Università degli Studi di Padova

FACOLTÀ DI SCIENZE STATISTICHE
Corso di Laurea Specialistica in Statistica e Informatica

Tesi di Laurea

Effetto dell'ozono sulla salute umana: un approccio basato sull'utilizzo delle concentrazioni orarie

Relatore:
Prof.ssa Monica Chiogna

Laureando:
Filippo Da Re

Anno Accademico 2008–2009

Indice

Introduzione **1**

1 Base di partenza **5**

1.1 Misura dell'esposizione 5

1.2 Effetto dell'ozono 9

2 I dati: un'analisi preliminare **13**

3 Preparazione dei Modelli **29**

3.1 Espansione dei dati giornalieri 29

3.2 Formulazione dei Modelli Additivi 32

3.2.1 Costruzione 32

3.2.2 La scelta dei GAM 33

3.3 Core Models 35

3.3.1 Ricalibrazione 40

4 Inserimento dell'Ozono **47**

4.1 I Modelli finali 48

4.2 Risultati sull'Ozono 55

5 Conclusioni **61**

A L'utilizzo di R **65**

Elenco delle figure

1.1 Esempio della distribuzione di Ozono nell'arco di una giornata . . . 8

2.1 Media giornaliera di Ozono negli anni 1998-2003 nella città di Milano 18
2.2 Valori giornalieri di O_3 per i tre mesi estivi dal '98 al '03 19
2.3 Valori orari di O_3 per i tre mesi estivi dal '98 al '03 20
2.4 Rilevazioni orarie di Ozono relative alla giornata dell'11 Agosto 2003 21
2.5 Distribuzione oraria della quantità di O_3 22
2.6 Numero di ricoveri legati a problemi di tipo respiratorio, registrati a Milano nel periodo indicato 23
2.7 Andamento del PM_{10} e dell' O_3 nell'arco dei sei anni considerati . 25
2.8 Temperatura giornaliera e livello di Ozono 26
2.9 Distribuzione rispetto ai giorni della settimana e di festa di Ozono e del numero di ricoveri. I giorni della settimana sono codificati con 0 = Domenica; 1 = Lunedì; 27
2.10 Numero di ricoveri, livello di Ozono e Temperatura divisi nei sei anni dello studio 28

3.1 Concentrazione di Ozono 30

4.1 Grafici delle spline di lisciamento nel primo modello. In senso orario da in alto a sinistra: s(Tempo), s(Temperatura) e s(Temperat.-rit.). 52
4.2 Analisi dei residui di devianza per i due modelli: grafico Quantile-Quantile e Istogramma delle frequenze 54
4.3 Grafico della spline di lisciamento di O_3 56
4.4 Valori lisciati dell'ozono suddivisi per ora 57

4.5 Effetto di Ozono e Temperatura-rit sulle previsioni del numero di ricoveri 59

Elenco delle tabelle

2.1 Valori di soglia per l'Ozono 17

3.1 Valori medi, divisi per ora, di ozono 30
3.2 Coefficienti parametrici, Core Model 1 37
3.3 Significatività approssimata dei termini lisciati, Core Model 1 37
3.4 Indici, Core Model 1 38
3.5 Coefficienti parametrici, Core Model 2 39
3.6 Significatività approssimata dei termini lisciati, Core Model 2 39
3.7 Indici, Core Model 2 39
3.8 Stime dei coefficienti $\beta_{\cdot}$ nei due Core Model 40
3.9 Indici relativi ai due Core Model 41
3.10 EDF per le variabili usate non parametricamente 42
3.11 Indicatori per le variabili utilizzate parametricamente 43
3.12 Calcolo delle costanti t_i di ricalibrazione 44

4.1 Indici relativi ai due Modelli definitivi 49
4.2 Fattori relativi al Modello "Giornaliero" 50
4.3 Fattori relativi al Modello "Orario" 51

Introduzione

Questo studio analizza la problematica relativa all'analisi della relazione tra inquinamento dell'aria e salute dell'uomo. In particolare, la tesi si concentra sull'ozono, per trovare un metodo significativo con cui trattarlo all'interno di uno studio sugli effetti causati dall'O_3 troposferico sulle condizioni di salute dell'uomo, rappresentata dal numero di ricoveri registrati dovuti a problemi alle vie respiratorie.
Diversi studi epidemiologici hanno già dimostrato l'effetto nocivo che le particelle di ozono nell'aria, respirate dall'uomo, hanno sul sistema respiratorio. Dunque lo scopo di questo lavoro è cercare di determinare un buon metodo di rappresentazione della concentrazione di O_3 in un modello che mette in relazione salute ed inquinamento. La motivazione di base che spinge a ricercare modi diversi di rappresentare la concentrazione dell'inquinante nei modelli è dovuta all'inefficacia di indici giornalieri quali la media, il massimo o la mediana.
Già diversi studi hanno dimostrato che modificando l'informazione relativa all'inquinante si possono ottenere risultati diversi.
Cercheremo quindi di verificare se, utilizzando le singole rilevazioni orarie, si riesca a dare il giusto peso all'ozono. Confronteremo due modelli additivi generalizzati, il primo contenente un indice giornaliero come la media, mentre nel secondo inseriremo tutti i dati orari per cercare di cogliere la grande diversità di valori che vengono misurati nell'arco di una giornata.
A tale scopo ci serviremo di modelli additivi generalizzati (GAM) di Poisson

per modellare il conteggio giornaliero di ricoveri ospedalieri, considerando solo quelli registrati per problemi alle vie respiratorie, in funzione di un insieme di variabili esplicative. Quest'ultime verranno divise in due gruppi: il primo sarà l'insieme di tutte le variabili confondenti, ovvero quelle che utilizzeremo per descrivere l'ambiente in cui lavoriamo ma che non saranno direttamente sotto studio, il secondo comprendente un'unica variabile riguardante l'ozono, oggetto di questa tesi.

Per verificare se un aumento di informazione sull'ozono, all'interno di un modello, comporti migliori risultati in uno studio salute-inquinamento, confronteremo tra di loro due modelli, simili tra loro, con due diverse specificazioni per la concentrazione dell'ozono. La prima sarà la media giornaliera di concentrazione di O_3, mentre la seconda utilizzerà le singole osservazioni orarie.

Lo studio e le analisi sviluppate in questa tesi sono state fatte su dati rilevati nella città di Milano nell'arco di tempo tra il 1998 e il 2003. Utilizzando le fonti Istat, abbiamo ottenuto il numero di ricoveri ospedalieri avvenuti nella città lombarda nel periodo di tempo indicato, dai quali abbiamo selezionato, considerando la classe di età e la tipologia di ricovero, i dati relativi ai ricoveri per problemi respiratori. Tutti i valori inerenti l'ozono e gli altri elementi atmosferici sono stati raccolti da ARPA Lombardia.

La tesi è suddivisibile in quattro parti delineate dai quattro capitoli principali.
Nel Capitolo 1 si introdurrà l'argomento dando uno sguardo generale ad articoli e lavori già pubblicati sull'argomento. In particolare, si sottolineerà la necessità di sopperire alla impossibilità di avere dati individuali sull'esposizione agli inquinanti aerei ripiegando su indici generali. Inoltre, si discuterà l'inefficacia degli indici giornalieri che non colgono l'andamento altalenante della concentrazione di ozono nell'arco di una giornata. Vedremo anche la proposta, per superare questo problema, avanzata da Chiogna e Pauli (2008) che suggeriscono l'uso di tre indici: l'intensità, la durata e l'esposizione notturna, che esprimono meglio la reazione dell'ozono all'attività solare e, di

conseguenza, rappresentano meglio i diversi valori registrabili nelle 24 ore giornaliere.
Nel Capitolo 2 si analizzeranno i dati raccolti per definire con maggiore accuratezza alcuni dettagli, utili in seguito per la formulazione dei modelli. In particolare, si delineeranno i soggetti da considerare, ovvero solo gli anziani di età superiore ai 75 anni, ed il periodo di studio, limitandoci solo ai mesi estivi, periodo in cui l'attività solare, e di conseguenza anche le concentrazioni di ozono, è maggiore. Inoltre, sarà dato un primo sguardo alla variabile ozono ed a tutte le altre considerate nello studio.
Nel Capitolo 3, andremo a preparare gli strumenti con i quali poi confronteremo i due modi di trattare l'ozono: sintesi giornaliera o valori orari. Qui, definiremo i modelli additivi generalizzati e le tecniche con cui tratteremo le singole variabili prese in considerazione, ad eccezione dell'ozono. Difatti formuleremo due modelli GAM di base (Core Models), aventi tutte le variabili esplicative esclusa quella relativa all'ozono. Avendo, infatti, tutte variabili formate da dati giornalieri, per poter stimare un modello con i valori orari di O_3, e poterlo successivamente confrontare con quello con la media giornaliera, dovremo espandere artificialmente tutti i dati; ovvero dovremo replicare 24 volte tutti i valori di tutte le variabili. Lo scopo dei due Core Model, calcolati uno sui dati giornalieri ed il secondo sui dati orari, sarà, appunto, quello di preparare delle basi calibrate su cui poi poter inserire i due indicatori dell'ozono e poterli confrontare.
Nel Capitolo 4 inseriremo l'ozono nei due Core Model, tramite media giornaliera e tramite osservazioni orarie, e vedremo le differenze tra i due metodi. Inoltre analizzeremo il modello calcolato con i dati orari soffermandoci ad osservare l'effetto dell'ozono.
Quello che ci aspettiamo è che aumentando l'informazione riguardante l'ozono, tramite l'utilizzo dei dati orari, si riesca a dar significato all'effetto che sappiamo che l'O_3 ha sulla salute dell'uomo.

In questa tesi tutte le analisi sono state compiute tramite il software statistico open source **R**. In particolare per la stima dei modelli GAM ci si è serviti della libreria **mgcv** sviluppata dal professore Simon N. Wood.

La stesura del testo e la sua gestione grafica sono state condotte tramite linguaggio LaTeX.

Capitolo 1

Base di partenza

Sono molti gli studi già compiuti sul tema della relazione tra problemi respiratori ed ozono, molti dei quali riportano un'associazione positiva tra i due elementi. Risulta ancora difficile, comunque, riuscire a misurare questo effetto, che, epidemiologicamente, è noto, ma che, statisticamente, è difficile da cogliere.

Nella preparazione di questa tesi, ci si è basati principalmente su due articoli - e su quelli ad essi connessi - che trattano questo argomento. Il primo lavoro (Chiogna e Bellini, 2002) introduce tre nuovi metodi di misurazione degli agenti inquinanti e li inserisce nello studio della relazione inquinamento-salute. Il secondo (Chiogna e Pauli, 2008) riguarda l'applicazione pratica dei tre indicatori descritti nel lavoro precedente e viene studiato, più in particolare, l'effetto ozono e la sua influenza sulla salute umana. In tale lavoro vengono utilizzati gli stessi dati impiegati in questa tesi, relativi alla città di Milano.

1.1 Misura dell'esposizione

L'effetto che gli inquinanti dell'aria hanno sui ricoveri ospedalieri o sui decessi dovuti a problemi respiratori è un tema in continuo approfondimento. Il punto di partenza comune degli studi su questo argomento è basato sulla analisi della serie storica dei pazienti considerati, modellata in funzione di

determinate covariate tra le quali quelle relative all'inquinamento; quest'ultime vengono intese come rappresentazione media giornaliera dell'esposizione a cui ogni soggetto è sottoposto.

L'idea base di questi modelli è quella di considerare per J giorni la variabile conteggio esprimente il numero di decessi o numero di ricoveri, $Y_j, j = 1, \ldots, J$, come una variabile casuale di Poisson, $Y_j \sim Poisson(\lambda_j)$, sul cui valore medio λ_j viene definito un modello di regressione di Poisson:

$$\log(\lambda_j) = \underline{\beta}' \underline{x}_j$$

con $\underline{x}_j$ vettore delle variabili esplicative indipendenti e $\underline{\beta}$ vettore dei parametri. La media giornaliera λ_j dipende dalle probabilità individuali θ_{ij} di morire o essere ricoverati in ospedale, con $i = 1, \ldots, n$ e n numero di soggetti considerati. Infatti, per ogni individuo i studiato nel giorno j, l'evento ricovero può essere considerato una realizzazione di una variabile casuale di Bernulli con probabilità θ_{ij}, che dipende da un insieme di covariate considerate, tra cui gli agenti atmosferici e quelli inquinanti.
Risulta evidente che, nella situazione ideale, dovremmo avere dei valori individuali di esposizione agli agenti inquinanti per modellare al meglio λ_j, ma questo non è possibile. La questione principale, dunque, è quella di determinare in quale modo sia meglio rappresentare l'esposizione all'inquinamento e come inserirla all'interno del modello di regressione.

La World Health Organization Guidelines for Air Quality ci fornisce la seguente definizione di esposizione giornaliera: *"L'esposizione totale giornaliera di un individuo all'inquinamento aereo è la somma dei contatti separati con l'inquinante vissuti dall'individuo mentre attraversa diverse condizioni, luoghi, ambienti durante il corso della giornata (luoghi di lavoro, casa, strada, ...). L'esposizione separata in ognuno di questi luoghi è calcolata dal prodotto tra la concentrazione degli inquinanti e il tempo speso nel luogo".*

È chiaro che cercare di rilevare la quantità di inquinamento aereo percepito da ogni individuo è un'operazione dispendiosa sia in termini di soldi che di tempo; inoltre è un'azione complicata da compiere. Quindi, l'utilizzo di rilevatori geografici di inquinamento è un espediente spesso utilizzato in

questi tipi di studi. Da questo ne derivano principalmente due problemi: il primo è che, così facendo, si deve rinunciare a un indice individuale per un indice più generale, che tralascia le differenze di locazione dei soggetti (ad esempio non risulta se l'individuo è in un luogo chiuso o aperto); il secondo è relativo alla necessità di scegliere un indice di sintesi per rappresentare la quantità di inquinamento rilevata, come per esempio la media giornaliera o la concentrazione massima.

A questo proposito, in Chiogna e Bellini (2002) gli autori propongono tre diversi indicatori per rappresentare l'esposizione a cui sono soggetti gli individui, da usare in alternativa agli indici più comuni. Due di essi, **intensità** (*severity*) e **durata** (*duration*), sono relativi al "tempo", mentre il terzo, **esposizione notturna** (*night exposure*), è connesso all'esposizione specifica individuale notturna.

L'idea alla base dei primi due indicatori nasce dalla necessità di sottolineare la differenza tra i giorni con alti picchi di inquinamento da quelli normali e per accentuare gli episodi in cui si sono registrati alti livelli di inquinamento per un lungo periodo di tempo. A tale proposito, in accordo con Abbey e Burchette (1996), gli autori introducono l'idea di soglia, ovvero di un livello fissato sopra il quale l'inquinante è definito avere un'alta concentrazione. Determinata questa soglia, che varia a seconda dell'inquinante, si può definire l'intensità come la differenza tra la massima concentrazione di inquinante registrata durante il giorno e il valore fissato della soglia, e la durata come la durata oraria del superamento della soglia.

La scelta del valore della soglia è lasciata a chi conduce le analisi. Un valore consigliato può essere quello di soglia determinata dagli enti nazionali (soglia d'informazione o d'allarme, differenti per ogni tipo di inquinante dell'aria), ma in alternativa si possono utilizzare la media, la mediana o il valore del terzo quartile calcolati sui dati rilevati.

L'idea alla base della costruzione dell'Intensità, della Durata e dell'Esposizione notturna è data in Figura 1.1. In essa sono rappresentati i valori orari della concentrazione d'ozono rilevati in una giornata tipica di Agosto[1]. Possiamo notare che un andamento così oscillante è sicuramente mal rappre-

[1] Il grafico è direttamente preso da Chiogna Pauli (2008).

sentato dalla media delle osservazioni, mentre un idea più precisa sembrano darcela i tre indicatori. Nel grafico, S rappresenta la soglia scelta, i è l'intensità (distanza tra la concentrazione massima e la soglia S), d è la durata (il periodo di durata durante il quale si hanno valori di ozono superiore alla soglia) e m è l'esposizione notturna (la media di inquinamento durante le ore notturne).

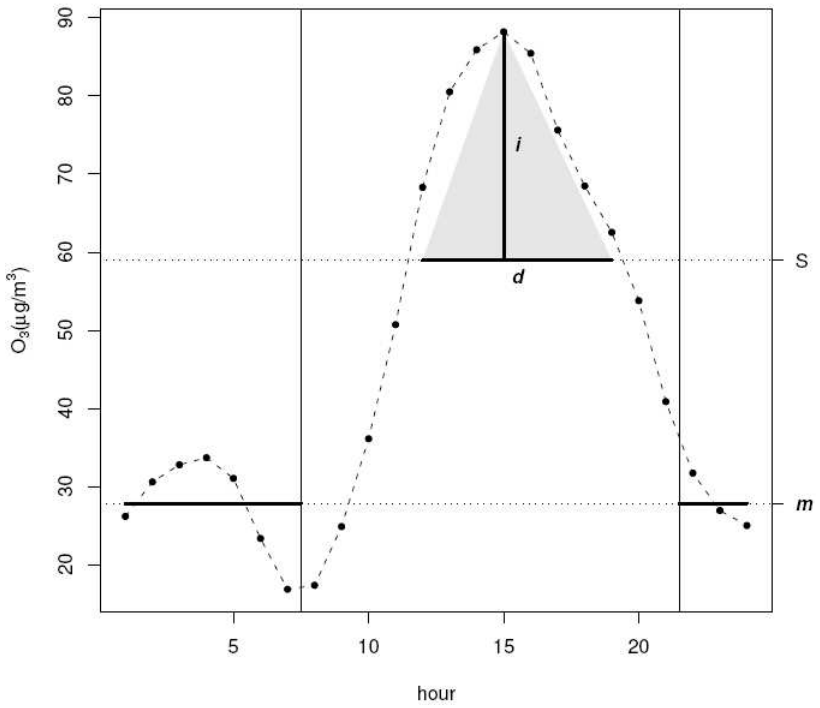


Figura 1.1: Esempio della distribuzione di Ozono nell'arco di una giornata

Per il calcolo dei valori di intensità e durata si utilizzano solo i dati delle ore diurne, calcolando una media sullo spazio tra i vari valori rilevati nelle diverse centraline di monitoraggio dislocate nei differenti punti geografici. Di conseguenza, entrambi gli indicatori non possono essere considerati come indici individuali di esposizione. La scelta di utilizzare solo i valori ottenuti durante le ore del giorno è dovuta al fatto che le massime concentrazioni di inquinamento si verificano durante le ore centrali della giornata, che sono quelle in cui le persone trascorrono maggior tempo all'aria aperta, per strada e al lavoro e che, dunque, sono quelle in cui i soggetti sono a maggior contatto

con gli agenti inquinanti.
Descritti i tre nuovi indici, nell'articolo preso in considerazione, gli autori cercano, dunque, di analizzare gli effetti dell'uso di indicatori di sintesi giornaliera invece delle esposizioni individuali nelle stime dei modelli di regressione.

In conclusione, la proposta fatta in questo articolo (di Chiogna e Bellini (2002)), è quella di adoperare un modello di Poisson, utilizzando il numero di decessi o di ricoveri come variabile conteggio. Per poter rappresentare la relazione salute–inquinamento dell'aria, vengono inseriti nel modello, come variabili esplicative, oltre a tutti i fattori ambientali e climatici, degli indicatori d'inquinamento alternativi. Al posto di impiegare semplici sintesi dell'inquinante, come la media giornaliera, ed essendo impossibilitati ad usare un effettivo indice individuale, vengono proposti tre indicatori: l'Intensità, la Durata e l'Esposizione Notturna. Questi tre elementi hanno il compito di descrivere l'esposizione a cui sono sottoposti i soggetti durante il giorno e la notte, sottolineando, in particolare, le giornate con alta concentrazione dell'inquinante sotto osservazione e quelle con alti livelli per lunghi periodi.

1.2 Effetto dell'ozono

Per vedere più nel dettaglio questi tre indicatori d'inquinamento, introdotti nel paragrafo precedente, ci si può riferire all'articolo di M. Chiogna e F. Pauli (2008).
In questo lavoro, gli autori utilizzano i tre indici, **intensità**, **durata** ed **esposizione notturna**, sui dati relativi alla città di Milano, inerenti ad inquinamento, condizioni atmosferiche ed ai soggetti considerati (decessi e ricoveri legati a problemi alle vie respiratorie). Nel lavoro, vengono utilizzati dei modelli additivi generalizzati per lo studio della relazione salute ed inquinamento dell'aria.

Questo articolo è di grande rilevanza per questa tesi, per alcuni importatanti fattori, di seguito richiamati.

- I dati utilizzati sono i medesimi impiegati per questo lavoro. Per le

analisi svolte in questa tesi si è utilizzato, almeno in parte, il medesimo dataset impiegato nello studio qui descritto.

- I modelli utilizzati (additivi generalizzati) sono gli stessi adottati in questa tesi.

- L'utilizzo dei tre indicatori di inquinamento è limitato allo studio dell'ozono. Nell'articolo in discussione, come in questa tesi, il tema centrale è quello di trovare un indicatore alternativo alla media giornaliera per poter cogliere la relazione tra salute ed ozono; non generalizzando su tutti i tipi di inquinamento dell'aria, ma concentrando l'attenzione solo su questo inquinante.

La specifica descrizione del dataset e la spiegazione di determinate scelte fatte nell'articolo e riprese in questa tesi vengono rimandate al capitolo successivo. Sottolineiamo solo che, in questo articolo, gli autori considerano il numero di ricoveri limitatamente a persone anziane di età superiore ai 75 anni; il periodo di studio è ristretto tra il 1995 e i 2003 ed ai soli mesi estivi.

Gli autori impostano lo studio della relazione tra salute e concentrazione di ozono formulando un modello additivo generalizzato (GAM).
Detta $Y_{k,t}$ la variabile che conta il numero di ricoveri in ospedale nel giorno t per la classe di età k (vengono utilizzate due classi: 75–89 anni e over–90), si assume:

$$Y_{k,t} \sim Poisson(\lambda_{k,t}),$$

e si considera il modello per la relazione ricoveri–ozono:

$$\log(\lambda_{k,t}) = \beta_{0,k} + confondenti_{k,t} + ozono_t,$$

dove $confondenti_{k,t}$ è costituito dall'insieme di tutti gli elementi che vengono considerati fattori confondenti tra il numero di ricoveri e l'effetto dell'ozono; $ozono_t$ invece è la funzione della concentrazione di O_3 che ne misura l'effetto. Il modello sopra formulato sarà simile al modello che verrà utilizzato in questa tesi; il fattore che maggiormente muterà sarà la parte relativa all'ozono.

L'elemento di confondimento viene espresso dagli autori come:

$$confondenti_{k,t} = f(t_t) + g_k(T_t) + z_t + \gamma h_t + \alpha w_t,$$

con $f(t_t)$ funzione di lisciamento del tempo, $g_k(T_t)$ funzione di lisciamento specifica della classe di età, della media della temperatura dei tre giorni precedenti (T_t), z_t è la concentrazione media giornaliera di PM_{10}, h_t è un indicatore di vacanza–festa e w_t è un indicatore del giorno della settimana. Altri elementi e dettagli vengono rimandati al capitolo seguente.

Normalmente l'ozono viene rappresentato tramite un indicatore giornaliero della concentrazione, come la media o il massimo, e la sua presenza in un modello, come quello formulato precedentemente, è semplicemente $ozono_t = \beta o_t$, essendo o_t l'indicatore giornaliero scelto.
Nell'articolo considerato, gli autori cercano di sostituire questi indicatori "semplici" con le tre misure introdotte da Chiogna e Bellini (2002), per dare peso maggiore a quei giorni in cui si sono registrate alte concentrazioni di ozono e a quelli con prolungata presenza di valori elevati. Gli indici introdotti vengono inseriti in maniera lineare nel del modello, anche se, a causa dell'alta correlazione, al posto di utilizzare separatamente i e d, se ne utilizza il prodotto.
Lo scopo degli autori in questo articolo è quello di determinare se l'utilizzo di questi nuovi indicatori dell'ozono consenta di cogliere meglio l'effetto dell'inquinante all'interno della relazione con la salute dell'uomo. Per fare ciò, tengono fissa la componente $confondenti_{k,t}$ e stimano diversi modelli con diverse soluzioni per il fattore $ozono_t$. Le formulazioni considerate nello studio sono 61 e comprendono modelli con il semplice inserimento di indicatori "classici" dell'ozono, media e massimo, e modelli con l'utilizzo degli indicatori intensità, durata ed esposizione notturna, calcolati con tre livelli diversi di soglia e in diverse combinazioni, utilizzando anche valori ritardati dei giorni precedenti.

La selezione dei modelli viene fatta utilizzando criteri quali l'Un–Biased Risk Estimate (UBRE) e la Generalized Cross Validation (GCV) suggerite da Wood (2000) per il trattamento dei GAM.

Inoltre gli autori utilizzano il metodo Bootstrap, suggerito da Sauerbrei (1999), per la scelta del modello, replicando, tramite ricampionamenti bootstrap, la selezione del modello ed usando la frequenza di selezione come metodo di valutazione.

I risultati ottenuti, descritti nell'articolo, mostrano una maggiore significatività dei modelli in cui vengono utilizzati intensità e durata, rispetto a quelli in cui viene inserita la media o il massimo giornaliero. Il metodo bootstrap sottolinea che tutti i modelli più frequentemente selezionati contengono la componente dell'ozono, sostenendo l'idea che questo elemento inquinante influenzi effettivamente la salute umana. Inoltre, osservando la significatività dei parametri relativi ad intensità e durata, rispetto a quelli inerenti a media e massimo, calcolata in modelli differenti, risulta evidente che sintesi giornaliere della concentrazione di ozono che riescano ad esprimere sia l'alta concentrazione sia la persistenza di questo fenomeno esprimano meglio l'influenza dell'inquinante rispetto a degli indicatori troppo riassuntivi come la media e il massimo.
Questi risultati sottolineano la necessità di esprimere la variabilità dei valori di ozono rilevati nell'arco di una giornata. Da questo punto di partenza, in questa tesi, si cercherà di andare oltre i tre indici descritti, che già evidenziano la significatività dell'effetto dell'ozono rispetto agli indici giornalieri, proponendo l'inserimento di valori orari dell'inquinante per cercare di cogliere appieno l'andamento fluttuante della concentrazione di ozono.

Capitolo 2

I dati: un'analisi preliminare

I dati analizzati provengono da due datasets distinti e raccolgono le informazioni[1] inerenti all'inquinamento rilevato nella zona di Milano nel periodo che va dal 01 Gennaio 1998 al 30 Dicembre 2003.
Il primo dataset, oltre al numero di ricoveri giornalieri avvenuti per problemi alle vie respiratorie, riporta i valori medi "giornalieri" di inquinamento rilevati nella zona e nel periodo indicati. Sono dunque contenute, in questo dataset, informazioni relative ai vari inquinanti ed alle condizioni atmosferiche (temperatura, umidità, velocità del vento, ...) tramite i valori medi calcolati in ogni giorno nel periodo sotto studio.
Il secondo insieme di dati (dataset), invece, è costituito semplicemente dai valori raggiunti dagli inquinanti nel periodo 1998-2003 rilevati da tre centraline nella zona di Milano (via Juvara, Parco Lambro e Verziere). Delle tre rilevazioni per ogni inquinante, una per centralina, si è ottenuta una media oraria ed è con queste rilevazione "orarie" che cercheremo di verificare in seguito le ipotesi formulate nell'introduzione. Quello che segue è una breve descrizione delle variabili che verranno utilizzati nella stima dei modelli.

OZONO
Sicuramente la variabile di maggior interesse per questo studio è l'ozono e per questo gli dedicheremo un maggior spazio rispetto agli altri elementi.

[1] Tutti i dati inerenti agli inquinanti e alle condizioni atmosferiche sono stati raccolti da ARPA Lombardia (Agenzia Regionale per la Protezione dell'Ambiente)

L'ozono[2] è un gas bluastro che nella nostra atmosfera è presente in due diversi strati con due funzionalità e conseguenze distinte. Al 90% è presente nella stratosfera (tra i 10 e i 50 km di altezza) e in questa zona funge da importante filtro per i raggi UV (l'attuale problema del buco dell'ozono si riferisce ad un impoverimento di O_3 in questa zona e di conseguenza all'assottigliamento della fascia protettiva contro i raggi UV). La formazione dell'ozono stratosferico ha luogo per la maggior parte a più di 30 Km di altezza. Qui le radiazioni UV con lunghezza d'onda inferiore ai 242 nm dissociano l'ossigeno molecolare in ossigeno atomico che, per la sua reattività, si combina rapidamente con una molecola di ossigeno originando l'ozono

$$O + O_2 \Longrightarrow O_3.$$

A loro volta le molecole di ozono che si formano nel corso di questa reazione assorbono le radiazioni solari con lunghezza d'onda compresa fra 240 e 340 nm, e questo ne provoca la fotolisi che libera un atomo ed una molecola di ossigeno

$$O_3 \Longrightarrow O_2 + O.$$

In definitiva questi processi instaurano un equilibrio dinamico che mantiene la concentrazione di ozono pressoché costante e che permette di schermare più del 90% delle pericolose radiazioni UV provenienti dal sole.
Il secondo strato atmosferico in cui è presente l'ozono è la troposfera (fino ai 10km di altezza); qui è presente, in maniera naturale, per effetto della circolazione atmosferica o per effetto di scariche elettriche durante i temporali. Una maggiore presenza di ozono nella troposfera, che lo rende un insidioso inquinante secondario, è dovuta alla reazione di inquinanti atmosferici principali prodotti da gas inquinanti emessi dalle automobili, dalle industrie, dalle raffinerie, che reagiscono in presenza della luce solare (smog fotochimico). Una delle sorgenti principali tra gli inquinanti primari è data dal biossido di azoto che, come già detto, in presenza della luce solare dà origine per fotolisi all'ossigeno atomico (che produce l'ozono reagendo con l'ossigeno molecolare).

[2]Fonte: ARPA Lombardia; e da nonsoloaria.com

Dunque la produzione di ozono da parte dell'uomo è indiretta, ma comunque determinante. I gas precursori dell'ozono difatti vengono immessi dall'uomo tramite processi di combustione civile e industriale e da processi che utilizzano o producono sostanze chimiche volatili, come solventi e carburanti.

Gli effetti che l'ozono ha sull'organismo[3] umano sono in relazione alla concentrazione della sostanza nell'aria, alla durata di esposizione (al numero di ore passate all'aria aperta), alla ventilazione polmonare durante l'esposizione (quindi ad eventuali sforzi legati al lavoro o ad attività sportive) e ad eventuali condizioni di salute dei soggetti esposti.
I disturbi causati all'organismo umano da un'eccessiva esposizione sono soprattutto a carico di:
- occhi (aumento della sensibilità e fenomeni irritativi);
- sistema respiratorio (tosse, irritazione alla gola e ai polmoni, riduzione delle funzioni polmonari e sensazione di oppressione al torace, "Fiato corto");
- sistema cardiocircolatorio (tachicardia e aumento del rischio in soggetti cardiopatici);
- sistema immunitario (aumento di sensibilità agli allergeni che provocano attacchi d'asma e maggiori possibilità di insorgenza di episodi acuti in soggetti asmatici).

Quattro gruppi di persone, sono particolarmente sensibili all'ozono.

- `bambini`. I bambini sono il gruppo a più alto rischio per una esposizione ad ozono, perchè essi trascorrono gran parte delle vacanze estive all'aperto, impegnati in attività fisiche intense. I bambini hanno anche maggiori probabilità di sviluppare l'asma o altre malattie respiratorie. L'asma è la malattia cronica più comune nei bambini e può essere aggravata da una esposizione all'ozono;

- `adulti che fanno attività fisica all'aperto`. Adulti in buona salute che fanno attività fisica all'aperto sono considerati un gruppo

[3]Fonte: Servizio sanitario regionale Emilia-Romagna, ARPA Emilia-Romagna

"sensibile" perché sono più esposti all'ozono, rispetto a popolazione meno attiva;

- `persone con malattie respiratorie, come ad esempio l'Asma`: Non c'è certezza che l'ozono causi asma o altre malattie respiratorie croniche, ma queste malattie rendono i polmoni più vulnerabili agli effetti dell'ozono. Così gli individui che si trovano in queste condizioni risentono prima degli effetti dell'ozono e a concentrazioni più basse rispetto agli individui meno sensibili;

- `anziani`. A tutt'oggi, vi sono alcune evidenze che indicano che gli anziani o le persone con malattie cardiache hanno un´aumentata sensibilità all'ozono. Comunque, come altri adulti, le persone anziane possono essere ad alto rischio se soffrono di malattie respiratorie o se sono attivi all'aperto, o se sono particolarmente suscettibili.

A causa degli effetti dell'ozono sull'uomo confermati da numerosi studi epidemiologici, la normativa europea, e a cascata quella italiana, hanno regolamentato la valutazione delle concentrazioni di tale inquinante. Il Decreto Legislativo 183/04, che recepisce la Direttiva 2002/3/CE, introduce le definizioni di:

- `soglia di informazione`: livello oltre il quale vi è un rischio per la salute umana in caso di esposizione di breve durata per alcuni gruppi particolarmente sensibili della popolazione;

- `soglia di allarme`: livello oltre il quale vi è un rischio per la salute umana in caso di esposizione di breve durata e raggiunto il quale devono essere adottate le misure previste dall'articolo 5, comma3 che prevede l'adozione di azioni a breve termine;

- `obiettivo a lungo termine`: concentrazione di ozono nell'aria al di sotto della quale si ritengono improbabili, in base alle conoscenze scientifiche attuali, effetti nocivi diretti sulla salute umana e sull'ambiente nel suo complesso. Tale obiettivo è conseguito nel lungo periodo, purché sia realizzabile mediante misure proporzionate, al fine di fornire un'efficace protezione della salute umana e dell'ambiente.

Nella tabella che segue sono indicati questi tre valori soglia fissati[4].

Obiettivo a lungo termine per la protezione della salute umana	Media su 8 ore massima giornaliera nel periodo di un anno civile	$120\mu g/m^3$
Soglia di informazione	Media di 1 ora	$180\mu g/m^3$
Soglia di allarme	Media di 1 ora	$240\mu g/m^3$

Tabella 2.1: Valori di soglia per l'Ozono

Di tutte le informazioni sull'ozono sopra citate, alcune in particolare sono state prese in considerazione in questo studio per ottenere migliori risultati. Di queste, certe verranno analizzate in seguito parlando specificatamente delle altre variabili osservate, come i soggetti considerati (solo gli anziani), la temperatura e i fattori inerenti al giorno della settimana.
Un altro fattore importante è legato a quanto è stato detto sulla relazione tra l'ozono e i raggi solari. È chiaro per gli studiosi che la presenza nociva di ozono per l'essere umano si verifica maggiormente nei periodi dell'anno in cui l'attività solare è più intensa.
Il grafico in Figura 2.1 rappresenta il valor medio giornaliero di ozono nei 6 anni indicati come periodo dello studio. Da qui è evidente l'andamento periodico dell'ozono, caratterizzato dalla ciclicità annuale in cui i picchi maggiori sono nei mesi estivi, quelli con maggiore attività solare. Notiamo, inoltre, che la media di concentrazione di ozono nel periodo estivo (i mesi di Giugno, Luglio e Agosto) è di $78,29\mu g/m^3$ mentre quella per i restanti mesi è solo di $25,98\mu g/m^3$.
Osservando quanto visto nel grafico e considerando i valori medi nella stagione estiva e nel resto dell'anno si è deciso, basandosi anche sulla medesima scelta fatta in Chiogna e Pauli(2008), di utilizzare in questo studio solo i tre mesi estivi per ogni anno preso in considerazione. Quindi tutte le analisi che seguiranno saranno fatte solo sui mesi Giugno, Luglio ed Agosto tra il 1998 e il 2003.
Come già spiegato nell'introduzione, lo scopo di questo lavoro è provare ad utilizzare osservazioni orarie dell'ozono per saggiare la significatività di un

[4]ARPA Veneto, (Agenzia Regionale per la Protezione dell'Ambiente)

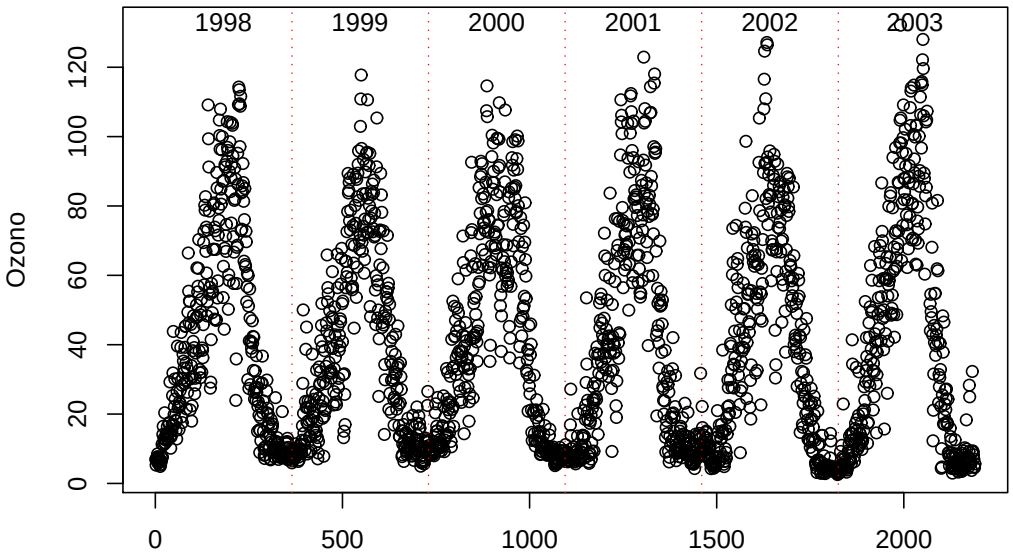


Figura 2.1: Media giornaliera di Ozono negli anni 1998-2003 nella città di Milano

nesso con la salute dell'uomo. Di conseguenza, è stato tenuto in considerazione un secondo dataset nel quale sono inseriti i valori orari di ozono registrati da tre centraline collocate a Milano. Nel grafico riportato in Figura 2.2

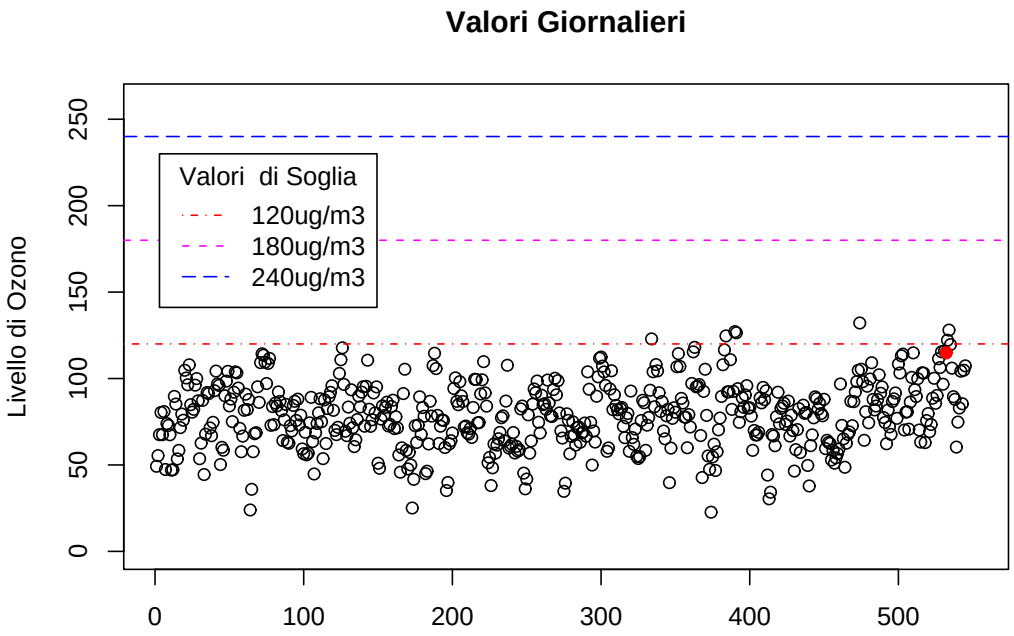


Figura 2.2: Valori giornalieri di O_3 per i tre mesi estivi dal '98 al '03

si riportano le osservazioni giornaliere (medie) di ozono per i tre mesi estivi nei sei anni tenuti in considerazione. Nel secondo grafico (Figura 2.3) sono rappresentate le rilevazioni orarie del medesimo inquinante. In entrambi i grafici sono state aggiunte le linee relative alla soglia di allarme, di informazione e quella relativa all'obiettivo a lungo termine precedentemente definite. Possiamo notare come siano pochi i valori, considerando la media giornaliera, che superino il valore di $120\mu g/m^3$ della soglia dell'obiettivo prefissato a lungo termine e che non ci siano valori che raggiungano, nell'arco dei sei anni, le soglie di informazione o d'allarme. Tutt'altra idea invece l'abbiamo osservando il secondo grafico in cui sono rappresentate le singole osservazioni orarie.

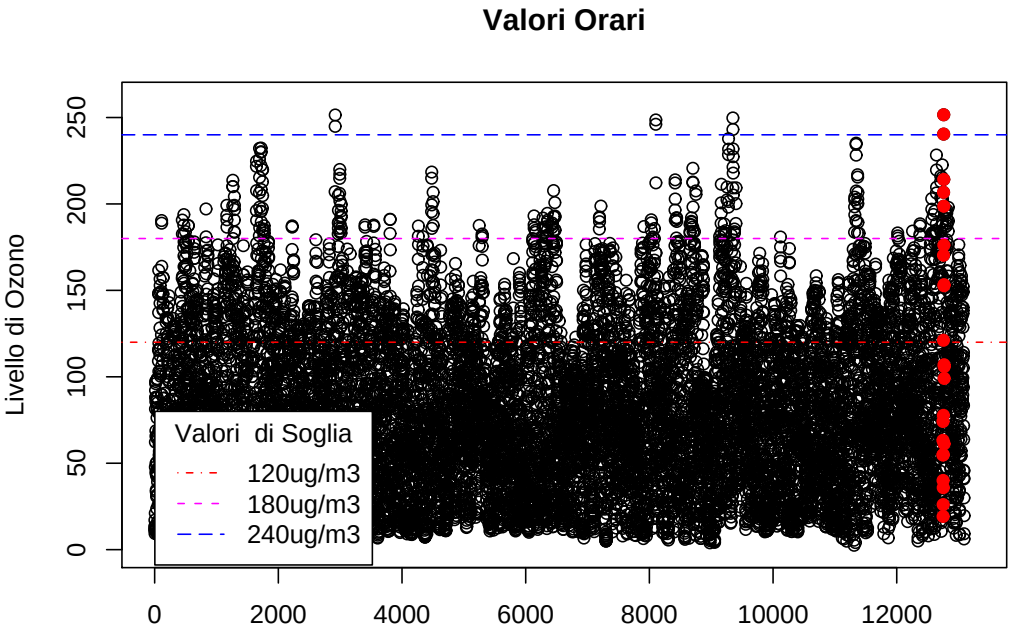


Figura 2.3: Valori orari di O_3 per i tre mesi estivi dal '98 al '03

Prendiamo ad esempio le osservazioni relative ad un singolo giorno (i punti rossi nei due grafici Figura 2.2 e Figura 2.3, inerenti ai dati raccolti l'11 Agosto 2003) riportate nel grafico in Figura 2.4. Notiamo un andamento cre-

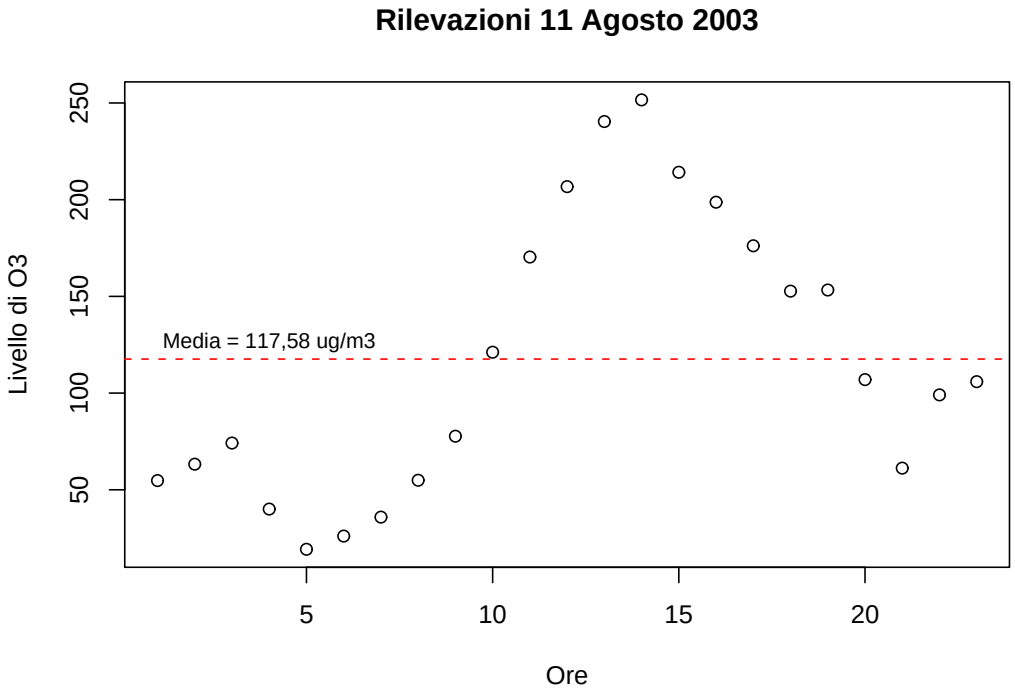


Figura 2.4: Rilevazioni orarie di Ozono relative alla giornata dell'11 Agosto 2003

scente con apice durante le ore centrali del giorno e un abbassamento durante le ore notturne. La media di queste osservazioni, e di conseguenza il valore utilizzato nel primo grafico dei due precedenti, è di $117,58 \mu g/m^3$, sotto la prima delle tre soglie, mentre il massimo è $251,6 \mu g/m^3$, oltre la soglia d'allarme. Possiamo pensare che la via che cercheremo di percorrere nei modelli che seguiranno, non utilizzando più un indice sintetico dell'ozono quale la media, ma tutte le osservazioni orarie, possa cogliere questi andamenti orari. Il grafico in Figura 2.5 rappresenta la distribuzione oraria, sui i sei anni, dell'ozono. Ogni diagramma a scatola è relativo ad ogni ora e mostra che quanto evidenziato nell'esempio riportato è valido per tutti i giorni presi in considerazione; ovvero che durante le ore centrali del giorno si verifica un

netto innalzamento dei valori di ozono, non colto da un indice sintetico quale la media. Questo aumento durante le ore centrali è confermato da quanto è

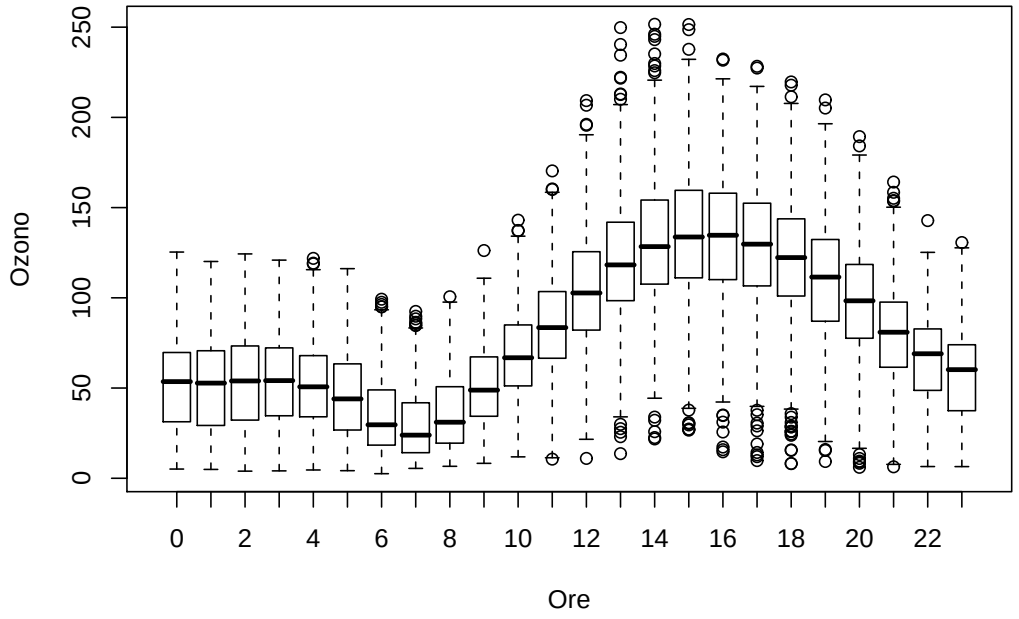


Figura 2.5: Distribuzione oraria della quantità di O_3

stato detto in precedenza sull'ozono e sulla sua reazione nei momenti di forte attività solare.

NUMERO RICOVERI

Come già delineato nell'introduzione, lo scopo dei modelli che stimeremo in seguito è valutare se l'ozono sia un fattore penalizzante per la salute dell'uomo. Per fare ciò, ci serviremo della variabile numero di ricoveri quale variabile risposta nei modelli esplicitati nei prossimi capitoli. Si è detto, parlando dell'ozono, che i soggetti maggiormente a rischio sono i bambini, persone che fanno attività fisiche, persone con precedenti problemi respiratori e gli anziani. In questo studio si è tenuto conto solo di questi ultimi, perché altamente a rischio. La variabile risposta utilizzata è, quindi, il numero

di ricoveri dovuti a problemi respiratori registrati dal sistema ospedaliero di Milano[5], per pazienti con età maggiore ai 75 anni residenti a Milano. Tutti gli altri gruppi di ricoveri, compreso quella per cause accidentali, non sono stati tenuti in considerazione nel conteggio.
In Figura 2.6 sono rappresentati il numero di ricoveri registrati nei sei anni di interesse. Da questa immagine possiamo cogliere il trend crescente di rico-

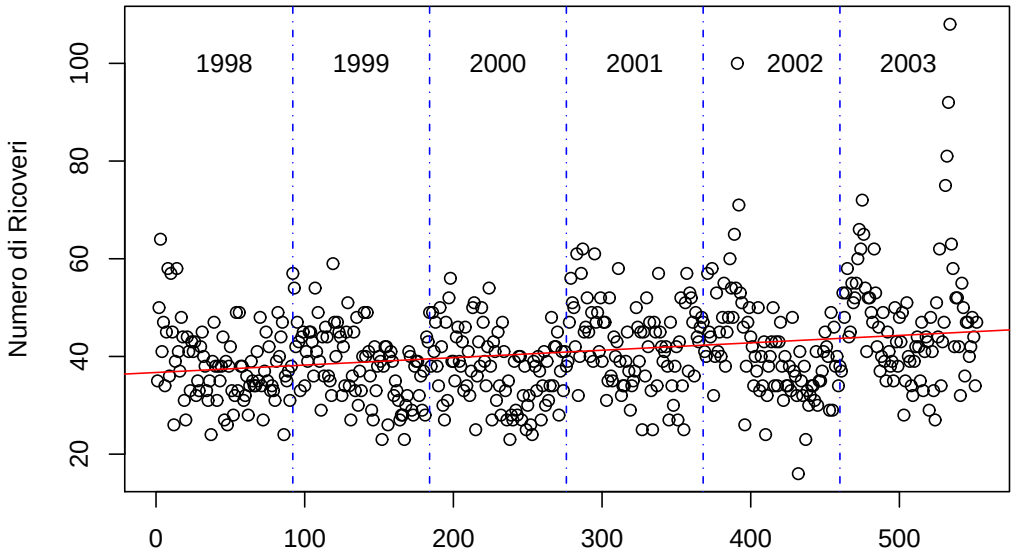


Figura 2.6: Numero di ricoveri legati a problemi di tipo respiratorio, registrati a Milano nel periodo indicato

veri, identificato dalla linea continua, che evidenzia un aumento progressivo col passare del tempo. Questo fattore "tempo" dovrà essere tenuto in considerazione nella formulazione del modello.

PM10

Molti sono gli inquinanti ritenuti nocivi per la salute dell'uomo e per questo

[5]Dati ottenuti dall'Istituto nazionale di statistica ISTAT

monitorati tramite le stazioni di rilevamento. I principali sono: Monossido di Carbonio (CO), Biossido di Azoto (NO_2), Ozono (O_3), Polveri PM_{10}, Biossido di Zolfo (SO_2) e Benzene (C_6H_6).
Di questi, solo due sono considerati in questa tesi e sono: l'oggetto principale dello studio, l'ozono, e le polveri sottili o PM_{10}. Non si è voluto utilizzare tutti gli altri elementi visto che, come evidenziato nell'introduzione, lo scopo di questo lavoro non è trovare un modello col quale fare delle previsioni future, ma piuttosto solo quello di trovare un metodo alternativo e significativo per trattare i valori dell'ozono. Inoltre il PM_{10} è spesso l'unico incluso, a differenza degli altri elementi, come effetto confondente in studi come questo in cui si modellano relazioni tra salute e inquinamento(Bell et al., Ito et al., Levy et al.(2005))[6].
Con PM10 viene identificato l'insieme di tutte le particelle solide o liquide che restano in sospensione nell'aria. Il particolato sospeso totale rappresenta un insieme estremamente eterogeneo di sostanze la cui origine può essere primaria (emesse come tali) o derivata (da una serie di reazioni fisiche e chimiche). Le particelle di dimensioni maggiori (diametro $> 10\mu m$, da qui pm10) hanno un tempo medio di vita nell'atmosfera che varia da pochi minuti ad alcune ore e hanno la possibilità di essere aerotrasportate per una distanza massima di 1-10 Km.
Le maggiori concentrazioni di pm10 si trovano nei centri abitati e trafficati e i picchi più alti di questo inquinante sono, all'esatto opposto dell'ozono, nei periodi invernali, quando sono più frequenti le condizioni di ristagno degli agenti inquinanti. Nella Figura 2.7 è sottolineato questo aspetto: i picchi dei valori dell'ozono corrispondono agli abbassamenti di PM_{10}, e viceversa.

TEMPERATURA
Un altro fattore da tenere in considerazione al momento della formulazione del modello è la temperatura. L'utilizzo di questa variabile è comune in studi come questo in cui si analizza lo stato di salute di persone di età avanzata. Inoltre, come detto precedentemente, l'ozono è legato all'attività solare e di conseguenza ai periodi con alte temperature. Il grafico in Figura 2.8 eviden-

[6]Articoli pubblicati su *Epidemiology 16(4)*

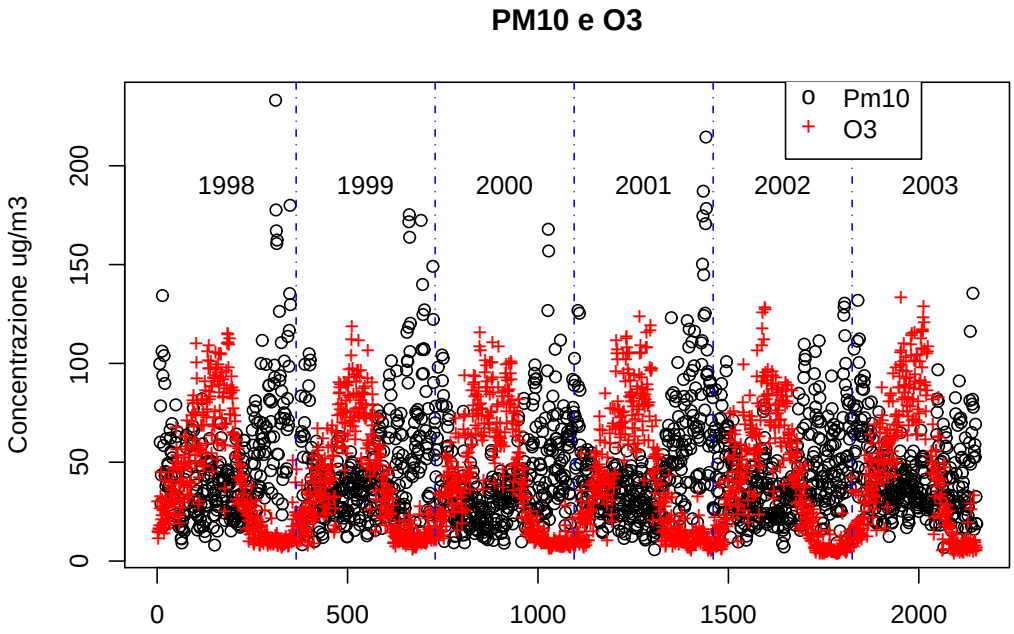


Figura 2.7: Andamento del PM_{10} e dell' O_3 nell'arco dei sei anni considerati

zia questo legame mostrando la corrispondenza tra la crescita di un fattore e quella dell'altro. È inoltre ragionevole ritenere che la temperatura abbia un

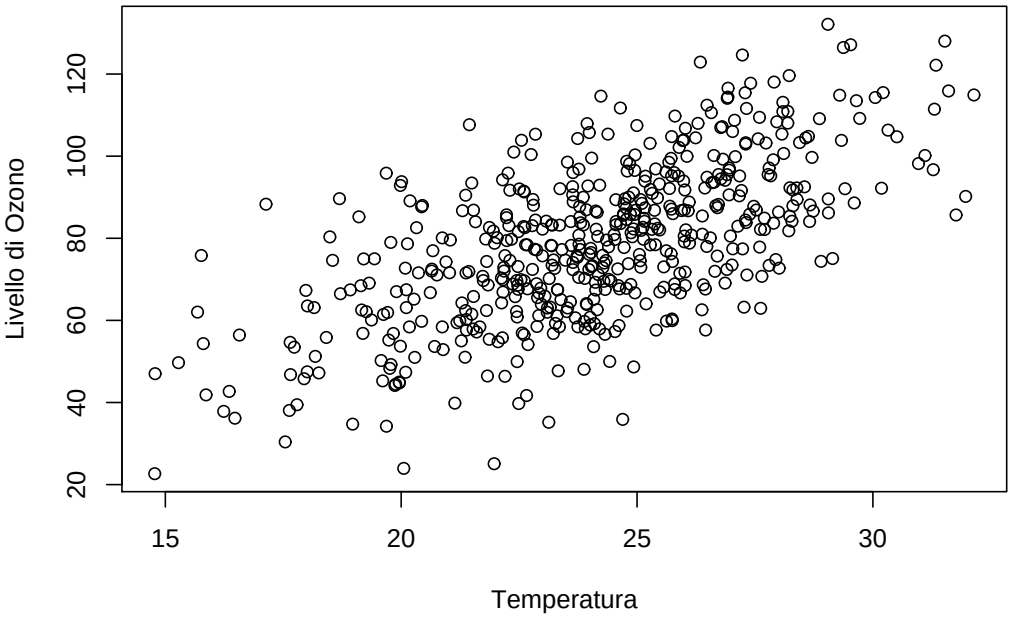


Figura 2.8: Temperatura giornaliera e livello di Ozono

effetto sulla salute che si prolunga al di là del giorno stesso. Per tenere conto di questo aspetto si inserirà nel modello una variabile che indicherà la media della temperatura dei tre giorni precedenti.

WDAY E FESTA

Gli ultimi due fattori che verranno tenuti in considerazione sono degli identificativi dei giorni della settimana e di giorno di festa. Questi indicatori possono essere utili, in particolare per identificare i giorni in cui le persone, nel nostro caso gli anziani milanesi, siano stati propensi a stare all'aria aperta e quindi essere maggiormente esposti all'ozono, oppure per identificare quei giorni in cui le emissioni di agenti inquinanti siano maggiori.

A prima vista (Figura 2.9) non sembra esserci un reale effetto di questi due

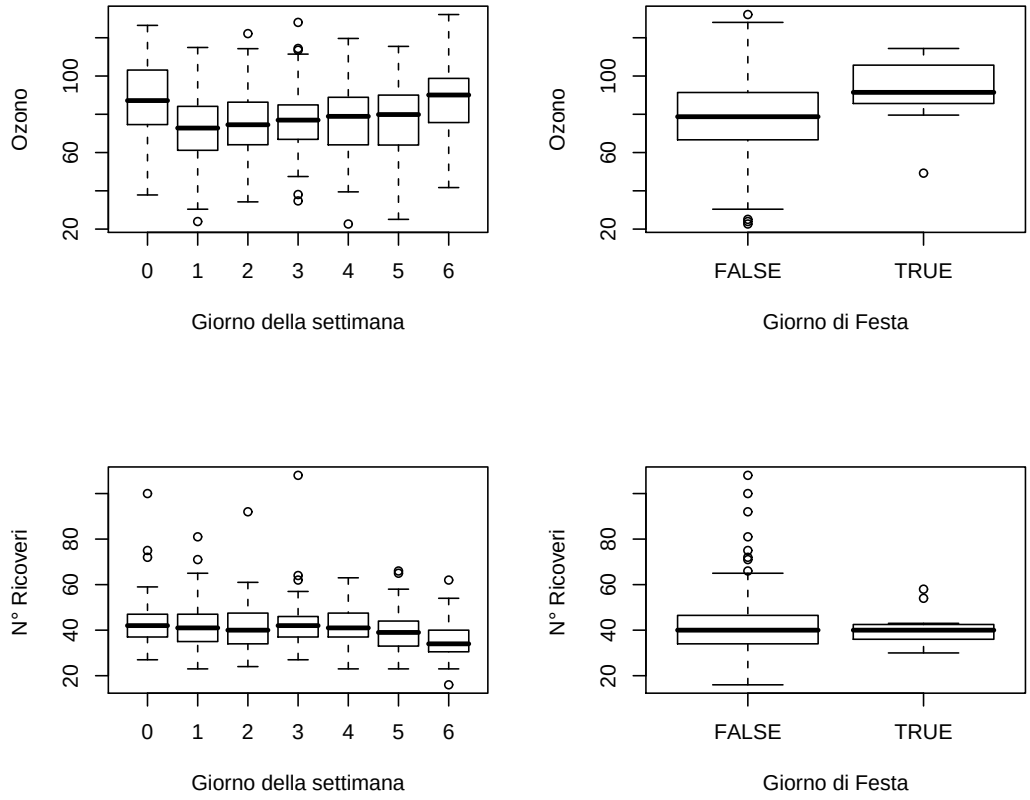


Figura 2.9: Distribuzione rispetto ai giorni della settimana e di festa di Ozono e del numero di ricoveri. I giorni della settimana sono codificati con 0 = Domenica; 1 = Lunedì; ...

fattori né sul numero di ricoveri né sulla quantità di ozono. Comunque è di prassi inserire, al momento della formulazione dei modelli, questi componenti.

ANNO (TEMPO)
Come già accennato parlando della variabile relativa al numero di ricoveri, dovremmo tenere in considerazione il fattore tempo e della differenza, anche se pur minima, di rilevazioni tra un anno ed un altro.

Dai tre grafici in Figura 2.10, in cui sono rappresentati i diagrammi a scatola

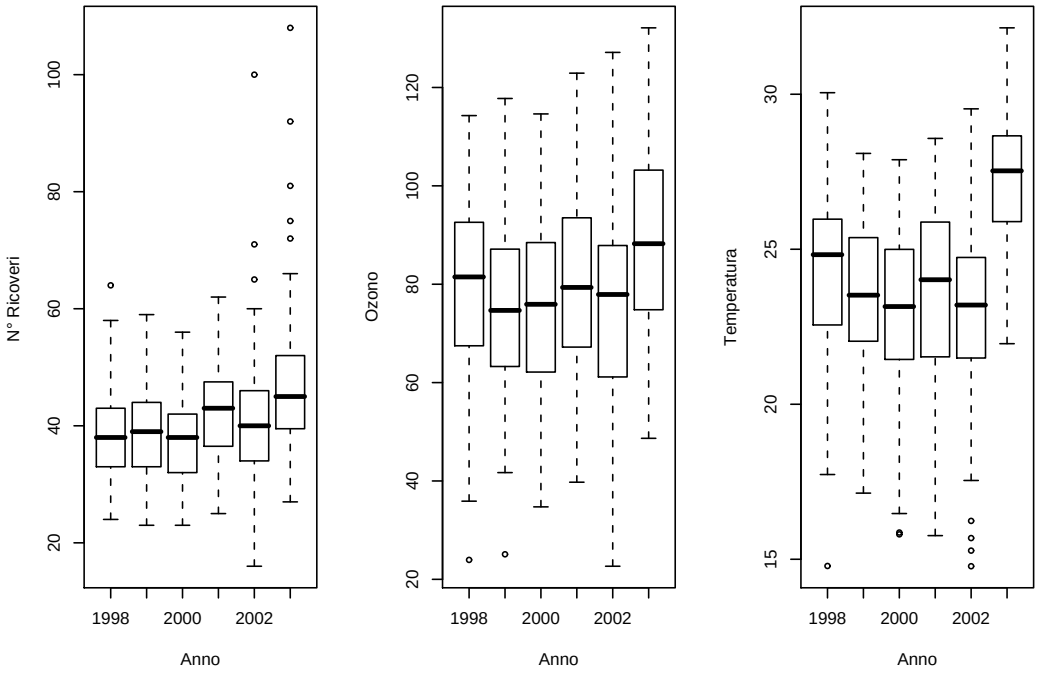


Figura 2.10: Numero di ricoveri, livello di Ozono e Temperatura divisi nei sei anni dello studio

divisi per anno di numero ricoveri, ozono e temperatura, notiamo in particolare che nell'anno 2003 tutti i parametri hanno registrato valori mediamente più alti.

Capitolo 3

Preparazione dei Modelli

Passeremo ora, utilizzando i dati appena descritti, a cercare di identificare una specificazione adeguata per esprimere la relazione tra l'ozono e i ricoveri dovuti a problemi respiratori. Per farlo, dovremmo decidere in quale modo utilizzare le informazioni relative alla quantità di inquinante O_3 registrato.

Abbiamo visto nell'analisi esplorativa dei dati, che la concentrazione di ozono varia drasticamente nelle diverse ore del giorno. In particolare, osservando la Tabella 3.1 e la Figura 3.1 si nota che durante le ore centrali della giornata, al crescere dell'attività solare, si verifica un aumento di ozono che decresce nel finire del giorno. Sarà questa osservazione il punto di partenza per il lavoro che segue.

3.1 Espansione dei dati giornalieri

Lo scopo di questa tesi è verificare se, utilizzando le singole osservazioni orarie, è possibile riuscire a rendere significativo, all'interno di un modello di regressione, il fattore che rappresenta l'ozono. L'influenza che questo inquinante ha sulla salute umana è noto, ma indicatori comunemente usati quali la media o il massimo giornaliero non sempre sono soddisfacenti per lo studio di questa relazione. Studi già condotti hanno mostrato che, aumentando l'informazione inserita nei modelli, si arriva spesso a migliori risultati. In particolare, lo studio condotto da M. Chiogna e F. Pauli (2008), già de-

Ora	Media	Ora	Media
0	51.33	12	104.06
1	51.45	13	120.78
2	53.76	14	130.47
3	53.96	15	135.02
4	51.89	16	133.39
5	46.35	17	128.21
6	35.53	18	120.12
7	30.19	19	109.49
8	36.45	20	96.76
9	51.37	21	79.40
10	68.52	22	65.82
11	85.55	23	56.79

Tabella 3.1: Valori medi, divisi per ora, di ozono

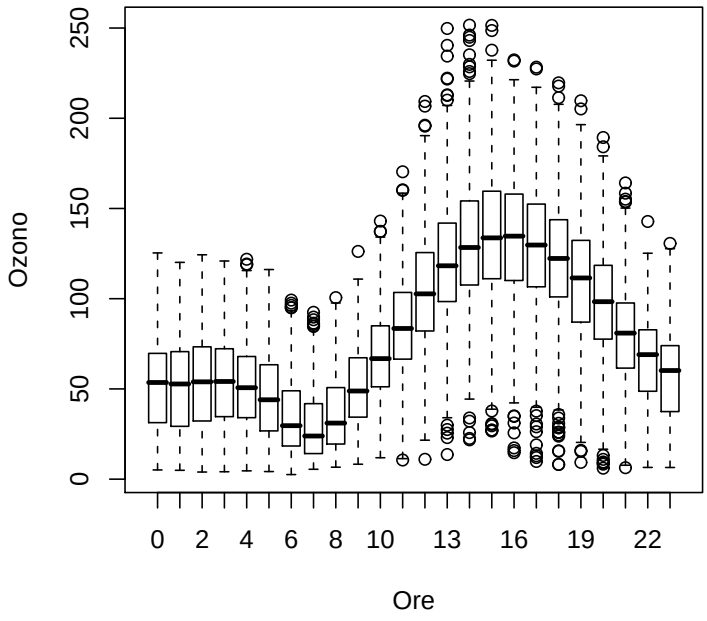


Figura 3.1: Concentrazione di Ozono

scritto nel paragrafo 1.2, sui dati inerenti alla città di Milano, usati anche per questa tesi, ha mostrato che utilizzando indici quali l'intensità, la durata e l'esposizione notturna, che maggiormente colgono l'andamento altalenante delle osservazioni di ozono nell'arco della giornata, si ottengono risultati più soddisfacenti.

In particolare si prenderà un modello basato su valori giornalieri per tutte le variabili esplicative e su valori orari di ozono.

Il primo ostacolo da superare in questo procedimento riguarda la disparità di rilevazioni che così facendo ci ritroviamo. Utilizzando un modello di regressione lineare a titolo esemplificativo, quello che ci si prefigge è, infatti, riuscire a stimare un modello del tipo:

$$Y_j = \beta_0 + \beta_1 x_{j,1} + \beta_2 x_{j,2} + \cdots + \beta_p x_{j,p}(h) + \epsilon_j; \tag{3.1}$$

$j = 1, \ldots, J$; $h = 1, \ldots, 24$; dove mettiamo in relazione una variabile risposta (Y_j), con un insieme di $p-1$ variabili indipendenti misurate su scala giornaliera (J osservazioni) e con una variabile, $x_{j,p}(h)$, misurata su base oraria h. Infatti, per tutte le variabili considerate, esclusa quella relativa all'ozono, possediamo la media giornaliera, avendo così 552 osservazioni per variabile (92 osservazioni per trimestre estivo per i 6 anni dello studio); dell'ozono, però, vogliamo utilizzare le singole osservazioni orarie per ogni giorno, disponendo così di $552 \times 24 = 13248$ osservazioni. Quindi, per poter stimare in seguito i modelli necessari, si è costruito un nuovo dataset "espanso", in cui ogni singola osservazione è stata replicata ventiquattro volte. In termini pratici, presa una qualsiasi covariata k considerata ($k \neq p$), ogni valore $x_{j,k}$, $j = 1, \ldots, J$, con $J = 552$ numero di giorni, viene replicato 24 volte ottenendo $x_{j,k}(h)$, con $h = 1, \ldots, 24$; ovvero $x_{j,k}(h) = x_{j,k}\ \forall h$.
Usando questo artificio otteniamo 13248 osservazioni per tutte le variabili e possiamo stimare i modelli necessari.

3.2 Formulazione dei Modelli Additivi

La scelta del tipo di modello per la verifica delle nostre ipotesi ricade nella classe dei modelli additivi generalizzati. Quello che si cercherà di fare è di formulare e calcolare un primo Core Model, un modello–nucleo, che conterrà tutte le variabili confondenti selezionate e ritenute utili. Per questo modello si utilizzeranno le medie giornaliere osservate (i dati originali). Il secondo passo sarà quello di creare un secondo Core Model, impiegando la stessa formulazione e gli stessi elementi del primo modello, ma utilizzando i dati "espansi". Formulati i due modelli, e calibrati tra di loro, si inseriranno in essi le due variabili relative all'ozono, rispettivamente la media giornaliera e le osservazioni orarie, per poter confrontare i due metodi.

3.2.1 Costruzione

La formulazione del modello di partenza è simile a quella introdotta nel paragrafo 1.2, anche se in questa tesi non verranno utilizzati modelli stratificati per età ma verranno presi in considerazione solo gli anziani over–75 anni.
Considerando il numero di ricoveri come variabile risposta, è facile impostare il problema con $Y_j, j = 1, \ldots, J$ (J numero di giorni) variabile conteggio e quindi:

$$Y_j \sim Poisson(\lambda_j)$$

e di conseguenza il modello additivo che andremo a calcolare sarà esprimibile:

$$\log \lambda_j = \beta_0 + confondenti_j + ozono_j. \tag{3.2}$$

Esattamente come nel caso descritto nel Capitolo 1, in $confondenti_j$ sono raccolte tutte le variabili che utilizziamo per delineare il problema e che servono per creare la base per poi studiare la relazione tra Y_j e $ozono_j$, quindi tra il numero di ricoveri e le quantità di O_3 rilevata.
Gli elementi di confondimento considerati ed inseriti sono:

$$confondenti_j = s_1(t_j) + s_2(temp_j) + s_3(temp_j^{lag}) + p_j + f_j + g_j. \tag{3.3}$$

Le tre funzioni $s_1(\cdot), s_2(\cdot), s_3(\cdot)$, sono funzioni di lisciamento opportune, che vedremo in dettaglio in seguito, e sono relative al tempo (t_j), alla media giornaliera della temperatura ($temp_j$) e alla media della temperatura dei tre giorni precedenti ($temp_j^{lag}$). Le altre tre variabili considerate descrivono la concentrazione media giornaliera di PM_{10} (p_j), il fattore festa (f_j) e il fattore relativo al giorno della settimana (g_j).

All'interno di $ozono_j$ viene inserito il valore registrato di O_3, ma a seconda di che indicatore utilizzeremo verrà espresso diversamente:

$$\begin{cases} o_j & \text{nel caso di indice giornaliero (media),} \\ o_j(h) & \text{nel caso di valori orari.} \end{cases}$$

3.2.2 La scelta dei GAM

La scelta di utilizzare i modelli additivi generalizzati (GAM Generalized Additive Models) è dovuta al fatto che questi strumenti di analisi si prestano bene a studi in cui bisogna tener conto di più fattori, di più variabili esplicative. Ma la caratteristica di maggior rilievo per questo studio è la possibilità di trattare le variabili esplicative all'interno dei modelli in modo parametrico o in termini non parametrici, attraverso appositi lisciatori. Questo ci permette di raggiungere maggiore flessibilità nella modellazione.
Ne è un esempio pratico il modo in cui vengono trattati gli elementi inseriti in $confondenti_j$ nella formula (3.3): i primi tre sono inseriti non parametricamente tramite le tre funzioni di lisciamento, mentre gli ultimi tre fattori, relativi al PM_{10}, a "festa" e a "giorno", sono trattati linearmente.
Di conseguenza, il modello che andremo a stimare, senza considerare l'ozono, sarà espresso nel modo seguente:

$$\log \lambda_j = \beta_0 + \beta_1 p_j + \beta_2 f_j + \beta_3 g_j + s_1(t_j) + s_2(temp_j) + s_3(temp_j^{lag}). \quad (3.4)$$

Essendo f_j e g_j variabili categoriali con, rispettivamente, due e sette livelli, i parametri β_2 e β_3 rappresentano, sinteticamente, i vettori di parametri relativi alle variabili indicatrici con cui i due fattori vengono codificati.

Le funzioni di regressione $s_w(x)$, usate in questa tesi, sono mezzi molto utili per descrivere le variabili in quanto sono strumenti che esprimono l'andamento dei valori considerati. Sono particolarmente efficaci anche se inserite all'interno di un modello di regressione nello studio della relazione tra due o più variabili. La struttura delle spline di regressione (il tipo di funzioni utilizzate in questa tesi) $s_w(x)$ è esprimibile come:

$$s_w(x) = \sum_{i=1}^{k_w} \beta_i \cdot b_{1i}(x), \tag{3.5}$$

dove x è la variabile da lisciare, k_w è la dimensione della base utilizzata nella spline w per rappresentare il termine lisciato, β_i sono i parametri all'interno della funzione e $b.(\cdot)$ è la funzione di base. Un esempio di funzione $b.(\cdot)$ è un polinomio di terzo grado; quindi lo spline di regressione formato da questi polinomi di terzo grado è detto spline cubico e la sua composizione è data dall'unione di k_w polinomi di terzo grado uniti tra di loro in k nodi, precedentemente fissati.
Dunque la procedura per la creazione di uno spline di regressione, differente per ogni variabile a cui si vuole applicare questa tecnica, consiste nel fissare k nodi e di disporli tra i valori della variabile d'interesse. Successivamente, scelto il tipo di funzione di base (ad esempio il polinomio di terzo grado), si applica questa funzione ad ogni intervallo tra due nodi facendo in modo che le due funzioni di base, aventi il nodo k_i in comune, si congiungano correttamente, ovvero che tutte le derivate di $s_w(x)$ siano continue in k_i.
Una tecnica del genere dipende molto dalla scelta della base e dei nodi; un'alternativa, adottata nei modelli additivi ed utilizzata in questa tesi, è quella di utilizzare degli spline di lisciamento che comportano l'inserimento di un parametro di penalizzazione. Passando ad un approccio non parametrico si utilizza per la stima il criterio dei minimi quadrati penalizzati:

$$\|y - X\beta\|^2 + \lambda \int_0^1 [s''(x)]^2 dx, \tag{3.6}$$

dove l'integrale delle derivata seconda della funzione penalizza il modello

mentre λ rappresenta il parametro di lisciamento;. Per maggiori dettagli sulla stima di λ tramite la convalida incrociata e per altri dettagli sugli spline di lisciamento si rimanda a Wood (2006).

Sono molte le possibili combinazioni che si possono provare per definire la spline di lisciamento più opportuna per la variabile studiata. Il considerare indipendentemente tutte le variabili da lisciare o accorparle in un'unica funzione, il tipo di base e la sua dimensione, che tipologia di spline utilizzare, penalizzata o di regressione, sono solo alcune opzioni selezionabili.
La scelta effettuata in questa tesi, per le tre variabili relative al tempo, alla temperatura ed alla temperatura ritardata, è stata di lisciare separatamente questi fattori e di utilizzare per tutti e tre delle spline cubiche con dimensione di base pari a 10.
La decisione di utilizzare le spline cubiche al posto delle thin-plate spline, o delle p-spline, oppure delle spline prodotto tensoriale o di altri tipi, è stata presa dopo aver provato diverse combinazioni ed alla fine si è optato per queste, che risultano anche più semplici nel calcolo computazionale. La scelta, invece, di utilizzare il valore di default (proposto dal software R, per dettagli si veda l'Appendice) di 10, per la dimensione della base, è legata alla grande arbitrarietà associata a questa opzione, tenendo anche conto che la dimensione della base k rappresenta il limite superiore dei gradi di libertà associabili alla variabile lisciata all'interno del modello additivo, limite pari a $k-1$. Associare ad una variabile da lisciare una dimensione di base troppo piccola rischia di limitare l'informazione contenuta nelle osservazioni, ma conferirle una dimensione troppo grande rischia di attribuire a quel fattore anche della variabilità presente nei dati ma non relativa a quella variabile, andando, probabilmente, a modificare l'interpretazione del problema. Si è voluto, dunque, evitare quest'ultima complicazione.

3.3 Core Models

Lo scopo principale dell'utilizzo dei due Core Models, che impieghiamo in questa tesi, è semplicemente quello di avere una base su cui poter confrontare

le due tecniche di misurazione dell'ozono. Confronti di questo tipo sono spesso utilizzati in epidemiologia in quanto permettono, come nel nostro caso, di predisporre e controllare tutti i fattori che influenzano il confronto.

Nel nostro caso l'uso dei Core Models è utile per poter confrontare l'utilizzo della media giornaliera o dei valori orari per lo studio dell'ozono. Quello che noi facciamo, non è stimare due modelli distinti, almeno per quanto riguarda i Core Models, ma semplicemente stimare lo stesso modello prima sui dati giornalieri e poi sui dati espansi per le ventiquattro ore. Dunque i due modelli saranno del tipo:

$$\text{Core Model 1:} \qquad \log \lambda_j = \beta_0 + \beta_1 p_j + \beta_2 f_j + \beta_3 g_j + s_1(t_j) + s_2(temp_j) + s_3(temp_j^{lag}); \quad (3.7)$$

$$\text{Core Model 2:} \qquad \log \lambda_j(h) = \beta_0 + \beta_1 p_j(h) + \beta_2 f_j(h) + \beta_3 g_j(h) + s_1(t_j(h)) + s_2(temp_j(h)) + s_3(temp_j^{lag}(h)); \quad (3.8)$$

$j = 1, \ldots, J$; $h = 1, \ldots, 24$. Quindi i due modelli sono stimati sugli stessi valori, ma il secondo utilizza ogni singolo dato replicato 24 volte.Il secondo modello è predisposto per l'inserimento dei valori orari dell'ozono.

La stima dei due modelli additivi generalizzati di Poisson con funzione legame logaritmica ($\log(\cdot)$) è stata effettuata tramite il software statistico R (maggiori dettagli nell'Appendice). I risultati computazionali dei due modelli sono riassunti nelle Tabelle 3.2, 3.3 e 3.4 per il Core Model 1, mentre per il Core Model 2 nelle Tabelle 3.5, 3.6 e 3.7.

Andando ad osservare la Tabella 3.2, dove sono raccolte le stime dei parametri relative alle variabili inserite linearmente nel modello, abbiamo una prima visione di come si comportano le variabili confondenti nei riguardi della variabile risposta relativa al numero di ricoveri. Il valore delle stime dei parametri β non è direttamente utilizzabile in quanto i parametri sono

	Stima	Std. Error	z value	Pr(>|z|)
β_0 (intercetta)	3.692701	0.027804	132.814	< 2e-16 ***
PM_{10}	0.003421	0.000774	4.420	9.86e-06 ***
Festa (=T)	0.042155	0.050202	0.840	0.40106
Giorno (=Lun)	-0.050193	0.026295	-1.909	0.05628 .
Giorno (=Mar)	-0.094633	0.027229	-3.475	0.00051 ***
Giorno (=Mer)	-0.078468	0.027762	-2.826	0.00471 **
Giorno (=Gio)	-0.087292	0.027556	-3.168	0.00154 **
Giorno (=Ven)	-0.149530	0.027396	-5.458	4.81e-08 ***
Giorno (=Sab)	-0.244439	0.027483	-8.894	< 2e-16 ***

Tabella 3.2: Coefficienti parametrici, Core Model 1

	edf	Ref.df	Chi.sq	p-value
s(Tempo)	7.452	7.952	47.29	1.29e-07 ***
s(Temperatura)	4.617	5.117	19.81	0.0015 **
s(Temperat.-rit.)	6.260	6.760	52.32	3.86e-09 ***

Tabella 3.3: Significatività approssimata dei termini lisciati, Core Model 1

stimati all'interno del predittore lineare del modello di Poisson, ma possiamo comunque interpretalo. Notiamo che l'effetto del PM_{10}, come ci aspettavamo, influisce "positivamente" sulla variabile risposta, ovvero all'aumentare della concentrazione di polveri aumenta anche il numero previsto di ricoveri. La stessa interpretazione può essere data al parametro di Festa, anche se questo risulta essere non significativo.

Spiegazione opposta invece può essere data per i parametri relativi ai giorni della settimana. Preso come valore base la domenica, i restanti parametri rappresentano lo scostamento da questo valore. Risulta così che il massimo numero di ricoveri avvenga la domenica e che ci sia una diminuzione col passare progressivo dei giorni fino al sabato che sembra essere il giorno col minor numero di ricoveri.

Delle tre variabili considerate parametricamente solo il parametro relativo a Festa risulta non significativo, ovvero nel test di verifica $H_0 : \beta_i = 0$ contro

R^2_{adj}	Dev. spiegata	UBRE	N. osservazioni
0.343	37.3%	0.5858	488

Tabella 3.4: Indici, Core Model 1

$H_1 : \beta_i \neq 0$ l'unico parametro per cui si accetta l'ipotesi H_0 è quello inerente alle festività. Non si è comunque voluto togliere la variabile seguendo la pratica corrente che prevede l'inserimento dei fattori confondenti a prescindere dalla loro significatività.

Per quanto riguarda le variabili trattate non parametricamente tramite le funzioni di lisciamento, spline cubiche, notiamo che non abbiamo un modo diretto per interpretarle, come avveniva per le variabili precedenti attraverso la stima dei coefficienti β, ma abbiamo solo la stima dei gradi di libertà associati alla variabile e la sua significatività (Tabella 3.3). Una possibile via sarebbe quella di osservare graficamente la spline stimata ma rimandiamo questa analisi al capitolo successivo in cui tratteremo il modello finale includente l'ozono. La stima dei gradi di libertà di ogni stimatore è data dalla scelta di k, dimensione della base, e dal fattore di lisciamento delineato per le spline di lisciamento ((3.6); Wood (2006)).
Per la stima del parametro di lisciamento nelle spline e per la valutazione del modello, ed il confronto con altri di diversa complessità, si possono usare due criteri: il criterio della convalida incrociata generalizzata (GCV, Generalized Cross Validation) e la stima di rischio non distorta (UBRE, Un-Biased Risk Estimate), definiti come:

$$GCV = \frac{nD}{(n - gdl)^2}; \tag{3.9}$$

$$UBRE = \frac{D}{n} + 2s\frac{gdl}{n - s}; \tag{3.10}$$

dove D rappresenta la devianza, n il numero di osservazioni, s il parametro di scala e gdl i gradi di libertà effettivi del modello, compresi quelli relativi alle variabili lisciate. Se il parametro di scala s è noto, conviene usare l'UBRE. Nel nostro caso, avendo un modello di Poisson, $s = 1$, quindi adopereremo

quest'ultimo criterio.
Il valore dell'UBRE non ci fornisce un metodo per selezionare un modello o per verificarne l'adeguatezza ma viene ampiamente usato per il confronto con altri modelli. Guardando alla sua definizione, risulta comunque preferibile avere un valore basso per questo indice.

	Stima	Std. Error	z value	Pr(>|z|)
β_0 (intercetta)	3.6970362	0.0057354	644.599	< 2e-16 ***
PM_{10}	0.0032826	0.0001603	20.483	< 2e-16 ***
Festa (=T)	0.0338572	0.0104655	3.235	0.00122 **
Giorno (=Lun)	-0.0480695	0.0054004	-8.901	< 2e-16 ***
Giorno (=Mar)	-0.0902252	0.0056128	-16.075	< 2e-16 ***
Giorno (=Mer)	-0.0772168	0.0057141	-13.513	< 2e-16 ***
Giorno (=Gio)	-0.0934877	0.0056831	-16.450	< 2e-16 ***
Giorno (=Ven)	-0.1538867	0.0056374	-27.298	< 2e-16 ***
Giorno (=Sab)	-0.2437999	0.0056497	-43.153	< 2e-16 ***

Tabella 3.5: Coefficienti parametrici, Core Model 2

	edf	Ref.df	Chi.sq	p-value
s(Tempo)	8.922	9.422	1279.0	<2e-16 ***
s(Temperatura)	8.977	9.477	675.5	<2e-16 ***
s(Temperat.-rit.)	8.954	9.454	1421.1	<2e-16 ***

Tabella 3.6: Significatività approssimata dei termini lisciati, Core Model 2

R^2_{adj}	Dev. spiegata	UBRE	N. osservazioni
0.385	38.2%	0.4585	11712

Tabella 3.7: Indici, Core Model 2

Nelle Tabelle 3.5 e 3.6 sono raccolti i valori inerenti alle variabili del secondo Core Model. Le cose dette per il primo modello analizzato sono valide anche per questo. L'interpretazione dei coefficienti delle variabili inserite linearmente e degli indici del modello (Tabella 3.7) non cambia.

Dato che lo scopo della formulazione di questi modelli additivi era quello di poter avere una base simile in cui poter inserire le osservazioni dell'ozono, possiamo fare principalmente tre osservazioni per confrontare il secondo Core Model col primo:

- gli indici di devianza, R^2 e UBRE discostano poco tra i due modelli;
- le stime dei coefficienti parametrici nei due modelli, raggruppate nella tabella (3.8), risultano molto simili tra loro; questo fattore, prevedibile data l'uguaglianza di valori utilizzati per i due gruppi di stime, ci permette un primo paragone tra i due modelli;
- i gradi di libertà delle variabili lisciate e gli standard error delle stime dei coefficienti β risultano differenti.

Tutte le differenze osservate tra i due modelli sono riconducibili al diverso numero dei dati impiegati nelle stime. In particolare, per le differenze del terzo punto, cercheremo una costante per poter calibrare i due modelli.

	Core Model Giornaliero	Core Model Orario
β_0 (intercetta)	3.692701	3.6970362
PM_{10}	0.003421	0.0032826
Festa (=T)	0.042155	0.0338572
Giorno (=Lun)	-0.050193	-0.0480695
Giorno (=Mar)	-0.094633	-0.0902252
Giorno (=Mer)	-0.078468	-0.0772168
Giorno (=Gio)	-0.087292	-0.0934877
Giorno (=Ven)	-0.149530	-0.1538867
Giorno (=Sab)	-0.244439	-0.2437999

Tabella 3.8: Stime dei coefficienti $\beta_.$ nei due Core Model

3.3.1 Ricalibrazione

Come abbiamo visto nel paragrafo precedente, i due Core Model risultano simili pressoché in tutti i loro aspetti. La stima dei parametri β, raccolta nella Tabella 3.8, ci mostra che l'artificio di espandere i dati per le $24h$ non

comporta grossi cambiamenti nella stima dell'influenza sulla variabile risposta delle variabili inserite nel modello. Lo scopo del formulare dei Core Model confrontabili era dovuto alla necessità di prepararci una basa equilibrata su cui poi poter studiare l'effetto dell'ozono in maniera più chiara.

Come era facile immaginare, dato che il calcolo è avvenuto sugli stessi valori, i due modelli sembrano già presentare caratteristiche ed indici simili, quindi sembrano essere già pronti per l'inserimento dei valori dell'ozono nelle due forme sotto studio: la media giornaliera e i valori orari.
Tuttavia, se le stime dei parametri sono molto simili, non si può dire la stessa cosa degli errori standard relativi ai parametri e, di conseguenza, della significatività di questi ultimi. Inoltre, anche i gradi di libertà stimati per i termini non parametrici ed alcuni altri fattori del modello sembrano essere diversi.

Tale differenza, è unicamente riconducibile all'artificio che abbiamo utilizzato per preparare il secondo Core Model, ottenuto replicando 24 volte le osservazioni del primo. Così facendo, abbiamo mantenuto i valori osservati variabili ed abbiamo artificialmente aumentato la numerosità: da 488 osservazioni impiegate nel primo modello (equivalente ai 552 valori giornalieri raccolti meno i valori mancanti) si è passati a 11712 valori. La conseguenza principale di questa operazione di "espansione" manuale dei dati è proprio la differenza nelle quantità che ora andremo ad analizzare.

La questione di base, infatti, è che, aumentando manualmente la numerosità n, diminuiamo artificialmente la variabilità del nostro problema. L'evidenza di questo fatto è reperibile, per esempio, sugli indici di bontà di adattamento dei modelli.

	Core Model Giornaliero	Core Model Orario
R^2_{adj}	0.343	0.385
Dev. spiegata	37.3%	38.2%
UBRE	0.5858	0.4585

Tabella 3.9: Indici relativi ai due Core Model

Nella Tabella 3.9 sono raccolti i tre indici calcolati dal software R per

la valutazione del modello. Tutti e tre mostrano, come era prevedibile, un miglioramento tra il modello giornaliero e quello orario, però la differenza, in tutti e tre, non sembra essere così netta. Infatti, l'aver espanso artificialmente la numerosità del campione su cui sono state fatte le stime non sembra comportare una netta differenza tra i modelli e quindi l'aumento di n non influisce particolarmente sulla determinazione dei tre indici posti nella tabella.
L'effetto più importante, invece, può essere visto sugli indici di variabilità degli stimatori.

Variabile	Core Model Giornaliero	Core Model Orario
Tempo (t_j)	7.45	8.92
Temperatura ($temp_j$)	4.62	8.98
Temper.-rit. ($temp_j^{lag}$)	6.26	8.95

Tabella 3.10: EDF per le variabili usate non parametricamente

L'utilizzo all'interno di un modello additivo, come i nostri Core Model, di variabili inserite non parametricamente utilizzando degli spline di regressione, comporta la stima di un certo numero di gradi di libertà da associare alla variabile lisciata. Nel nostro caso, per tutte e tre le variabili trattate non parametricamente, abbiamo utilizzato degli spline cubici come lisciatori, fissando a $k = 10$ la base degli spline. Così facendo abbiamo posto come limite massimo per i gradi di libertà stimati (EDF, estimated degrees of freedom) $k - 1 = 9$. Notiamo, osservando la Tabella 3.10, che i gradi di libertà stimati per il primo Core Model, prima colonna, sono inferiori a quelli ottenuti nel secondo, che addirittura risultano essere contenuti nella stima dal limite massimo fissato di 9.

Ancora più evidente è quanto traspare dai valori raccolti nella Tabella 3.11 dove sono stati elencati, per ogni variabile indipendente, gli standard error, nella prima colonna, i valori della statistica test e il p-value relativo, per la significatività del parametro, nella seconda e terza colonna.
Tralasciando la terza colonna, in cui notiamo solamente qualche differenza numerica e che il parametro relativo alla variabile Festa e quello per Gior-

	Std.Error		z-value		p-value	
Variabile	CM Gior.	CM Orar.	CM Gior.	CM Orar.	CM Gior.	CM Orar.
β_0	0.027804	0.0057354	132.814	644.599	$<$ 2e-16	$<$ 2e-16
PM_{10}	0.000774	0.0001603	4.420	20.483	9.86e-06	$<$ 2e-16
Festa (=T)	0.050202	0.0104655	0.840	3.235	0.40106	0.00122
Giorno (=Lun)	0.026295	0.0054004	-1.909	-8.901	0.05628	$<$ 2e-16
Giorno (=Mar)	0.027229	0.0056128	-3.475	-16.075	0.00051	$<$ 2e-16
Giorno (=Mer)	0.027762	0.0057141	-2.826	-13.513	0.00471	$<$ 2e-16
Giorno (=Gio)	0.027556	0.0056831	-3.168	-16.450	0.00154	$<$ 2e-16
Giorno (=Ven)	0.027396	0.0056374	-5.458	-27.298	4.81e-08	$<$ 2e-16
Giorno (=Sab)	0.027483	0.0056497	-8.894	-43.153	$<$ 2e-16	$<$ 2e-16

Tabella 3.11: Indicatori per le variabili utilizzate parametricamente

no(=Lun) diventano significativi nel secondo modello, passiamo ad osservare le prime due colonne.

Per studiare l'influenza dell'espansione artificiale dei dati sulle stime del modello ci basta osservare una delle prime due colonne della tabella in analisi. La relazione tra l'errore standard e il valore della statistica test è tale che ci permette di guardare una sola di esse; difatti, essendo $t = (\hat{\beta}_j - 0)/se$, ed avendo visto che le stime dei β. (Tabella 3.8) non si discostano di molto tra i due modelli, analizzare le differenze tra gli standard error o tra i valori della statistica test è indifferente.

Guardando gli errori delle stime notiamo che l'aumento di n comporta una drastica diminuzione della stima della varianza. Difatti, tutti gli standard error nel Core Model "giornaliero" risultano molto più elevati rispetto ai corrispondenti valori nel Core Model "orario". Questa diminuzione di errore provocata artificialmente dal nostro stratagemma modifica la significatività dei parametri del modello. Dunque, prima di procedere all'inserimento dell'ozono, dobbiamo trovare una costante con la quale ricalibrare i valori della statistica test del modello stimato sui valori espansi.

Si è deciso di reperire tale costante di ricalibrazione dai due Core Model precedenti, previo inserimento della variabile relativa all'ozono, dato che è inevitabile che un successivo inserimento di questa modificherebbe i valori in analisi. Quindi nel primo Core Model è stata aggiunta la variabile o_j,

media giornaliera di ozono, mentre nel secondo Core Model è stata aggiunta la variabile $o_{j,h}$ che non rappresenta i valori orari dell'ozono, rappresentati da $o_j(h)$ e che utilizzeremo solo in seguito, ma semplicemente è la media giornaliera precedentemente utilizzata, replicata per le $24h$.
La costante di ricalibrazione è semplicemente reperibile dividendo i valori delle statistiche nei due modelli. Nel seguito, essa verrà calcolata come media dei rapporti tra i valori delle statistiche test dei due modelli:

$$t = \sum_{i=0}^{m} \left(\frac{z - value_i^2}{z - value_i^1} \right) / m \tag{3.11}$$

dove i valori al numeratore appartengo al secondo modello ,modello "Orario", ed i valori al denominatore sono del primo modello, modello "Giornaliero", e con $m - 1$ numero di parametri relativi alle variabili trattate parametricamente nei modelli.

z-value:	M Gior.	M Orar.	t_i
β_0	66.56	323.65	4.86
PM_{10}	4.26	19.83	4.65
Festa (=T)	0.72	2.69	3.74
Giorno (=Lun)	-1.47	-6.99	4.75
Giorno (=Mar)	-2.96	-13.81	4.66
Giorno (=Mer)	-2.48	-11.93	4.81
Giorno (=Gio)	-2.74	-14.45	5.28
Giorno (=Ven)	-5.01	-25.27	5.04
Giorno (=Sab)	-8.89	-43.11	4.85
$\overline{O}_3$	1.11	4.77	4.28

Tabella 3.12: Calcolo delle costanti t_i di ricalibrazione

Dalla Tabella 3.12 risulta che $t = \sum t_i/m = 4.69$ con varianza di 0.18. La varianza di 0.18 è da imputare al fatto che si stanno utilizzando dei GAM, cosicché la presenza di funzioni di lisciamento non permette una stima di t più accurata. Per verificare che la variabilità della costante di ricalibrazione è dovuta alla stima delle componenti non parametriche, le stesse costanti sono state calcolate utilizzando le stime ottenute da un modello lineare generaliz-

zato (GLM) stimato sui medesimi dati e con la stessa formulazione. La stima di t, in questo caso, è risultata essere di 4.75 con varianza approssimabile a 0. Questo ha confermato la sensatezza della procedura per il calcolo di t e causa della varianza della sua stima.

Cercare di ricalibrare anche i gradi di libertà associati alle variabili lisciate non è possibile in quanto non c'è modo per rapportare le due procedure di stima e così ottenere un parametro di ricalibratura.

Ora che abbiamo la costante desiderata, almeno per i termini parametrici, questa verrà utilizzata in seguito per ricalibrare i modelli finali in cui inseriremo l'ozono orario.

Capitolo 4

Inserimento dell'Ozono

Fino ad ora, in questa tesi, abbiamo preparato gli strumenti per poter arrivare a studiare l'effetto dell'ozono ed in particolare l'influenza che esso ha sulla salute dell'uomo. Partendo dall'analisi descrittiva e dallo studio di diversi articoli sull'argomento, siamo giunti a definire quali variabili e quali fattori, relativi alla questione in analisi, inerenti ad informazioni sulle condizioni ambientali-atmosferiche o ai soggetti da considerare, dovessero essere impiegati in questo lavoro.
Abbiamo definito, successivamente, lo strumento che adopereremo per verificare questa relazione tra i ricoveri e le concentrazioni di O_3, ovvero un modello additivo generalizzato di Poisson. Da questa base abbiamo stimato due modelli, i Core Model: il primo utilizzando i dati contenenti i valori giornalieri per ogni variabile, il secondo utilizzando i medesimi valori del primo, ma espansi per 24 volte, a simulare osservazioni orarie. Trovata la costante t per calibrare i due modelli tra di loro, abbiamo così ottenuto una base in cui poter inserire, rispettivamente nel primo e nel secondo modello, i valori medi giornalieri di ozono e i valori orari della concentrazione dell'inquinante sotto studio.
Procederemo ora ad inserire l'O_3 all'interno dei modelli e valuteremo la bontà del metodo utilizzato per il trattamento delle osservazioni dell'inquinante.

4.1 I Modelli finali

I due modelli definitivi che andremo infine a stimare hanno l'espressione seguente:

$$\text{Modello "Giornaliero":} \qquad \log \lambda_j = \beta_0 + \beta_1 p_j + \beta_2 f_j + \beta_3 g_j + s_1(t_j) + s_2(temp_j) + s_3(temp_j^{lag}) + \beta_4 o_j; \quad (4.1)$$

$$\text{Modello "Orario":} \qquad \log \lambda_j(h) = \beta_0 + \beta_1 p_j(h) + \beta_2 f_j(h) + \beta_3 g_j(h) + s_1(t_j(h)) + s_2(temp_j(h)) + s_3(temp_j^{lag}(h)) + s_4(o_j(h)); \quad (4.2)$$

dove nel primo modello, è stato aggiunto, rispetto al Core Model, il fattore relativo alla media giornaliera di ozono (o_j), mentre nel secondo sono stati inseriti i singoli valori orari($o_j(h)$) di O_3.
La scelta del modo in cui i due tipi di indice vengono inseriti nei due modelli è chiara ed è simile a come sono state trattate precedentemente le altre variabili nei Core Model ed ora nei modelli definitivi (comunque basati sui primi). Si è deciso di impiegare ed inserire diversamente i dati sull'ozono differenziando la tecnica nei due modelli. Nel modello "Giornaliero", le medie giornaliere di ozono vengono trattate normalmente e tramite l'utilizzo del coefficiente β_4 vengono inserite parametricamente e linearmente nel modello additivo. Nel modello "Orario", invece, si è scelto di percorrere la via non parametrica. Quest'ultima scelta è stata fatta allo scopo di utilizzare la maggiore flessibilità di una spline per cogliere con maggiore effetto la grande variabilità di valori riscontrati nell'arco delle giornate estive milanesi sotto studio.
L'idea di utilizzare le due modalità differenti è stata pensata proprio alla luce della differenza di valori raccolti per i due modelli. Impiegare una spline anche per i dati giornalieri avrebbe solo convogliato nella variabile o_j molta variabilità dei dati non associabile alla media giornaliera di ozono. Viceversa

l'aver utilizzato le osservazioni orarie di O_3 nel modello "Orario" in maniera lineare non avrebbe colto a pieno l'andamento altalenante dei valori. Inoltre, la presenza di molte osservazioni comporta spesso la difficoltà di utilizzare un approccio parametrico al problema, dato che è difficile ottenere il vero modello di rappresentazione del caso sotto studio. Non abbiamo questa difficoltà utilizzando un approccio non parametrico, tramite le spline, lasciando i dati più liberi in una struttura meno rigida.
La scelta della spline di lisciamento ricade, come per le altre variabili trattate analogamente, su una spline cubica con dimensione di base pari a 10. Esattamente come per le impostazioni relative a s_1, s_2 e s_3, si è optato per una spline cubica dopo diverse prove; mentre la preferenza per il valore della base pari a 10 è dovuta al non voler associare troppa variabilità dei dati nella variabile dell'ozono.

Gli indici relativi ai due modelli stimati sono raccolti nelle tabelle che seguono. In primo luogo, osserviamo la Tabella 4.1 relativa agli indici dei due modelli. Come per il confronto tra i due Core Model, non diamo molto peso alle differenze presenti tra gli indici di bontà di adattamento; il secondo modello sembra essere migliore ma non ci baseremo su questi valori per determinare se l'utilizzo di dati orari sia o no appropriato.
Maggior rilievo verrà dato alle altre tabelle del paragrafo. Sono stati raccolti i risultati ottenuti sul modello "Giornaliero" in Tabella 4.2, mentre quelli relativi al modello "Orario" nella Tabella 4.3.

	Modello "Giornaliero"	Modello "Orario"
R^2_{adj}	0.344	0.386
Dev. spiegata	37.4%	38.4%
UBRE	0.58777	0.45623

Tabella 4.1: Indici relativi ai due Modelli definitivi

Tralasciando per ora l'ozono, l'interpretazione dei parametri relativi alle variabili trattate linearmente è simile, per entrambi i modelli, a quella data per i Core Model. La stima dei coefficienti β ci suggerisce che al crescere della concentrazione di PM_{10} cresca anche il valore atteso del numero di ricoveri

	Coefficienti parametrici			
	Stima	Std. Error	z value	Pr(>|z|)
β_0 (intercetta)	3.6403982	0.0546939	66.559	< 2e-16 ***
PM_{10}	0.0033205	0.0007793	4.261	2.04e-05 ***
Festa (=T)	0.0363768	0.0504836	0.721	0.47118
Giorno (=Lun)	-0.0407512	0.0276852	-1.472	0.14103
Giorno (=Mar)	-0.0849324	0.0286403	-2.965	0.00302 **
Giorno (=Mer)	-0.0709585	0.0286345	-2.478	0.01321 *
Giorno (=Gio)	-0.0785561	0.0286725	-2.740	0.00615 **
Giorno (=Ven)	-0.1417942	0.0282901	-5.012	5.38e-07 ***
Giorno (=Sab)	-0.2441349	0.0274743	-8.886	< 2e-16 ***
$\overline{O}_3$	0.0006286	0.0005638	1.115	0.26490

	Termini lisciati			
	edf	Ref.df	Chi.sq	p-value
s(Tempo)	7.442	7.942	46.59	1.74e-07 ***
s(Temperatura)	4.515	5.015	20.08	0.00122 **
s(Temperat.-rit.)	6.316	6.816	48.88	1.97e-08 ***

Tabella 4.2: Fattori relativi al Modello "Giornaliero"

e lo stesso vale per il fattore festa (in una giornata di festa il valore atteso della variabile risposta cresce rispetto ad una giornata normale). In maniera opposta, rispetto alla giornata di domenica, il valore atteso λ_j di Y_j decresce nei diversi giorni della settimana.

I valori dei test relativi ai coefficienti β nel secondo modello, Tabella 4.3, sono stati ricalibrati, come era stato detto nel paragrafo precedente, per compensare l'aumento della numerosità artificiale condotta per poter utilizzare i dati orari dell'ozono. Quindi, gli errori standard e i valori dei test (z-value), per la significatività dei parametri, siano stati ricalibrati utilizzando la costante t ottenuta in precedenza. Per ricalibrare i valori dei test basta dividerli per la costante t; per ricalibrare gli errori standard basta moltiplicarli per la stessa costante t.
Fatto questo procedimento abbiamo ottenuto i parametri corretti, relativi ai coefficienti. Successivamente sono stati ricalcolati anche i livelli di si-

	Coefficienti parametrici			
	Stima	Std. Error	z value	Pr(>|z|)
β_0 (intercetta)	3.695701	0.02706599	136.539232	< 2e-16 ***
PM_{10}	0.003303	0.00075509	4.373561	1.22e-05 ***
Festa (=T)	0.033423	0.04932004	0.677612	0.489
Giorno (=Lun)	-0.047042	0.02547608	-1.846482	0.0642 .
Giorno (=Mar)	-0.089749	0.02646098	-3.391898	0.00069 ***
Giorno (=Mer)	-0.076714	0.02691122	-2.850533	0.0043 **
Giorno (=Gio)	-0.092627	0.02677052	-3.459915	5.4e-04 ***
Giorno (=Ven)	-0.152599	0.02655009	-5.747761	< 2e-16 ***
Giorno (=Sab)	-0.243553	0.02651726	-9.184435	< 2e-16 ***

	Termini lisciati			
	edf	Ref.df	Chi.sq	p-value
s(Tempo)	8.917	9.417	1263.4	< 2e-16 ***
s(Temperatura)	8.978	9.478	683.2	< 2e-16 ***
s(Temperat.-rit.)	8.952	9.452	1407.4	< 2e-16 ***
s($O_3(h)$)	7.773	8.273	41.2	2.45e-06 ***

Tabella 4.3: Fattori relativi al Modello "Orario"

gnificatività osservati relativi ai test sulla significatività dei parametri, ovvero i test $H_0 : \beta_i = 0$ contro $H_1 : \beta_i \neq 0$. Il suddetto test bilaterale di valore z si distribuisce approssimativamente come $N(0,1)$, quindi si ha $\alpha^{oss} = 2min(\Phi(z - value); 1 - \Phi(z - value))$ con Φ funzione di ripartizione della Normale. Risulta quindi, come era avvenuto nei Core Model, che il parametro relativo a festa e quello relativo al primo giorno della settimana non risultano significativi.

Per riuscire a capire come vengono trattate le variabili lisciate, ci avvaliamo dei grafici riprodotti in Figura 4.1. I grafici sono relativi al primo modello, rimandando nel paragrafo successivo l'analisi dell'ozono lisciato.
La prima cosa che notiamo è che tutte e tre le curve, costruite mettendo sull'asse delle ascisse i valori della variabile ordinati in maniera crescente e sull'asse dell'ordinate i valori lisciati della variabile, sono centrate in 0. Infatti

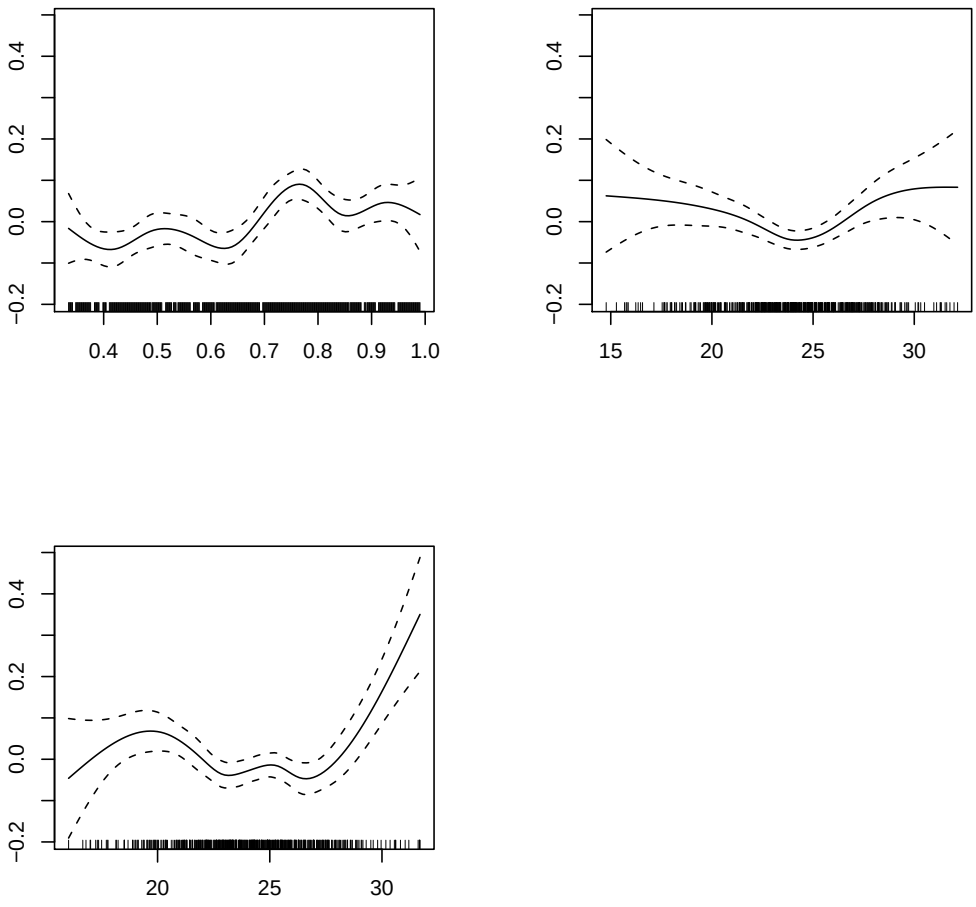


Figura 4.1: Grafici delle spline di lisciamento nel primo modello. In senso orario da in alto a sinistra: s(Tempo), s(Temperatura) e s(Temperat.-rit.).

in un modello additivo generalizzato del tipo:

$$g(E[Y|x_1, \ldots, x_p]) = \beta_0 + \sum_{j=1}^{p} s_j(x_j),$$

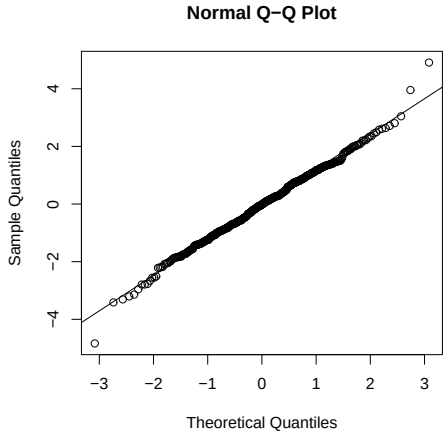
in cui vengono sommate le spline relative alle diverse variabili esplicative considerate per evitare problemi di identificabilità del modello, viene posto il vincolo che ogni s_j abbia valori centrati sullo 0 (Azzalini e Scarpa (2004)):

$$\sum_{i=1}^{n} s_j(x_{ij}) = 0.$$

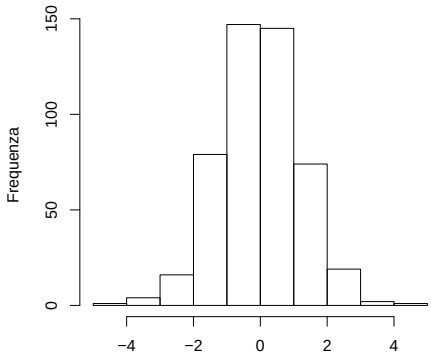
Ad ogni curva stimata è aggiunta la corrispondente banda di variabilità, che assume un ruolo simile a quello dell'intervallo di confidenza per la stima della funzione $s.(\cdot)$. Queste curve costituiscono la stima delle spline di lisciamento per tutte e tre le variabili trattate non parametricamente nel modello costruito con i dati giornalieri (le curve ottenute per il secondo modello sui dati orari non cambieranno di molto). Tutte e tre le curve rappresentano l'andamento dei dati registrati, riuscendo ad inserire nel modello l'effetto di questi valori con maggiore libertà.

Prima di analizzare più nel dettaglio i risultati inerenti l'ozono, passiamo ad osservare velocemente i residui del modello. In Figura 4.2 sono riportati i grafici relativi ai residui di devianza calcolati sui modelli stimati.

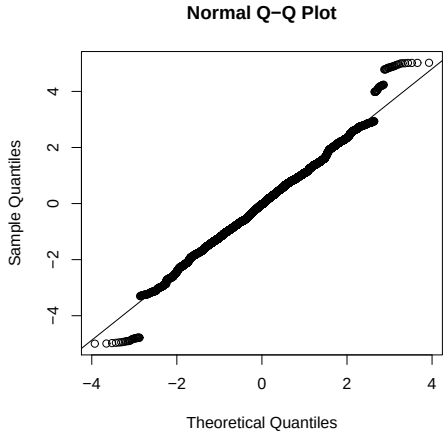
Per entrambi i modelli, i grafici sembrano suggerirci che l'assunzione di normalità dei residui sembra abbastanza soddisfatta. Concentrando l'attenzione principalmente sui secondi due grafici, quelli relativi al modello con i dati espansi e con l'ozono orario, notiamo un andamento soddisfacentemente lineare (nel grafico Quantile-Quantile), ma con la presenza di problemi sulle code. L'istogramma ci conferma quanto detto prima: l'assunzione di normalità dei residui sembra soddisfatta ma sono molti i residui presenti sulle code del grafico. Questo fattore è facilmente imputabile all'espansione dei dati e alla composizione del modello additivo tramite la combinazione di variabili trattate parametricamente e non.



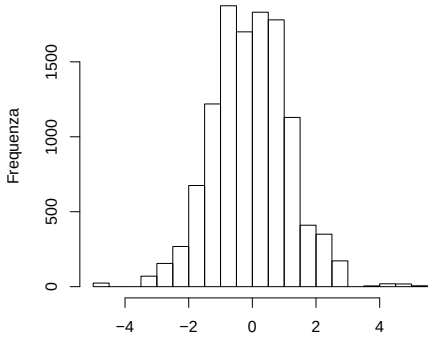
(a) Modello "Giornaliero"



(b) Modello "Giornaliero"



(c) Modello "Orario"



(d) Modello "Orario"

Figura 4.2: Analisi dei residui di devianza per i due modelli: grafico Quantile-Quantile e Istogramma delle frequenze

Nonostante lo scopo di questa tesi non sia trovare un modello "perfetto" per i dati analizzati, ma piuttosto quello di verificare l'utilità dell'inserimento dei dati orari dell'ozono, l'analisi dei residui ci suggerisce che la formulazione alla fine scelta è comunque abbastanza soddisfacente.

4.2 Risultati sull'Ozono

Passando ora a parlare dell'oggetto di studio di questa tesi, la prima cosa da ricordare è la diversa natura dell'utilizzo della variabile O_3 nei due modelli. Come avevamo già detto ed esplicitato nelle formule (4.1) e (4.2), l'ozono compare tra le variabili impiegate parametricamente nel primo modello (Tabella 4.2) e tra quelle utilizzate non parametricamente nel secondo (Tabella 4.3).

Osservando il primo modello, ci concentriamo principalmente su due elementi: la stima del coefficiente β e la sua significatività. Inserendo le medie giornaliere della concentrazione di O_3, rilevate nella città di Milano nei mesi estivi tra il 1998 e il 2003, notiamo che la stima del coefficiente risulta essere $\beta_4 = 0.00063$. Un coefficiente positivo implica che il crescere del valore medio delle concentrazione di ozono comporta un aumento del numero atteso di ricoveri per problemi alle vie respiratori. Questo risultato è atteso ed, in particolare, rispecchia quanto è noto a livello teorico sull'effetto dell'inquinante ozono sulla salute dell'uomo. Inoltre, e di maggior rilievo per questa tesi, notiamo la non significatività del parametro. Con un livello di significatività osservato di 0.2649 non si rifiuta l'ipotesi di nullità dell'effetto, quindi si ottiene che l'effetto dell'ozono non è statisticamente significativo.
In conclusione, il modello "Giornaliero" ci suggerisce che l'effetto dell'ozono, valutato tramite la media giornaliera, segnala un aumento del numero di ricoveri all'aumentare della concentrazione di O_3, ma tale effetto non risulta essere significativo.

Nel secondo modello, la variabile che esprime le osservazioni orarie dell'ozono è inserita non parametricamente. Come abbiamo detto, si è voluto utilizzare una spline cubica per esprime al meglio l'andamento dei valori. La prima cosa che possiamo notare è la significatività del lisciatore. A differenza

del caso precedente, ora risulta che l'effetto dell'ozono, sotto forma di rilevazioni orarie lisciate, è statisticamente significativo. L'interpretazione diretta del fattore è difficile, in quanto i valori della variabile si trovano all'interno della funzione di lisciamento. Come per le altre variabili trattate non para-

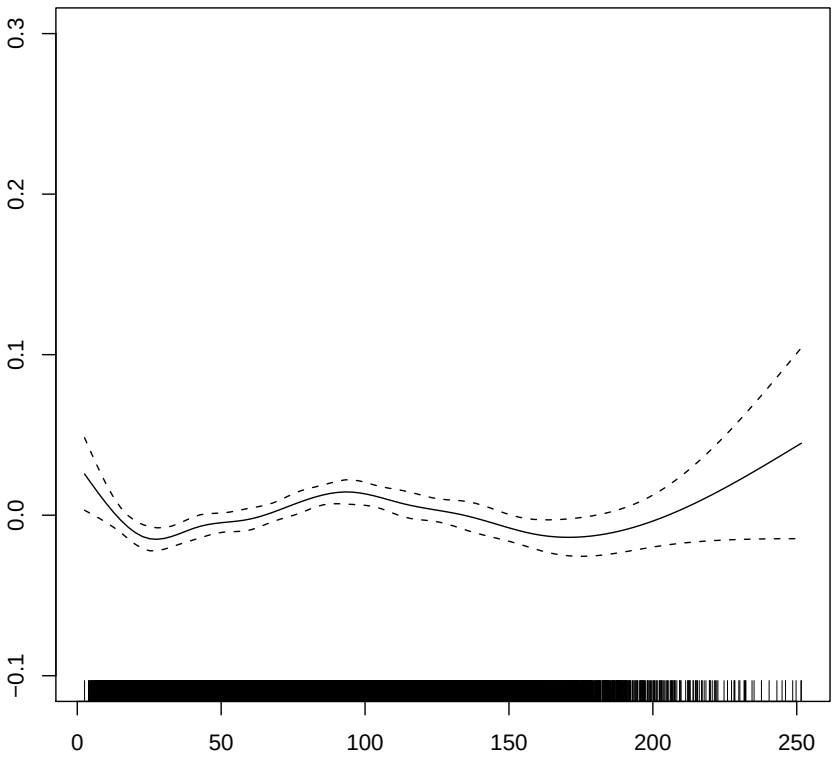


Figura 4.3: Grafico della spline di lisciamento di O_3

metricamente, osserviamo in Figura 4.3 la stima della funzione di lisciamento dell'ozono. Come avevamo detto per la Figura 4.1, anche in questo grafico la curva rappresentata costituisce il modo in cui l'effetto dei dati orari della concentrazione di ozono sono stati inseriti all'interno del modello.

Andiamo ora ad osservare l'andamento lisciato dell'ozono all'interno del

modello. Abbiamo già visto che l'elemento s($O_3(h)$) risulta significativo all'interno del modello e che i gradi di libertà stimati risultano essere 7.773. Cercheremo ora di verificare se la funzione di lisciamento sia riuscita, all'interno del modello additivo, a cogliere l'andamento dei valori orari.

Il grafico in Figura 4.4 è stato ottenuto prendendo i valori lisciati della

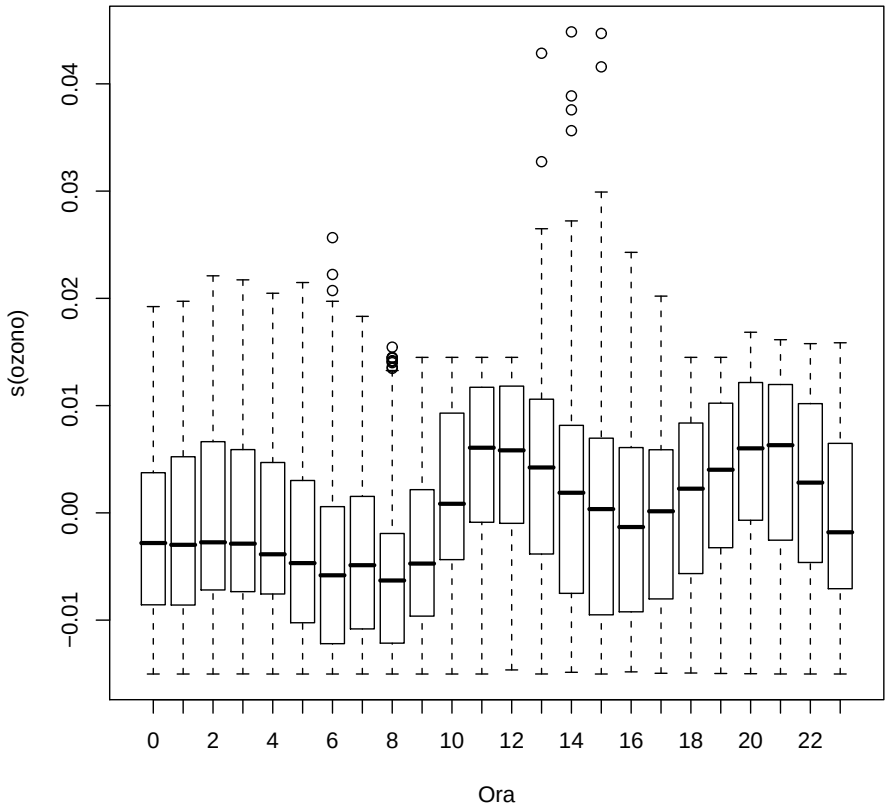
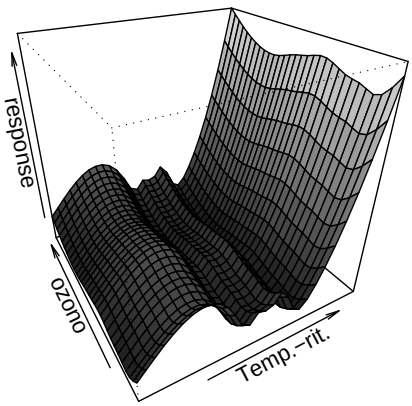


Figura 4.4: Valori lisciati dell'ozono suddivisi per ora

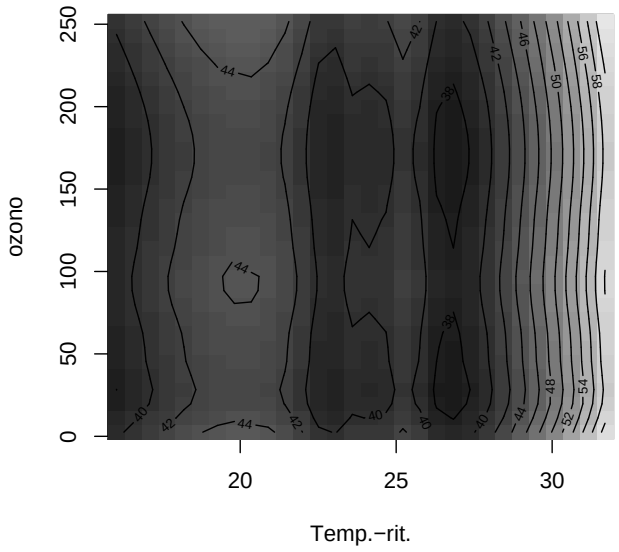
concentrazione oraria di ozono divisi per ora. Ogni diagramma a scatola rappresenta il gruppo di valori lisciati di O_3 per quella determinata ora. Il risultato ideale che potevamo ottenere era un grafico simile a quello riportato in Figura 3.1 in cui veniva sottolineata la crescita di concentrazione dell'inquinante durante le ore centrali con un abbassamento verso il termine

del giorno ed un minimo nelle prime ore del mattino (circa tra le 06:00 e le 08:00). I risultati ottenuti in Figura 4.4, invece, rappresentano quanto è stato colto dal nostro GAM.
Tenendo conto che tra i due grafici va confrontato l'andamento e non i valori, dato che uno rappresenta la concentrazione effettiva rilevata mentre il secondo è formato dai valori lisciati centrati in zero, possiamo notare che la funzione di lisciamento sembra cogliere abbastanza bene l'andamento orario dell'ozono. In particolare, constatiamo che l'andamento per le prime ore (circa dalle 00:00 alle 11:00) è pressoché uguale tra i due grafici, con un minimo attorno alle 06:00 del mattino e con un rialzamento dei valori al passare delle ore e all'avvicinarsi delle ore centrali in cui l'attività solare è maggiore. Riscontriamo inoltre che, durante le ore pomeridiane, la spline non sembra crescere gradatamente per poi decrescere verso le ore serali come nel grafico relativo ai valori reali, ma pare avere un calo attorno alle 16:00 ed un altro definitivo attorno alle tarde ore della sera (22:00-23:00), come nel primo grafico. Questa unica differenza tra i due grafici, nelle prime ore pomeridiane, è accompagnata, in Figura 4.4, dalla presenza di alcuni valori anomali, e da un'ampia estensione dei baffi dei diagrammi a scatola relativi alle ore pomeridiane (13:00-17:00). Queste ultime osservazioni ci fanno pensare che l'utilizzo delle osservazioni orarie, e l'impiego di spline cubiche per rappresentarle all'interno di un modello additivo generalizzato, sia una soluzione valida per rappresentare l'effetto che l'ozono ha sulla salute degli esseri umani.

Per osservare l'effetto dell'ozono sulla variabile risposta possiamo guardare i grafici in Figura 4.5. In queste immagini si è voluto riportare, come esempio, l'effetto congiunto dell'ozono e della temperatura media, calcolata sui valori dei tre giorni precedenti l'osservazione normale, sul valore predetto dal modello del numero di ricoveri. Abbiamo voluto utilizzare per l'esempio la temperatura, in quanto è ben noto l'effetto che questa ha sull'ozono.
Quello che si nota, graficamente nel primo grafico e numericamente nel secondo, è che la curva lisciata dell'ozono si alza al crescere della temperatura. In particolare ai valori più alti di temperatura corrisponde una crescita drastica del numero di ricoveri. A questo proposito, osservando le curve di livello,



(a) Grafico in 3 Dimensioni



(b) Curve di Livello

Figura 4.5: Effetto di Ozono e Temperatura-rit sulle previsioni del numero di ricoveri

notiamo che dai 27° inizia un innalzamento delle curve relative dell'ozono, che comportano una crescita del valore predetto della variabile risposta. Da un valore di 40 ricoveri predetti si arriva ad un livello di 58-60 degenze legate a problemi respiratori.

Tutti questi risultati ci consentono di pensare che l'inserimento di valori orari dell'ozono comporti reali miglioramenti nel modello additivo ed inoltre ci portano a decidere che la tecnica di lisciare, tramite spline, le rilevazioni di O_3 sia una tecnica appropriata per affrontare il problema in questione.

Capitolo 5

Conclusioni

Il modello finale, ottenuto ed elaborato nell'ultimo capitolo, per lo studio della relazione tra ozono e salute dell'uomo, mette in evidenza che la scelta di impiegare le rilevazioni orarie dell'ozono porta a buoni risultati rispetto all'utilizzo di indici giornalieri.
Le analisi compiute si sono focalizzate sulla concentrazione oraria di O_3 per riuscire a cogliere l'effetto ben noto in termini medici sulla salute dell'uomo, ed in particolare sui problemi legati alle vie respiratorie. Per farlo abbiamo utilizzato, come esempio di studio, i dati raccolti nella città di Milano nei periodi estivi tra il 1998 e il 2003.
Presa come variabile risposta il conteggio di ricoveri legati a problemi respiratori, abbiamo stimato un modello additivo generalizzato nel quale abbiamo inserito un insieme di variabili esplicative confondenti, ovvero variabili che delineano il problema e che "precisano" il caso sotto studio. Successivamente, abbiamo inserito l'ozono, prima parametricamente utilizzando le medie giornaliere, poi non parametricamente, tramite una funzione di lisciamento, impiegando le rilevazioni orarie.
Anche altre vie sono state provate impiegando altri indici come il massimo e la mediana delle osservazioni, come indici giornalieri; inoltre si sono osservati i risultati in Chiogna e Pauli (2008) relativi a indici quali l'intensità, la durata e l'esposizione notturna che già fornivano maggiore informazione rispetto ad un singolo indicatore giornaliero.

Il confronto tra tutte queste soluzioni ci porta ad affermare che il modo più efficiente per trattare l'ozono è quello di sfruttare al massimo la diversità di valori che si ottengono durante l'arco di una giornata. Difatti abbiamo visto, graficamente e numericamente, quanto l'intensità dell'attività solare influisca sui livelli di concentrazione di ozono portando a forti disparità di valori di concentrazione dell'inquinante tra un'ora ed un'altra delle 24 ore giornaliere. Riuscire a cogliere al massimo questa disparità di livelli di concentrazione è la via migliore per cogliere l'effetto dell'ozono sulla salute umana.

In conclusione, siamo riusciti a verificare in questa tesi, utilizzando i dati sull'inquinamento di Milano, che un metodo efficace, per ovviare al problema di trovare un buon indice per l'ozono, è quello di incrementare le informazioni impiegate relative all'inquinante. Ovvero, abbandonando l'utilizzo di indici giornalieri, come la media o il massimo, ed adottando indici più specifici che riescano a cogliere maggiormente la grande diversità di valori registrabili nell'arco delle 24 ore giornaliere, si riesce ad esprimere in maniera significativa l'effetto che l'O_3 ha sull'uomo. Partendo dalla media, sintesi giornaliera delle informazioni sull'ozono, siamo passati ad aumentare l'informazione sull'inquinante sotto studio tramite gli indici suggeriti da Chiogna e Pauli (2008): intensità, durata ed esposizione notturna. Già in questo passaggio, si sono ottenuti dei miglioramenti importanti nello studio dell'O_3 e della sua relazione con i problemi respiratori dell'uomo. Infine siamo passati ad aumentare maggiormente l'informazione, impiegando tutte le rilevazioni orarie registrate e, tramite appositi lisciatori, le abbiamo inserite in un modello additivo.

$$\begin{aligned}\log \lambda_j(h) = \beta_0 + \beta_1 p_j(h) + \beta_2 f_j(h) + \beta_3 g_j(h) + \\ s_1(t_j(h)) + s_2(temp_j(h)) + s_3(temp_j^{lag}(h)) + s_4(o_j(h));\end{aligned} \tag{5.1}$$

Il modello finale di questa tesi è quello posto in (5.1). In esso abbiamo inserito l'ozono considerando tutte le sue rilevazione registrate ad ogni ora ed abbiamo trattato questi dati tramite una spline cubica. Con questa operazione abbiamo potuto verificare definitivamente che le migliorie apportate da quest'ultimo modo di trattare l'ozono risultano essere significative al fine di rappresentare le concentrazioni di ozono nella relazione tra la salute

respiratoria umana e l'inquinamento atmosferico.

Appendice A

L'utilizzo di R

Per tutte le analisi svolte in questa tesi ci si è serviti del software statistico Open Source **R**.
Lo studio dei dati della città di Milano sull'inquinamento aereo e le analisi grafiche e descrittive sono state fatte con i comandi base forniti dal software e non saranno riportati qui di seguito.
Lo scopo di questa sezione della tesi è quello di sottolineare e riportare le tecniche utilizzate, computazionalmente, per la stima dei modelli additivi.
Per ottenere dei GAM si è visto che bisogna poter usufruire di metodi diversi da quelli che necessiteremmo utilizzando dei modelli più classici come quelli lineari. La necessità di impiegare spline e funzioni di lisciamento varie all'interno del modello, e tutti i calcoli che ne derivano, esige strumenti adatti.
Una caratteristica di rilievo di **R** è quella di poter utilizzare delle librerie che implementano le capacità del software permettendo l'impiego di di comandi adeguati per le diverse necessità e finalità da soddisfare. Per la stima e lo studio dei modelli additivi generalizzati abbiamo la possibilità di usufruire di tre librerie: la *gss*, scritta da Chong Gu, la *gam*, di Trevor Hastie, e la *mgcv*, sviluppata da Simon Wood.
Per questa tesi ci siamo serviti dalla libreria sviluppata da Simon N. Wood: la **mgcv**. Questa libreria utilizza delle procedure, molto simili a quelle implementate in *gam*, per stima di modelli additivi basandosi sull'utilizzo di spline di regressione penalizzate (di lisciamento), con selezione automatica

del parametro di lisciamento tramite GCV o UBRE.

Tralasciando i Core Model, passiamo direttamente mostrare la formulazione dei due modelli di finali: quello "Giornaliero" e quello "Orario".

```
conf.G<-paste("s(tempo.est,bs='cr',k=10)+s(temmean,bs='cr',k=10)
              +s(temmeanlagged,bs='cr',k=10)+pm10.AVE
              +as.factor(festa)+as.factor(wday)")
ozone.G<-paste("o3.AVE")
finale.G<-paste("n~",conf.G,"+",ozone.G)
mod.G<-gam(formula(finale.G),family=poisson,data=giornalieri.estate)

conf.O<-paste("s(tempo.est,bs='cr',k=10)+s(temmean.e,bs='cr',k=10)
              +s(temmeanlagged.e,bs='cr',k=10)+pm10.AVE.e
              +as.factor(festa.e)+as.factor(wday.e)")
ozone.O<-paste("s(o3_or,bs='cr',k=10)")
finale.O<-paste("n.e~",conf.O,"+",ozone.O)
mod.O<-gam(formula(finale.O),family=poisson,data=orari.estate)
```

Le variabili utilizzate sono suddivise tra confondenti ed ozono e poi unite nella formula finale. In entrambi i casi la formulazione effettiva avviene tramite il comando `gam` in cui notiamo la specificazione del dataset, l'inserimento della formula e, soprattutto, la definizione della famiglia del modello. Per questi casi è stata imposta la scelta della famiglia Poisson che prevede, di default, l'impiego della funzione legame logaritmica.
La caratteristica principale della formulazione è legata al modo in cui vengono inserite le variabili all'interno della formula. I due modelli, a parte l'ozono, sono uguali e vediamo che nel secondo vengono inserite le medesime variabili del primo con l'aggiunta del `.e` che indica l'espansione per le 24 ore della variabile.
Maggiore attenzione richiedono i comandi per trattare le variabili non parametricamente. Prendiamo, ad esempio la variabile relativa al tempo:

```
s(tempo.est,bs='cr',k=10).
```

Il comando `s()` indica l'utilizzo di una spline di lisciamento per la variabile in esame. Al suo interno viene specificata la base desiderata e la tipologia di spline. Con `k=10`, poniamo a dieci il numero della dimensione di base e

con `bs='cr'` selezioniamo l'utilizzo di una spline cubica come funzione di lisciamento.
Per la verifica della bontà di adattamento dei modelli ci si è serviti dei comandi:

```
summary.gam(mod.O)
gam.check(mod.O)
qqnorm(residuals.gam(mod.O,type='deviance'))
hist(residuals.gam(mod.O,type='deviance')).
```

Con il primo comando, otteniamo l'output del modello stimato,con le stime dei parametri e la loro significatività, i termini non parametrici e gli indici del modello. Con i restanti tre comandi, abbiamo analizzato i residui di devianza del modello.
Per la costruzione del grafico di verifica (Figura 4.4) costruita usando i valori lisciati dell'ozono e la divisione oraria, ci si è serviti dei seguenti comandi, partendo dal modello "Orario" presentato precedentemente (`mod.O`):

```
raw <- mod.O$model[mod.O$smooth[[4]]$term]
xx2 <- sort(mod.O$model[mod.O$smooth[[4]]$term][[1]])
dat2 <- data.frame(x = xx2)
names(dat2) <- mod.O$smooth[[4]]$term
X2 <- PredictMat(mod.O$smooth[[4]], dat2)
first <- mod.O$smooth[[4]]$first.para
last <- mod.O$smooth[[4]]$last.para
p <- mod.O$coefficients[first:last]
fit2 <- X2 %*% p
pd.item2 <- list(fit = fit2, dim = 1, x = xx2, raw = raw[[1]])

ordine<-order(mod.O$model[mod.O$smooth[[4]]$term][[1]])
ora.ord<-mod.O2$model$ora[ordine]
s_oz.ora<-data.frame(ora.ord,fit2)
s_oz.ora

plot(factor(s_oz.ora$ora.ord),s_oz.ora$fit2).
```

Tramite questi comandi, partendo dalle stime effettuate dal comando `gam`, si sono calcolati i valori orari dell'ozono lisciati tramite la spline cubica e si

sono suddivisi nelle diverse ore giornaliere.
Infine, per la creazione degli ultimi grafici (Figura 4.5) in cui viene rappresentato l'effetto che le spline stimate, relative all'ozono e alla temperatura ritardata, hanno sulla variabile risposta, ci si è serviti dei comandi seguenti:

```
vis.gam(mod.0,view=c("temmeanlagged.e","o3_media"),
phi=30,theta=-30,color="gray",type='response')

vis.gam(mod.0,view=c("temmeanlagged.e","o3_media"),
plot.type='contour',color="gray",type='response')
```

Il primo produce il grafico in tre dimensioni, il secondo quello con le curve di livello.

Bibliografia

[1] Abbey, D.E, Burchette, R.J. (1996). Relative power of alternative ambient air pollution metrics for detecting chronic health effects in epidemiological studies. *Environmetrics* **7**, 453–470.

[2] Azzalini, A., Scarpa, B. (2004). *Analisi Dei Dati E Data Mining.* Springer.

[3] Bell, M., Dominici, F., Samet, J. (2005). A Meta-Analysis of Time-Series Studies of Ozone and Mortality With Comparison to the National Morbidity, Mortality, and Air Pollution Study. *Epidemiologys* **16(4)**, 436–445.

[4] Chiogna, M., Bellini, P. (2002). Alternative air pollution measures for detecting short-term health effects in epidemiogical studies. *Environmetrics* **13**, 55–69.

[5] Chiogna, M., Pauli, F. (2008). Short term ozone effects on morbidity for the city of Milano, Italy, 1996-2003. *Working Paper Series* **2**, January 2008.

[6] Hastie, T.J., Tibshirani, R.J. (1990). *Generalized additive models.* Chapman & Hall.

[7] Iacus, S., Masarotto, G. (2003). *Laboratorio Di Statistica Con R.* Wiley & Son.

[8] Ito, K., De Leon, S., Lippmann, M.(2005). Associations Between Ozone and Daily Mortality: Analysis and Meta-Analysis. *Epidemiologys* **16(4)**, 446–457.

[9] Levy, J., Chemerynski, Sarnat, J.(2005). Ozone Exposure and Mortality: An Empiric Bayes Metaregression Analysis. *Epidemiologys* **16(4)**, 458–468.

[10] Martuzzi, M., Mitis, F., Iavarone, I., Serinelli, M. (2006). Health Impact Of PM_{10} And Ozone In 13 Italian Cities. *World Health Organization.*

[11] Medina-Ramo, M., Zanobetti, M., Schwartz, J. (2006). The Effect of Ozone and PM10 on Hospital Admissions for Pneumonia and Chronic Obstructive Pulmonary Disease: A National Multicity Study. *American Journal of Epidemiology* **163(6)**, 579–588.

[12] Sauerbrei, W. (1999). The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics* **48**, 313–329.

[13] Wasserman, Larry (2006). *All of nonparametric models.* McGraw-Hill.

[14] Wood, S.N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society (B)* **62(2)**, 413–428.

[15] Wood, S.N. (2001). mgcv: GAMs and Generalized Ridge Regression for R. *R News* **1(2)**, 20–25.

[16] Wood, S.N. (2006). *Generalized additive models an introduction with R.* Chapman & Hall.