APAT
Agenzia per la protezione
dell'ambiente e per i servizi tecnici
Dipartimento Stato dell'Ambiente e Metrologia Ambientale

CONTROLLO, ANALISI ED ELABORAZIONI DEI DATI DI PM_{10} PROVENIENTI DALLE STAZIONI DI MONITORAGGIO ED ARCHIVIATI NEL DATABASE BRACE

Dott. Pietro Marinelli

Tutor: Anna Maria Caricchia

Co-Tutor: Alessandro Di Menno Di Bucchianico, Alessandra Gaeta

Ottobre 2007

Indice

Riassunto

Il presente lavoro ha come obiettivo l'analisi dei dati di concentrazione di PM_{10} in un'ottica statistico-probabilistica con riferimento all'anno 2001.

Partendo dall'esigenza di effettuare la validazione dei dati in questione, con l'obiettivo di renderli disponibili al pubblico, è stata affrontata la questione del controllo di qualità dei dati.

La numerosità elevata dei dati disponibili ha consentito di poter effettuare la ricerca di una funzione di densità di probabilità a partire dalla distribuzione delle frequenze relative osservate. Tale funzione di densità di probabilità è rappresentata da una distribuzione Gamma, confermata ulteriormente da un'analisi di regressione che ha prodotto un indice di bontà dell'accostamento prossimo all'unità.

A partire dalla distribuzione trovata sono stati analizzati i vantaggi derivanti dall'applicazione della distribuzione Gamma nel controllo di qualità dei dati in particolare nella determinazione delle soglie per definire i dati sospetti.

Abstract

This work aims to analyse the concentration of PM_{10} from a statistic-probabilistic point of view with reference 2001 data. The work starts with the data validation and deals with the issue of quality control.

The high number of available data has made it possible to search for the probability density function starting by the distribution of frequencies.

It is easy to conclude that a Gamma distribution fits very well with PM10 data and this intuition is confirmed by a regression analysis that produces an index of goodness almost equal to 1.
Referring to the opportunities given by that probability distribution in quality control we have found a lower bound and an upper bound that define suspect data.

1. PM10: Generalità

Il presente lavoro si prefigge l'obiettivo di realizzare un'analisi statistica della concentrazione di PM_{10} sul territorio italiano.

Per materiale particolato aerodisperso si intende l'insieme delle particelle atmosferiche solide e liquide aventi diametro aerodinamico (d_a) variabile fra 0,01 e circa 100 μm (Marconi, 2003). Le particelle più grandi di 10 μm sono in genere polveri derivanti dall'erosione o da processi industriali, vengono depositate al suolo in tempi piuttosto brevi e sono responsabili di fenomeni di inquinamento su scala spaziale abbastanza ridotta. Le particelle con diametro aerodinamico inferiore o uguale ai 10 μm, e quelle con diametro aerodinamico inferiore o uguale a 2,5 μm, sono l'oggetto della maggior parte degli studi sull'inquinamento atmosferico e vengono comunemente identificate nelle classi PM_{10} e $PM_{2,5}$. Esse sono caratterizzate da lunghi tempi di permanenza in atmosfera e possono quindi essere trasportate a distanze anche molto grandi dal punto di emissione, hanno una natura chimica particolarmente complessa e variabile, sono in grado di penetrare nell'albero respiratorio umano e di avere quindi effetti negativi sulla salute (Brunekreef, 2002).

Parte di queste sostanze vengono emesse in atmosfera già sotto forma di particolato (i cosiddetti aerosol primari) mentre le altre derivano da processi chimico-fisici che si realizzano fra altre specie inquinanti (aerosol secondari).

Queste polveri disperse nell'aria possono avere sia un'origine naturale (per esempio, l'erosione dei venti sulle rocce, le eruzioni vulcaniche, l'autocombustione di boschi e foreste) sia antropica (per esempio, il traffico autoveicolare, l'usura del manto stradale e combustioni di vario genere). Di origine antropica sono anche molte delle sostanze su cui si basano i fenomeni di inquinamento secondario e che portano alla formazione di particelle di piccola granulometria (come, per esempio, il biossido di zolfo, gas, che in determinate condizioni viene ossidato a solfato, particolato).

Le proprietà e gli effetti delle particelle aerodisperse sono strettamente legati alle loro dimensioni: la velocità di sedimentazione e il loro tempo di permanenza nell'atmosfera, come pure la loro deposizione all'interno dei polmoni e l'effetto di dispersione della luce, dipendono da questo parametro.

Se si studia la distribuzione del numero di particelle in funzione del loro diametro aerodinamico, si trova che la maggior parte di esse sono piuttosto piccole, con dimensioni inferiori a 0,1 μm, mentre le particelle con diametro aerodinamico maggiore di 0,1 μm sono in numero inferiore ma rappresentano la gran parte del volume (e la gran parte della massa) del materiale particellare atmosferico.

Per identificare le particelle si usano generalmente tre differenti convenzioni: la classificazione modale, basata sulla distribuzione per ampiezza e sui meccanismi di formazione; la classificazione rispetto al taglio, basata sull'efficienza di taglio del dispositivo di campionamento e la classificazione dosimetrica, basata sulla capacità di accesso alle differenti parti dell'apparato respiratorio. Tra queste tre, la più usata è generalmente la classificazione modale che prevede, in termini estremamente sintetici, due frazioni principali dette 'fine' e 'grossolana'. Nella frazione fine sono contenute tutte le particelle con diametro aerodinamico inferiore a 2÷2,5 μm, mentre nella grossolana quelle con diametro aerodinamico maggiore. I valori 2,5 e 10 μm sono invece legati ai sistemi di taglio, a loro volta basati su considerazioni di tipo dosimetrico.

Tabella 1.1 - Costituenti e sorgenti del particolato fine ($d_a < 2,5$ μm)

Sorgenti				
Primarie			**Secondarie**	
Specie	**Naturale**	**Antropica**	**Naturale**	**Antropica**
SO_4^{2-}	Spray marino	Uso di combustibili fossili	Ossidazione di SO_2 e H_2S emessi negli incendi e dai vulcani	Ossidazione di SO_2 dovuto all'impiego di combustibili fossili
NO_3^-	---	Emissioni di autoveicoli, combustioni	Ossidazione di NO_x prodotto dal suolo, da incendi boschivi e dalla luce	Ossidazione di NO_x dovuto all'impiego di combustibili fossili
Minerali	Erosione delle rocce	Polveri fuggitive, strade, agricoltura e silvicoltura	---	---
NH_4^+	---	---	Emissione di NH_3 da parte di animali selvatici	Emissione di NH_3 da parte di allevamenti animali, acque di scarico, terreni fertilizzati, autoveicoli
Carbonio organico	Incendi boschivi	Combustione di legna, cottura di cibi, emissioni di autoveicoli, usura di pneumatici, emissioni industriali	Ossidazione di idrocarburi emessi dalla vegetazione (terpeni), incendi boschivi	Ossidazione di idrocarburi emessi dagli autoveicoli, combustione di legna
Carbonio elementare	Incendi boschivi	Combustione di legna, cottura di cibi, emissioni di autoveicoli, emissioni industriali	---	---
Metalli		Uso di combustibili fossili, usura di freni, siderurgia	---	---
Bioaerosol	Virus, batteri	---	---	---

(EPA, 2002)

Tabella 1.2 - Costituenti e sorgenti del particolato grossolano (d_a > 2,5 μm)

Sorgenti				
Primarie			**Secondarie**	
Specie	**Naturale**	**Antropica**	**Naturale**	**Antropica**
Minerali	Erosione delle rocce	Polveri volatili, strade, agricoltura e silvicoltura	---	---
Metalli	Erosione, residui organici	---	---	---
Ioni	Spray marino	Spargimento di sale	---	---
Carbonio organico	---	Usura dell'asfalto e dei pneumatici	---	---
Residui Organici	Frammenti di piante e insetti	---	---	---
Bioaerosol	Pollini, funghi, spore, agglomerati di batteri	---	---	---

(EPA, 2002)

2. Normativa di riferimento

La normativa italiana sulla qualità dell'aria si inquadra in una legislazione europea che tutti i paesi membri sono tenuti a rispettare. Di conseguenza si è avuta una standardizzazione dei riferimenti, dei limiti e dei metodi di misura.

La normativa italiana ha subito numerosi cambiamenti nel corso degli anni. A partire dalla prima legge del 1966 sono subito emerse le difficoltà che derivano dal dover individuare criteri e metodi per gestire un problema così importante e complesso.
Tra l'altro, va segnalata la mancanza di riferimenti espliciti alla tutela dell'ambiente nella Costituzione Italiana. Fatta eccezione per i riferimenti ai diritti *inviolabili dell'uomo*, alla *tutela del paesaggio* e al *diritto alla salute*, va constata la totale assenza di riferimenti alla questione ambientale.

La prima legge ad occuparsi esplicitamente di problemi legati all'inquinamento dell'aria è, quindi, proprio quella del 1966.

Tralasciando alcuni fondamentali passaggi avvenuti dal 1966 in poi (riportati per completezza nella tabella qui sotto) passiamo a tempi più recenti. È del 1996, infatti, l'emanazione da parte del parlamento europeo della nuova direttiva quadro sulla qualità dell'aria 96/62/CE (recepita dall'Italia con D.L. N°351 del 1999) con la quale si è avuto un riordinamento completo del quadro normativo.

Da questa direttiva scaturiscono i limiti attuali per il PM10 che è l'oggetto della nostra analisi. Per la precisione, la normativa stabilisce i limiti di 50 $\mu g/m^3$ per la media sulle 24 ore e 40 50 $\mu g/m^3$ per la media annua.

Tabella 2.1 - Cronologia della legislazione ambientale con particolare riferimento al particolato atmosferico

LEGGI	**DATA**	**ARGOMENTO**
RIFERIMENTI GENERALI		
Art. 674 *Codice Penale* R. D. N° 1398	19/10/1930	Getto pericoloso di cose
Art. 216 *Testo unico delle Leggi Sanitarie* R. D. N° 1265	27/07/1934	Lavorazioni insalubri
Art. 844 *Codice Civile* R. D. N° 262	16/03/1942	Immissioni
Art. 2; 9; 32 *Costituzione della Repubblica Italiana*	22/12/1947	Diritti inviolabili dell'uomo; tutela del paesaggio; tutela della salute
N° 615 *Provvedimenti contro l'inquinamento atmosferico*	13/07/1966	Inquinamento transfrontaliero, inquinamento atmosferico (prima legge italiana sull'inquinamento atmosferico)
D.P.R. N° 322 *Regolamento per l'esecuzione della L. 13 luglio 1966, n. 615, recante provvedimenti contro l'inquinamento atmosferico, limitatamente al settore dell'industria*	15/04/1971	Inquinamento atmosferico
LEGGI ITALIANE CHE SI OCCUPANO DI PARTICELLE		
D.P.C.M. *Limiti massimi di accettabilità delle concentrazioni e di esposizione relativi ad inquinanti dell'aria nell'ambiente esterno*	28/03/1983	Inquinamento atmosferico, inquinamento ambientale
D.P.R. N° 203 *Attuazione delle direttive CEE numeri 80/779, 82/884, 84/360 e 85/203 concernenti norme in materia di qualità dell'aria, relativamente a specifici agenti inquinanti, e di inquinamento prodotto dagli impianti industriali*	24/05/1988	Attuazione direttive comunitarie, inquinamento atmosferico, impianti industriali
D.M. *Criteri per l'elaborazione dei piani regionali per il risanamento e la tutela della qualità dell'aria*	20/05/1991	Risanamento atmosferico, Qualità dell'aria, protezione ambientale

D.M. *Norme tecniche in materia di livelli e di stati di attenzione e di allarme per gli inquinanti atmosferici nelle aree urbane, ai sensi del D.P.R. n. 203 e del D.M. 20 maggio 1991*	15/04/1994	Normativa tecnica, aree urbane, inquinamento atmosferico
D.M. *Aggiornamento delle norme tecniche in materia di limiti di concentrazione e di livelli di attenzione e di allarme per gli inquinamenti atmosferici nelle aree urbane e disposizioni per la misura di alcuni inquinanti di cui al decreto ministeriale 15 aprile 1994*	25/11/1994	Aree urbane, inquinamento atmosferico, normativa tecnica
D. L. N° 351 *Attuazione della direttiva 96/62/CE in materia di valutazione e di gestione della qualità dell'aria ambiente*	4/08/1999	Qualità dell'aria
D. M. N° 60 Recepimento della direttiva 1999/30/CE concernente i valori limite di qualità dell'aria ambiente per il biossido di zolfo, il biossido di azoto, gli ossidi di azoto, le particelle e il piombo e della direttiva 2000/69/CE relativa ai valori limite di qualità dell'aria ambiente per il benzene ed il monossido di carbonio	2/04/2002	Qualità dell'aria

La tabella 2.2, estratta dall'allegato III del D.M. 60/02, riporta i valori limite per la protezione della salute umana, insieme al margine di tolleranza, alle modalità di riduzione di tale margine e alla data di entrata in vigore dei valori limite.

Tabella 2.2 - Particolato sospeso PM_{10}, valori limite per la protezione della salute umana (allegato III del D.M. 60/02)

	Periodo di mediazione	**Valore limite**	**Margine di tolleranza**	**Data alla quale il valore limite deve essere raggiunto**
Valore limite di 24 h	24 ore	50 $\mu g/m^3$ da non superare più di 35 volte per anno civile	50% del valore limite, pari a 25 $\mu g/m^3$, all'entrata in vigore della direttiva 99/30/CE (19/7/99). Tale valore è ridotto il 1° gennaio 2001 e successivamente ogni 12 mesi, secondo una percentuale annua costante, per raggiungere lo 0% il 1° gennaio 2005	1° gennaio 2005
Valore limite annuale	Anno civile	40 $\mu g/m^3$	20% del valore limite, pari a 8 $\mu g/m^3$, all'entrata in vigore della direttiva 99/30/CE (19/7/99). Tale valore è ridotto il 1° gennaio 2001 e successivamente ogni 12 mesi, secondo una percentuale annua costante, per raggiungere lo 0% il 1° gennaio 2005	1° gennaio 2005

3. La Banca dati BRACE

La banca dati BRACE (Banca dati Relazionale Atmosfera Clima Emissioni) fa parte degli strumenti realizzati nell'ambito della rete SINA[1] (Sistema Informativo Ambientale) per favorire la condivisione delle informazioni ambientali e territoriali e migliorarne la fruibilità.

La Banca dati BRACE nasce dalla necessità di adempiere alle esigenze dettate dalla normativa europea in tema di qualità dell'aria: da una parte le Decisioni 97/101/CE e 2001/752/CE, "Exchange of Information" (EoI), che instaurano uno scambio reciproco di informazioni e di dati provenienti dalle reti e dalle stazioni di monitoraggio dell'inquinamento atmosferico negli Stati Membri; dall'altra, la Direttiva 96/62/CE che pone le basi in materia di valutazione e gestione della qualità dell'aria ambiente, seguita da una serie di Direttive attuative, tra cui in particolare la Direttiva 2002/3/EC relativa all'ozono in aria ambiente.

All'APAT spetta lo sviluppo e la gestione del Sistema Informativo, tramite il quale si realizza la raccolta e l'organizzazione sistematica delle informazioni. L'utilizzo di tecnologie informatiche per la conservazione e organizzazione delle informazioni ne facilita enormemente la gestione. Inoltre, il Sistema Informativo offre un importante supporto alle strategie decisionali che in questo caso riguardano le politiche inerenti le problematiche ambientali .

La banca dati BRACE contiene le informazioni sui metadati (reti, stazioni e analizzatori utilizzati per il monitoraggio della qualità dell'aria) e i relativi dati di concentrazione degli inquinanti. Tali informazioni sono raccolte a livello locale dai Punti Focali Regionali (Regioni, Province, Comuni, ARPA/APPA) e successivamente trasmesse all'APAT, che a sua volta effettua la trasmissione a livello europeo secondo lo schema di Fig.1.

[1] A livello europeo, la rete del SINA è integrata nella rete EIOnet (Environment Information and Observation network) dell'Agenzia Europea dell'Ambiente (EEA), di cui rappresenta il nodo italiano (National Focal Point).

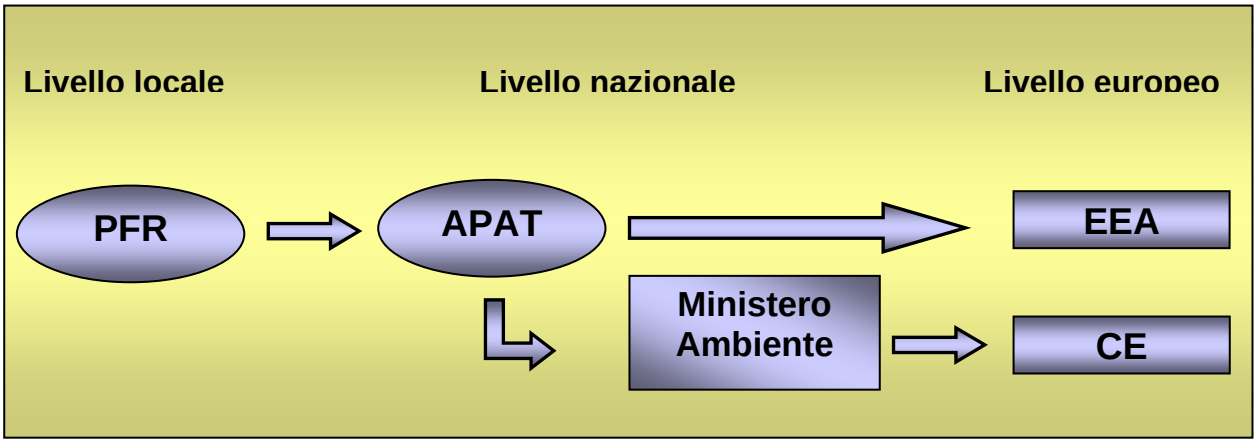

Fig.1 Schema del flusso di informazioni

3.1 Architettura del sistema BRACE: database e applicativi collegati

Il sistema BRACE consta di tre database e tre applicativi informatici:

Database:

- · PFR (database caricamento dati);
- · DBCON (database controllo/elaborazione dati da parte di APAT);
- · DBWEB (dati consolidati e accesso al pubblico).

Applicativi:

- · Winair (per la trasmissione ad APAT via web dei dati da parte dei gestori e/o PFR)
- · Conair (per il controllo e l'elaborazione dei dati da parte di APAT)
- · Brace (sito web di consultazione)

PFR è il database sul quale i PFR o fornitori dei dati caricano i dati di qualità dell'aria mediante l'applicativo web Winair disponibile all'indirizzo www.winair.sinanet.apat.it.

DBCON è il database per il controllo e le elaborazioni dei dati. L'applicativo, disponibile all'indirizzo www.brace.sinanet.apat.it/starold/winair_custom.avvio, viene utilizzato da APAT per effettuare le validazioni dei dati e tutte le elaborazioni richieste dalla normativa (calcolo di statistiche e superamenti).

DBWEB è il database dei dati controllati e consolidati, disponibili per il pubblico: attualmente sono presenti i dati di concentrazione degli inquinanti e le elaborazioni (statistiche e superamenti) dal 2002 al 2006, consultabili all'indirizzo www.brace.sinanet.apat.it/.

4. Controllo della qualità dei dati

I metadati e i dati di qualità dell'aria caricati dal gestore nella piattaforma PFR, al termine della procedura di trasferimento nella piattaforma di controllo DBCON, sono sottoposti ad una procedura di controllo di qualità che avviene tramite un apposito modulo dell'applicativo Conair.

La procedura di controllo di qualità dei dati restituisce un flag (FLG_VALIDITA) per ogni record con valore N (nullo), S (sospetto), V (valido). Il valore così ottenuto viene sottoposto ad un altro step di controllo, ad opera del gestore che conferma o smentisce il dato, popolando un altro flag (FLG_GESTORE), che permette al dato di essere utilizzato per il calcolo degli indicatori.

I dati N sono quelli che presentano dei valori palesemente errati superiori ad una soglia prefissata; per il PM10 tale soglia è pari a 600 $\mu g/m^3$.

Nel grafico seguente (Fig.2) riportiamo l'esempio di un valore superiore alla soglia di 600 $\mu g/m^3$, registrato a Roma nella stazione di Piazza E.Fermi (codice stazione BRACE 1205813), il 16/8/2001 alle ore 14:00. Dal grafico di Fig.2 si tratta di un valore avulso dall'andamento della serie il che fa pensare ad un errore nella rilevazione piuttosto che a un picco verificatosi realmente.

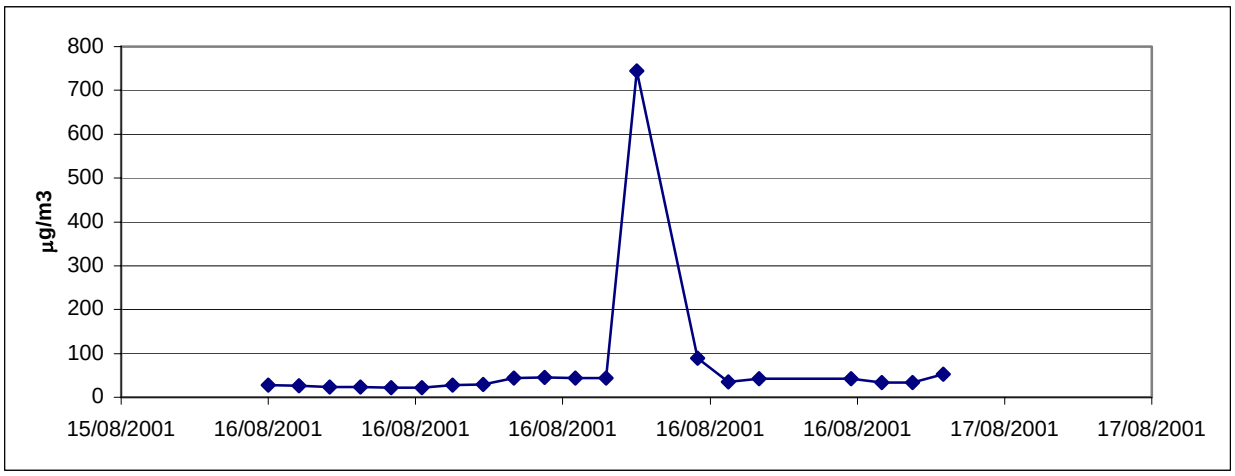

Fig.2 Dati di concentrazione oraria di PM_{10} nella stazione P.zza Fermi di Roma, registrati il giorno 16/08/2001

Alcuni dati che non superano la soglia vengono comunque classificati come sospetti. Si tratta di dati che presentano anomalie e che vanno verificati dal gestore prima di essere inseriti nel dataset dal quale verranno ricavati gli indicatori o qualunque tipo di informazione.

Riportiamo di seguito un esempio di dato classificato come sospetto, rilevato a Palermo nella stazione Castelnuovo (codice stazione BRACE 1908208), il giorno 22/2/2001 . Benché il valore più alto sia inferiore alla soglia di 600 $\mu g/m^3$, nel contesto della serie, esso appare anomalo. Al dato in questione è stato dunque assegnato il flag_gestore NULL, motivo per cui il valore non è stato considerato nell'aggregazione dei dati da valori orari a valori medi giornalieri e nelle elaborazioni statistiche successive.

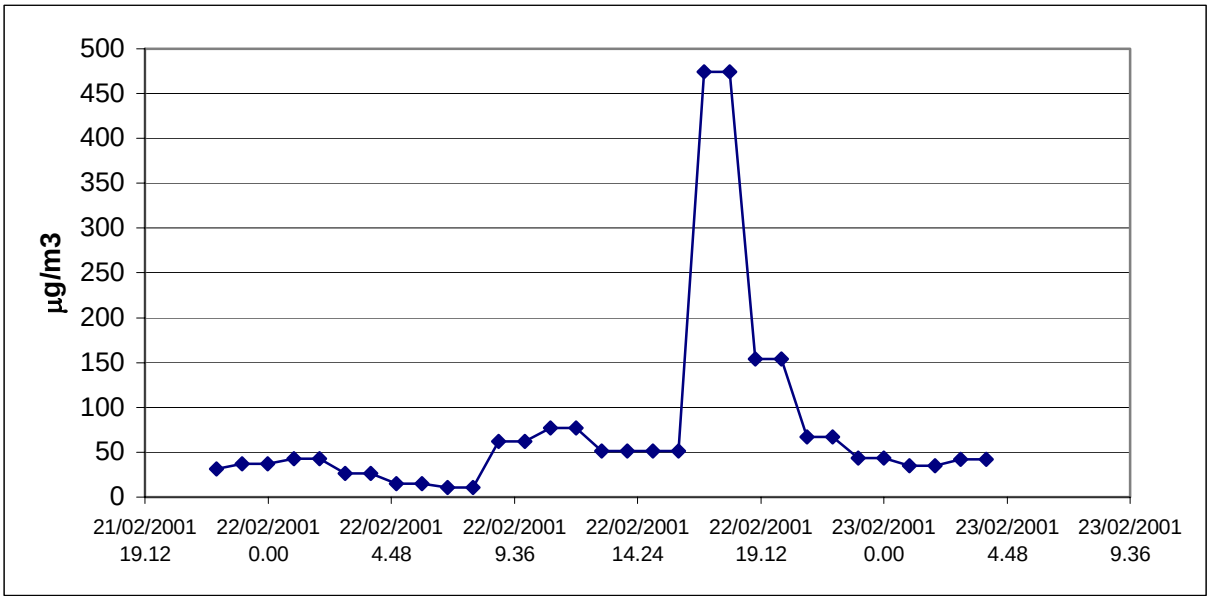

Fig.3 Dati di concentrazione oraria di PM_{10} nella stazione Castelnuovo di Palermo, registrati il giorno 22/02/2001

Il metodo statistico considerato ha come obiettivo quello di mantenere la coerenza nel tempo dei momenti della distribuzione di frequenza statistica; la qualità dell'aria descritta da una stazione di monitoraggio è infatti completamente descritta dalla distribuzione statistica dei dati osservati sull'intervallo di tempo di riferimento. Un'ipotesi iniziale è che la distribuzione di frequenza dei dati sia pressoché la stessa in periodi diversi, purché sufficientemente lunghi; su tale ipotesi si basa parte del metodo di "segnalazione" applicato alle serie. L'identificazione di dati anomali è una procedura molto delicata perché rischia di rilevare come anomali valori che sono semplicemente poco probabili. L'obiettivo è identificare valori che sono errati con una certa probabilità (alta) e non eliminare i valori estremi validi.

I passi dell'algoritmo atto alla segnalazione delle anomalie, inseriti nella procedura sono i seguenti:

A) Individuazione della media e della deviazione standard dell'insieme dei dati
B) Individuazione degli outlier dei dati.

I dati di concentraizione degli inquinanti seguono molto spesso la distribuzione di tipo gamma[2] e la distribuzione delle famiglie lognormali[3] o comunque asimmetriche, per cui il valore della media e della deviazione standard non sono sufficienti ad identificare delle possibili anomalie dei dati. La variabile casuale gamma in particolare viene utilizzata nei problemi di teoria delle file d'attesa.
Alla base della procedura di determinazione dei valori sospetti c'è il concetto di "outlier".

Outlier: Osservazione/i che sono generate da meccanismi diversi rispetto a quelli che generano la distribuzione di riferimento. In esso i valori sono diversi ed occorre indagare la loro diversità, forniscono informazioni sul fenomeno da cui sono differenziati e su fenomeni non considerati. Gli elementi outlier indicano che la struttura dei dati contiene più strutture al suo interno che occorre separare: un esempio può essere fornito da un campione di dati composto da due sotto-insiemi relativi al mattino e sera che vanno indagati come sotto-campioni separati. L'aspetto interessante è dato dal valore di separazione del mondo degli outlier rispetto ai dati della distribuzione di riferimento.

4.1 Metodi per la determinazione degli outlier

Tradizionalmente i valori al di fuori di una distribuzione derivano da test statistici basati sul rapporto fra le varianze dei dati dove progressivamente vengono eliminati elementi della serie ordinata di valori.

Di seguito si descrive un algoritmo che spiega come si arriva all'eliminazione degli outlier:
1) Si calcola la varianza

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

[2] La funzione di densità di probabilità della v.c. gamma è :

$$\gamma(x) = \frac{L(b)}{a_b e_{-ax} x_{b-1}} (0 < x < \infty, a > 0, b > 0)$$

[3] La funzione di densità della v.c. lognormale è

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{[\ln(x) - \mu]^2}{2\sigma^2}\right)$$, dove $0 < x < \infty$, $\sigma > 0$

2) Si elimina il max di x_i e si ricalcala la varianza su n-1 dati

3) Si calcola il rapporto K fra la varianza n-1 e la varianza n

4) Se $K \leq 0.7$ vai a 2 altrimenti esci e memorizza il valore massimo come limite di tutti gli outlier.

Fissare una soglia per il rapporto tra le varianze porta a determinare gli outlier in funzione di quanto diminuisce la varianza quando questi valori vengono esclusi dalla serie di dati. Si avrà una serie di dati che vengono eliminati finché non c'è una diminuzione significativa della varianza. A questo punto si giunge alla determinazione dell'outlier.
Per identificare il valore soglia si procede al calcolo del seguente valore:

VS=1.2*media(A,B)

Dove A è il valore al di fuori del 99.97% dei dati considerati della distribuzione normale (pari a 3.49 σ) e B è il valore dell'outlier ottenuto con il metodo precedente.

Durante l'esecuzione delle procedure dell'interfaccia gestionale per il calcolo dei superamenti e delle statistiche, si possono comunque incontrare di volta in volta situazioni ambigue o apparentemente anomale da approfondire andando a monitorare le serie temporali dei dati grezzi dai quali sono stati calcolati.

Dopo aver effettuato l'analisi per apporre il flag_gestore ai dati sospetti, il calcolo delle statistiche è condizionato dalla presenza di un certo numero di dati validi (per la media e la mediana il 50% dei dati validi mentre per il 95° percentile, il 99,5° percentile ed il valore massimo, il 75%). Successivamente al controllo dei dati, vengono lanciate le procedure di aggregazione e di calcolo degli indicatori così come previsto dalla normativa.

4.2 PM_{10}: Algoritmi di calcolo degli indicatori previsti dalla normativa

Riferimento normativo: D.M. 60 del 02/04/0)

Dato richiesto: media su 24h (protezione salute umana)

Arrotonda a *0* decimali.

SE: f dati grezzi = 1 ora

Aggrega da **1 ora** ad **1 giorno**

ALTRIMENTI

SE: f dati grezzi = 2 ore

Aggrega da **2 ore** ad **1 giorno**

ALTRIMENTI

SE: f dati grezzi = 3 ore

Aggrega da **3 ore** ad **1 giorno**

Arrotonda a *0* decimali.

Calcola superamenti **50 μg/m3**.

Verifica limite superamenti **35 volte**.

Dato richiesto: media su 1 anno (protezione salute umana)

Verifica numerosità dati >= del **50 % su anno.**

Verifica rapporto numerosità dati estate/inverno e inverno/estate > di **2**.

Arrotonda a *0* decimali.

SE: f dati grezzi = 1 ora

Aggrega da **1 ora** ad **1 giorno**

ALTRIMENTI

SE: f dati grezzi = 2 ore

Aggrega da **2 ore** ad **1 giorno**

ALTRIMENTI

SE: f dati grezzi = 3 ore

Aggrega da **3 ore** ad **1 giorno**

Arrotonda a *0* decimali.

Verifica numerosità dati >= del **50 % su anno.**

Verifica rapporto numerosità dati estate/inverno e inverno/estate > di **2**.

Calcola valore medio su **anno**

Arrotonda a *0* decimali.

Calcola superamenti **40 μg/m3**.

Dato richiesto: media su 24 h - Soglie di valutazione superiore e inferiore

Dai dati giornalieri ricavati:

Calcola superamenti **30 μg/m3**. *[soglia superiore]*

Verifica limite superamenti **7** volte.

Calcola superamenti **20 µg/m3**. *[soglia inferiore]*

Verifica limite superamenti **7** volte.

Dato richiesto: media su anno - Soglie di valutazione superiore e inferiore

Dai dati annuali ricavati:

SINAnet Riferimento: SISTEMA BRACE.doc

SISTEMA BRACE 10-03-2005 Page 9 of 13

Calcola superamenti **14 µg/m3**. *[soglia superiore]*

Calcola superamenti **10 µg/m3**. *[soglia inferiore]*

Riferimento normativo: EoI - Decisioni 1997/101/CE e2001/752/CE

Dato base: valori giornalieri delle concentrazioni.

Aggregare i dati se il dettaglio di misura è inferiore ad un giorno.

Arrotonda a *0* decimali.

Dato richiesto: media e mediana (50° percentile) .

Verifica numerosità dati >= del **50 % su anno.**

Verifica rapporto numerosità dati estate/inverno e inverno/estate > di **2**.

Calcola valore medio su **anno**.

Arrotonda a *0* decimali.

Calcola **50° percentile**.

Dato richiesto: 98° percentile, 99.9° percentile, massimo

Verifica numerosità dati >= del **75 % su anno.**

Verifica rapporto numerosità dati estate/inverno e inverno/estate > di **2**.

Calcola **98° percentile**.

Calcola **99,9° percentile**.

Trova **massimo**.

5. Procedure alternative al controllo di qualità dei dati

Vista la disponibilità di numerosi dati abbiamo deciso di procedere alla ricerca di una distribuzione di probabilità che descrivesse la concentrazione di PM_{10} e alla successiva stima dei parametri che la caratterizzano.

L'utilità della distribuzione di probabilità risiede nella possibilità di trovare una soglia per la determinazione dei valori Suspect senza dover svolgere tutti i passi dell'algoritmo riportato nel paragrafo precedente.

In pratica avendo una distribuzione di riferimento è possibile fissare una probabilità e determinare l'intervallo di valori che assume la variabile con quella probabilità. Trovato l'intervallo si fissano le soglie corrispondenti agli estremi dell'intervallo trovato.

In Fig.4 si riporta la distribuzione di frequenza dei dati di concentrazione di PM10 relativi all'anno 2001.

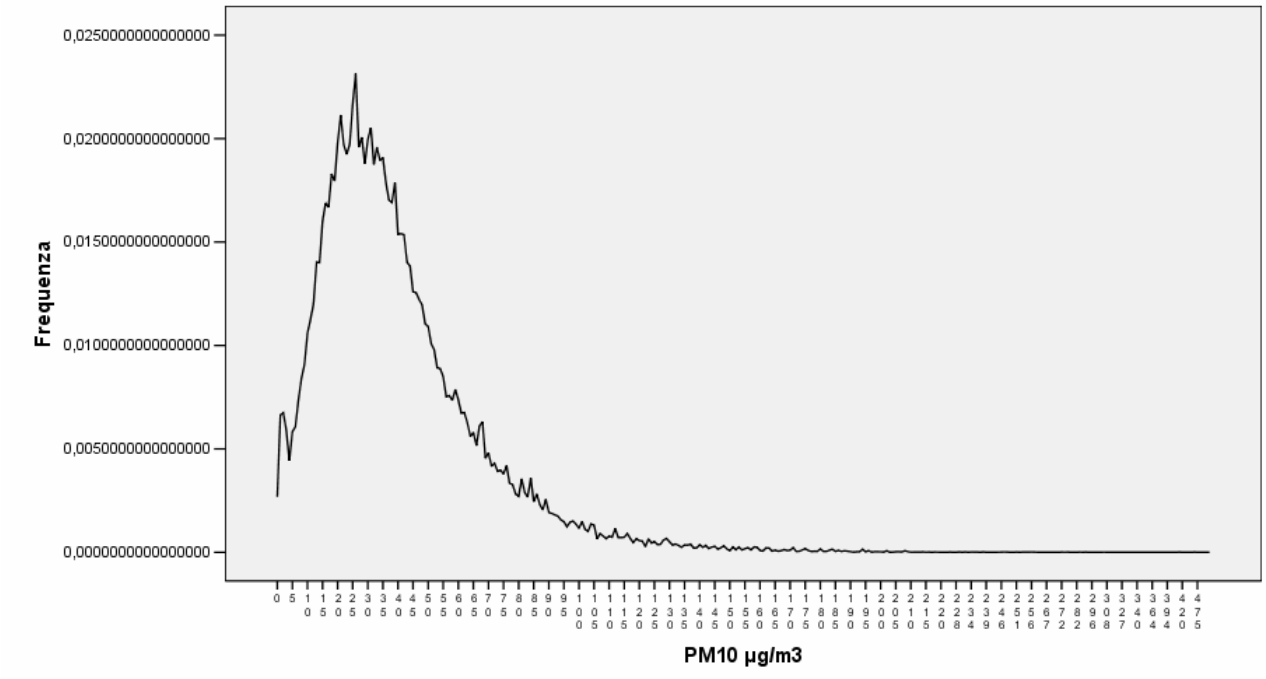

Fig.4 Anno 2001: distribuzione di frequenza dei dati di concentrazione di PM_{10}

Già nel paragrafo precedente si era accennato al fatto che i dati di concentrazione degli inquinanti seguono andamenti asimmetrici e dalla distribuzione di frequenza, ricavata da questi dati, non solo troviamo una conferma a quanto detto, ma osserviamo che, per la precisione, tra tutte le distribuzioni asimmetriche quella che meglio li rappresenta è una funzione di distribuzione Gamma.

A questo punto si procede alla stima dei due parametri che definiscono la distribuzione con il metodo dei momenti.

Con il metodo dei momenti otteniamo i seguenti parametri:

$\alpha = 2.2$

$\lambda = 0.07$

Partendo dalle stime dei parametri ottenuti con il metodo dei momenti, si è effettuata la ricerca di nuovi parametri con una regressione dei dati su una Gamma. L'algoritmo iterativo inizia con le stime ottenute con il metodo dei momenti.

Così facendo otteniamo due nuovi valori:

$\alpha = 2.745$

$\lambda = 0.07$

L'indice di bontà dell'accostamento è: $R^2 = 0.985$.

Di seguito si riportano alcune tabelle riepilogative dei risultati (Tabb.5.1-5.3).

Tab.5.1 Iteration History(b)

Iteration Number(a)	Residual Sum of Squares	Parameter	
		S	SC
1.0	,001	2,200	,070
1.1	,000	2,326	,056
2.0	,000	2,326	,056
2.1	,000	2,645	,067
3.0	,000	2,645	,067
3.1	,000	2,728	,070
4.0	,000	2,728	,070
4.1	,000	2,743	,070
5.0	,000	2,743	,070
5.1	,000	2,745	,070
6.0	,000	2,745	,070
6.1	,000	2,745	,070
7.0	,000	2,745	,070
7.1	,000	2,745	,070
8.0	,000	2,745	,070
8.1	,000	2,745	,070

Derivatives are calculated numerically.
a Major iteration number is displayed to the left of the decimal, and minor iteration number is to the right of the decimal.
b Run stopped after 16 model evaluations and 8 derivative evaluations because the relative reduction between successive residual sums of squares is at most SSCON = 1,00E-008.

Tab.5.2 Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
S	2,745	,033	2,681	2,810
SC	,070	,001	,068	,072

Tab. 5.3 ANOVA(a)

Source	Sum of Squares	df	Mean Squares
Regression	,013	2	,007
Residual	,000	307	,000
Uncorrected Total	,014	309	
Corrected Total	,010	308	

Dependent variable: V2
a R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = ,985.

Ritorniamo ora al problema di determinare una soglia per identificare i valori sospetti e suggeriamo una possibile procedura:

- Definisco l'evento E nel seguente modo:

E = "*Il valore della concentrazione x di PM_{10} è compreso in un intervallo $a \leq x \leq b$*"

- Fisso una probabilità P(E) = 0,9866

- Pongo $P(x \geq b) = 0,9999$

- Pongo $P(x \leq a) = 0,0133$

- Determino **a** e **b**

- Definisco come suspect tutte le concentrazioni x tali che x<a, x>b.

Per ricavare i valori di **a** e **b** abbiamo utilizzato il software MATLAB.

La figura seguente rappresenta la finestra del software utilizzato, con il quale è possibile modificare i parametri della funzione di ripartizione della distribuzione Gamma.
In questo caso stabiliamo il valore della probabilità e i parametri A (parametro di forma) e B (inverso del parametro di scala); in questo modo otteniamo nella casella X il valore 191.5621 $\mu g/m^3$, che corrisponde all'estremo superiore **b** dell'intervallo cercato. In particolare abbiamo che $P(x \geq 191.5651)=0,9999$.

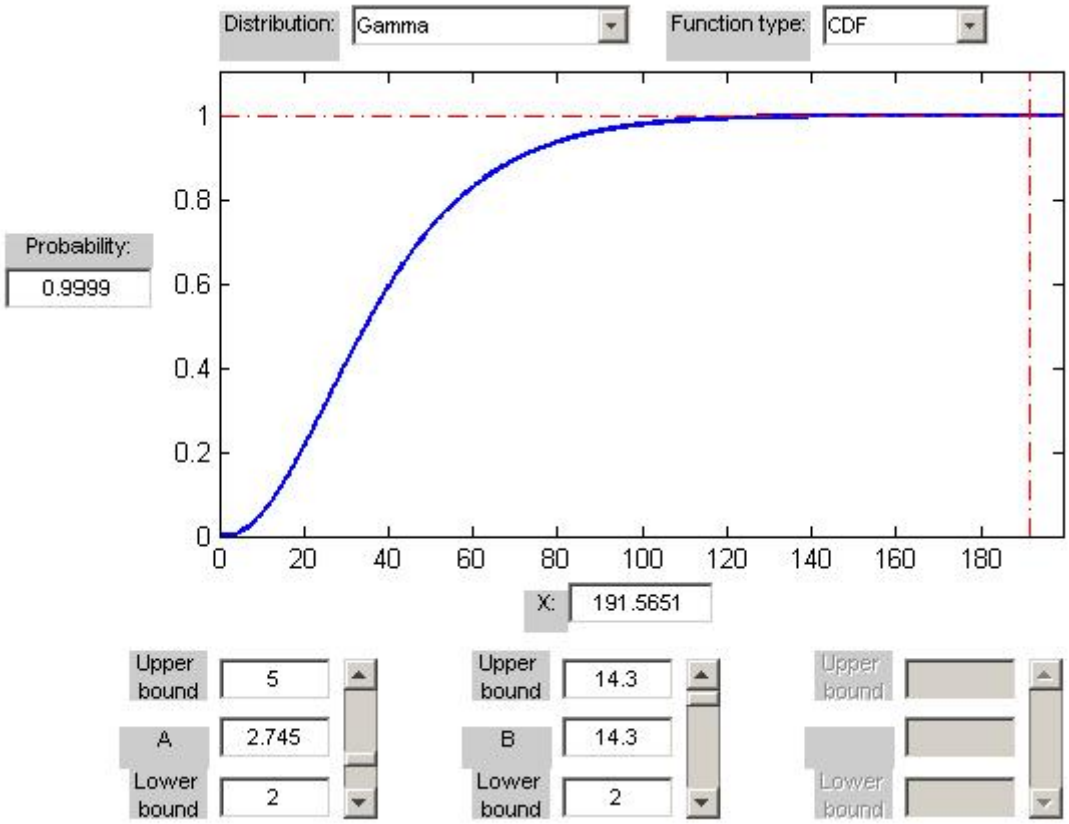

Fig. 5 Determinazione dell'estremo superiore b dell'intervallo cercato

Allo stesso modo procediamo per la determinazione dell'estremo inferiore **a** tale che $P(x \leq a)=0{,}0133$.

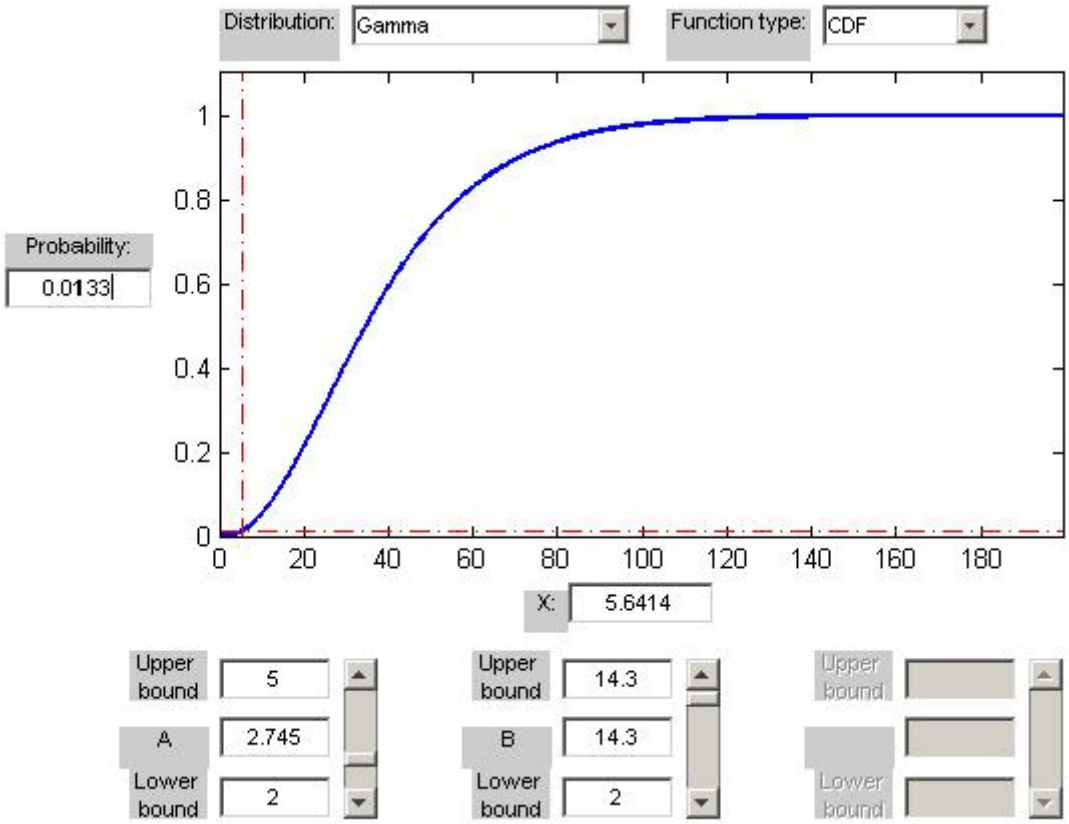

Fig. 6 Determinazione dell'estremo inferiore a dell'intervallo cercato

L'intervallo ottenuto è quindi il seguente:

$$5.6\ \mu g/m^3 \leq x \leq 191.6\ \mu g/m^3$$

5.1 Analisi dei risultati

Con la procedura implementata attualmente nel database vengono segnalati i valori "sospettosamente alti" mentre vengono ignorati i valori "sospettosamente bassi". Infatti, si ottengono, per i dati di concentrazione di PM_{10} del 2001, circa 150 valori sospetti mentre confrontando con la procedura che utilizza la funzione di distribuzione Gamma, i valori sospetti ottenuti sono 15203, di cui 527 sono maggiori di 192 e i restanti sono minori o uguali a 6; di questi, 1089 sono uguali a 0 e 2546 a1.

5.2 La distribuzione Gamma e il PM_{10}

La distribuzione Gamma è una distribuzione asimmetrica. I parametri che la descrivono sono 2 e non corrispondono come in altre distribuzioni alla media e alla varianza. Si tratta infatti di un parametro di forma α e un parametro di scala λ.

La densità della distribuzione gamma è la seguente

$$f(x) = \frac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$$

Alla luce di quanto detto vanno fatte alcune considerazioni sul comportamento del PM_{10}, in particolare riguardo alle conclusioni a cui si perviene quando si registra un cambiamento negli indicatori più comuni, quali media e varianza.
Nella distribuzione Gamma la media e la varianza sono così definite:

-Media $\mu = \frac{\alpha}{\lambda}$

-Varianza $\sigma^2 = \frac{\alpha}{\lambda^2}$

Al contrario di quanto succede per la distribuzione Normale, nella distribuzione Gamma, media e varianza non sono i parametri della distribuzione ma sono funzione dei parametri.
Quindi ai fini di una corretta valutazione del comportamento del PM_{10} va osservato che le conseguenze di una diminuzione della media di concentrazione del PM_{10} possono essere determinate da un cambiamento del parametro di forma o di scala.
Ad esempio, allo scopo di descrivere la metodologia utilizzata, abbiamo studiato la concentrazione di PM_{10} per il 2001 e il 2006 (di cui riportiamo sotto le relative distribuzioni) e abbiamo ottenuto i seguenti risultati:

Per l'anno 2001, la funzione Gamma è definita dai seguenti parametri:
α =2.745
λ =0.07
R^2 =0.985,

con una media del campione pari a 39.68 e una media teorica $\mu = \frac{\alpha}{\lambda}$ =38.57.

Per l'anno 2006 la funzione Gamma è definita dai seguenti parametri:

α=2.324

λ=0.072

R^2=0.99,

con una media dal campione pari a 34.9 una media teorica $\mu = \frac{\alpha}{\lambda}$=32.3.

E' importante osservare che la bontà dell'accostamento è migliore nel 2006 rispetto al 2001, grazie alla maggiore disponibilità di dati per l'anno 2006.

Nelle figure seguenti (Fig.7) si riportano le funzioni di distribuzione rispettivamente negli anni 2001 e 2006.

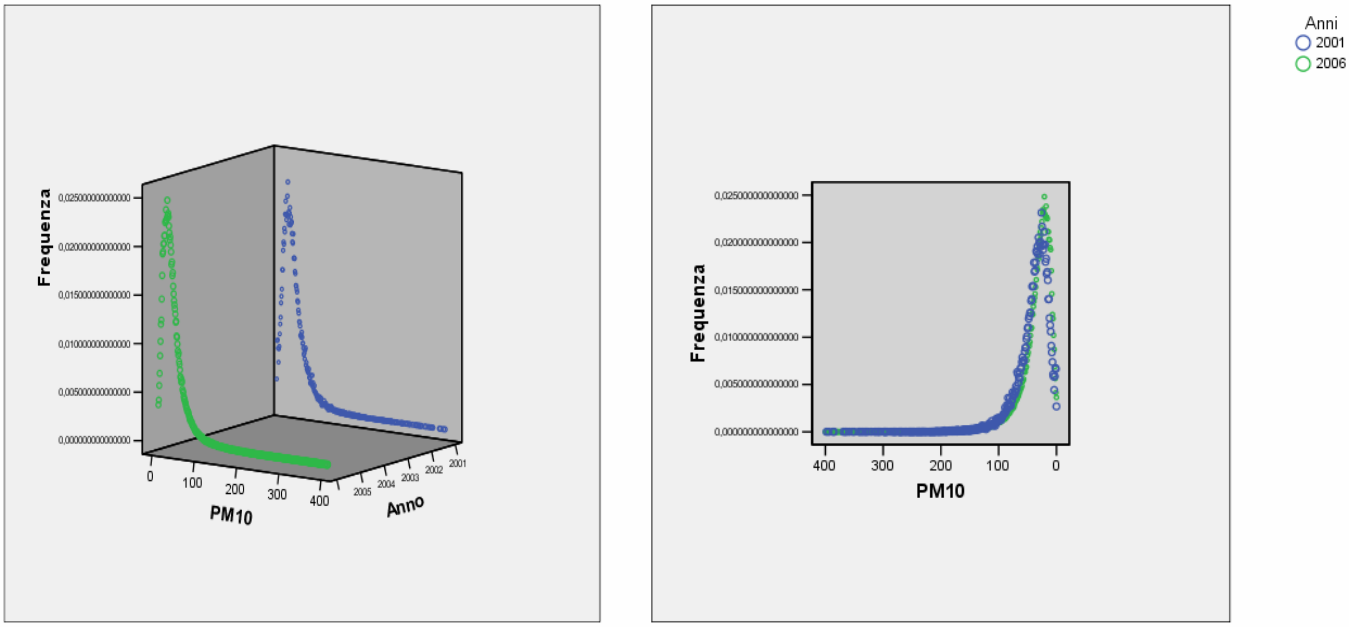

Fig. 7 Andamento della funzione di distribuzione dei dati di PM10 per gli anni 2001 e 2006

Applicando una rotazione al grafico si evince che il cambiamento della media dal 2001 al 2006, non sposta la distribuzione come accade con la funzione Normale ma è il cambiamento del parametro di forma α a schiacciare leggermente la funzione.

Tutto questo ovviamente ha conseguenze sui superamenti del valore medio giornaliero di 50 $\mu g/m^3$.

Calcoliamo ora la probabilità di superare il valore di 50 $\mu g/m^3$ negli anni 2001 e 2006.

Per l'anno 2001 si ha:

$\alpha = 2.745$

$\lambda = 0.07$

$R^2 = 0.985$

In base a questi valori otteniamo la funzione di ripartizione riportata in Fig.8.

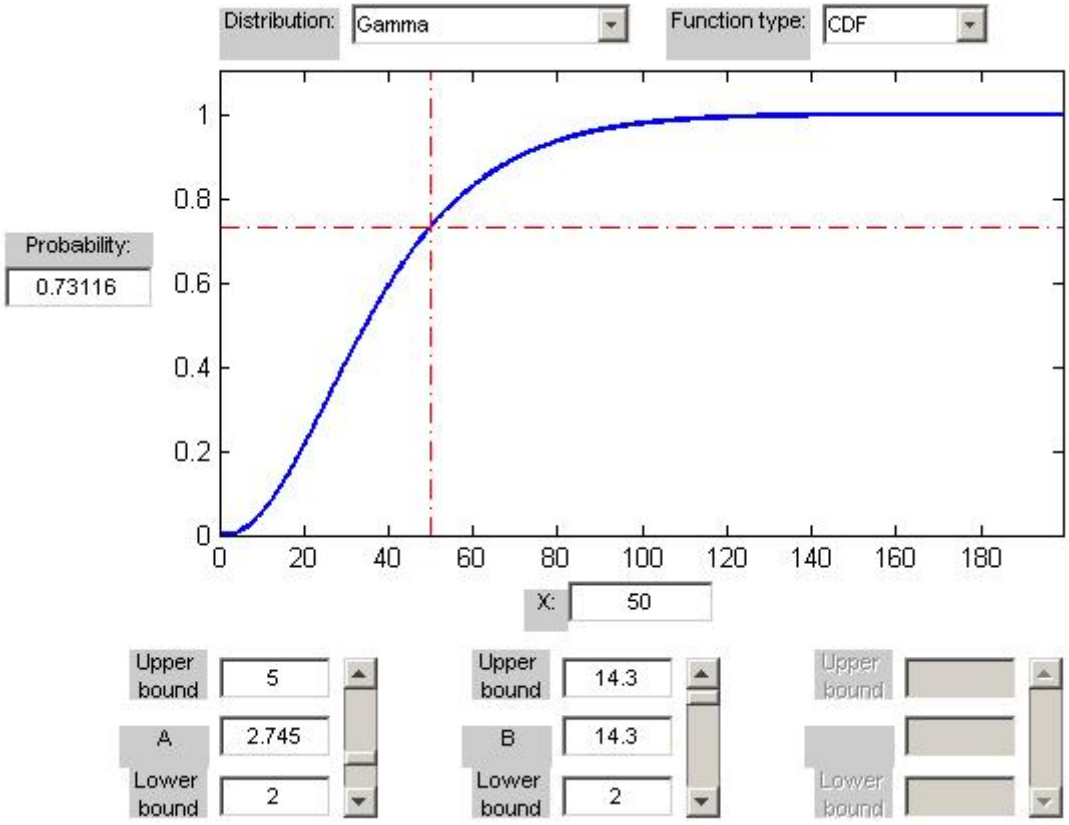

Fig.8 Funzione di ripartizione per l'anno 2001

Il valore B è dato dal rapporto $1/\lambda$ cioè in questo caso 1/0.055=18.2.

In questo caso la probabilità di superare la soglia dei 50 $\mu g/m^3$ è pari a:

$P(X>50\ \mu g/m^3)=1-0.73116 = 0.26884$

Per l'anno 2006 si ha:

α=2.324

λ=0.072

R^2=0.99

In base a questi valori otteniamo la funzione di ripartizione riportata in Fig.9.

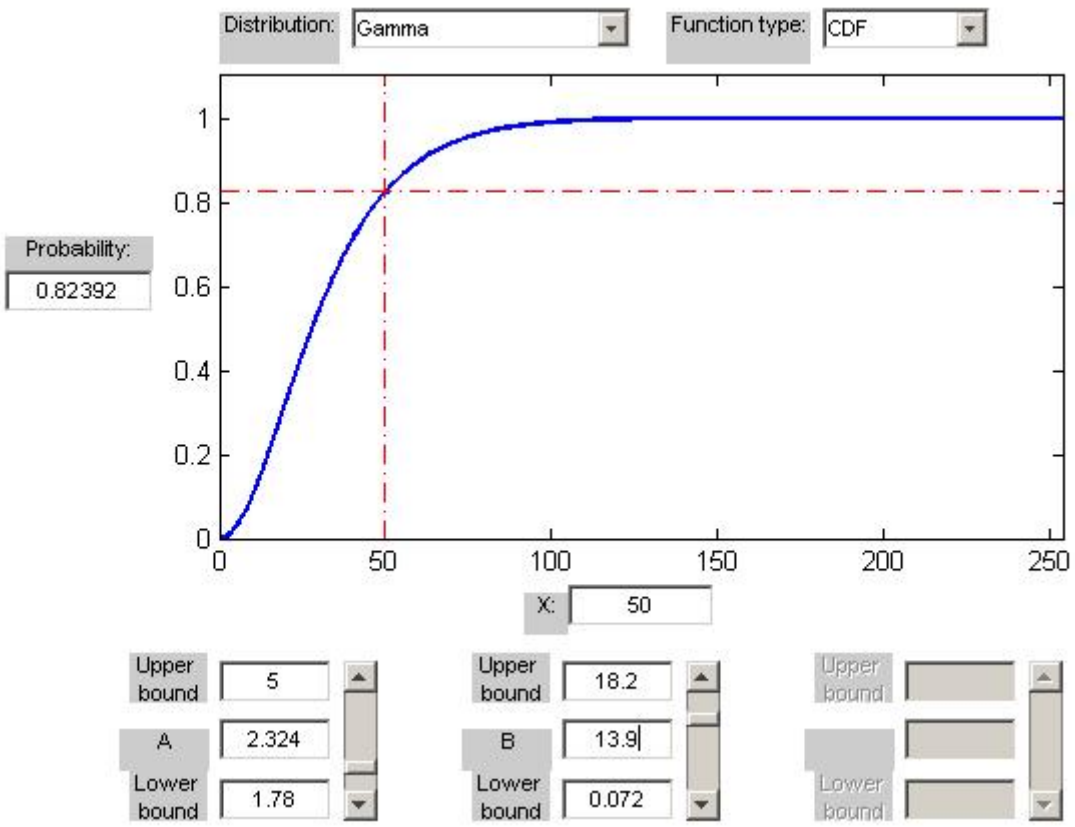

Fig.9 Funzione di ripartizione per l'anno 2006

Osserviamo che la probabilità di superare il valore limite giornaliero di PM_{10} per l'anno 2006, diminuisce:

P(X>50 $\mu g/m^3$)=1-0.82392=0.17608

Ipotizziamo adesso che il cambiamento della media non sia dovuto al parametro di scala λ ma a quello di forma α, ovvero si sia passati da una media del campione pari a 39.68 e una media teorica $\mu = \frac{\alpha}{\lambda}$=38.57 a una media del campione pari a 34.9 e una media teorica $\mu = \frac{\alpha}{\lambda}$=32.3 e che questo cambiamento sia dovuto a un cambiamento del parametro di forma.

Dall'analisi delle funzioni di distribuzione del 2001 e del 2006, fissiamo λ=0.055.

Risolvendo una banale equazione abbiamo che $\mu = \frac{\alpha}{\lambda} = 32.3 = \frac{\alpha}{0.055}$ il che implica che α=32.3*0.055=1.78.

Nell'ipotesi che il cambiamento sia dovuto a una modifica del parametro di forma con il parametro di scala costante si ha:

α=1.78

λ=0.055,

la media teorica rimane $\mu = \frac{\alpha}{\lambda}$=32.3.

In Fig.10 si riporta l'andamento della funzione di ripartizione con il cambiamento dovuto al parametro di forma.

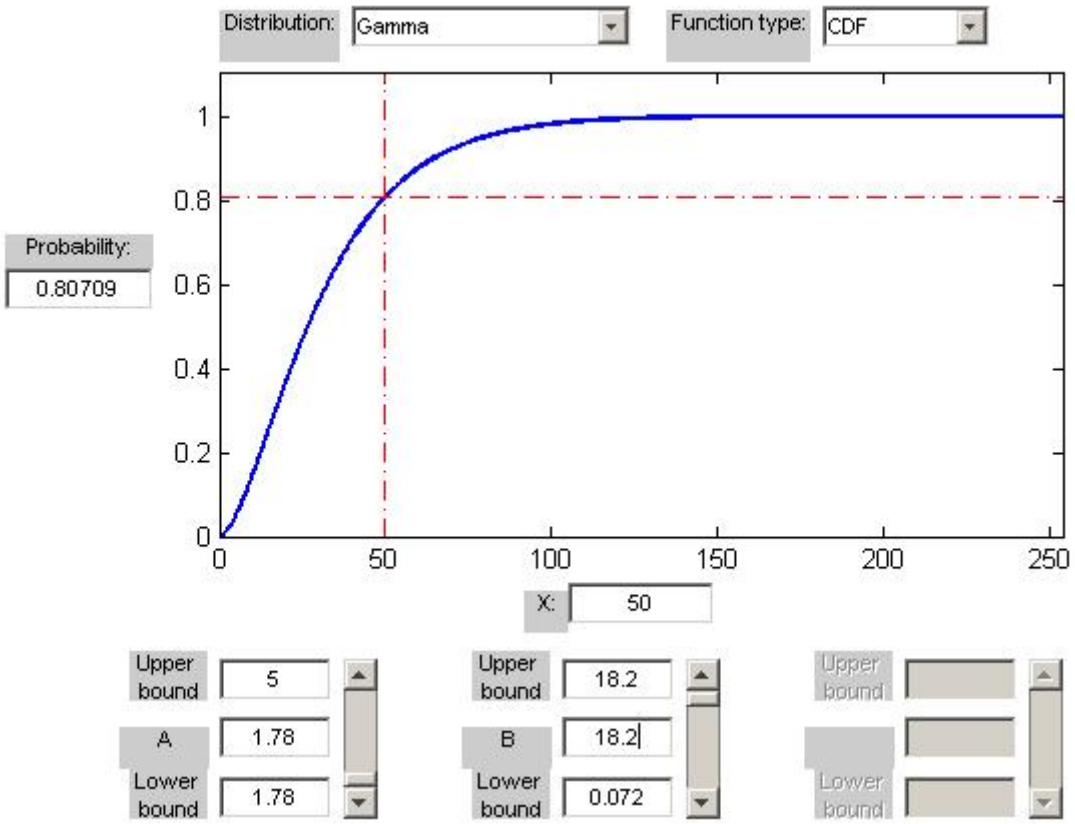

Fig.10 Andamento della funzione di ripartizione con cambiamento del parametro di forma, per l'anno 2006

La probabilità calcolata in questa seconda ipotesi è data da:

P(X>50 $\mu g/m^3$)=1-0.80709=0.19291

ed è diversa da quella calcolata per il 2006 P(X>50 $\mu g/m^3$)=1-0.82392=0.17608.

In figura 11 si mostra come il cambiamento del parametro di forma, possa influenzare la funzione densità di probabilità per i dati di concentrazione di PM_{10}.

Tale cambiamento potrebbe essere caratteristico di stazioni appartenenti a tipologie differenti (*traffico*, *fondo*, *industriale*) o a differenti zone (*urbana*, *suburbana*, *rurale*); ragion per cui, la funzione di distribuzione analizzata potrebbe essere uno strumento utile per studi futuri sulla caratterizzazione delle stazioni di monitoraggio di PM_{10}.

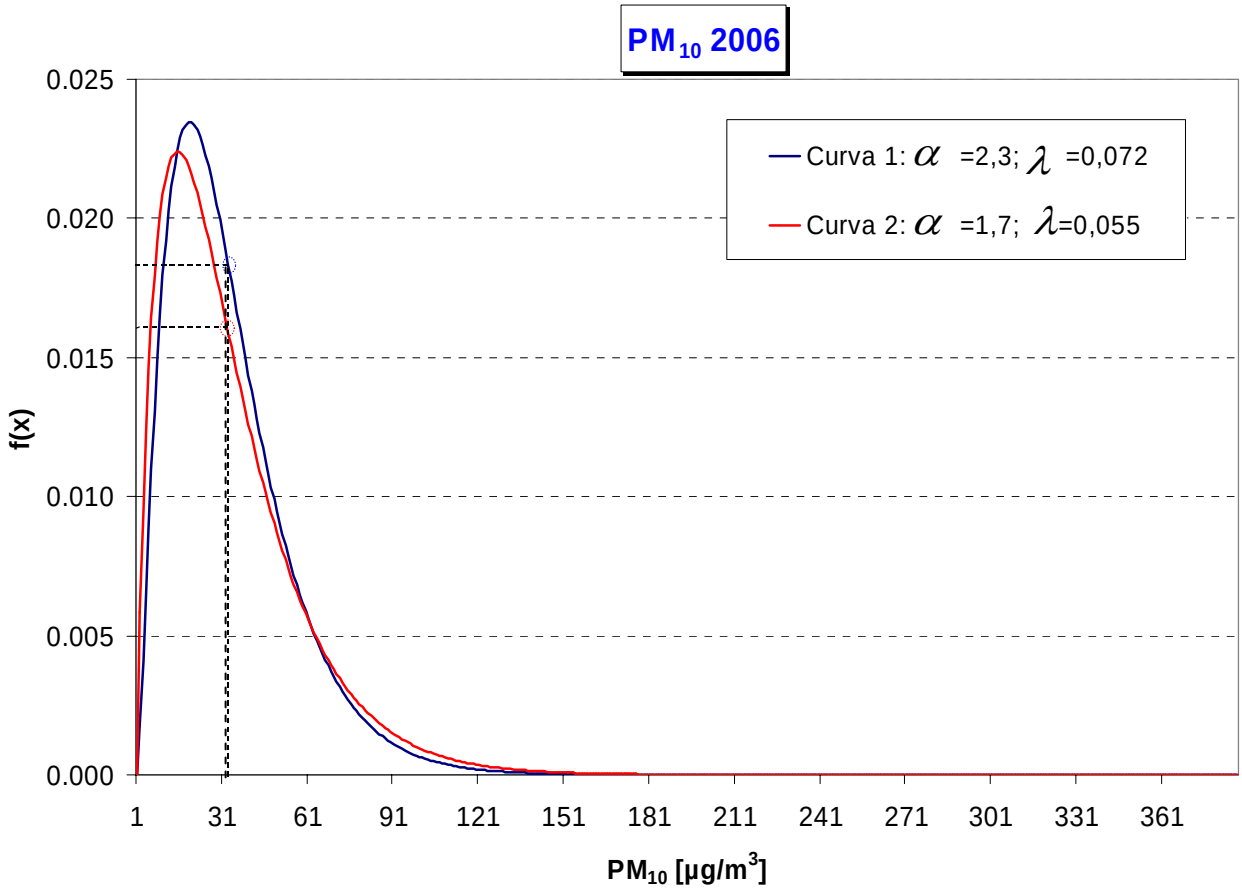

Fig.11 Confronto fra le funzioni di distribuzione del PM10 nell'anno 2006

6. Conclusioni

L'analisi di tipo statistico dei dati di concentrazione di PM_{10} con particolare riferimento all'anno 2001, consente di trarre le seguenti conclusioni:

- La funzione di distribuzione *Gamma* caratterizza meglio l'andamento delle concentrazioni di PM_{10} rispetto alla funzione di distribuzione *Normale*;
- La procedura suggerita, con l'utilizzo della funzione di distribuzione *Gamma*, per la determinazione dei valori *suspect*, consente di individuare un numero maggiore di dati anomali;
- La funzione di distribuzione *Gamma*, attraverso i suoi parametri caratteristici α e λ, potrebbe essere un utile strumento per caratterizzare stazioni appartenenti a tipologie differenti (*traffico*, *fondo*, *industriale*) o a differenti zone (*urbana*, *suburbana*, *rurale*), nonché per realizzare un confronto fra gli andamenti rilevati in stazioni dello tipo.