, è il valore dei dati raccolti al tempo t_i, $m(t)$ è la funzione sottostante e $\varepsilon_i \sim N\left(0, \sigma^2\right)$. L'obiettivo è trovare una stima appropriata per la funzione ignota $m(t)$ tramite l'adattamento di un modello parametrico o non parametrico ai dati (t_i, x_i), per $i = 1, \ldots, n$.

Il secondo passo del metodo consiste nel costruire un intervallo di confidenza

(IC) puntuale. Per un α fissato, gli estremi dell'IC al tempo t_i sono:

$$[\widehat{x_i} - z_{\alpha/2} S_i(pred), \widehat{x_i} + z_{\alpha/2} S_i(pred)] \tag{1.5}$$

dove $S_i(pred) = MSE + s_i^2(\widehat{x_i})$ e $z_{\alpha/2}$ è il quantile $\alpha/2$ di una Normale Standard. ll dato registrato al tempo t_i è identificato come outlier se cade al di fuori degli estremi dell'IC calcolati in 1.5.
Questo metodo proposto non utilizza un parametro di soglia globale, ma tramite la stima di un intervallo di confidenza puntuale riesce efficacemente a identificare sia outlier locali che globali.

1.2.2 Metodi di identificazione di sottosequenze outlier

Come illustrato nel sottoparagrafo 1.1.4, in alcune applicazioni può essere di interesse rilevare un insieme di punti consecutivi che si comportano in modo anomalo.
Rilevare questo tipo di outlier è più impegnativo rispetto ai punti anomali. Il problema principale risiede nella lunghezza della sottosequenza di outlier. Esistono due tipi di metodi di rilevamento: la maggior parte utilizzano sottosequenze di lunghezza fissa prespecificata dall'utente, mentre altri utilizzano sottosequenze di lunghezza variabile. Nel caso in cui la lunghezza sia fissa, il numero di sottosequenze analizzate dipenderà dalla lunghezza stessa. Un ulteriore aspetto che si può rilevare sono le sottosequenze outlier periodiche. Queste ultime sono sottosequenze anomale che si ripetono nel tempo.
Uno dei possibili metodi applicabili a questa categoria di outlier sono quelli basati sulla costruzione di modelli di previsione. L'obiettivo è costruire un modello di previsione che tramite l'informazione sui dati passati possa essere in grado di fare previsioni.

Definizione 1.3 (Metodo basato sul modello di previsione)**.** Data una serie

storica univariata X, identifichiamo una sottosequenza di outlier S se:

$$\sum_{i=p}^{p+n-1} |x_i - \widehat{x_i}| > \delta \tag{1.6}$$

dove $S = x_p, \ldots, x_{p+n-1}$ è la sottosequenza osservata, $\widehat{S} = \widehat{x}_p, \ldots, \widehat{x}_{p+n-1}$ è la sottosequenza predetta e δ è la soglia prefissata.

1.3 Serie storiche multivariate

Negli ultimi anni, grazie alla maggiore efficienza dei calcolatori, è aumentata la mole di dati che necessita di essere analizzata nel minor tempo possibile. In molti contesti, i dati sono raccolti in modo continuo (in streaming) e vi è la necessità di marcare un dato come anomalo oppure no non appena esso viene registrato.
Il rilevamento degli outlier nelle serie storiche multivariate è più impegnativo rispetto al contesto univariato. Definiamo una serie storica multivariata come:

Definizione 1.4 (serie storiche multivariate)**.** Una serie temporale multivariata $\mathbf{X} = \{\mathbf{x}_t\}_{t \in T}$ è definita come un insieme ordinato di vettori k-dimensionali, ognuno dei quali è registrato in un tempo specifico $t \in T \subseteq \mathbb{Z}^+$ e consiste in k osservazioni a valori reali, $\mathbf{x}_t = (x_{1t}, ..., x_{kt})$.

Quindi, $\forall j \in \{1, \ldots, k\}$, $X_j = \{x_{jt}\}_{t \in T}$ è una serie storica univariata e ciascuna osservazione $x_{jt} \in \mathbf{x}_t$ è una realizzazione di una variabile casuale dipendente dal tempo X_{jt} in $\mathbf{X}_t = (X_{1t}, \ldots, X_{kt})$. Le variabili X_j possono avere lunghezze differenti. L'aspetto più importante da sottolineare è che nelle serie storiche multivariate abbiamo più di una variabile dipendente dal tempo. Di conseguenza, ogni variabile dipende non solo dai suoi valori passati, ma, possibilmente, anche dalle altre variabili.
I metodi per rilevare i valori anomali nelle serie storiche multivariate sono molteplici. In questa tesi ne vengono esplicitati tre, i quali differiscono in

base a come trattano la dimensionalità della serie.

Il primo consiste nell'applicare i metodi univariati ad ogni singola serie storica. In questo caso, l'identificazione di un eventuale valore anomalo nella serie storica univariata X_j ad uno specifico tempo $t \in T \subseteq \mathbb{Z}^+$ è univariato perchè riguarderà un singolo elemento del vettore $\mathbf{x}_t$. Quest'approccio permette di utilizzare i numerosi ed efficienti algoritmi presenti in letteratura citati nel paragrafo 1.2. Tuttavia, la rilevazione di outlier tramite tecniche univariate quando la serie in ingresso è multivariata porta a non tenere in considerazione la struttura di correlazione tra le serie storiche X_η e X_λ, $\forall \eta, \lambda \in \{1, \ldots, k\}$ con $\eta \neq \lambda$, e ciò, porta ad una perdita notevole di informazione.

La seconda strategia consiste nell'applicare preliminarmente una riduzione della dimensionalità tramite l'analisi delle componenti principali (PCA) e successivamente identificare gli outlier tramite un algoritmo univariato o bivariato applicato al nuovo insieme di variabili trasformate. Proseguendo per questa via, si ottiene un nuovo insieme ridotto di variabili incorrelate in grado di rappresentare le caratteristiche dei dati senza perdita di informazione. Gli outlier che si indentificano sul nuovo set di variabili trasformate sono multivariati. Nel sottopragrafo 1.3.3 viene riportato un esempio.

La terza possibilità consiste nell'applicare direttamente algoritmi multivariati senza effettuare alcuna trasformazione preliminare dei dati. Nei sottoparagrafi 1.3.1 e 1.3.2 vengono presentati alcuni metodi di indentificazione multivariati.

Infine, quando si trattano serie multivariate può essere di interesse rilevare intere serie storiche (univariate) che mostrano un comportamento anomalo. Nel contesto dell'identificazione degli outlier, questo aspetto rappresenta la maggiore differenza tra le serie temporali univariate e multivariate.

1.3.1 Punti outlier nelle serie storiche multivariate

Partendo dalla Definizione 1.4, un punto anomalo in una serie storica multivariata può interessare una variabile (singolo punto univariato x_{jt}) o più variabili (punto multivariato, vettore $\mathbf{x}_t$ al tempo t).
Il principale metodo per identificare punti anomali nelle serie storiche multivariate si basa sull'identificazione di un adeguato modello in grado di cogliere le caratteristiche intrinseche della serie. Non di rado si considerano modelli non parametrici basati sulle B-Spline o, sulle serie di Fourier nel caso di dati periodici, in grado di cogliere con maggior flessibilità l'andamento della serie.

Definizione 1.5. Data una serie storica multivariata $\mathbf{X}$, identifichiamo un vettore di punti $\mathbf{x}_t$ come outlier al tempo $t \in T \subseteq \mathbb{Z}^+$ se la differenza tra il valore osservato e quello stimato è maggiore di una soglia prefissata δ:

$$\|\mathbf{x_t} - \widehat{\mathbf{x}}_\mathbf{t}\| > \delta \tag{1.7}$$

dove $\mathbf{x_t} = (x_{1t}, \ldots, x_{kt})$ e $\widehat{\mathbf{x}}_\mathbf{t} = (\widehat{x}_{1t}, \ldots, \widehat{x}_{kt})$ sono rispettivamente i vettori dei valori osservati e stimati al tempo t.

Non di rado si preferiscono parametri di soglia δ dinamici.

1.3.2 Sottosequenze outlier nelle serie multivariate

Questo tipo di outlier può essere rilevato tramite l'adattamento di un modello di previsione in grado di catturare la dinamica della serie utilizzando i dati passati e quindi fare previsioni sul futuro. Le sottosequenze che si discostano da tali previsioni sono marcate come outlier.

Definizione 1.6 (Metodo basato sul modello di previsione)**.** Data una serie storica multivariata $\mathbf{X}$, identifichiamo una sottosequenza multivariata $\mathbf{S} = \mathbf{x}_p, \ldots, \mathbf{x}_{p+n-1}$ di lunghezza n come anomala se:

$$\sum_{i=p}^{p+n-1} \|\mathbf{x}_i - \widehat{\mathbf{x}}_i\| > \delta \tag{1.8}$$

dove $S = \mathbf{x}_p, \ldots, \mathbf{x}_{p+n-1}$ è la sottosequenza osservata, $\widehat{S}$ è il suo valore predetto, e δ è la soglia selezionata.

Questa tecnica rileva sottosequenze di outlier allineati temporalmente che influenzano più variabili.

1.3.3 Serie storiche outlier

È di interesse esplicitare in modo formale cosa si intende per un'intera serie storica che manifesta un comportamento anomalo in un contesto multivariato.

Definizione 1.7 (Serie storiche anomale)**.** Dato un insieme di serie storiche $\mathbf{X} = \{\mathbf{x}_t\}_{t \in T}$, una serie outlier univariata $X_j \in \mathbf{X}$, per $j \in \{1, \ldots, k\}$, è una serie storica il cui comportamento si discosta significativamente dal resto delle serie storiche presenti in $\mathbf{X}$.

La letteratura non fornisce numerosi metodi per questo problema. I principali algoritmi sviluppati per il rilevamento di intere serie anomale in ambito multivariato si basano sul clustering e sull'applicazione di metodi univariati dopo una riduzione della dimensionalità.
Hyndman (2015) sviluppa un'interessante metodologia la cui applicazione è stata usata dall'azienda Yahoo per rilevare server difettosi. Dal punto di vista dell'analisi dei dati, ciò significa un monitoraggio continuo di grandi volumi di dati di serie temporali per rilevare potenziali guasti o anomalie nel minor tempo possibile.
Il primo passo del metodo consiste nel selezionare τ caratteristiche dalle k serie temporali. L'insieme delle caratteristiche selezionate ha l'obiettivo di catturare il comportamento globale della serie come la presenza di picchi, trend, stagionalità, linearità, *etc.*
Dopo aver costruito il vettore di caratteristiche per ogni serie temporale si applica la PCA per identificare le prime due componenti principali. Una

componente principale è una combinazione delle variabili originali dopo una trasformazione lineare. Solitamente, è sufficiente considerare solamente le prime due componenti in quanto sono in grado di cogliere la maggior parte della variabilità dei dati.
Nel secondo passo del metodo proposto, le serie temporali anomale sono identificate in base alla loro deviazione dalla regione di massima densità nello spazio delle prime due componenti principali.
Lo scopo finale è selezionare γ serie anomale, con $\gamma \in \{1, \ldots, k\}$. Questo metodo può essere esteso ad altri contesti di dati ed ha il vantaggio di non dover selezionare un parametro di soglia. Inoltre, è molto flessibile in quanto permette all'utente di scegliere quali caratteristiche includere nella PCA in base all'andamento della serie storica multivariata.

1.4 La scelta dell'algoritmo di rilevamento

In generale, i principali algoritmi di identificazione degli outlier provengono da questi tre rami:

- Metodi statistici;
- Metodi di Machine learning;
- Metodi che usano reti neurali (Deep learning).

In questa tesi sono stati presentati e approfonditi solo metodi statistici. La serie temporale può essere influenzata dalla presenza di trend, stagionalità, ciclicità. Alcuni algoritmi possono avere buone performance nel rilevare outlier nelle serie storiche affette dalla presenza di trend, ma possono essere inadatti quando nella serie è presente una componente periodica. Analizzare preliminarmente le componenti della serie temporale è di fondamentale interesse sul metodo da scegliere. Per questo motivo, si è preferito optare per una classificazione basata sul tipo di outlier che si desidera identificare,

piuttosto che sul metodo parametrico o non parametrico applicabile.
Un ulteriore aspetto da tenere in considerazione è il costo computazionale dell'algoritmo, specialmente quando i dati sono raccolti in streaming e/o nelle serie storiche multivariate.
L'applicazione dell'algoritmo non può essere sufficiente per comprendere perché una o più osservazioni mostrino un comportamento anomalo.
Nel Capitolo 2 viene presentato un metodo non parametrico.

Capitolo 2

Un approccio non parametrico

L'obiettivo di questo capitolo è presentare un metodo non parametrico di rilevazione degli outlier nelle serie storiche univariate. La scelta di questo approccio è collegata al tipo di dati analizzati nel Capitolo 3.
Il metodo è stato sviluppato da Čampulová (2018c).

2.1 Perchè scegliere un approccio non parametrico?

La regressione non parametrica è un metodo di lisciamento utile a stimare la funzione di regressione a partire da dati "rumorosi". I metodi di regressione non parametrica descritti in letteratura sono molteplici: la regressione kernel, la regressione polinomiale locale, le splines di regressione, le splines di lisciamento, le regressione basata sull'espansione in serie di Fourier, la trasformata wavelet e molti altri (Čampulová, 2018b).
In questa tesi, ci concentreremo sul metodo basato sul kernel con parametro di lisciamento locale (bandwidth). Vista la difficoltà in molti ambiti nel trovare un insieme finito di parametri in grado di cogliere interamente la forma funzionale dei dati, è preferibile lasciare che siano i dati a scegliere quale funzione si adatta meglio a loro senza le restrizioni che un modello parametrico impone. Nello specifico, le concentrazioni di PM_{10} in una determinata area

sono influenzate da un numero elevato di variabili, nella maggior parte dei casi difficili da rilevare o sconosciute. Inoltre, esiste una correlazione significativa tra le concentrazioni di PM_{10} e i vari inquinanti presenti nell'ambiente (Masiol, 2013).
In questo contesto, l'adattamento di un modello parametrico può soffrire di eccessiva complessità in quanto deve tenere in considerazione un numero elevato di fattori.
Ricercando in letteratura, Campulova (2018a) ha adattato un modello parametrico con l'obiettivo di costruire un modello di previsione per le concentrazioni orarie di PM_{10}. Viene utilizzato un Modello Lineare Generalizzato con distribuzione Gamma della variabile risposta e funzione di link inversa. Le variabili esplicative significative introdotte nel modello per stimare la concentrazione di PM_{10} all'ora t sono state: la concentrazione di NO_2 rilevata all'ora t, la concentrazione di PM_{10} rilevata all'ora $t-1$, l'intensità e la direzione del vento rilevata all'ora t.

2.2 Descrizione del metodo

L'algoritmo di identificazione degli outlier proposto in questa tesi è basato su tre passi. Di seguito vengono illustrati gli obiettivi di ognuno dei passi dell'algoritmo:

1. Ottenere un vettore di residui lisciati;
2. Individuare i punti nella serie in cui la variabilità dei residui cambia;
3. Identificare i residui anomali in ogni segmento.

2.2.1 Dai dati grezzi ai residui

L'obiettivo dell'analisi di regressione non parametrica è stimare una funzione di risposta (o curva) ignota a partire dai dati osservati.

Assumiamo che la variabile ignota Y_i sia stata misurata in n istanti di tempo discreti differenti $t_1, t_2, \ldots, t_n$ che appartengono all'intervallo $[a, b]$, per $i = 1, \ldots, n$.

Il modello di regressione eteroschedastico può essere scritto come (Herrmann, 1997):

$$Y_i = m(t_i) + \sigma(t_i)\,\varepsilon_i \tag{2.1}$$

dove m denota la funzione di regressione ignota, ε_i sono gli errori casuali indipendenti e identicamente distribuiti ($i.i.d$) con media zero e varianza unitaria, mentre $\sigma(t_i)$ è la funzione di deviazione standard che esprime la variabilità di Y_i.

Il principale scopo che ci porta ad utilizzare la regressione kernel è che tale metodologia stima il trend dei dati in un determinato istante temporale discreto come media pesata delle osservazioni "rumorose" nell'intorno del punto. I pesi dipendono dalla scelta della funzione kernel, mentre la quantità di osservazioni utilizzate per la media dipende dalla scelta del parametro di lisciamento. Il kernel risulta molto più efficiente rispetto al metodo delle medie mobili, che attribuisce lo stesso peso a tutte le osservazioni "vicine" del punto in cui si vuole stimare la funzione (Čampulová, 2018b).

Per la stima della funzione di regressione $m(\cdot)$ si è deciso di utilizzare lo stimatore di Gasser-Muller (1984) definito come:

$$\widehat{m}(t, h_t) = \sum_{i=1}^{n} Y_i \int_{l_{i-1}}^{l_i} \frac{1}{h_t} K\left(\frac{t-u}{h_t}\right) du, x \in [a + h_x, b - h_x] \tag{2.2}$$

dove $K(\cdot)$ denota la funzione kernel di ordine $(0, k)$, $h_t = h(t)$ è il parametro di lisciamento locale chiamato bandwidth nel punto t e gli estremi di integrazione sono $l_0 = a$ e $l_i = 0.5(t_i + t_{i+1})$ per $i = 1, \ldots, n-1$, $l_n = b$.

Per implementare lo stimatore di Gasser-Muller (2.2) è necessario fornire, oltre ai dati, il parametro di bandwidth h_t e la funzione nucleo $K(\cdot)$ che si vuole implementare. Tuttavia, quest'ultima non costituisce un elemento

critico, e tendenzialmente funzioni nucleo differenti producono risultati grossomodo simili purché abbiano un supporto finito. In questa tesi si adotta la funzione kernel di Epanechnikov con supporto $[-1, 1]$. Essa è una funzione di densità di probabilità simmetrica attorno all'origine e che soddisfa le due condizioni:

$$K(x) > 0 \tag{2.3}$$

$$\int K(z)\, dz = 1 \tag{2.4}$$

La parte critica nella regressione non parametrica risiede nella scelta della larghezza di banda, che sia essa globale o locale. L'obiettivo è trovare il giusto *trade-off* distorsione-varianza.

Un'indicazione diretta della scelta del parametro di lisciamento globale è visibile nella *Fig. 2.1.* In generale, si nota che abbassando il valore di h si ottiene una curva più "vicina" al comportamento dei dati ed inevitabilmente più irregolare. Se $h \to 0$ si abbatte la distorsione, ma questo porta la varianza della stima a $+\infty$. Al contrario, all'aumentare di h otteniamo una curva sempre più liscia, ma allo stesso tempo aumenta la possibilità di sottostimare eventuali "picchi". Se $h \to +\infty$ si abbatte la varianza, ma esplode la distorsione.

La scelta di un parametro di lisciamento locale $h(t)$, che dipende dal punto t, permette una maggiore flessibilità in quanto attua un diverso grado di lisciamento in base al comportamento locale dei dati. Lo stimatore può adattarsi alla struttura delle funzioni di regressione, lisciando maggiormente nelle parti piatte della curva, e meno nelle parti con presenza di picchi.

Nel contesto della rilevazione degli outlier, tale metodo ha un effetto particolarmente positivo poiché ci permette di cogliere il comportamento locale dei dati e, se presenti, a rilevare outlier contestuali (locali).

Quando si applica uno stimatore di regressione non parametrico è di interesse quantificare due tipi di errori: locale e globale (Herrmann, 1997). Come

criterio per l'errore locale viene utlizzato l'errore quadratico medio (MSE) definito:

$$MSE\left(\widehat{m}\left(t;h_t\right)\right) = \mathbb{E}\left(\widehat{m}\left(t;h_t\right) - m\left(t\right)\right)^2 \tag{2.5}$$

Come misura globale, si utilizza l'errore quadratico medio integrato (MISE) rispetto a una funzione di peso w:

$$MISE(\widehat{m}) = \mathbb{E}\int_a^b \omega\left(t\right)\left\{\widehat{m}\left(t;h_t\right) - m\left(t\right)\right\}^2 dt \tag{2.6}$$

Nel Capitolo 3 viene utilizzato un metodo iterativo (*plug-in*) per stimare la larghezza di banda locale h_t.

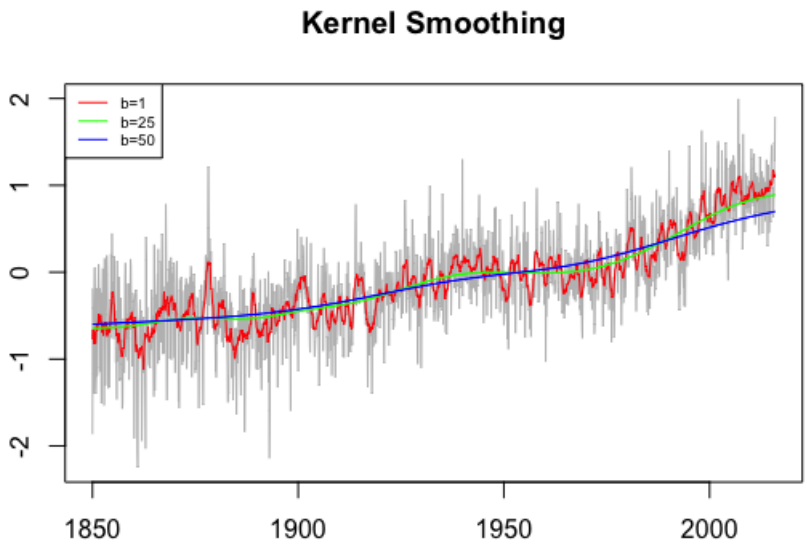

***Fig. 2.1**: Esempio di lisciamento con funzione kernel con parametro di lisciamento globale; h=1(rosso), h=25(verde), h=50(blu)*

L'obiettivo del primo passo è ottenere i residui $R(x_i) = Y_i - \widehat{m}(t_i)$, per $i = 1, \ldots, n$.

2.2.2 Analisi dei Punti di Cambiamento

Quando si analizzano le serie temporali, l'ipotesi che le statistiche di sintesi non subiscano variazioni può essere poco realistica. Per superare questo problema, possiamo identificare un insieme di punti di cambiamento, tra i quali le proprietà statistiche della serie rimangono costanti.

È di interesse sottolineare che questo metodo non consiste nel trovare gli istanti temporali a cui corrispondono valori più elevati, ma nell'identificare gli istanti temporali in cui si verificano cambiamenti bruschi e duraturi. Il metodo si pone tre obiettivi principali: identificare eventuali cambiamenti nella sequenza di variabili casuali osservate, stimare il numero di cambiamenti e la loro posizione (Costa, 2016). Generalmente si vogliono identificare cambiamenti in media, variabilità, o in entrambi.

In letteratura sono presenti varie tecniche (parametriche e non parametriche) in grado di rilevare un solo punto di cambiamento o punti di cambiamento multipli. Inoltre, esistono due diverse modalità di rilevamento: online e offline. Nel rilevamento dei punti di cambiamento online, i dati vengono raccolti e analizzati in tempo reale con lo scopo di identificare i cambiamenti il più rapidamente possibile. Il rilevamento dei cambiamenti offline, che verrà applicato in questa tesi, riguarda l'analisi retrospettiva di un insieme di dati storici e ha l'obiettivo di stimare con precisione il numero e la posizione dei cambiamenti.

Per applicare l'Analisi dei Punti di Cambiamento (Change Point Detection) è necessario avere una sequenza ordinata di dati $y_{1:n} = (y_1, \ldots, y_n)$. Questa assunzione contribuisce a rendere questo metodo molto diffuso nell'ambito delle serie storiche.

Definizione 2.1 (Change-point analysis)**.** Consideriamo una sequenza di variabili casuali indipendenti $X_1, X_2, \ldots, X_n$ con funzione di distribuzione di probabilità $F_1, F_2, ..., F_n$, rispettivamente. Siamo interessati a verificare l'ipotesi nulla:

$$\begin{cases} H_0 : F_1 = F_2 = \ldots = F_n \\ H_1 : F_1 = \ldots = F_{\tau_1} \neq F_{\tau_1+1} = \ldots = F_{\tau_m} \neq F_{\tau_m+1} = \ldots = F_n \end{cases}$$

dove $1 < \tau_1 < \tau_2 < \ldots < \tau_m < n$, $m \in \{1, \ldots, n-1\}$ è il numero ignoto di punti di cambiamento e $\tau_1, \tau_2, ..., \tau_m$ sono le posizioni ignote da stimare. Inoltre, poniamo $\tau_0 = 0$ e $\tau_{m+1} = n$. Pertanto, gli m punti di cambiamento dividono i dati in $m + 1$ segmenti e il j-esimo segmento contiene i dati $y_{(\tau_{j-1}+1:\tau_j)}$.

Il secondo passo del metodo presentato consiste nell'applicare l'Analisi dei Punti di Cambiamento ai residui $R(x_i)$, per $i = 1, \ldots, n$.
È di interesse stimare il numero ignoto di punti di cambiamento $\widehat{m}$, nonchè la loro posizione $\widehat{\tau}_1, \widehat{\tau}_2, ..., \widehat{\tau}_{\widehat{m}}$ in cui la variabilità dei residui $R(x_1), \ldots, R(x_n)$ cambia.
Un ulteriore aspetto da tenere in considerazione è il costo computazionale. In questo caso, si è deciso di sviluppare l'algoritmo Pruned Exact Linear Time (PELT), che è stato dimostrato avere un costo computazionale $O(n)$. In aggiunta, grazie alla maggiore velocità e accuratezza nel rilevare la posizione e il numero di punti di cambiamento, si preferisce questo metodo basato sulla programmazione dinamica rispetto ad altri metodi come la segmentazione binaria (Wambui, 2015). Infatti, Il metodo PELT, essendo un metodo parametrico, richiede l'assunzione che i residui $R(x_1), ..., R(x_n)$ siano osservazioni casuali indipendenti e che seguano una distribuzione normale all'interno dei segmenti omogenei.
Nel caso in cui non si possa assumere la normalità dei residui all'interno di ogni segmento, si può optare per l'applicazione del metodo non parametrico E-Divise. L'obiettivo è stimare $\widehat{m} + 1$ segmenti omogenei con varianza costante al loro interno (ma eterogenea tra i diversi segmenti), il cui generico $\widehat{S}_j = \left(R\left(x_{\widehat{\tau}_{j-1}}\right), \ldots, R\left(x_{\widehat{\tau}_j}\right)\right)$, per $j = 1, \ldots, \widehat{m} + 1$.

2.2.3 Identificazione dei residui anomali

In questo sottoparagrafo vengono descritti tre differenti strumenti (disuguaglianza di Chebyshev, test di Grubbs, quantili della distribuzione normale) utili per l'individuazione dei residui anomali all'interno di ogni segmento omogeneo $S_j = \left(R\left(x_{\widehat{\tau}_{j-1}+1}\right), \ldots, R\left(x_{\widehat{\tau}_j}\right)\right)$, $j = 1, \ldots, \widehat{m}+1$.
Prima di applicare le procedure di rilevamento basate sul test di Grubbs e i quantili della distribuzione normale è necessario verificare l'assunzione che i residui all'interno di ogni segmento S_j seguano, almeno approssimativamente, una distribuzione normale. Inoltre, se la curva di regressione stimata non parametricamente è in grado di cogliere adeguatamente l'andamento della serie, è ragionevole assumere che i residui siano indipendenti. Indubbiamente, è consigliabile verificare queste due assunzioni in quanto la violazione potrebbe portare a delle conclusioni fuorvianti e, in questo particolare contesto, a identificare osservazioni normali come anomale (falsi outlier).
Se non è plausibile l'ipotesi di normalità nel segmento S_j, $j = 1, \ldots, \widehat{m}+1$, si può tentare di trasformare i residui applicando una trasformazione di Box-Cox. In questo caso, i residui sui singoli segmenti omogenei vengono trasformati in variabili casuali $U\left(x_{\widehat{\tau}_{j-1}+1}\right), \ldots, U\left(x_{\widehat{\tau}_j}\right)$, $j = 1, \ldots, m+1$.
Tuttavia, se i residui trasformati mostrano ancora deviazioni dall'ipotesi di normalità, allora possiamo optare per un metodo non parametrico.
La disuguaglianza di Chebyshev dimostra che la maggior parte dei valori della distribuzione si raggruppano intorno alla media della distribuzione (Amidan, 2005).

Teorema 2.1 (disuguaglianza di Chebyshev)**.** *Se una variabile casuale ξ ha varianza finita, allora , $\forall \lambda > 1$, si ha*

$$\mathbb{P}(|\xi - \mathbb{E}(\xi)| \geq \lambda \sigma(\xi)) \leq \frac{1}{\lambda^2} \tag{2.7}$$

dove $\mathbb{V}(\xi) = \sigma^2(\xi)$.

Ad esempio, si ottiene un limite superiore di 1/9 per la probabilità di osservare variazioni casuali che superano la media di 3**deviazione standard*. Ci aspettiamo che un insieme di residui non venga classificato come outlier se sono distribuiti in modo casuale intorno a zero.

L'obiettivo è costruire un intervallo a partire dal Teorema 2.1 e marcare come outlier i residui al di fuori dell'intervallo. Gli estremi dell'intervallo considerato sono $\mu \pm Ls_j$, dove μ è la media dei residui (approssimativamente pari a zero), s_j è la deviazione standard campionaria dei residui appartenenti al segmento S_j, per $j = 1, \ldots, \widehat{m} + 1$, e L è un'opportuna costante.

La scelta della costante L rappresenta una parte critica dell'analisi in quanto è legata al numero di outlier rilevati. In generale, minore è il valore di L, più ci aspettiamo che una percentuale maggiore di residui non cada nell'intervallo e quindi sia marcata come valore anomalo.

La costante L può essere scelta globalmente oppure localmente. In questa relazione, si preferisce selezionare la costante L_j per ogni segmento.

Il secondo metodo di identificazione degli outlier che viene trattato si basa sui quantili della distribuzione normale. Questo metodo è una generalizzazione dell'approccio non parametrico basato sulla disuguaglianza 2.7. Gli estremi dell'intervallo sono $\pm u_{\alpha/2}\widehat{s}_j$, per $j = 1, \ldots, \widehat{m} + 1$, dove $u_{\alpha/2}$ è il quantile di livello nominale $\alpha/2$ di una Normale Standard.

Inoltre, deve essere rispettata la seguente uguaglianza :

$$\mathbb{P}\left[R\left(x_i\right) \in \left[-u_{\alpha/2}\widehat{s}_j, +u_{\alpha/2}\widehat{s}_j\right]\right] = 1 - \alpha \tag{2.8}$$

con $\alpha \in [0, 1]$.

La scelta del parametro α rappresenta un aspetto cruciale dell'analisi. Esso può essere scelto secondo i criteri esposti precedentemente per la scelta della costante L.

L'ultimo metodo proposto per l'identificazione dei residui anomali è stato sviluppato da Grubbs (Urvoy, 2014).

Definizione 2.2 (Test di Grubbs (1950))**.** È di interesse verificare il seguente sistema di ipotesi:

$$\begin{cases} H_0\text{: Non ci sono outlier nell'insieme di dati considerato} \\ H_1\text{: C'è esattamente un outlier nell'insieme di dati considerato} \end{cases}$$

La statistica test è la seguente:

$$T = \max_{i=1,\dots,n} \frac{\left|Y_i - \overline{Y}\right|}{S} \tag{2.9}$$

L'ipotesi nulla (no outlier) è rifiutata ad un livello di significatività α se:

$$T > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/2n,n-2}}{N-2+t^2_{\alpha/2n,n-2}}} \tag{2.10}$$

dove $t_{(\alpha/2n,n-2)}$ rappresenta il quantile di livello $\alpha/(2n)$ di una *t-student* con $n-2$ *g.d.l* ed n è la numerosità del campione.

Nell'Equazione 2.9 sostituiamo agli $Y_1, \dots, Y_n$ i residui $R(x_1), \dots, R(x_n)$ e alla deviazione standard S la deviazione standard campionaria dei residui s_j calcolata per ogni segmento S_j. Possiamo assumere che $\overline{Y} \approx 0$.
L'applicazione del test di Grubbs avviene ricorsivamente. Se l'i-esimo residuo in valore assoluto $|R(x_i)|$ assume il valore massimo tra tutti i residui considerati, cioè $|R(x_i)| > |R(x_j)|$ per $i, j = 1, \dots, n$, $i \neq j$, e l'ipotesi nulla nell'equazione 2.10 viene rifiutata ad un livello di significatività $\alpha = 5\%$, allora il residuo $i-esimo\ R(x_i)$ viene classificato come outlier e rimosso dall'insieme dei residui. Successivamente viene considerato il secondo residuo con valore assoluto maggiore e viene ripetuta la verifica d'ipotesi.
L'applicazione ricorsiva del test di Grubbs su ogni segmento S_j, per $j = 1, \dots, m+1$, termina quando l'Equazione 2.10 è falsa, e ciò denota che il residuo con valore assoluto maggiore tra il set dei residui considerati non è classificato come outlier.

Infine, le misure della variabile originale Y_i, $i = 1, \ldots, n$, corrispondenti ai residui anomali rilevati con uno dei tre metodi vengono identificate come outlier.

2.3 Validità del metodo

Il metodo descritto risulta particolarmente efficiente quando non è possibile assumere l'ipotesi di omoschedasticità sui residui. Inoltre, non prevede alcuna trasformazione dei dati. Il secondo passo del metodo proposto ha un forte impatto sul rilevamento degli outlier. Infatti, applicare i metodi di rilevamento in ogni finestra della serie ci permette di identificare gli outlier locali.
Nel Capitolo 3 applichiamo questa procedura per rilevare le concentrazioni anomale di PM_{10}.

Capitolo 3

Un'analisi sui PM_{10} nella città di Padova

In questo capitolo viene presentata una breve descrizione della serie storica relativa ai PM_{10} e, successivamente, viene svolta un'analisi basata sul metodo presentato nel Capitolo 2.
Le analisi sono state svolte con il linguaggio di programmazione **R**. Le funzioni utilizzate per sviluppare la metodologia proposta nel Capitolo 2 sono presenti nella libreria **envoutliers** (Čampulová, 2022).
L'obiettivo dell'analisi è identificare le concentrazioni anomale di PM_{10}.

3.1 I PM_{10}

L'inquinamento atmosferico è una miscela sospesa nell'aria di particelle solide e liquide che variano per numero, dimensione, forma e composizione chimica. Queste particelle possono rimanere sospese nell'aria per numerosi giorni e trasformarsi ulteriormente. I principali inquinanti atmosferici sono:

- monossido di carbonio (CO);
- ossidi di azoto (NO_x);

- ossidi di zolfo (SO_x);
- idrocarburi (C_xH_y);
- particelle sospese (PTS);
- ozono (O_3).

I PM_{10} (materiale particolato aerodisperso di dimensione inferiore a 10 μm) costituiscono una delle componenti dell'inquinamento su cui vi è maggior attenzione. Ciò è dovuto al fatto che numerosi studi epidemiologici hanno identificato un'associazione tra le concentrazioni di PM_{10} e un incremento sia della mortalità che dei ricoveri per malattie cardiorespiratorie.
Possiamo definire i PM_{10} come un problema "transfrontaliero": a causa dell'effetto dei venti e della morfologia del territorio, il particolato è in grado di invadere Paesi differenti da quelli di produzione rendendo ancora più difficile fare una previsione sulle concentrazioni a breve e a lungo termine in una determinata area.
A causa di questo effetto, alcune aree del Nord Italia scarsamente popolate o rurali possono essere caratterizzate dalla presenza di elevate concentrazioni di PM_{10}. Focalizzandoci sulla regione Veneto, sono state rilevate concentrazioni di PM_{10} particolarmente elevate nella stagione estiva in alcune zone di alta montagna in provincia di Belluno (BL). Questo risultato è atipico in quanto le concentrazioni di PM_{10} sono solitamente più basse nel periodo estivo rispetto all'inverno. La causa di questo comportamento è strettamente associata alla temperatura. Infatti, nel periodo invernale la neve tende ad intrappolare le particelle inquinanti e, successivamente, a rilasciarle con l'incremento delle temperature a partire dal mese di aprile (Masiol, 2013).
Nel contesto della pianura Padana, la presenza delle catene montuose (Alpi e Appennini) e del mare Adriatico creano una schermatura. Ciò causa un indebolimento dei venti e un minor numero di perturbazioni atmosferiche. Lo scarso ricambio d'aria, l'elevata industrializzazione e la presenza di città

densamente abitate in un'area circoscritta contribuiscono a rendere il Bacino Padano una delle aree più inquinate d'Europa (Schiavon, 2003).

3.1.1 Strumenti di rilevamento dei PM_{10}

I dati dei PM_{10} sono raccolti dalle stazioni di monitoraggio distribuite sul territorio nazionale. Gli enti che si occupano di analizzarli sono le Agenzie Regionali per la Protezione dell'Ambiente (ARPA).
Il metodo manuale utilizzato per l'analisi è detto gravimetrico. Tale metodo richiede una fase preventiva di pesata del filtro, la fase di campionamento e una fase successiva di pesatura. Quest'ultima operazione deve avvenire in condizioni di umidità e temperatura controllate. Le particelle vengono raccolte su un filtro al quarzo di 47 mm.
L'unità di misura con la quale vengono misurate le concentrazioni di PM_{10} è microgrammi per metro cubo $\mu g/m^3$.

3.1.2 Limiti di legge e impatto sulla salute

Il parametro di valutazione della concentrazione di PM_{10} in Italia è la media giornaliera. Il valore limite giornaliero è di 50 $\mu g/m^3$. Questo valore non dovrebbe essere superato per più di 35 volte nel corso dell'anno solare. Per quanto concerne la media annua, essa non dovrebbe essere superiore a 40 $\mu g/m^3$.
E' di interesse valutare sia la media annua, in quanto ci dà una misura dell'esposizione media della popolazione, nonché il numero di giorni di superamento della soglia, che ci fornisce una misura dell'esposizione a picchi di concentrazione in un breve periodo. È rilevante sottolineare che la World Health Organization (WHO) reputa pericolosa per la salute pubblica una media annua maggiore di 20 $\mu g/m^3$.
Sebbene tutto il PM_{10} sia respirabile, ovvero sia in grado di penetrare all'interno delle vie respiratorie, la frazione con diametro minore come il PM_1

(materiale particolato aerodisperso di dimensione inferiore a 1 μm) ha la capacità di penetrare all'interno del circolo sanguigno provocando maggiori rischi per la salute. In letteratura sono presenti numerosi studi svolti in varie aree del mondo che dimostrano una correlazione positiva tra le elevate concentrazioni di inquinanti in un luogo e un aumento della mortalità e dei ricoveri ospedalieri nei soggetti fragili, bambini e anziani.

È di interesse citare lo studio svolto da Scapellato (2009) su un gruppo di pazienti affetti da asma nella città di Padova, monitorati dal 2004 al 2006. I dati sono stati raccolti rilevando, tramite un appropriato strumento fornito ai pazienti, la concentrazione individuale media giornaliera di PM_{10} in vari giorni dell'anno. Successivamente, le concentrazioni individuali sono state confrontate con i valori "outdoor" ottenuti dalle stazioni di monitoraggio site in Mandria (PD) e in Arcella (PD). Un confronto tra le concentrazioni individuali e "outdoor" ha mostrato che le prime sono risultate circa il doppio in tutti i giorni della rilevazione, con picchi di 330 $\mu g/m^3$. Infine, tramite la costruzione di un modello ad effetti casuali, si è scoperto che l'esposizione individuale è stata influenzata dalle concentrazioni "outdoor", dal fumo (sia attivo che passivo), dalla temperatura e dalla stagione dell'anno. Quindi, possiamo concludere che la popolazione può essere esposta a picchi elevati di concentrazioni di PM_{10} in un breve lasso di tempo (ore) e che le concentrazioni rilevate dalle due stazioni di monitoraggio sottostimano la reale esposizione della popolazione Padovana.

3.1.3 Fattori che influenzano la concentrazione dei PM_{10}

Le concentrazioni di PM_{10} sono influenzate da un complesso mix di fonti antropiche e naturali. È difficile stabilire in una particolare area quale sia la percentuale di ognuna delle due fonti.

Le fonti di emissioni di tipo antropico sono riconducibili a sei settori princi-

pali:

- Trasporti;
- Centrali Termoelettriche;
- Industria;
- Domestico e Terziario;
- Agricoltura;
- Allevamento.

Le fonti principali di tipo naturale sono riconducibili a:

- Incendi;
- Spray marino;
- Eruzioni vulcaniche;
- Polveri del deserto.

Questo mix di fonti varia notevolmente nel tempo (variabilità interannuale e stagionale) e il loro effetto su una particolare area dipende dall'intensità e dalla distanza tra la zona di origine e la centrale di monitoraggio.
Un'ulteriore distinzione da fare in merito all'origine dei PM_{10} è distinguerli tra **primario** e **secondario**. Possiamo parlare di PM_{10} **primario** quando esso viene emesso in atmosfera direttamente sotto forma solida o liquida e la sorgente può essere naturale o antropica. Il PM_{10} **secondario** si forma direttamente in atmosfera tramite reazioni chimico-fisiche tra gli inquinanti gassosi. Una fonte secondaria impropria di PM_{10} è la **risospensione**. Non possiamo considerarla una vera e propria fonte di PM_{10} in quanto non genera nuovo particolato, ma permette al particolato depositato sul suolo di tornare in circolazione. Le cause che producono questo fenomeno possono essere di

origine naturale (come il vento) o antropiche (come il traffico) (Schiavon, 2003).

È di interesse spiegare l'effetto che hanno le polveri del deserto provenienti dal Nord Africa e dal Medio Oriente sulle concentrazioni medie giornaliere ed annue di PM_{10} in Italia (Barnaba, 2022). Quanto questa fonte influenzi le concentrazioni annue di PM_{10} in un'area, dipende prettamente dal numero di giorni caratterizzati dalla presenza di polveri in atmosfera e dall'intensità del fenomeno. In aggiunta, a causa della **risospensione** nelle aree urbane caratterizzate da intenso traffico, il contributo di questa fonte sulle concentrazioni di PM_{10} è maggiore rispetto alle aree rurali. L'Italia Settentrionale è caratterizzata da un minor numero di giorni in cui sono presenti polveri del deserto in atmosfera, tuttavia, quando tale fenomeno si verifica, ciò avviene con maggiore intensità. Inoltre, se da un lato le polveri del deserto provocano un aumento contenuto della concentrazione media annua di PM_{10} pari a 1 $\mu g/m^3$, dall'altro il contributo medio giornaliero può arrivare a 12 $\mu g/m^3$. Questa fonte, indubbiamente, contribuisce in modo significativo all'incremento del numero di giorni in cui viene superata la soglia critica giornaliera.

Le variabili meteorologiche sono di fondamentale importanza rispetto ai livelli di inquinamento atmosferico. La loro influenza gioca un ruolo fondamentale nell'incremento o meno dei livelli di concentrazione di PM_{10} in un determinato periodo dell'anno. Le principali sono :

- direzione ed intensità del vento;
- pressione atmosferica (mbar);
- temperatura al suolo (°C);
- umidità relativa (%);
- radiazione solare (Wm^{-2}).

Nella stagione invernale, a causa delle basse temperature, l'utilizzo da parte della popolazione di stufe a pellet, camini, etc. è responsabile dell'incremento delle concentrazioni di PM_{10} (Masiol, 2013). In aggiunta, le condizioni di stabilità atmosferica e il ridotto rimescolamento degli strati inferiori favoriscono l'accumulo di particolato a pochi metri dal suolo.

3.2 Analisi dei dati: Mandria (PD) nel 2021

In questo paragrafo, viene analizzata una serie storica univariata sulle concentrazioni di PM_{10} nell'aria, a questo scopo utilizzerò il metodo proposto nel Capitolo 2.

Il dataset è stato gentilmente fornito dall'ARPAV.[1]

I dati provengono dalla stazione di monitoraggio sita in un quartiere di Padova denominato Mandria (*Fig.3.2.1*) e si riferiscono all'anno 2021. Il quartiere è ubicato nella zona Sud del Comune di Padova ed è caratterizzato dalla presenza di una vasta zona agricola attraversata dal fiume Bacchiglione. Inoltre, la stazione di monitoraggio è lontana (450 metri) da qualsiasi strada principale.

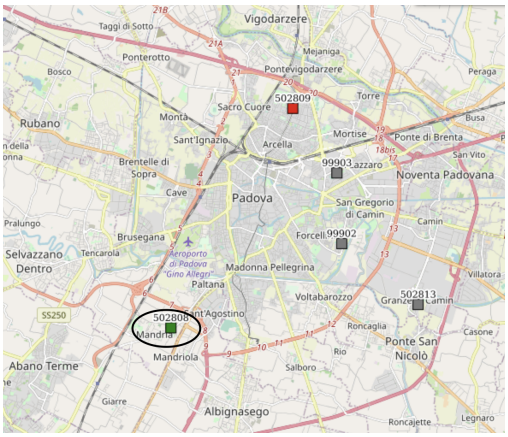

Fig. 3.2.1 *: Carta politica della città di Padova e dintorni. Il quadratino verde nell'ellisse nero rappresenta la stazione di monitoraggio di interesse.*

[1] http://www.arpa.veneto.it/

3.2.1 Presentazione dei dati grezzi

La generica Y_i rappresenta la concentrazione media dell'i-esimo giorno, $i = 1, \ldots, n$. La dimensione del campione è $n = 365$.

Non sono state rilevate le concentrazioni medie giornaliere Y_i in sei giorni. I dati mancanti sono da imputare a guasti e/o ricalibrazione del sistema di monitoraggio. Questi valori sono stati sostituiti calcolando la media tra l'osservazione del giorno precedente e successivo. Nella *Tabella 3.2.1* vengono riportate le date e i valori delle concentrazioni di PM_{10} che sono state calcolate seguendo questa strategia.

data(gg/mm/aaaa)	valore sostituito($\mu g/m^3$)
31/03/2021	51 $\mu g/m^3$
04/04/2021	15 $\mu g/m^3$
06/04/2021	10 $\mu g/m^3$
26/07/2021	19 $\mu g/m^3$
27/08/2021	11 $\mu g/m^3$
17/11/2021	34 $\mu g/m^3$

***Tabella 3.2.1**: Concentrazioni medie di PM_{10} calcolate per i valori mancanti*

3.2.2 Analisi esplorative

La concentrazione media annuale di PM_{10} nel 2021 è stata pari a 27.86 $\mu g/m^3$. Il numero di giorni che è stata superata la soglia di allerta (50 $\mu g/m^3$) è pari a 52. Questi giorni, con concentrazioni particolarmente elevate di PM_{10}, sono situati in due distinti periodi dell'anno: il primo, dal 1 gennaio al 2 aprile e, il secondo, tra il 19 ottobre e la fine del periodo di osservazione. L'andamento della serie storica rappresentata in *Fig. 3.2.2* evidenzia una non stazionarietà in media nel periodo di osservazione. L'ipotesi di non stazionarietà è stata confermata dal test formale di Dickey-Fuller. Per ciò

che concerne la variabilità, la serie è chiaramente eteroschedastica.
Inoltre, è evidente la presenza di un trend e un effetto stagionale confermato della funzione di autocorrelazione (*Fig. 3.2.3*). L'**ACF** mostra correlazioni marcate a *lag* elevati e viene definito processo a memoria lunga. In questo tipo di processi, l'**ACF** converge molto lentamente a zero e ciò implica che la dipendenza tra osservazioni successive decade lentamente all'aumentare della distanza temporale tra le stesse.

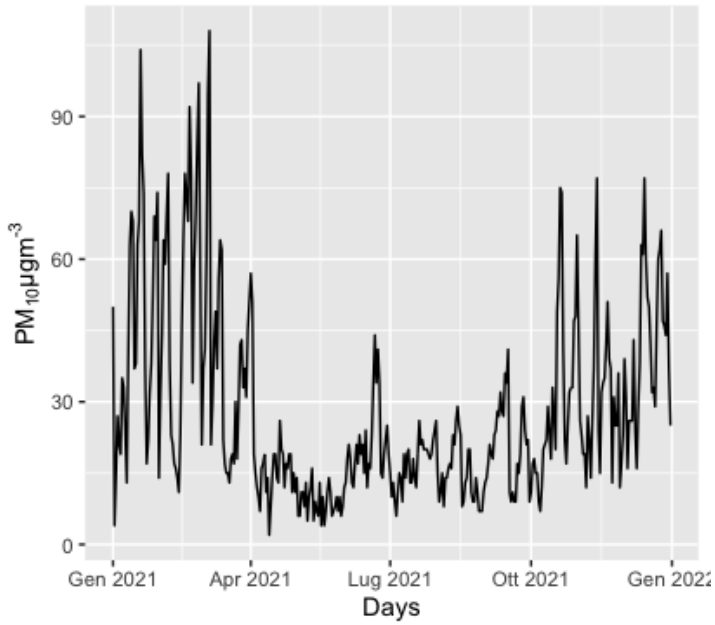

***Fig. 3.2.2**: Serie storica annuale (2021) sulle concentrazioni medie giornaliere di PM_{10} provenienti dalla centralina sita in Mandria (PD)*

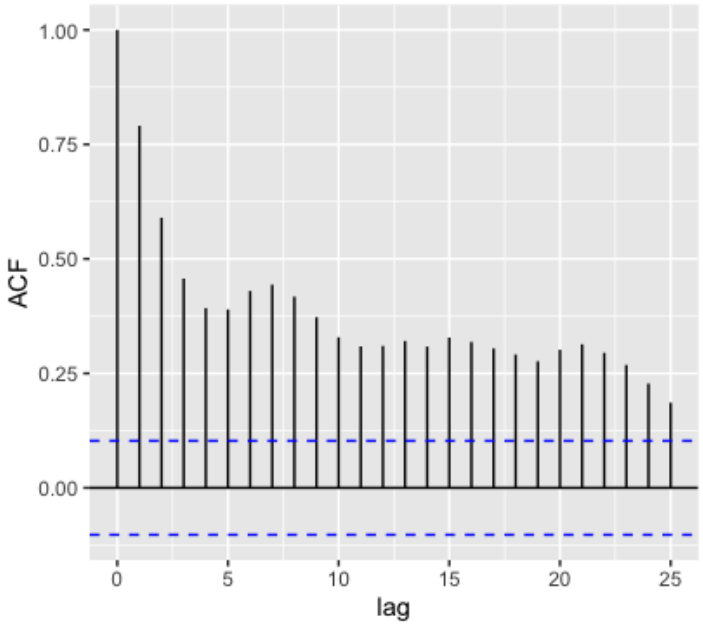

***Fig. 3.2.3**: Funzione di autocorrelazione (**ACF**) dei dati originali*

È di interesse analizzare il comportamento della variabile Y nei vari mesi dell'anno *(Tabella 3.2.2)*. Dall'analisi dei boxplot si nota una forte eterogeneità tra le concentrazioni mensili (in *Fig. 3.2.4*). In particolare, la mediana dei mesi invernali e autunnali è significativamente maggiore rispetto ai mesi estivi. La stessa eterogeneità tra i due periodi dell'anno (mesi invernali e estivi) è presente per la variabilità. Ad esempio, la differenza tra il terzo e il primo quartile delle "scatole" dei boxplot per i mesi di marzo risulta essere molto più grande rispetto ai mesi di maggio e giugno. L'analisi tramite boxplot segnala la presenza di valori anomali per i mesi di marzo, aprile, giugno, ottobre e novembre.

Mese	Media ($\mu g/m^3$)	Mediana ($\mu g/m^3$)	Standard deviation	Range ($\mu g/m^3$)
Gennaio	44.09	38	24.77	4 - 104
Febbraio	51.07	57.05	26.25	11 - 97
Marzo	38.64	37	22.48	13 - 108
Aprile	17.4	15.5	11.04	2 - 57
Maggio	8.87	8	3.19	4 - 16
Giugno	21.1	19	8.43	12 - 44
Luglio	16.77	18	5.05	6 - 26
Agosto	14.58	14	6.15	7 - 29
Settembre	21.26	21	8.79	9 - 41
Ottobre	30.12	23	18.44	7 - 75
Novembre	29.93	26	14.73	12 - 77
Dicembre	42.03	42	15.83	16 - 77

***Tabella 3.2.2**: Distribuzione delle concentrazioni di PM_{10} nel 2021, per mese dell'anno*

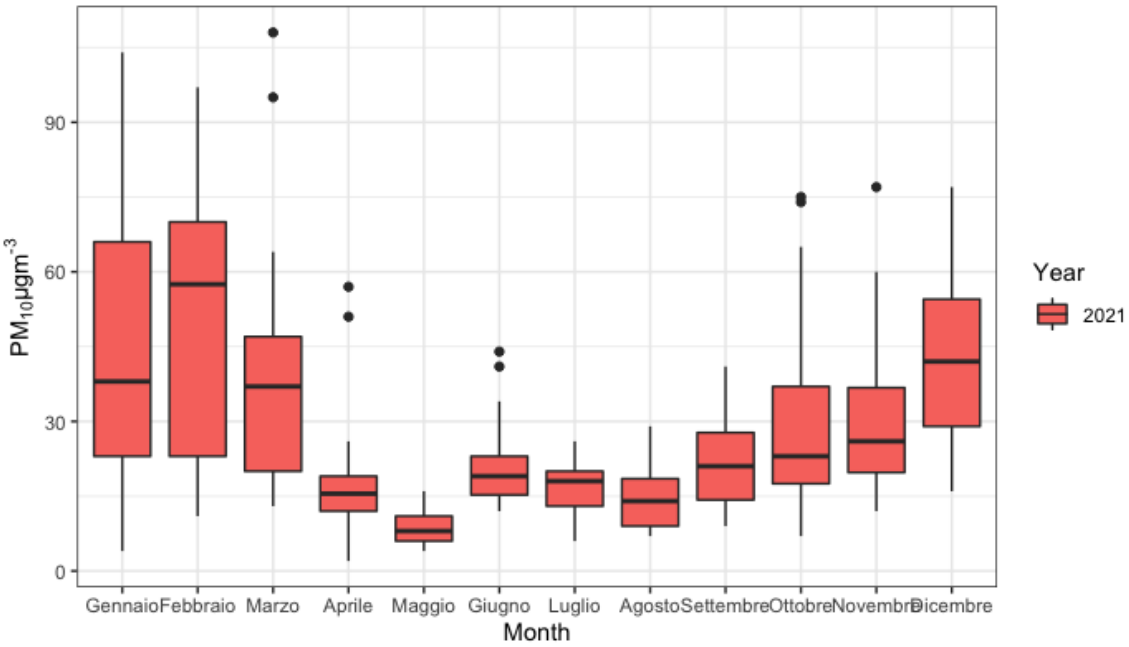

***Fig. 3.2.4**: Boxplot mensili delle concentrazioni di PM_{10} nell'anno 2021*

La serie delle concentrazioni di PM_{10} (*Fig. 3.2.2*) mostra un forte decremento delle concentrazioni di PM_{10} tra marzo-aprile e successivamente un forte incremento delle concentrazioni tra ottobre e novembre. Il fattore che maggiormente influenza la differenza tra i mesi invernali ed estivi è senza alcun dubbio legato all'utilizzo dei riscaldamenti (stufe a pellet, camini, etc.). Questi periodi coincidono con le variazioni più significative della temperatura dell'aria. Secondo la normativa nazionale, gli impianti di riscaldamento non alimentati a gas naturale possono essere accesi nel Comune di Padova dal 15 ottobre al 15 aprile, per un massimo di 14 ore giornaliere. Si può notare un forte calo delle concentrazioni di PM_{10} al di fuori di questo intervallo temporale. Per completezza, la *Fig. 3.2.5* riporta la serie storica delle temperature medie giornaliere nella città di Padova nel 2021.

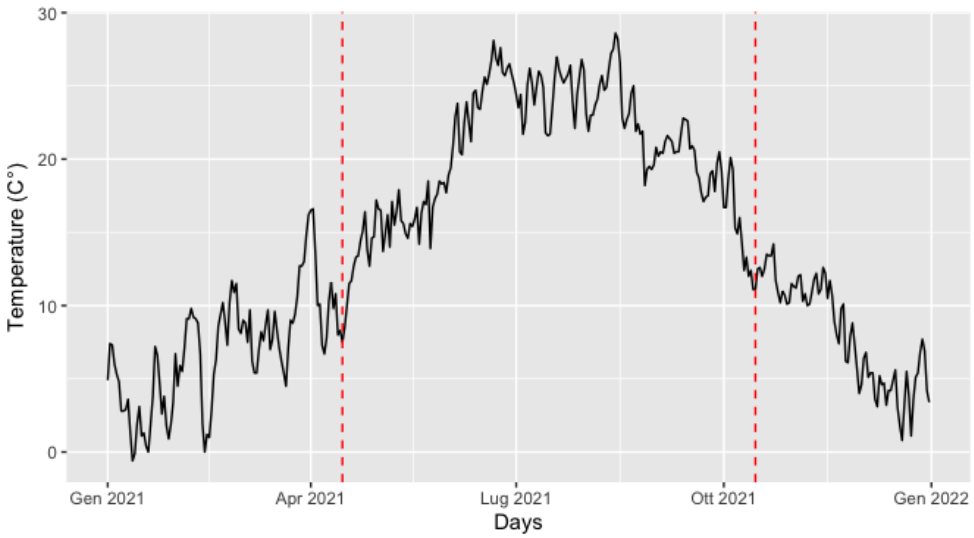

***Fig. 3.2.5**: Serie Storica delle temperature medie giornaliere nella città di Padova(PD) nell'anno 2021. Le linee rosse identificano i giorni 15 aprile e 15 ottobre.*

3.2.3 Sviluppo dell'algoritmo

La presenza di una struttura di correlazione per la serie storica Y_i, nonché la forte eterogeneità in media e variabilità tra i vari mesi dell'anno, portano allo sviluppo di un metodo di identificazione dei valori anomali basato sui residui e non direttamente sulla variabile originale osservata.

In primo luogo, sviluppiamo il primo passo dell'algoritmo (sottoparagrafo 2.2.1) che consiste nello stimare una funzione di regressione non parametrica $\widehat{m}(x_i)$. È stata utilizzata la funzione kernel di Epanechnikov.

Per cogliere maggiormente la variabilità e la forma dei dati viene applicato uno smoothing con parametro di lisciamento locale (h_{x_i}). Come si può notare in *Fig. 3.2.6*, otteniamo una curva di regressione maggiormente irregolare in grado di cogliere sufficientemente bene l'andamento dei dati bilanciando il trade off distorsione-varianza.

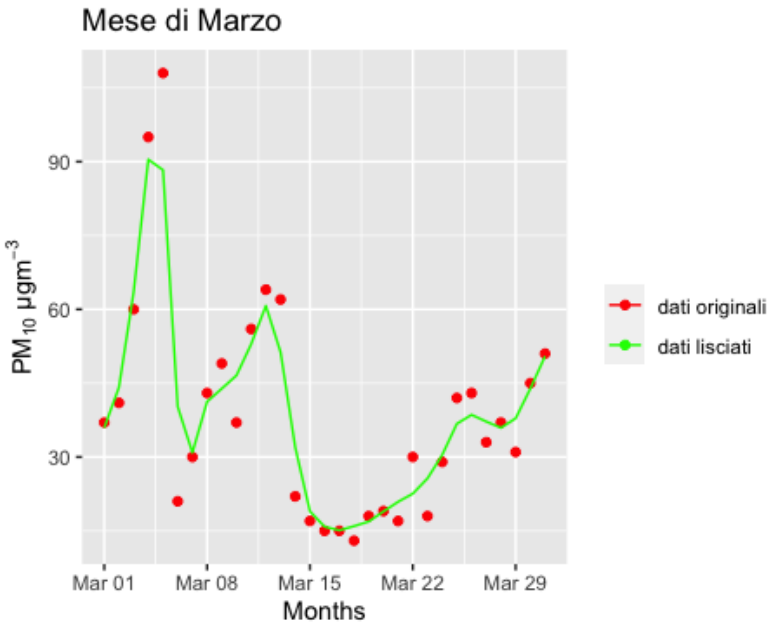

***Fig. 3.2.6**: Concentrazioni di PM_{10} nel mese di marzo 2021, e relativa stima della funzione di regressione $\widehat{m}(x_i)$ con parametro di lisciamento locale $h(t)$.*

L'obiettivo finale del primo passo dell'algoritmo è ottenere un insieme di residui $R(x_1), \ldots, R(x_n)$ come mostrato in *Fig. 3.2.7*.
Si può notare che i residui con valori maggiori sono presenti nei mesi di gennaio e febbraio a cui corrispondono le concentrazioni di PM_{10} più elevate. Possiamo asserire che i residui sono stazionari in media (≈ 0), ma, allo stesso tempo, non è possibile assumere l'ipotesi di omoschedasticità.
La funzione di autocorrelazione dei residui (*Fig. 3.2.8*) è influenzata dalla larghezza di banda h_t della regressione kernel. Tuttavia, possiamo assumere una correlazione trascurabile per $lag > 3$. Ciò implica che possiamo considerare i residui sufficientemente distanti come indipendenti.
Questa assunzione è fondamentale per l'Analisi dei Punti di Cambiamento.

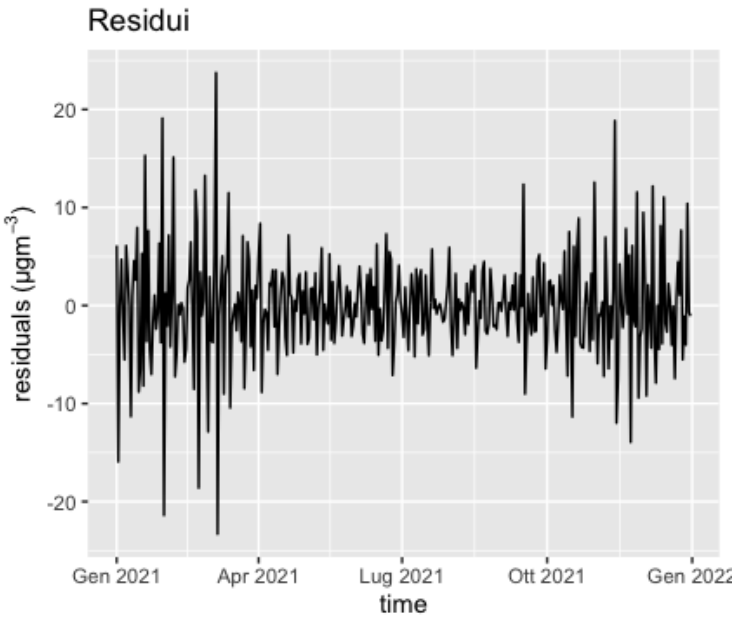

***Fig. 3.2.7**: Serie storica dei residui del modello*

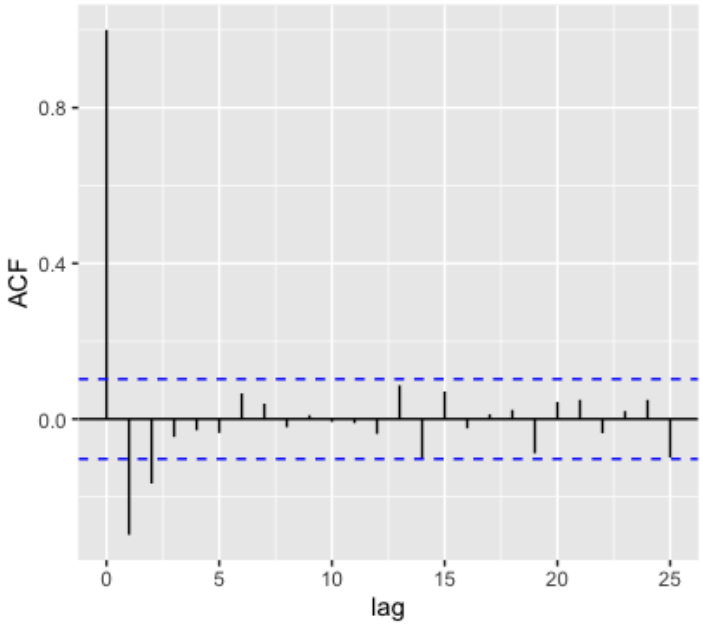

***Fig. 3.2.8**: Funzione di autocorrelazione empirica dei residui*

Lo scopo del secondo passo è applicare l'Analisi dei Punti di Cambiamento sui residui come descritto nel sottoparagrafo 2.2.2.

È stata verificata la normalità dei residui su ogni segmento S_j e ciò ha reso possibile l'applicazione dell'algoritmo PELT. Inoltre, è stato impostato un vincolo per cui ogni segmento S_j deve contenere almeno 20 osservazioni.

Come mostrato in *Fig. 3.2.9*, il metodo PELT individua due punti di cambiamento τ_1,τ_2 che dividono i residui in tre segmenti S_1, S_2, S_3.

I due punti di cambiamento stimati $\widehat{\tau_1}, \widehat{\tau_2}$ corrispondono al 1 Aprile 2021 e al 14 Settembre 2021.

Il fatto che sia plausibile l'ipotesi di normalità sui residui per ogni segmento S_j, $j = 1, 2, 3$, rende possibile sviluppare il rilevamento degli outlier sia tramite il test di Grubbs che i quantili della normale senza effettuare trasformazioni dei residui. Allo stesso tempo, applicheremo anche l'approccio di rilevamento non parametrico basato sulla disuguaglianza di Chebyshev.

La funzione **KRDetect.outliers.changepoint(. . .)** calcola le quantità descritte.

In *Tabella 3.2.3*, viene riportato sia il valore critico stimato L_j per la disuguaglianza di Chebyshev, che le stime del parametro α_j per il metodo basato sui quantili della distribuzione normale. Il numero di outlier identificati dipende fortemente dalla scelta dei parametri α_j e L_j.
Per il test di Grubbs, fissiamo il livello di significatività pari a α=5%, $\forall S_j$, $j = 1, 2, 3$.

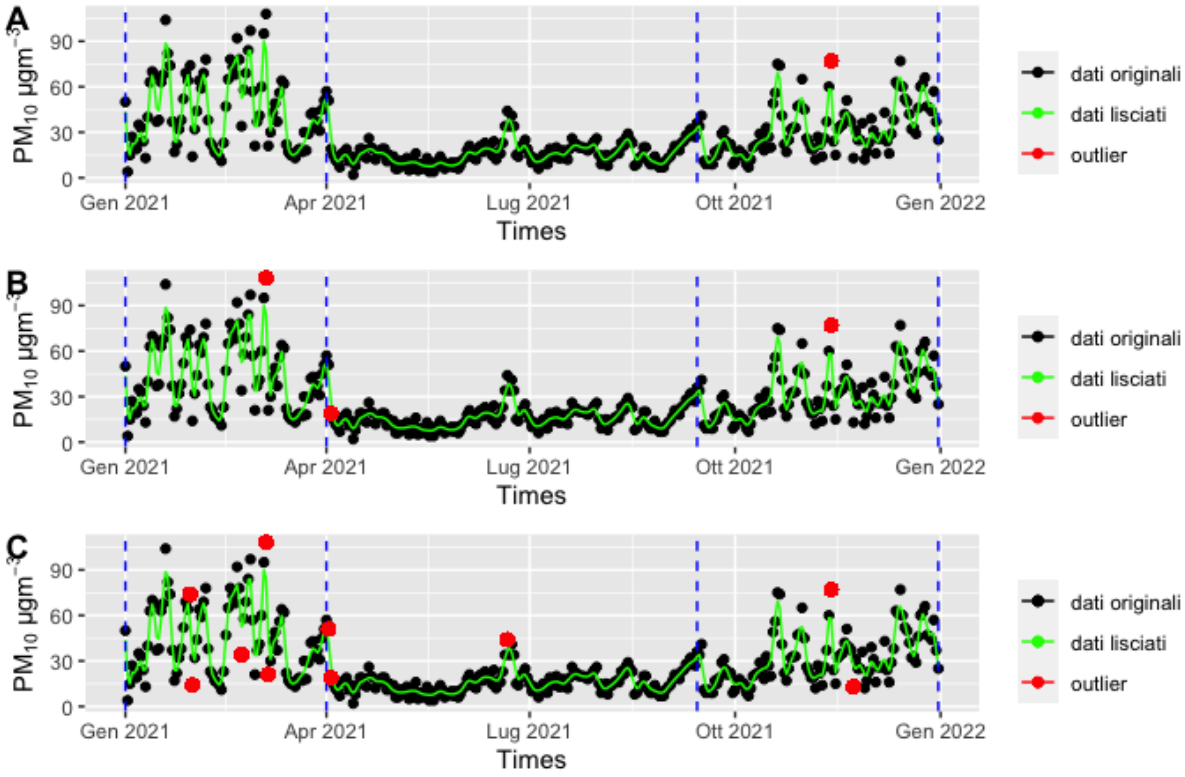

Fig. 3.2.9*: Concentrazione di PM_{10} osservate Y_i e outlier rilevati tramite test di Grubbs(A), disuguaglianza di Chebyshev(B) e quantili della normale(C). Le linee blu dividono i segmenti S_j con variabilità omogenea*

	L_j	α_j	$u_{\alpha/2}$
Segmento 1	3.059062	0.02	2.3263
Segmento 2	2.811252	0.02	2.3263
Segmento 3	3.312959	0.02	2.3263

***Tabella 3.2.3**: Stima dei parametri L_j e α_j, per ogni segmento S_j*

Nella *Tabella 3.2.4* sono riportati i residui identificati come outlier con i tre approcci. Si può notare che un unico residuo viene identificato come anomalo da tutti e tre i metodi e corrisponde alla concentrazione media del giorno 2021-11-13.

***Tabella 3.2.4**: Concentrazioni osservate rilevate come outlier dai tre metodi*

Test di Grubss

S_j	data	concentrazione osservata
S_3	2021-11-13	77 $\mu g/m^3$

Disuguaglianza di Chebyshev

S_j	data	concentrazione osservata
S_1	2021-03-05	108 $\mu g/m^3$
S_2	2021-04-03	19 $\mu g/m^3$
S_3	2021-11-13	77 $\mu g/m^3$

Quantili della distribuzione normale

S_j	data	concentrazione osservata
S_1	2021-01-30	74 $\mu g/m^3$
S_1	2021-01-31	14 $\mu g/m^3$
S_1	2021-02-22	34 $\mu g/m^3$
S_1	2021-03-05	108 $\mu g/m^3$
S_1	2021-03-06	21 $\mu g/m^3$
S_2	2021-04-02	51 $\mu g/m^3$
S_2	2021-04-03	19 $\mu g/m^3$
S_2	2021-06-21	44 $\mu g/m^3$
S_3	2021-11-13	77 $\mu g/m^3$
S_3	2021-11-23	13 $\mu g/m^3$

3.2.4 Confronto sui PM_{10} negli anni 2020-2021

In Appendice A è stata svolta l'analisi considerando le concentrazioni di PM_{10} nel 2020 provenienti dalla stessa stazione di monitoraggio.

Da un confronto sulle concentrazioni tra i due anni considerati emergono due aspetti salienti. La concentrazione media annua di PM_{10} dal 2020 al 2021 è diminuita del 12.85%, mentre il numero di giorni che eccedono la soglia limite è passato da 80 a 52. Un confronto tra i boxplot nei due anni mostra che la variabilità nei vari mesi del 2020 è particolarmente analoga a quella del 2021. Infatti, l'Analisi dei punti di Cambiamento effettuata sui dati del 2020 porta a dividere i residui in tre segmenti S_j. Il primo punto di cambiamento τ_1 individuato nei due anni si differenzia di un giorno; il secondo punto di cambiamento τ_2 viene stimato nel 2021 con sedici giorni di anticipo rispetto al 2020.

Conclusioni

In questa tesi viene applicato l'algoritmo sviluppato da Čampulová (2018c) con l'obiettivo di identificare le concentrazioni anomale di PM_{10}. Il metodo sviluppato è applicabile a dataset provenienti da molteplici aree di ricerca ed è particolarmente performante quando le misurazioni sono raccolte in modo continuo con alta risoluzione temporale. Sfortunatamente, non è possibile identificare automaticamente la reale causa degli outlier rilevati. Nel contesto dei dati ambientali, le osservazioni anomale possono derivare da malfunzionamenti del sistema di rilevazione, errori nel caricamento dei dati oppure da un comportamento anomalo della variabile osservata. Solo un esperto del settore può essere in grado di comprendere l'origine del valore anomalo rilevato. Lo scopo del metodo proposto è fornire una procedura automatica per l'identificazione dei valori anomali e permettere agli esperti di concentrarsi su un insieme ristretto di osservazioni. La rilevazione degli outlier può dimostrarsi una sfida ardua, ma in questa tesi è stato mostrato che, grazie ad un numero elevato di metodi, si può affrontare questo problema. Infine, studi futuri sono necessari per implementare i metodi presenti in letteratura, specialmente quando la serie storica è multivariata.

.1 Appendice A

In questa appendice si analizzano i dati racccolti nel 2020 sulle concentrazioni PM_{10} dalla centralina della Mandria(PD). La dimensione del campione è n=366 (anno bisestile).

La concentrazione media annuale calcolata nel 2020 è stata pari a 31.97 $\mu g/m^3$. Il numero di giorni che è stata superata la soglia di allerta di 50 $\mu g/m^3$ è pari a 80. Dalla *Fig. A.2* la serie storica mostra tre segmenti con variabilità differente.

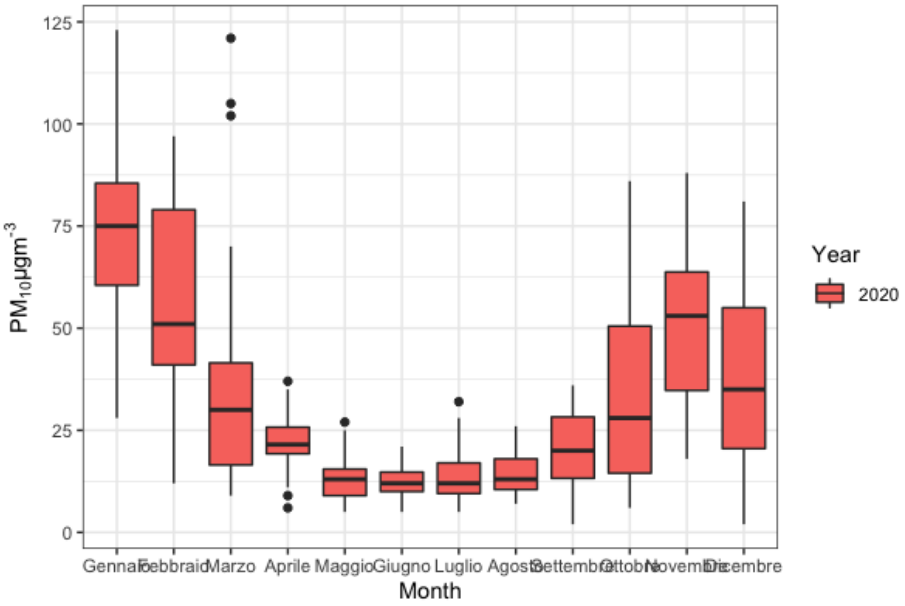

***Fig. A.1**: Boxplot mensili delle concentrazioni di PM_{10} nell'anno 2020*

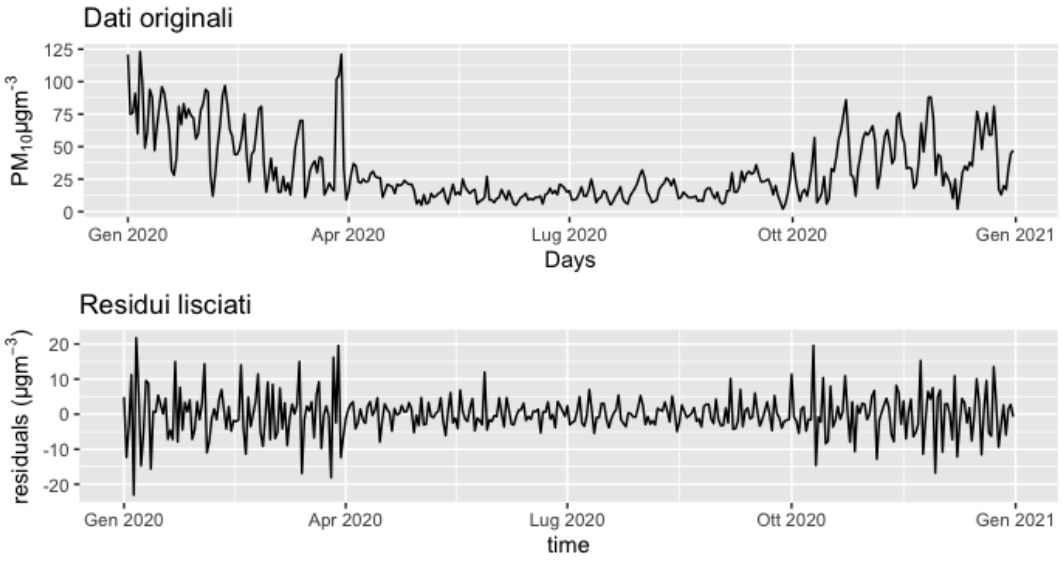

***Fig. A.2**: Serie originale e residui lisciati sulle concentrazioni di PM_{10} provenienti dalla centralina sita in Mandria(PD) ,2020*

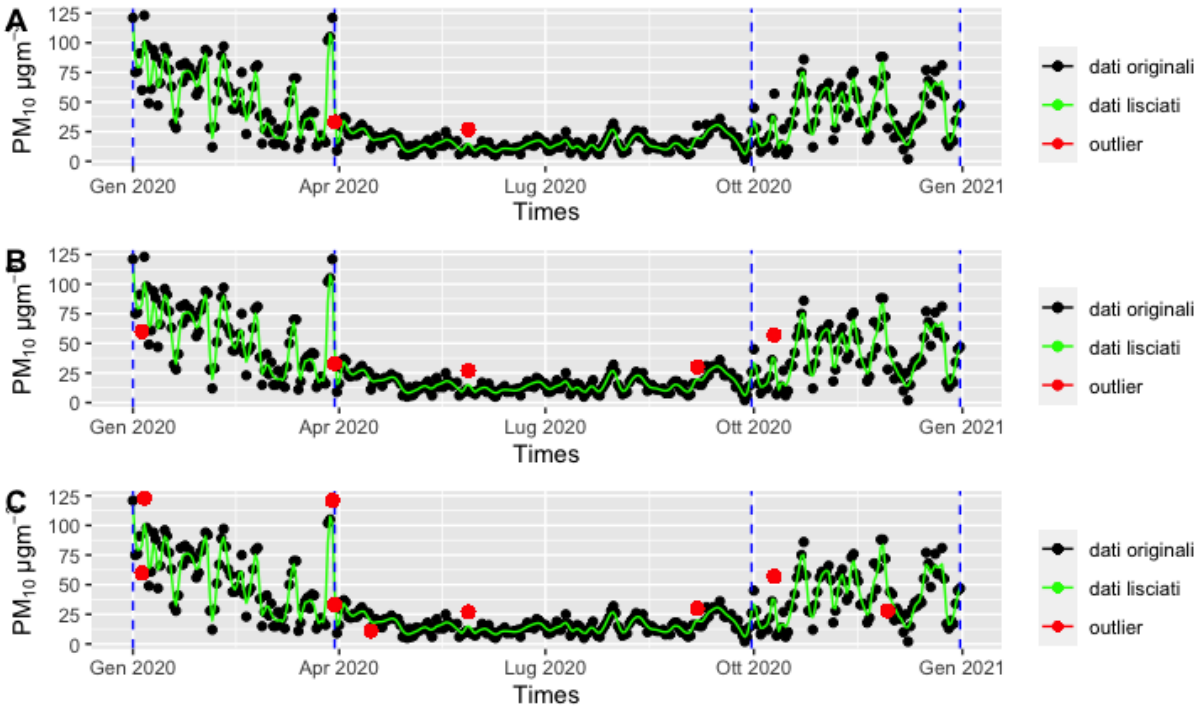

***Fig. A.3**: Concentrazione di PM_{10} osservate Y_i e outlier rilevati tramite test di Grubbs(A), disuguaglianza di Chebyshev(B) e quantili della*

normale(C). Le linee blu dividono i segmenti S_j con variabilità omogenea

L'applicazione del metodo **PELT** porta all'individuazione di due punti di cambiamento $\widehat{\tau}_1, \widehat{\tau}_2$ corrispondenti al giorno 30 Marzo 2020 e 30 Settembre 2020. Infine, i due outlier identificati tramite il test di Grubbs vengono rilevati anche dagli altri due metodi.

***Tabella A.1**: Concentrazioni osservate rilevate come outlier dai tre metodi*

Test di Grubss

S_j	data	concentrazione osservata
S_1	2020-03-30	33 $\mu g/m^3$
S_2	2020-05-28	27 $\mu g/m^3$

Disuguaglianza di Chebyshev

S_j	data	concentrazione osservata
S_1	2020-01-05	60 $\mu g/m^3$
S_1	2020-03-30	33 $\mu g/m^3$
S_2	2020-05-28	27 $\mu g/m^3$
S_2	2020-09-06	30 $\mu g/m^3$
S_3	2020-10-10	57 $\mu g/m^3$

Quantili della distribuzione normale

S_j	data	concentrazione osservata
S_1	2020-01-05	60 $\mu g/m^3$
S_1	2020-01-06	123 $\mu g/m^3$
S_1	2020-03-29	121 $\mu g/m^3$
S_1	2020-03-30	33 $\mu g/m^3$
S_2	2020-04-15	11 $\mu g/m^3$
S_2	2020-05-28	27 $\mu g/m^3$
S_2	2020-09-06	30 $\mu g/m^3$
S_3	2020-10-10	57 $\mu g/m^3$
S_3	2020-11-29	28 $\mu g/m^3$

Bibliografia

[1] Amidan, B., Ferryman, T., Cooley, S. (2005), *Data outlier detection using the Chebyshev theorem.* IEEE Aerospace Conference, pp. 3814-3819.

[2] Barnaba, F., Romero, N.A., Bolignano, A., Basart, S., Renzi, M. (2022), *Multiannual assessment of the desert dust impact on air quality in Italy combining PM10 data with physics-based and geostatistical models.* Environment International, Vol. 163.

[3] Blazquez-Garcìa, A., Conde, A., Mori, U., Lozano, J. (2021) *A review on outlier/anomaly detection in time series data.* ACM Computing Surveys, Vol. 54(3), pp. 1-33.

[4] Čampulová, M., Grochová, L., Michálek, J. (2018a), *Outlier detection in PM_{10} aerosols by generalised linear model.* In: Proceedings of the International Conference of Numerical Analysis and Applied Mathematics. American Institute of Physics (AIP), Melville.

[5] Čampulová, M. (2018b), *Comparison of methods for smoothing environmental data with an application to particulate matter PM10.* Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis.

[6] Čampulová, M., Michálek, J., Mikuška, P., Bokal, P. (2018c), *Nonparametric algorithm for identification of outliers in environmental data.* Journal of Chemometrics, Vol. 32(5).

[7] Čampulová, M., Čampula, R., Holešovskýc, J. (2022), *An R package for identification of outliers in environmental time series data.* Environmental Modelling and Software.

[8] Chen, H., Huang, C. (2014), *The use of a CUSUM residual chart to monitor respiratory syndromic data.* IIE Transactions, Vol. 46(8), pp. 790-797.

[9] Chen, J., Li, W., Lau, A., Cao, J., Wang, K. (2010), *Automated load curve data cleansing in power systems.* IEEE Transactions on Smart Grid, Vol. 1(2), pp. 213-221.

[10] Costa, M., Gonçalves, A M., Teixeira, L. (2016), *Change-point detection in environmental time series based on the informational approach.* Electronic Journal of Applied Statistical Analysis, Vol. 9(2), pp. 267-296.

[11] Herrmann, E. (1997), *Local bandwidth choice in kernel regression estimation.* Journal of Computational and Graphical Statistics, Vol. 6(1), pp. 35-54.

[12] Hodge,V., Austin, J. (2004), *A survey of outlier detection methodologies.* Artificial intelligence review, Vol. 22(2), pp. 85-126.

[13] Hyndman, R., Wang, E., Laptev, N. (2015), *Large-scale unusual time series detection.* IEEE international conference on data mining workshop.

[14] Lai, K.H., Zha, D., Xu, J., Zhao, Y., Wang, G., Hu, X. (2021), *Revisiting time series outlier detection: Definitions and benchmarks.* Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

[15] Masiol, M., Formenton, G., Pasqualetto, A., Pavoni, B. (2013), *Seasonal trends and spatial variations of PM_{10}-bounded polycyclic aromatic hydro-*

carbons in Veneto Region, Northeast Italy. Atmospheric Environment, Vol. 79, pp. 811-821.

[16] Scapellato, M. L., Canova, C., De Simone, A., Carrieri, M., Maestrelli, P., Simonato, L., Bartolucci, G. B. (2009), *Personal PM_{10} exposure in asthmatic adults in Padova, Italy: seasonal variability and factors affecting individual concentrations of particulate matter.* International journal of hygiene and environmental health, Vol. 212(6), pp. 626-636.

[17] Schiavon, M. (2003), *Osservatorio sui PM_{10} 2003.* Legambiente Padova.

[18] Urvoy, M., Autrusseau, F. (2014), *Application of Grubbs test for outliers to the detection of watermarks.* Proceedings of the 2nd ACM workshop on Information hiding and multimedia security, pp. 49-60.

[19] Wambui, G.D., Waititu, A., Wanjoya, A. (2015), *The Power of the Pruned Exact Linear Time(PELT) test in multiple changepoint detection.* American Journal of Theoretical and Applied Statistics, Vol. 4(6), pp. 581-586.