Indice

Riassunto

Scopo del presente studio è lo sviluppo ed il confronto di alcuni modelli di dipendenza, atti a spiegare al meglio il comportamento dell'ozono, misurato nell'area urbana di Udine presso tre stazioni di rilevamento della qualità dell'aria, descritte nel primo capitolo, sulla base della conoscenza dei valori simultanei e precedenti dei restanti parametri misurati (variabili meteorologiche ed altri inquinanti).

L'analisi ha riguardato il periodo 2000 – 2005; per ciascun anno di dati, si è presa in considerazione la *Ozone season*, identificata nel periodo che va da aprile a settembre. Ci si concentra, in particolare, sul valore massimo della media trascinata di 8 ore per ogni giorno, rilevante ai sensi della normativa vigente, e sul massimo del valore orario giornaliero di ozono. Nel capitolo 2 viene presentata l'analisi preliminare dei dati; si stabiliscono, tra l'altro, alcuni standard sul formato dei dati, originariamente orari ma trasformati in giornalieri. Si è proceduto ad una verifica della qualità delle serie storiche, con particolare attenzione alla distribuzione dei dati mancanti ed alla presenza di discontinuità nelle serie. Sono state analizzate le distribuzioni delle variabili, utilizzando metodi grafici (boxplot ed istogrammi), indici di simmetria e di curtosi, test di normalità. I risultati delle analisi preliminari condotte, hanno tra l'altro suggerito di utilizzare la trasformazione radice quadrata per l'ozono. Infatti in questa maniera si ottiene una distribuzione maggiormente simmetrica che dovrebbe facilitare la modellazione successiva. È' stata eseguita un'analisi di tipo autoregressivo sulle serie disponibili, per riconoscere fino a che punto ogni variabile dovesse essere ritardata. Si è alla fine ritenuto appropriato considerare ritardi fino a

tre giorni nell'analisi dell'ozono, mentre i precursori e le variabili meteorologiche, sono state ritardate di un giorno.

Per ogni centralina, i dati relativi alle variabili che riassumono l'ozono sono stati raccolti in quattro gruppi (basso, medio, alto, altissimo) in base ai quantili della relativa distribuzione. Ove uno dei quantili fosse molto vicino al valore soglia previsto dalla legge (110 $\mu g/m^3$, la cui radice quadrata vale 10.49), è stato sostituito con tale valore, utile per un rapido riferimento alla normativa.

Nel terzo capitolo vengono illustrati i modelli statistici sperimentati: modelli lineari, alberi di classificazione e regressione e *Random Forest*.

Per ogni modello sono state eseguite delle previsioni, in un primo momento, legate ai singoli modelli:

- si sono sviluppati i modelli lineari sui primi cinque anni di dati, utilizzandoli quindi per prevedere i dati del sesto anno (2005);
- per gli alberi, le 1598 osservazioni sono state divise casualmente in due gruppi, *analisi*, comprendente 1298 osservazioni su cui è stato stimato l'albero e *test*, comprendente le restanti 300 osservazioni utilizzate per stimare la capacità previsiva del modello;
- per Random Forest, è stato stimato il modello usando solo le variabili più importanti allo scopo di abbassare l'errore di classificazione e aumentare la capacità previsiva del modello.

In seguito, al fine di confrontare le prestazioni di tutti i modelli tra di loro, viene sviluppata una previsione giorno per giorno dell'ultimo anno. I risultati ottenuti sono stati confrontati fra loro e con un modello *naive*, consistente nella ipotesi di persistenza del livello di ozono del giorno precedente e le percentuali di successo di ogni modello si possono vedere nella tabella seguente.

L'esplorazione dell'efficacia dei modelli è stata per alcuni aspetti limitata causa la mancanza di dati relativi agli indicatori della stabilità atmosferica o dell'intensità del traffico veicolare nell'area urbana di Udine.

Complessivamente per quanto riguarda il massimo giornaliero della media trascinata nelle otto ore di ozono:

- i modelli di tutti e tre i tipi segnalano l'ozono del giorno precedente come variabile essenziale. Radiazione solare globale e tempo di insolazione (eliofania) sono altamente correlate fra loro e vengono scelte alternativamente da tutti i modelli;
- *Random Forest* dà più peso rispetto agli altri anche all'ozono di due giorni precedenti, così come, più in generale, ad altre variabili ritardate;
- variabili come benzene, pressione ed umidità, che erano state segnalate dalla regressione lineare, tornano ad essere importanti in *Random Forest*;

Previsioni media trascinata "giorno per giorno"

Stazioni di :	via Cairoli	S.Osvaldo	via Manzoni
modello naive	55%	59%	59%
modello lineare	60%	71%	65%
alberi di classificazione	59%	59%	70%
Random Forest	62%	68%	70%

Previsioni massimo giornaliero "giorno per giorno"

Stazioni di:	via Cairoli	S.Osvaldo	via Manzoni
modello naive	54%	54%	50%
alberi di classificazione	56%	63%	60%
Random Forest	57%	62%	62%

- il benzene manifesta la sua importanza in tutti tre i modelli relativi ad Osvaldo. In Manzoni, stazione in cui viene rilevato, assume significatività solo nei modelli lineari e in *Random Forest*;
- per quanto riguarda il vento, la fascia oraria predominante, selezionata per tutte le stazioni, è quella 10-17. La sua importanza nel determinare valori alti di ozono si nota soprattutto negli alberi;
- anno e mese di riferimento erano importanti per il modello lineare in Osvaldo e Manzoni; non vengono utilizzati negli altri modelli, tranne che nel modello ad albero di Osvaldo.

Per quanto riguarda il massimo orario giornaliero diurno:

Le considerazioni appena svolte valgono, sostanzialmente, anche per la previsione di questa variabile; in generale, le percentuali di successo dei modelli, compreso il modello *naive*, sono inferiori, tranne che per l'albero di classificazione di Osvaldo.

Al giorno d'oggi, è ampliamente riconosciuto che le relazioni tra ozono, precursori e variabili meteorologiche sono complesse e non lineari. In generale, l'utilizzo operativo dei modelli ai fini predittivi richiede di disporre delle previsioni per alcune delle variabili esplicative. Dato il carattere preliminare di questo studio, si è ritenuto importante avviare il lavoro su variabili *misurate*, e non *previste*, per stabilire la massima efficacia previsiva ottenibile dai modelli. E' necessario attendersi dunque che, in un utilizzo operativo dei modelli qui sviluppati, le percentuali di successo siano destinate ad abbassarsi, a causa dell'incertezza aggiuntiva sulle previsioni delle variabili esplicative.

Nelle conclusioni della tesi sono stati inoltre indicati alcuni possibili sviluppi del lavoro. In particolare:

- creazione e previsione di un unico indice di concentrazione dell'ozono per l'area urbana di Udine;
- utilizzo di variabili esplicative, in particolare quelle meteorologiche, *previste* anziché *misurate*;
- esplorazione dei meccanismi NO/NO_2, riconsiderando le serie dei dati orari; introduzione di una variabile indicante il livello di ozono di background, da ottenere mediando opportunamente rilevazioni provenienti da un'area più vasta o acquisendo i risultati di modelli fotochimici fatti operare da atri Enti su scala di bacino padano, nazionale od europea;
- introduzione di variabili meteorologiche misurate in quota, per mezzo del radiosondaggio di Campoformido (UD) e introduzione di parametri di stabilità atmosferica (altezza dello strato di rimescolamento, ecc.), stimate da opportuni *processori* meteorologici;
- nei modelli di classificazione, introduzione di una matrice di pesi per i vari tipi di errore, in modo da ottimizzare l'efficacia nella previsione dei valori alti ed altissimi;
- sulla base della conoscenza dei fenomeni fisico-chimici: analisi mirate sui dati orari, sviluppo dei modelli su sottinsiemi dei dati originali, corrispondenti a particolari condizioni meteorologiche, introduzione di termini di interazione fra le variabili, esplorazione dei Generalized Additive Models (GAM).

Ai fini dell'utilizzo dei modelli per un supporto agli operatori nella *validazione* dei dati, infine, è opportuno condurre l'analisi delle serie storiche dei dati *orari*. L'utilizzo di modelli del tipo qui implementato richiede, in tal caso, di affrontare le complessità legate alla *ciclostazionarietà* delle serie storiche.

1 - Il problema

1.1 L'inquinamento atmosferico[1] da ozono

La qualità dell'aria in un'area è condizionata da diversi fattori, quali la densità di insediamenti (residenziali e produttivi), l'intensità e la congestione del traffico, le condizioni meteorologiche e morfologiche.

Alcuni dei principali e più diffusi inquinanti sono:

SO_2 Biossido di zolfo

NO_x Ossidi di azoto (Monossido e Biossido di azoto)

CO Monossido di carbonio

O_3 Ozono

BTX Benzene – Toluene – Xileni

COV Composti organici volatili

In questo lavoro si prenderà in considerazione, in particolare, l'inquinamento da ozono.

L'ozono al suolo non deve essere confuso con quello presente nell'alta atmosfera (ozonosfera), dove si forma per effetto delle radiazioni ultraviolette del sole sulle molecole di ossigeno: la presenza dell'ozonosfera risulta

[1] Con il termine "inquinamento atmosferico" si intende la modificazione della normale composizione dell'atmosfera dovuta alla presenza di una o più sostanze indesiderabili o estranee (inquinanti) che possono costituire un pericolo per la salute umana. L'origine di queste sostanze è spesso attribuibile ad attività umane (origine antropica) quali il traffico autoveicolare, l'utilizzo degli impianti termici, la presenza di insediamenti industriali o artigianali che impiegano svariati prodotti nei cicli produttivi.

essenziale per la vita sulla terra in quanto agisce da filtro per le radiazioni solari nocive; l'assottigliamento di questo strato protettivo (buco dell'ozono) viene tenuto sotto controllo a livello planetario.

Nella bassa atmosfera l'ozono è un inquinante secondario, in quanto non ha sorgenti proprie; si forma come prodotto di reazioni chimiche, innescate dalla radiazione solare, che coinvolgono inquinanti primari (i precursori, quali ossidi d'azoto e composti organici volatili) emessi dalle sorgenti antropiche, principalmente legate al traffico veicolare; l'ozono è legato ai precursori da reazioni fortemente non lineari.

Un ruolo fondamentale per determinare la concentrazione di ozono rivestono anche alcune variabili meteorologiche, quali l'irraggiamento solare, la temperatura, la direzione e velocità del vento, la stabilità dell'atmosfera.

E' un composto che presenta effetti irritanti per le vie respiratorie, dannoso soprattutto per bambini, anziani, persone che soffrono di asma o che sono sottoposte ad intensi sforzi fisici.

In *tabella 1.1* vengono brevemente richiamati i disposti normativi in materia di controllo di tale inquinante.

Per quanto attiene la problematica dell'inquinamento da ozono, un ruolo importante viene svolto dagli ossidi dell'azoto. Si considerano, in particolare, il monossido (NO) ed il biossido di azoto (NO2); essendo però la tossicità di quest' ultimo notevolmente superiore a quella del monossido, la normativa vigente prevede dei limiti per la protezione della salute umana solamente per il biossido di azoto. Esso è un gas irritante per occhi, naso e vie respiratorie e può combinarsi con l'emoglobina del sangue impedendone così il trasporto dell'ossigeno.

Elevate emissioni di NOx[2] derivano dalla combustione di combustibili fossili effettuata per produrre energia elettrica e per riscaldare gli edifici. In ambiente urbano la maggior parte delle emissioni di NOx deriva dal traffico autoveicolare.

Gli ossidi da azoto, una volta emessi in atmosfera, hanno un tempo medio di permanenza di circa 5 giorni ed i principali meccanismi di rimozione sono le precipitazioni, l'assorbimento su superfici liquide, l'assorbimento su particelle solide e le reazioni chimiche e biologiche.

[2] cioè di ossidi di azoto come somma di monossido e biossido

Tab. 1.1

Riferimento normativo	Periodo di mediazione	**Valore di riferimento**
DPCM 28-mar-83	valore limite: media oraria da non raggiungere più di una volta al mese	200 µg/m3
DM 16-mag-96	livello di attenzione: media oraria	180 µg/m3
	livello di allarme: media oraria	360 µg/m3
	livello per la protezione della salute : valore medio su 8 ore	110 µg/m3
DLgs 21-mag-04 n. 183	soglia di informazione: media oraria	180 µg/m3
	soglia di allarme: media oraria	240 µg/m3
	valore bersaglio per la protezione della salute: media massima giornaliera su 8 ore da non superare per più di 25 giorni per anno civile come media su 3 anni (dal 01/01/2010)	120 µg/m3

Avendo una permanenza lunga in atmosfera possono venire trasportati lontano dalle fonti di emissione.

Nella normativa viene preso in considerazione il Benzene, dannoso per l'uomo perché cancerogeno. Esso proviene, per i 2/3 delle emissioni, dal gas di scarico dei veicoli; oltre il 90% è prodotto nei centri urbani.

Nello studio dell'inquinamento da ozono, esso può essere considerato come tracciante delle emissioni da traffico e della presenza in atmosfera anche di altri COV.

1.2 La chimica dell'ozono nella troposfera

Nella troposfera, le principali reazioni di formazione e distruzione dell'ozono sono le seguenti (Baird C., 1997 (b)):

$$O_2 \; (+ UV\text{-}A) \rightarrow NO + O \qquad (1)$$

$$O + O_2 \rightarrow O_3 \qquad (2)$$

$$O_3 + NO \rightarrow O_2 + NO_2 \qquad (3)$$

Altre reazioni implicanti la partecipazione dell'ossigeno atomico nella troposfera non possono competere con la (2) a causa dell'abbondante concentrazione di O_2 che comporta un'elevata velocità di collisione e, quindi, di reazione fra l'ossigeno molecolare e quello atomico. Va anche osservato che il biossido d'azoto è la sola fonte di ossigeno atomico quantitativamente significativa.

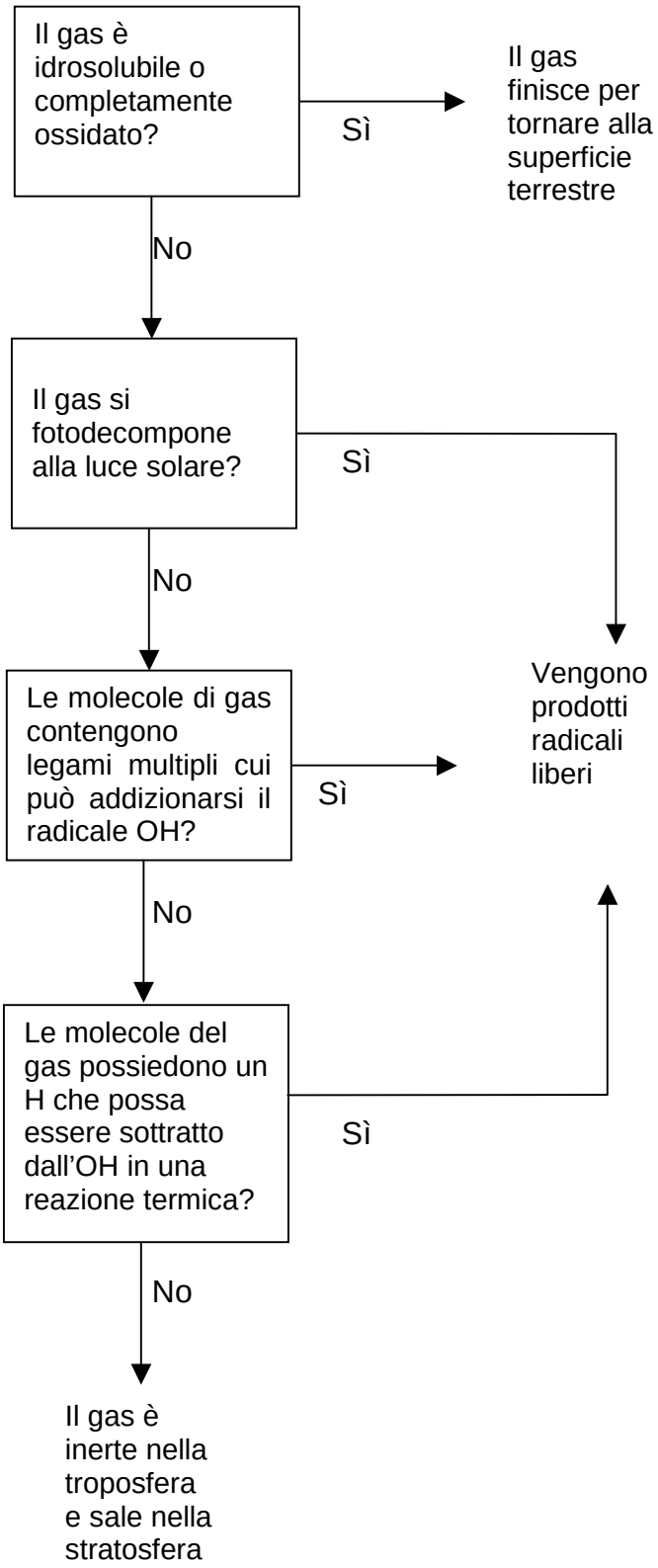
Poichè nella troposfera la concentrazione di O_2 è tale da poter essere considerata costante rispetto a tali processi, la concentrazione delle altre specie risulta governata da un equilibrio di tipo dinamico, dipendente, in particolare, dalla presenza di radiazione solare nello spettro UV-A.

Le reazioni 1, 2 e 3 costituiscono il cosiddetto "ciclo fotolitico" di NO_2; se l'equilibrio dinamico non viene perturbato da fattori esterni, le concentrazioni di ozono, NO_2 e NO entrano in poco tempo in uno stato stazionario e non subiscono notevoli variazioni. La quantità di ozono presente allo stato stazionario risulta direttamente proporzionale alla concentrazione di NO_2 ed inversamente correlata alla concentrazione di NO.La comprensione di come tale equilibrio venga alterato, in presenza di

un’atmosfera inquinata, richiede che si prendano in considerazione i processi di ossidazione di alcuni fra i principali gas presenti nella troposfera (figura 1.1). In questi meccanismi di reazione, infatti, si verifica l‘ossidazione di NO ad NO_2; l’aumento di concentrazione di NO_2 e la diminuzione della concentrazione di NO perturbano gli equilibri delle reazioni (1), (2) e (3) nella direzione della formazione di ozono (Baird C., 1997 (a)).

In particolare, è importante comprendere in tali reazioni il ruolo svolto dai radicali liberi e dai fenomeni di trasporto (Bolzacchini E., 2005).

Fig. 1.1


Il gas è idrosolubile o completamente ossidato?
Sì
Il gas finisce per tornare alla superficie terrestre
No
Il gas si fotodecompone alla luce solare?
Sì
No
Le molecole di gas contengono legami multipli cui può addizionarsi il radicale OH?
Sì
Vengono prodotti radicali liberi
No
Le molecole del gas possiedono un H che possa essere sottratto dall'OH in una reazione termica?
Sì
No
Il gas è inerte nella troposfera e sale nella stratosfera

1.2.1 I radicali liberi

I radicali liberi sono molecole neutre paramagnetiche, caratterizzate dalla presenza di un elettrone *spaiato* e, dunque, estremamente reattive.

a) il radicale ossidrile $OH^{\bullet}$:

Nella troposfera non inquinata, il radicale ossidrile si forma dalla decomposizione fotochimica delle tracce di ozono presente e dall'interazione degli atomi di ossigeno eccitati con il vapore acqueo:

$$\mathbf{O_3 \, (\, + \, UV\text{-}B \,) \rightarrow O_2 + O^*}$$

$$\mathbf{O^* + H_2O \rightarrow 2\ OH^{\bullet}}$$

Il radicale libero ossidrile è reattivo nei confronti di un'ampia varietà di altre molecole, fra cui gli idruri del carbonio, dell'azoto e dello zolfo, nonché di molte molecole contenenti doppi e tripli legami, come CO ed SO_2.

Solo di recente è stato compreso il ruolo svolto dal radicale ossidrile nella troposfera: esso inizia il processo di ossidazione di gran parte degli inquinanti presenti, consentendone quindi la rimozione.

Da attività biologiche e vulcaniche vengono costantemente immessi nell'atmosfera gas parzialmente ossidati, quali il monossido di carbonio ed il biossido di zolfo, ed altri in forma altamente ridotta, semplici composti dell'idrogeno quali il solfuro di idrogeno e l'ammoniaca.

La gran parte di questi gas viene gradualmente ossidata in aria, tuttavia nessuno di essi reagisce direttamente con l'ossigeno biatomico. Tutte le reazioni iniziano, invece, con l'attacco del radicale libero ossidrile $OH^{\bullet}$. Infatti, nonostante la concentrazione in aria di questa specie sia

notevolmente bassa, la sua reattività è molto elevata a causa della presenza di un elettrone spaiato; il radicale $OH^{\bullet}$ reagisce con molecole radicaliche o neutre in modo da associare l'elettrone libero con un altro elettrone di spin opposto e raggiungere così uno stato fondamentale energeticamente più favorevole;

b) i perossiradicali $HOO^{\bullet}$ e $CH_3OO^{\bullet}$:

Tali radicali devono il proprio nome alla presenza del legame O-O, del tipo di quello presente nei perossidi.

Un esempio di processo che porta alla formazione dell'idroperossido è la decomposizione fotochimica e successiva ossidazione della formaldeide:

$$H_2CO\ (+ UV\text{-}A) \rightarrow H^{\bullet} + HCO^{\bullet}$$

$$HCO^{\bullet} + O_2 \rightarrow CO + OOH^{\bullet}$$

Il radicale $CH_3OO^{\bullet}$ può invece formarsi per ossidazione del radicale metile:

$$CH_3^{\bullet} + O_2 \rightarrow CH_3OO^{\bullet}$$

Il più comune destino dei perossiradicali è quello di reagire con il monossido di azoto $NO^{\bullet}$:

$$HOO^{\bullet} + NO^{\bullet} \rightarrow OH^{\bullet} + NO^{\bullet}_2$$

$$CH_3OO^{\bullet} + NO^{\bullet} \rightarrow CH_3O^{\bullet} + NO^{\bullet}_2$$

E' proprio attraverso queste reazioni, presenti nelle catene di ossidazione degli idrocarburi e del metano, che la gran parte del monossido di azoto presente nella troposfera viene ossidato a biossido di azoto (figura 1.2).
La produzione di NO_2 a partire dall'NO perturba l'equilibrio della principale reazione di formazione dell'ozono nella troposfera (reazioni 1 e 2).

1.2.2 L'ossidazione del metano nella troposfera

La reazione è descritta, nel dettaglio dei suoi passaggi, dallo schema specifico allegato (figura 1.3), che consegue dal diagramma di flusso di figura 1.1; complessivamente, il bilancio è il seguente:

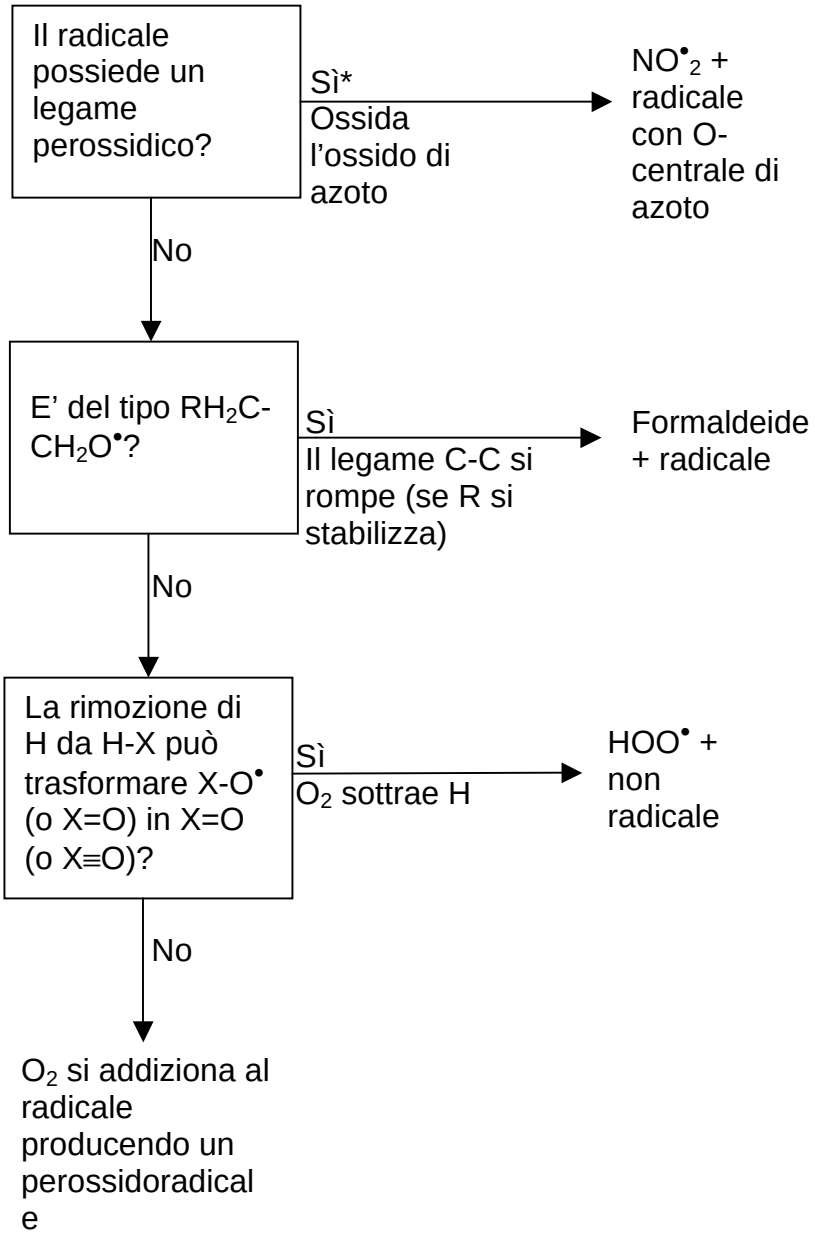
$$\mathbf{CH_4 + 5\ O_2 + NO^{\bullet} + 2\ OH^{\bullet}\ (\ +\ UV\text{-}A\ )\ \rightarrow\ CO_2 + H_2O + NO^{\bullet}{}_2 + 4\ HOO^{\bullet}}$$

Se, quindi, si considera che gli idroperossidi $HOO^{\bullet}$ prodotti vengono trasformati nuovamente in radicali $OH^{\bullet}$ per reazione con quattro molecole di $NO^{\bullet}$, si ottiene la reazione completa:

$$\mathbf{CH_4 + 5\ O_2 + \underline{5\ NO^{\bullet}}\ (\ +\ UV\text{-}A\ )\ \rightarrow\ CO_2 + H_2O + \underline{5\ NO^{\bullet}{}_2} + 2\ OH^{\bullet}}$$

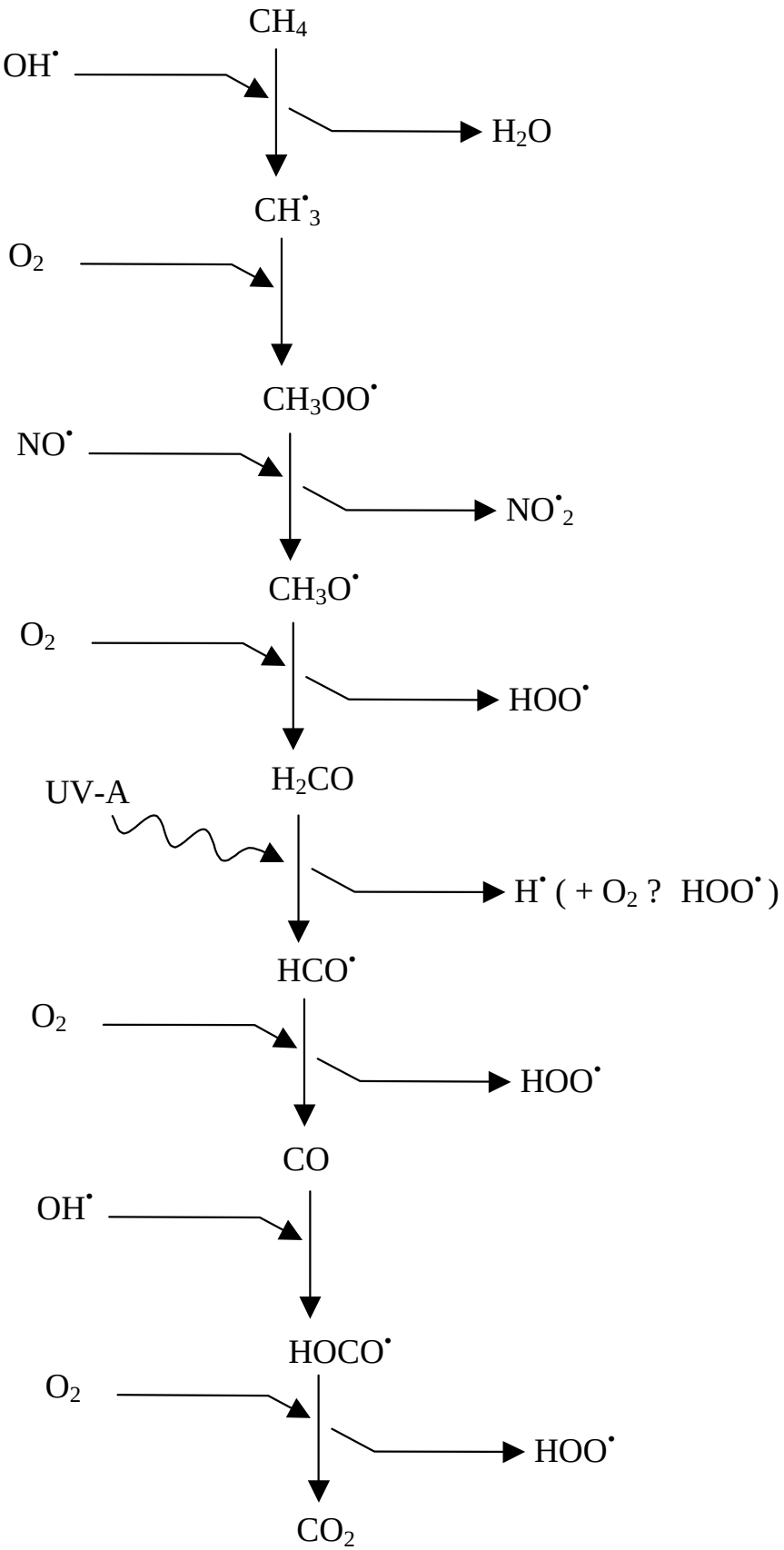
Complessivamente, dunque, si ha una produzione di biossido di azoto a partire dal monossido ed un conseguente aumento della concentrazione dei radicali ossidrile.

Fig. 1.2



* nelle condizioni in cui è significativamente presente l'ossido di azoto e non sono quantitativamente importanti le reazioni radicale + radicale

Fig. 1.3


CH4
OH•
H2O
CH3•
O2
CH3OO•
NO•
NO2•
CH3O•
O2
HOO•
H2CO
UV-A
H• (+ O2 ? HOO•)
HCO•
O2
HOO•
CO
OH•
HOCO•
O2
HOO•
CO2

1.2.3 *L'ossidazione degli idrocarburi nella troposfera*

Gli idrocarburi contenenti il legame C=C sono i più reattivi fra i composti organici volatili (COV). L'esempio più semplice è costituito dall'etene, $H_2C=CH_2$; la struttura degli idrocarburi più complessi deriva dallo scambio di ciascuno dei gruppi H con gruppi CH_3 o catene alchiliche variamente sostituite (es: $CH_3–CH_2–$), che vengono indicate convenzionalmente con il simbolo R (gruppi alchilici) (figura 1.4).

E' possibile allora considerare il bilancio complessivo del processo di ossidazione di un generico idrocarburo con doppio legame RHC=CHR:

$$\mathbf{RHC=CHR + 6\ O_2 + 5\ NO^{\bullet}\ (+\ UV) \rightarrow 5\ NO^{\bullet}_2 + 2\ RO^{\bullet} +\ OH^{\bullet} + 2\ CO + HOO^{\bullet}}$$

Se quindi si considera, analogamente a quanto visto per il metano, che il radicale $HOO^{\bullet}$ si trasforma nuovamente in $OH^{\bullet}$ per reazione con una molecola di $NO^{\bullet}$, si ottiene la reazione complessiva:

$$\mathbf{RHC=CHR + 6\ O_2 + \underline{6\ NO^{\bullet}}\ (+\ UV) \rightarrow \underline{6\ NO^{\bullet}_2} + 2\ RO^{\bullet} + 2\ OH^{\bullet} + 2\ CO}$$

L'esame della catena completa delle reazioni coinvolte mostra come, per ciascuna molecola RHC=CHR, siano prodotte due aldeidi RHC=O, che vengono poi decomposte fotochimicamente.

Di nuovo, sono d'interesse: l'aumento dei radicali $OH^{\bullet}$ presenti, che danno anche inizio alla catena di reazioni, e l'ossidazione del monossido di azoto.

1.2.4 Fasi avanzate dello smog fotochimico

In presenza di fenomeni di smog fotochimico, nelle fasi avanzate della giornata, allorché le concentrazioni dei radicali diventano elevate, le interazioni radicale-radicale non possono più essere trascurate.

Fig. 1.4

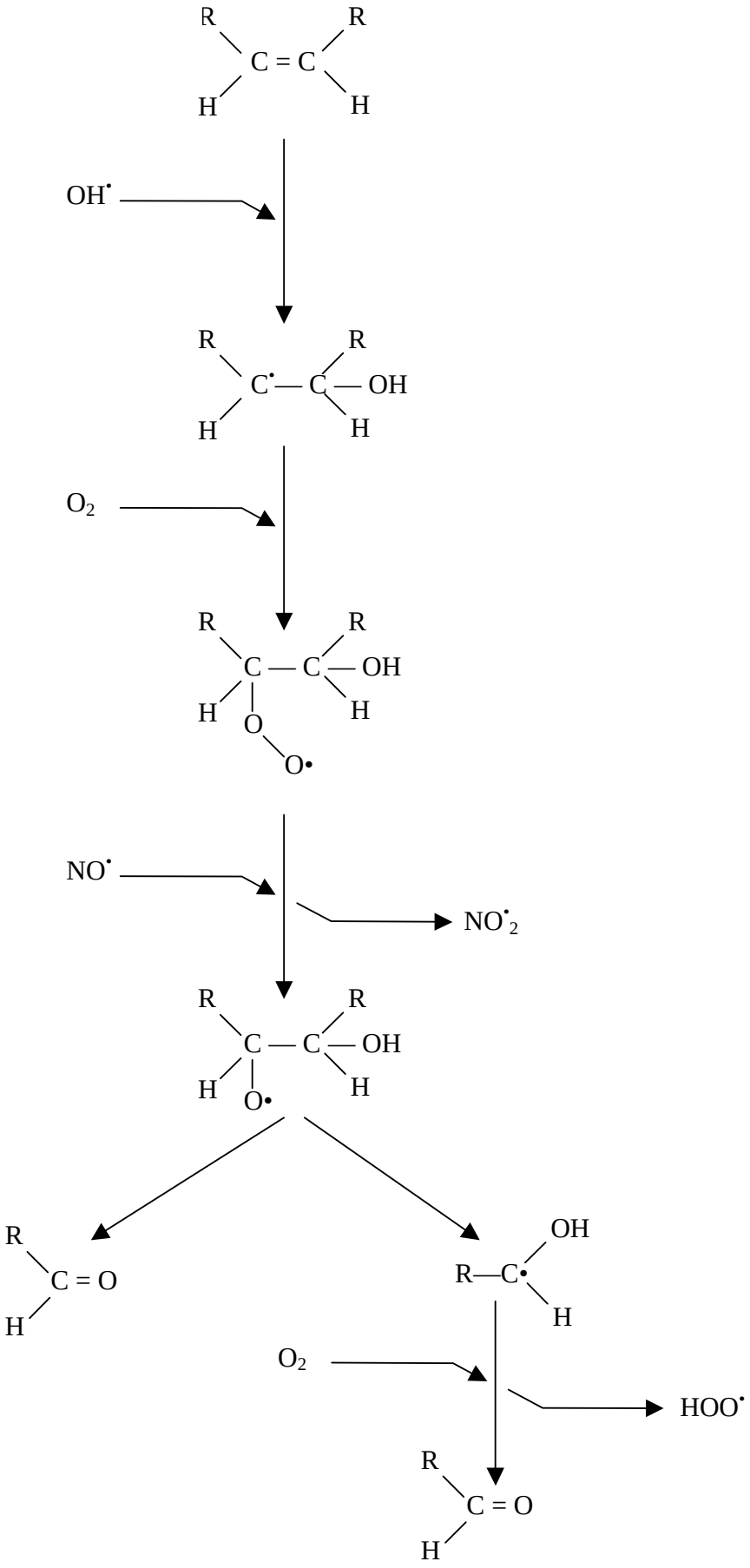
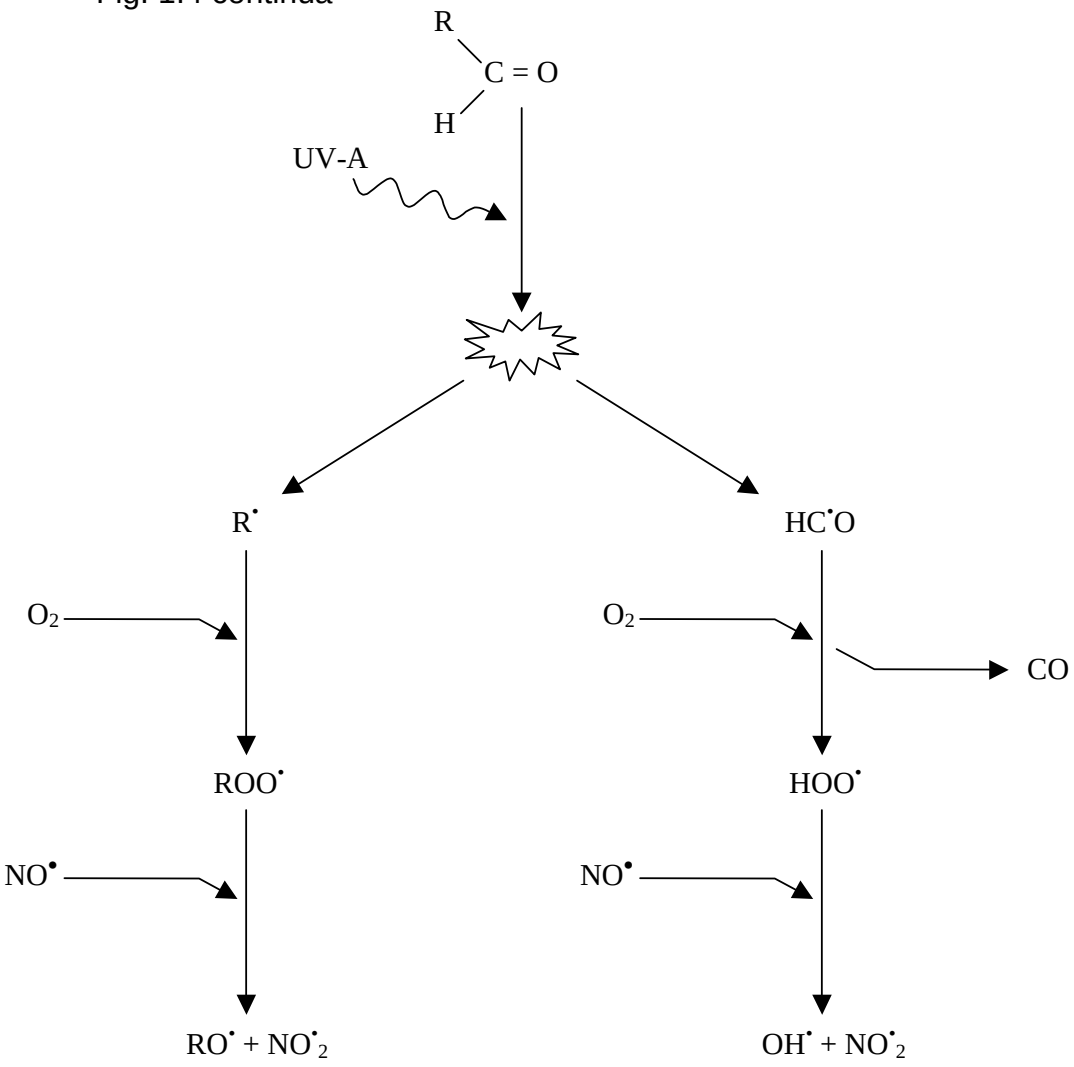

R
R
C = C
H
H
OH•
R
R
C• — C — OH
H
H
O2
R
R
C — C — OH
H
O
H
O•
NO•
NO•2
R
R
C — C — OH
H
O•
H
R
C = O
H
OH
R—C•
H
O2
HOO•
R
C = O
H

Fig. 1.4 continua



In generale, tali reazioni producono una molecola non radicalica.

La contestuale diminuzione e scomparsa della radiazione solare comporta la stabilizzazione di alcune specie, che altrimenti si troverebbero in forma dissociata: alcune di esse, in particolare, costituiscono un “pozzo” per il biossido di azoto. Si riportano di seguito alcune fra le principali reazioni nella troposfera:

a) sottrazione dei radicali ossidrile con la formazione di acido nitrico da biossido di azoto:

$$OH^{\bullet} + NO^{\bullet}_2 \rightarrow HNO_3$$

La persistenza in atmosfera di tale molecola è dell'ordine di alcuni giorni, prima di essere disciolta nelle precipitazioni o decomposta fotochimicamente.

Analogamente, dal monossido di azoto si ha la formazione dell'acido nitroso:

$$OH^{\bullet} + NO^{\bullet} \rightarrow HNO_2$$

L'acido nitroso, decomposto durante il giorno prevalentemente per via fotochimica, è invece stabile durante la notte.

b) interazione fra due radicali ossidrile o perossidrile e conseguente formazione di un importante ossidante, il perossido di idrogeno:

$$2\ OH^{\bullet} \rightarrow H_2O_2$$

$$2\ OOH^{\bullet} \rightarrow H_2O_2 + O_2$$

c) interazione fra un'aldeide ed un radicale ossidrile e successiva ossidazione del prodotto, R-C$^{\bullet}$=O:

$$RHCO + OH^{\bullet} \rightarrow RCO^{\bullet} + H_2O$$

$$RCO^{\bullet} + O_2 \rightarrow RCO^{\bullet}_3$$

Tale molecola ossida l'$NO^{\bullet}$, quando la concentrazione di quest'ultimo è elevata; viceversa, nel pomeriggio, quando la gran parte dell' $NO^{\bullet}$ è già stata ossidata, essa si addiziona all'$NO^{\bullet}_2$, producendo un nitrato.

Nel caso più comune, in cui il gruppo alchilico R è un -CH_3, il nitrato prodotto è il perossiacetilnitrato (PAN), irritante per gli occhi e velenoso per le piante:

$$CH_3\text{-}CO^{\bullet}_3 + NO^{\bullet}_2 + O_2 \rightarrow CH_3\text{-}CO_3\text{-}NO_2$$
(PAN)

d) in presenza di concentrazioni elevate di biossido di azoto ed ozono acquista importanza la reazione:

$$NO^{\bullet}_2 + O_3 \rightarrow O_2 + NO^{\bullet}_3$$

Durante il giorno, l'$NO^{\bullet}_3$ viene dissociato fotochimicamente ad $NO^{\bullet}$ ed O_2, ma durante la notte è stabile e può "catturare" l'$NO^{\bullet}_2$ presente nella troposfera secondo la reazione:

$$NO^{\bullet}_2 + NO^{\bullet}_3 \rightarrow N_2O_5$$

il pentossido di azoto così formato in presenza di vapore acqueo produce acido nitrico:

$$N_2O_5 + H_2O \rightarrow 2\ HNO_3$$

1.2.5 Interazione con gli idrocarburi aromatici

La chimica dei COV risulta particolarmente complessa per la varietà dei composti organici presenti nella troposfera e per la loro specifica reattività dovuta alla presenza di doppi e tripli legami, alla lunghezza delle catene carboniose, al carattere alifatico o aromatico (Rindone B., 2005). Come esempio di percorso di reazioni che coinvolgono i composti aromatici si prenda ad esempio il caso del toluene (figura 1.5): l'attacco del radicale ossidrile sul toluene può originare due diversi percorsi di reazione, denominati A e B.

Il percorso A prevede la ritenzione dell'anello aromatico mediante sottrazione da parte di $OH^{\bullet}$ di un protone radicalico dal metile toluenico per formare il corrispondente radicale benzilico, che in presenza di O_2 si ossida a benzaldeide ed alcool benzilico.

Il percorso B comporta, invece, l'attacco del radicale ossidrile alla posizione orto del toluene per formare un derivato radicalico che in presenza di O_2 può dare origine all'o-cresolo oppure, mediante l'ossidazione di NO ad NO_2 e l'apertura dell'anello aromatico, ad una di-aldeide insatura che, reagendo con $OH^{\bullet}$, si frammenta ulteriormente per produrre due molecole di ossalaldeide.

1.2.6 Il rapporto COV/NO_x ed i fenomeni di trasporto

Il diagramma di isoconcentrazione di figura 1.6 rappresenta il livello massimo di ozono raggiunto in presenza di diverse concentrazioni di composti organici volatili e NO_x, per un determinato tempo di irraggiamento (Bolzacchini, 2005).

Si possono sostanzialmente distinguere due andamenti in relazione alle caratteristiche dell'area di interesse ed al rapporto COV/NO_x:

Fig. 1.5

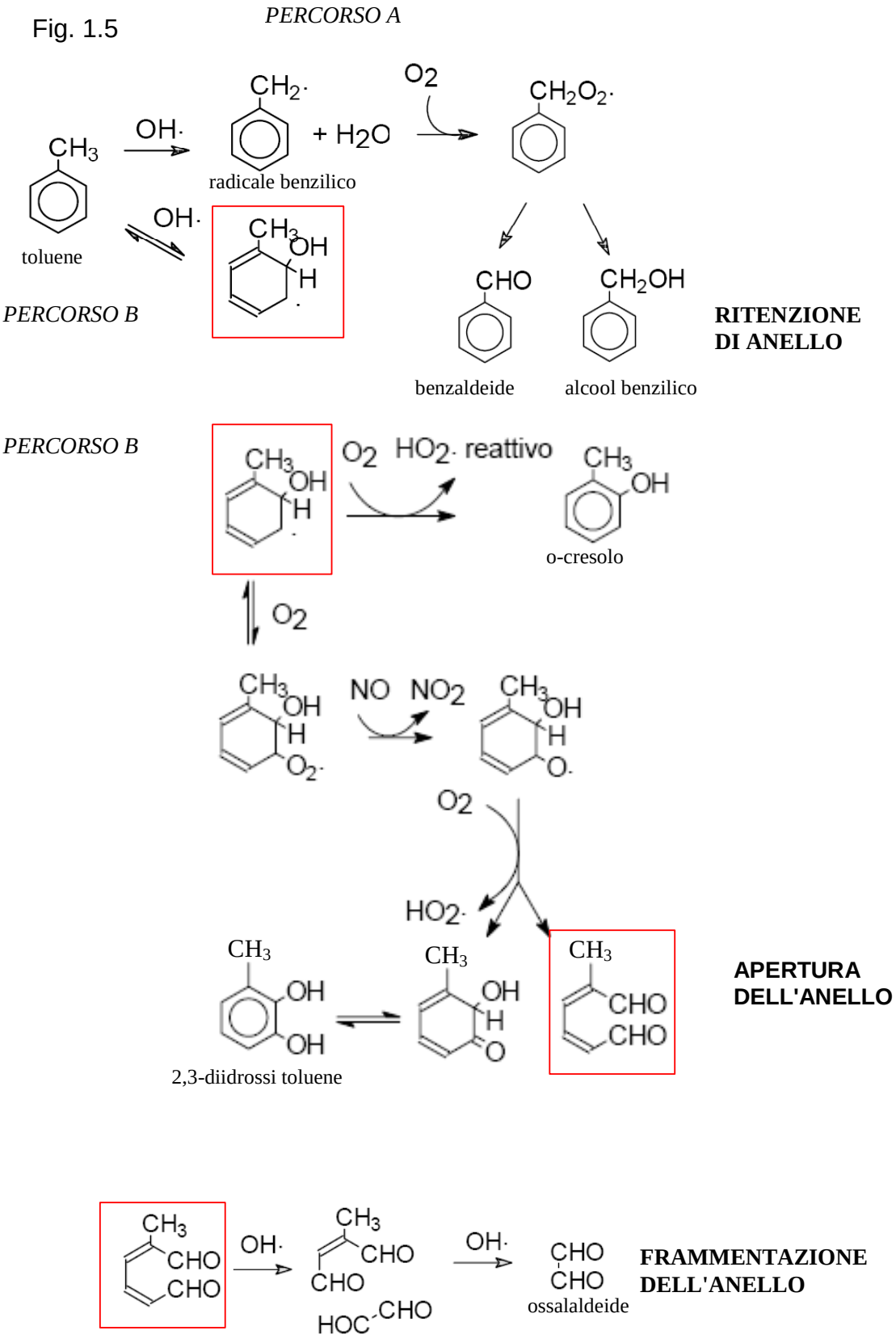
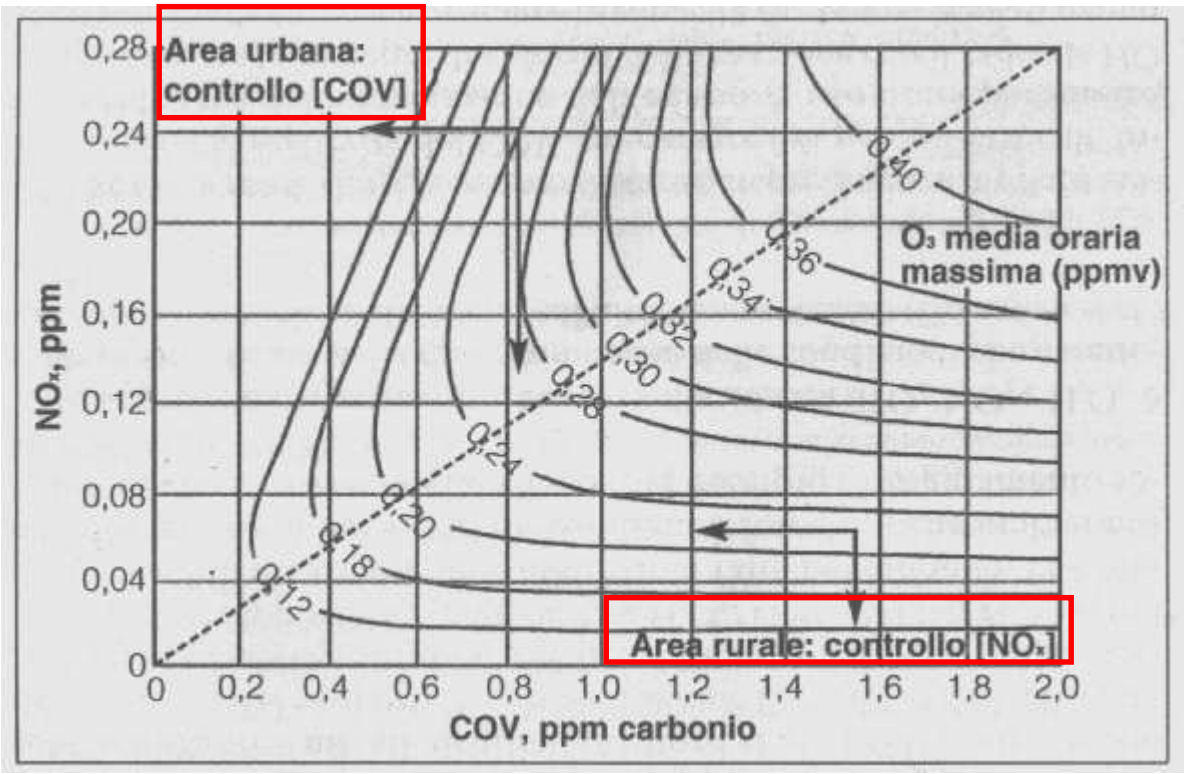

PERCORSO A
toluene
OH·
radicale benzilico
+ H2O
O2
CH2O2·
PERCORSO B
CHO
CH2OH
benzaldeide
alcool benzilico
RITENZIONE DI ANELLO
PERCORSO B
O2
HO2· reattivo
o-cresolo
NO
NO2
HO2·
2,3-diidrossi toluene
APERTURA DELL'ANELLO
ossalaldeide
FRAMMENTAZIONE DELL'ANELLO

Fig. 1.6



a) <u>aree urbane, di traffico intenso</u>: i valori massimi sono tendenzialmente più bassi. Ciò è dovuto al fatto che il 90% delle emissioni di NO_x è sotto forma di monossido di azoto. In prossimità della sorgente si ha, dunque, un maggiore effetto di rimozione dell'ozono per la reazione con NO (reazione 3).

Nel diagramma tali aree sono caratterizzate da un rapporto $[COV]/[NO_X]$=3-6.

Osservando le curve di livello, si nota come una diminuzione degli NO_x porti ad un effetto debole e non univoco sulla concentrazione dell'ozono;

viceversa, una diminuzione dei COV, comporta sempre un calo consistente.

Ciò dipende del fatto che, in queste condizioni ambientali, il fattore limitante per la produzione di ozono è la disponibilità di COV. Le reazioni tra NOx ed ozono, viceversa, sono favorite dalla sovrabbondante concentrazione degli stessi NOx, dovuti ai processi di combustione; inoltre, NO ed NO_2 svolgono ruoli antagonisti nella produzione/distruzione di ozono. Pertanto un calo complessivo delle loro concentrazioni non incide in modo marcato sulla concentrazione dell'ozono;

b) <u>aree extraurbane</u>: i fattori che influenzano queste aree sono diversi.

La maggiore intensità dei picchi nelle concentrazioni di ozono è dovuta alle minori emissioni di NO, che non permettono una rimozione efficace dell'ozono.

L'incidenza di eventuali masse d'aria inquinate, provenienti da aree urbane, determina in ogni caso la crescita del picco di formazione dell'ozono: gli NO_x che vengono trasportati, infatti, sono presenti soprattutto in forma di NO_2, a causa dell'avvenuta ossidazione dell'NO, e come tali contribuiscono in modo determinante alla formazione dell'ozono.

A ciò si deve il fatto che, per queste aree, in cui $[COV]/[NO_x]=20\text{-}40$, una diminuzione degli NO_x porta ad una sensibile riduzione dei livelli di ozono. In questo caso è effettivamente la bassa disponibilità di NO_2 che, attraverso le reazioni (1) e (2), limita la produzione dell'ozono. In presenza di elevate quantità di COV, invece, una diminuzione della loro concentrazione non comporta variazioni tali da produrre anche un decremento significativo della presenza di ozono.

I processi di trasporto sono, dunque, importanti, poiché favoriscono elevate concentrazioni di ozono nelle aree poste sottovento rispetto alla sorgente: gli episodi più intensi di inquinamento fotochimico si manifestano, perciò, soprattutto in area extraurbana.

Per il gran numero di reazioni che si verificano nell'aria inquinata, la dipendenza della produzione dell'ozono dalla concentrazione dei reagenti è assai complessa e senza l'uso di simulazioni è difficile prevedere la conseguenza netta di modeste diminuzioni delle concentrazioni degli inquinanti primari.

Semplificando, la chimica dell'ozono è caratterizzata dal rapporto COV/NO_x e dal rapporto NO_2/NO, diversi fra aree sorgente ed aree interessate dal trasporto degli inquinanti:

- ad aree con un alto rapporto COV/NOx (extraurbane) sono associati i massimi più intensi poiché, in seguito al trasporto, in esse viene liberato il potenziale di produzione di ozono accumulato con la formazione di NO_2 nelle aree sorgente;
- le aree sorgente (urbane, trafficate) sono caratterizzate da un rapporto COV/NO_x basso e la minore concentrazione di ozono deriva dalle abbondanti emissioni di NO. La disponibilità di COV è, invece, il fattore limitante nella produzione di ozono.

1.3 La rete di rilevamento della qualità dell'aria nell'area urbana di Udine

Nel caso della città di Udine si possono identificare nel traffico autoveicolare e negli impianti termici le due principali sorgenti di inquinamento atmosferico, stante la ridotta presenza di attività produttive nell'ambito del territorio comunale.

I concetti di salvaguardia e prevenzione ambientale non possono prescindere dal monitoraggio, il quale costituisce il punto di partenza per qualunque valutazione.

Il monitoraggio dell'inquinamento atmosferico, che è stato ben illustrato dal dott. Flavio Moimas, Responsabile della Rete di Rilevamento della Qualità dell'Aria del Dipartimento provinciale di Udine dell'ARPA FVG, si effettua mediante reti di centraline di rilevamento delle concentrazioni di inquinanti. Le centraline sono delle cabine a postazione fissa dislocate in alcuni punti strategici del territorio d'interesse ed ospitano rilevatori automatici degli inquinanti.

Il numero di centraline situate nella città di Udine e la loro disposizione, è stata scelta secondo le indicazioni contenute nel D.M. 20/05/1991, che individua appunto gli inquinanti da monitorare in ambito urbano e quattro tipologie di stazioni, da scegliere anche in base al numero di abitanti presenti nella città d'interesse.

Per nuclei urbani con un numero di abitanti inferiore a 500.000, è prevista l'installazione di sei stazioni, secondo i seguenti criteri:

- una stazione di TIPO A : stazione di riferimento localizzata in aree non direttamente interessate dalle sorgenti di emissione urbana (parchi, isole pedonali, ecc.) per la misura dell'inquinamento di fondo;
- due stazioni di TIPO B : stazione situata in zone ad elevata densità abitativa per la misura di inquinanti primari e secondari;
- due stazioni di TIPO C : stazione posizionata in vicinanza di strade ad elevato traffico per la misura degli inquinanti emessi direttamente dalle autovetture;

- una stazione di TIPO D : stazione collocata in periferia o in aree suburbane per la misura degli inquinanti fotochimici.

Nel caso di Udine, la principale sorgente di inquinamento atmosferico è proprio il traffico, quindi i principali punti di monitoraggio sono stati individuati nei nodi di connessione fra le direttrici di accesso alla città e l'anello della circonvallazione, integrando poi la rete secondo le indicazioni del Decreto.

Nell'analisi che segue, vengono analizzati i dati pervenuti da tre delle sei centraline installate, quelle che almeno negli ultimi sei anni hanno rilevato l'ozono:

- la stazione di via Cairoli può essere classificata di Tipo A, in quanto si trova nell'area verde attrezzata "G. Ambrosoli", un parco urbano nel cuore della città, non direttamente interessato dal traffico;
- la stazione di via Manzoni è di Tipo B/C; si trova infatti all'incrocio con via Crispi, in una zona interessata dai flussi di traffico interni alla città ed inserita in un'area densamente edificata;
- S.Osvaldo è classificata come come Tipo D: ubicata in via Pozzuolo, all'interno dell'area dell'Azienda Agraria "Servadei" dell'Università di Udine, si trova dunque in un'area verde, lontana da traffico e abitazioni, tendenzialmente sottovento al nucleo urbano.

Relativamente ai dati di interesse per il presente studio, tutte e tre queste stazioni misurano la concentrazione di ozono (O_3) e degli ossidi di azoto (NOx); in v. Manzoni si misura il Benzene, dal 2000.

Quest'ultima variabile è molto importante per valutare l'incidenza del traffico; con l'introduzione della marmitta catalitica, ha assunto in tal senso maggiore importanza rispetto al monossido di carbonio (CO).

La successiva figura 1.7 riporta la dislocazione delle centraline sulla mappa della città di Udine dove sono state evidenziate le principali direttrici di traffico e le aree maggiormente edificate.

La particolare delicatezza e precisione degli strumenti di misura automatici adottati fa sì che spesso si verifichino dei guasti, determinando la presenza di “buchi” nelle serie dei dati.

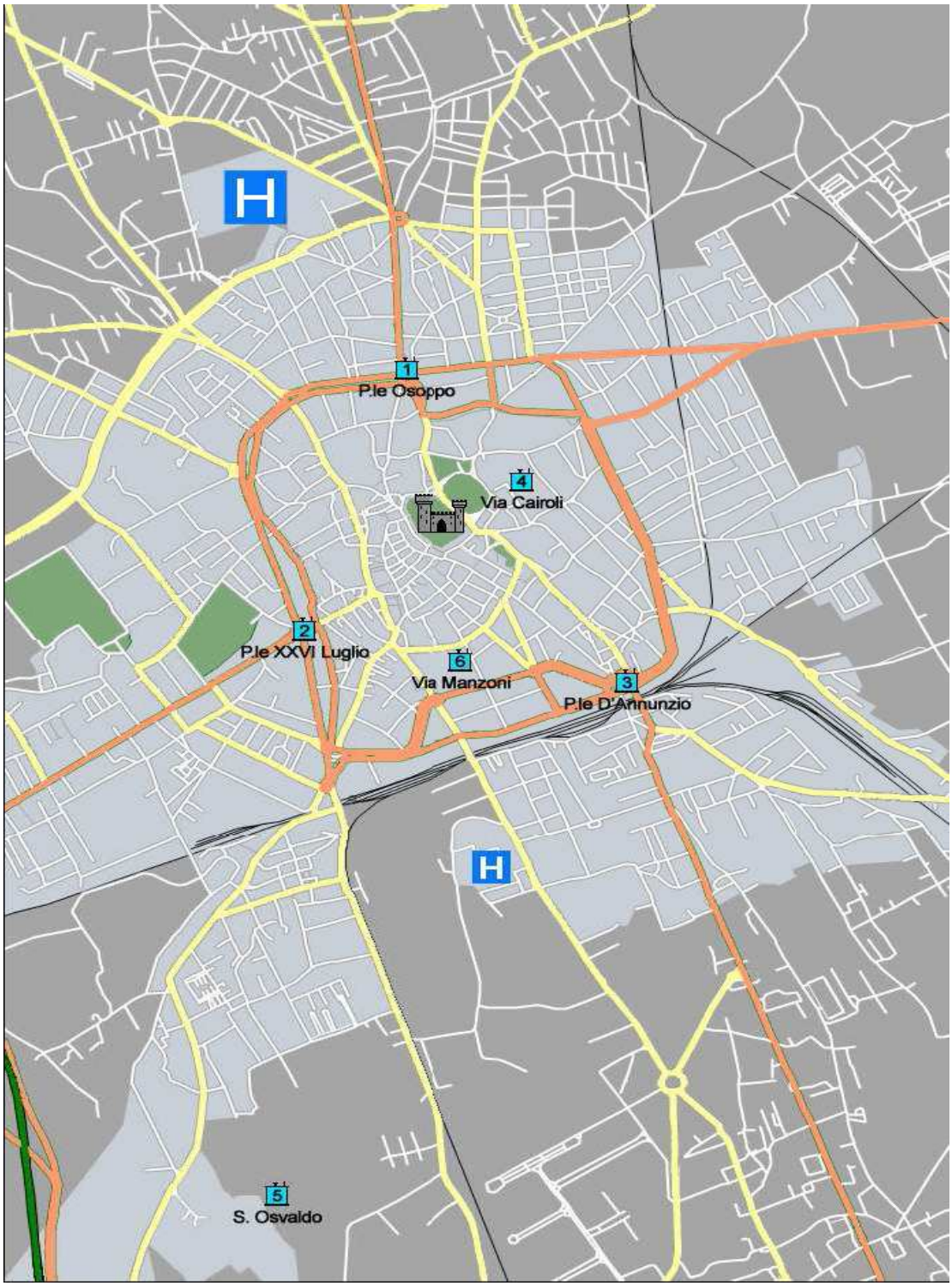
Oltre all’impoverimento delle serie storiche dovuto ai dati mancanti, la presenza di valori non attendibili costituisce un problema, in quanto può alterare le analisi statistiche, inducendo componenti sistematiche d’errore.

E’ stato accurato che la gran parte dei dati anomali trovati nelle serie seguono o precedono proprio i “buchi”; si collocano, cioè, in periodi appena precedenti o seguenti a quelli in cui il mal funzionamento degli strumenti arriva a causare la perdita assoluta del dato.

Tuttavia, il problema della qualità dei dati ottenuti da misure prese in automatico va ricondotto anche alla possibile presenza di errori che non si manifestino così chiaramente: per esempio, in fase di calibrazione o in una non corretta metodologia di prelievo.

La procedura di taratura automatica degli strumenti deve, inoltre, essere necessariamente semplice; essa consiste tipicamente nell’acquisizione dell’output in corrispondenza di due concentrazioni note dell’inquinante: una pari allo zero, l’altra ad un valore cosiddetto di *span*.

Fig. 1.7


H
1
P.le Osoppo
4
Via Cairoli
2
P.le XXVI Luglio
6
Via Manzoni
3
P.le D'Annunzio
H
5
S. Osvaldo

La curva di taratura, di solito lineare, viene quindi aggiustata in modo da passare per tali punti Nel caso dell'ozono, per esempio, la procedura di taratura che viene adottata garantisce l'accuratezza dei dati nell'intorno ad un valore di riferimento, 125 $\mu g/m^3$, prossimo ad una delle soglie previste dalla normativa. Più ci si allontana da tale valore, minore è l'accuratezza della misura.

Ciò accade, sia per com'è costruito lo strumento, sia per le decisioni assunte nella sua gestione: è particolarmente importante, infatti, garantire valori attendibili nell'intervallo di interesse previsto dalla normativa. Valori particolarmente bassi, viceversa, se pure possono essere altrettanto interessanti nello studio qui presentato, non hanno altrettanto rilievo ai fini del monitoraggio.

Alcuni analizzatori degli inquinanti sono estremamente stabili, mentre l'operazione di taratura giornaliera può essere resa incerta da perdite nelle bombole contenenti il gas campione e nelle condotte. Per cui, in alcuni casi, vengono preferite operazioni di taratura manuale, con frequenza settimanale o anche più dilazionate.

1.4 Inquadramento meteoclimatico

Il Friuli Venezia Giulia è situata alle medie latitudini dov'è molto marcato il contrasto tra le masse d'aria polare e tropicale (ARPA FVG, 2005; ARPA FVG, 2002). Tale contrasto genera frequentemente delle perturbazioni allo stato normale dell'atmosfera, influenzate fortemente dai rilievi e dalla loro disposizione rispetto alla circolazione delle masse d'aria. La presenza delle Alpi induce significativi cambiamenti della temperatura, dell'umidità e della direzione delle masse d'aria che interessano la regione. Molto importante è anche la presenza del mar Adriatico, poco profondo, e della laguna,

caratterizzata da considerevoli escursioni termiche. Quindi troviamo un clima continentale moderato con connotazione umida (secondo la classificazione di Koeppen), dettata dall'elevata piovosità, nell'alta pianura friulana e nella zona prealpina.

Oltre alle stazioni di rilevamento della qualità dell'aria, nel presente studio viene considerata una stazione meteorologica, ubicata anch'essa in località S. Osvaldo, dalla quale abbiamo tratto i dati relativi ai principali parametri meteo indispensabili per l'interpretazione dei valori di inquinamento atmosferico.

Essa è dotata di sensori per la misurazione di: Velocità e Direzione del vento, Pressione atmosferica, Temperatura, Umidità relativa, Pioggia, Radiazione solare, Tempo di irraggiamento.

1.5 Scopo del progetto

E' stato avviato uno studio sui valori di qualità dell'aria forniti dalle centraline di rilevamento, basato su analisi statistiche delle serie storiche dei dati rilevati, relativi alle concentrazioni degli inquinanti ed ai parametri meteo.

Ci si concentra qui, in particolare, sui due valori giornalieri di concentrazione di Ozono rilevanti ai sensi della normativa vigente: il valore massimo orario giornaliero ed il valore massimo giornaliero della media mobile sulle 8 ore.

Scopo dello studio è lo sviluppo ed il confronto di alcuni modelli di dipendenza, atti a spiegare al meglio il comportamento dell'ozono, sulla base della conoscenza dei valori simultanei e precedenti dei restanti parametri misurati (variabili meteorologiche ed altri inquinanti).

Vi è un'ampia letteratura scientifica di riferimento – in particolare a partire dai primi anni novanta - relativa all'applicazione di tali tecniche ai dati forniti dalle reti di rilevamento della qualità dell'aria.

In Italia, studi in tale campo sono stati condotti da varie Università ed Istituti di ricerca, in collaborazione con Enti locali ed Agenzie Regionali per la Protezione dell'Ambiente; alcuni di essi sono stati inseriti in progetti finanziati nell'ambito del V e VI Programma Quadro dell'Unione Europea.

I modelli statistici qui sperimentati appartengono alle categorie dei modelli lineari, degli alberi di classificazione e regressione e *Random Forest*.

Con il presente studio si stabiliscono, inoltre, alcuni standard sul formato dei dati e sull'analisi preliminare (per esempio: relativamente al trattamento dei dati mancanti) e sull'utilizzo del linguaggio "R", utili al fine di successivi lavori.

2 - I dati

2.1 I dataset originari

I dataset utilizzati per questo studio sono all'origine composti da dati orari e includono:

- Cairoli (stazione tipo A che si trova all'interno dell'area verde attrezzata G. Ambrosoli): serie storiche dei dati orari registrati dalla centralina di rilevamento della qualità dell'aria di via Cairoli;
- Osvaldo (tipo D, area coltivata dell'Azienda Agraria "Servadei" dell'Università): serie storiche dei dati orari registrati dalla centralina di rilevamento della qualità dell'aria di località S.Osvaldo;
- Manzoni (tipo B/C, zona caratterizzata dalla presenza di palazzi a molti piani): serie storiche dei dati orari registrati dalla centralina di rilevamento della qualità dell'aria di via Manzoni;
- Meteo: serie storiche dei dati orari registrati dalla centralina meteorologica di località S.Osvaldo.

L'analisi ha riguardato il periodo 2000 – 2005; si è infatti ritenuta sufficiente la quantità di dati disponibili in tale periodo. A fronte di ciò si è deciso di non estendere ulteriormente l'intervallo temporale d'interesse al fine di evitare l'influenza di variabili, come il traffico veicolare, per le quali non si dispone di adeguati indicatori e che potrebbero essere parecchio cambiati nel tempo.

2.2 Analisi preliminare dei dati

Si sono analizzate inizialmente le serie dei dati orari relativi alla concentrazione di ozono, ai suoi precursori ed alle variabili meteorologiche che, a partire dalla letteratura in materia, sono ritenute rilevanti.

Per ciascun anno di dati, si è presa in considerazione la *Ozone season* (Damon, 2001), identificata nel periodo che va da aprile a settembre.

Nella *figura 2.1*, l'andamento della media mensile dei valori massimi giornalieri della media trascinata sulle 8 ore relativa alla stazione di via Cairoli; considerato che la numerazione dei mesi va da 0 ad 11, dal grafico appare chiaramente la sensatezza nella definizione adottata per la *Ozone season*.

Si è proceduto ad una verifica della qualità delle serie storiche, con particolare attenzione alla distribuzione dei dati mancanti ed alla presenza di discontinuità o derive nelle serie.

Sono state analizzate le distribuzioni di queste variabili, utilizzando metodi grafici (boxplot ed istogrammi), indici di simmetria e di curtosi, test di normalità (Pastore, 1997).

Si sono studiati i loro andamenti, al fine di individuare i principali cicli (giornalieri, settimanali, stagionali, mensili) o tendenze.

Ciascuna variabile meteorologica presenta un numero di dati orari mancanti al più di poche unità, per ciascun anno considerato.

Le serie relative agli inquinanti sono, da questo punto di vista, di qualità inferiore.

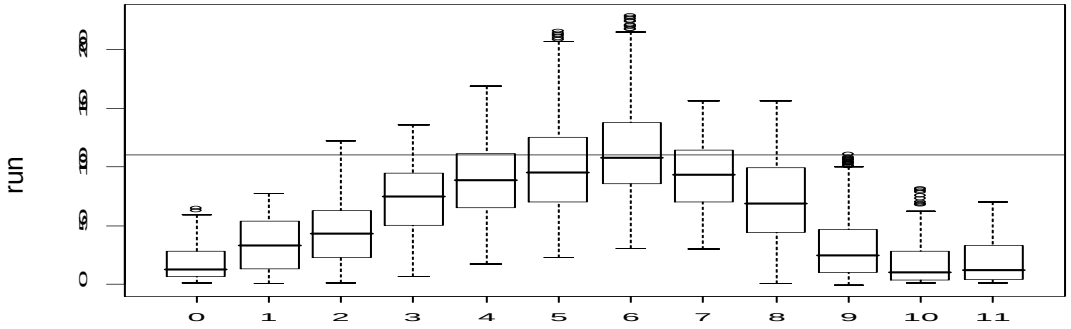
La distribuzione dei dati mancanti all'interno delle serie denota tipicamente il manifestarsi di problemi nel sistema di acquisizione dell'intera stazione, piuttosto che del singolo analizzatore (per esempio, problemi di alimentazione).

I dati mancanti si presentano infatti per lo più simultaneamente su tutte le serie e per periodi dell'ordine di alcuni giorni (tempi legati agli interventi manutentivi e di ripristino): ciò, in particolare, induce ad escludere l'opportunità di completare le serie per mezzo di interpolazioni.

Gli analizzatori di ozono e degli ossidi di azoto effettuano, inoltre, una procedura di calibrazione quotidiana, che comporta la perdita del dato delle ore 01:00 o delle 02:00 di ciascun giorno.

Per ciascuna variabile, sono stati scelti degli indici giornalieri rappresentativi (valori massimi, medî, mediani, ecc.), al fine di ottenere delle serie storiche di dati con frequenza giornaliera.

Fig. 2.1



Riguardo le concentrazioni dell'ozono e degli inquinanti precursori, le scelte sono legate alla normativa vigente.

Nel caso dell'ozono, si è inoltre provveduto a distinguere il periodo diurno da quello notturno, in considerazione del fatto che i massimi diurni e quelli notturni appaiono essere conseguenza di fenomeni fisico-chimici diversi.

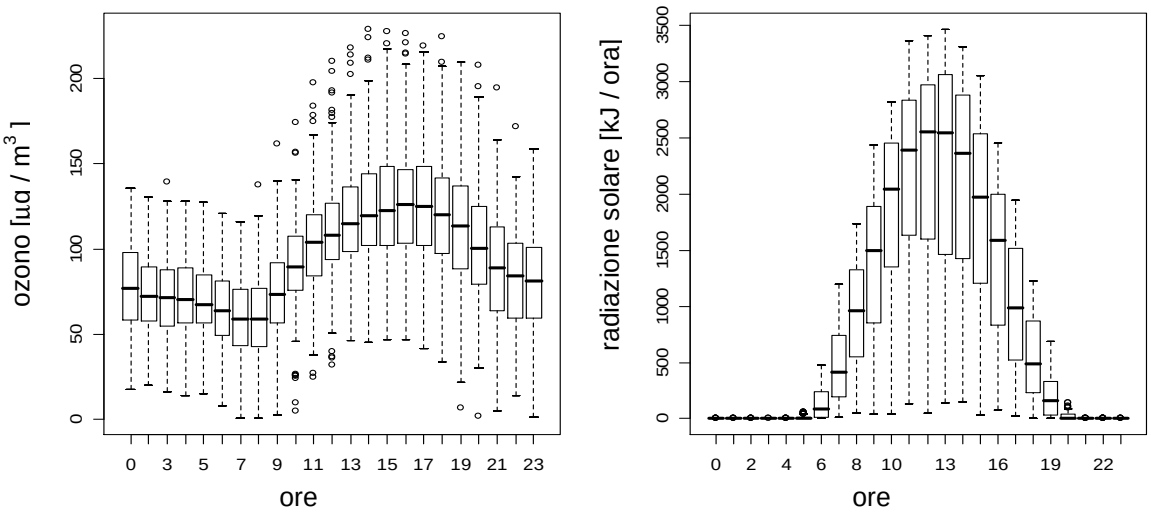
Dai dati dell'ozono si sono dunque estratti tre indici giornalieri, per ciascuna stazione:

- massimo della media trascinata di otto ore per ogni giorno;
- massimo del valore orario giornaliero;
- massimo del valore orario notturno.

Come detto, se il primo indice è stato scelto con riferimento alla normativa (Davis & al., 1999), la decisione nel distinguere massimo notturno e diurno è stata presa in quanto, essendo l'irraggiamento solare ad avere una notevole importanza nei picchi diurni, i picchi notturni (osservati verso l'una, due di notte) devono essere spiegati da variabili diverse. L'O_3 presenta una correlazione ritardata di due-tre ore con l'irraggiamento solare (*figura 2.2*); al fine di distinguere le ore diurne e notturne, quindi, si sono considerati rispettivamente i periodi 9 - 21 e 21 – 9.

Particolare attenzione è stata posta al trattamento dei dati mancanti: qualora in una giornata siano assenti più di 5 valori, il calcolo degli indici giornalieri restituisce a sua volta un dato mancante.

Fig. 2.2



Si è quindi cercata una trasformazione Box Cox tale da ottenere una migliore rappresentazione della concentrazione dell'ozono, soprattutto in termini di asimmetria e curtosi. Si ricorda che la trasformazione Box Cox consiste in

$$T(Y_t) = \begin{cases} \dfrac{Y_t^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln Y & \lambda = 0_t \end{cases}$$

La trasformazione logaritmica si ha dunque per $\lambda = 0_t$, mentre quella mediante radice quadrata si ha per $\lambda = \frac{1}{2}$.

PIù specificamente, molti dei modelli con cui si può cercare di *spiegare* le concentrazioni dell'ozono si fondano sulla minimizzione degli scarti quadratici; si può ritenere che la loro efficienza migliori con la gaussianità della distribuzione della variabile. Più Nei grafici si rappresentano per semplicità i dati della centralina di via Cairoli, in quanto le stesse analisi,

Tab. 2.1

	λ=1	λ=1/2	λ=0
curtosi	1.91	2.03	1.98
asimmetria	0.30	0.07	-0.19

ripetute per tutte tre le centraline, hanno consentito di verificare che viene sostanzialmente riprodotto lo stesso comportamento. Inoltre la centralina di via Cairoli è la più adatta all'analisi, in quanto ci sono pochissimi dati mancanti.

Nella *figura 2.3* vediamo rappresentati in ordine i grafici quantile-quantile per la trasformazione radice quadrata e logaritmica dei dati originali del massimo giornaliero della media trascinata nelle otto ore, rappresentati del terzo grafico.

Fig. 2.3

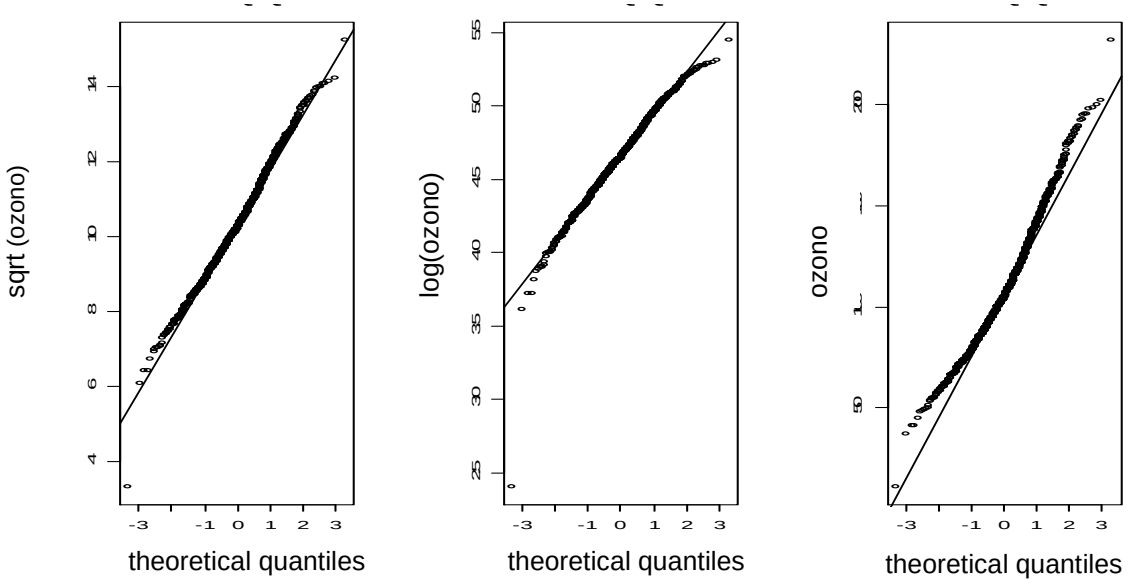
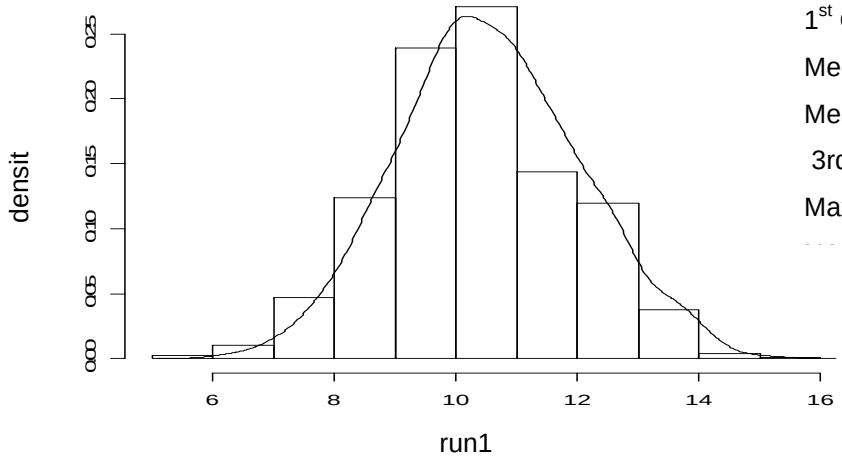


Fig. 2.4



Tab. 2.2

Cairoli	
Min	37.05
1st Qu.	91.16
Median	110.32
Mean	113.57
3rd Qu.	133.39
Max	231.75

La trasformazione radice quadrata appare essere la migliore, anche in termini di curtosi[3] e asimmetria[4] (*tabella 2.1*). Quindi, da questo punto in poi, verrà presa in considerazione la variabile trasformata, a cui ci si riferirà con il simbolo `run`.

Il punto anomalo in basso corrisponde al 15 maggio 2000, giorno a cui seguono molti dati mancanti: probabilmente il valore rilevato è dovuto a errori di rilevazione ed è quindi stato rimosso, sostituendolo con l'indicatore di *dato mancante*. Per vedere l'andamento totale dei dati a disposizione, osserviamo seguenti tabelle riassuntive (da evidenziare che le tabelle si riferiscono ai dati originali, il grafico a quelli trasformati).

[3] indicatore che misura in modo oggettivo quando una distribuzione è appiattita o appuntita (misura la densità dei dati attorno alla propria media). La gaussiana ha curtosi =3; si ha un valore minore di 3 se la distribuzione è platicurtica (molto appiattita), maggiore di 3 se è leptocurtica (piccata).

[4] indicatore che misura in modo oggettivo quando una distribuzione presenta evidenti sbilanciamenti verso destra o sinistra. Se la distribuzione è simmetrica l'indice assume valore 0; altrimenti, è positivo in caso di asimmetria positiva (la distribuzione ha la coda di sinistra lunga), negativo nel caso di asimmetria negativa

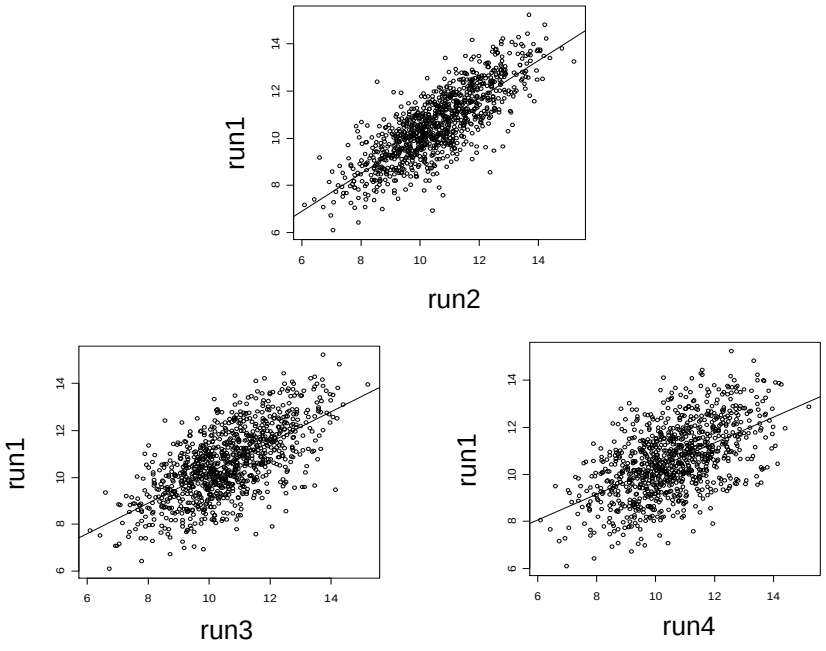
La *figura 2.4* rappresenta la distribuzione dei valori della variabile `run` per la stazione di via Cairoli.

La soglia per la tutela della salute umana, indicato nella normativa europea sull'inquinamento atmosferico da ozono che è in vigore in Italia, è di 110 µg/m3 per le misure medie di 8-ore, la cui radice quadrata è 10.49.

2.3 Autoregressione

È' stata eseguita una semplice analisi autoregressiva sulla serie, per riconoscere fino a che ritardo l'ozono dei giorni precedenti influenza l'ozono del giorno corrente. Per quanto riguarda i dati utilizzati nel presente studio, è apparsa da subito l'enorme importanza del livello di ozono del giorno precedente. In una prima fase è stato quindi analizzato un modello contenente l'ozono ritardato fino a sette volte e sono stati effettuati vari test di

Fig. 2.5



Tab. 2.3

	Massimo ozono diurno	Massimo ozono notturno
Min	31.53	8.58
1st Qu.	98.43	56.57
Median	119.8	74.57
Mean	123.66	74.71
3rd Qu.	145.52	91.61
Max	246.93	171.76
Asimmetria	0.03	-0.21
Curtosi	1.87	2.02
NA	574	583

verifica, quali analisi della varianza, test chi quadrato, analisi dell'autocorrelazione, per decidere il numero di ritardi sufficiente a spiegare meglio l'ozono del giorno corrente (*figura 2.5*).

Si è alla fine ritenuto appropriato considerare ritardi fino a tre giorni nell'analisi di questi dati; per questo motivo, il *dataframe* considerato per ciascuna stazione contiene le variabili run1 (variabile da spiegare), run2, run3, run4, in cui si introduce un ritardo da 0 a 3 giorni. Un'analisi analoga si è condotta relativamente al massimo notturno e diurno. Vediamo nella *tabella 2.3 e figura 2.6* i boxplot corrispondenti. Anche le restanti variabili, i precursori e le variabili meteorologiche, sono state ritardate di un giorno.

Tab. 2.4

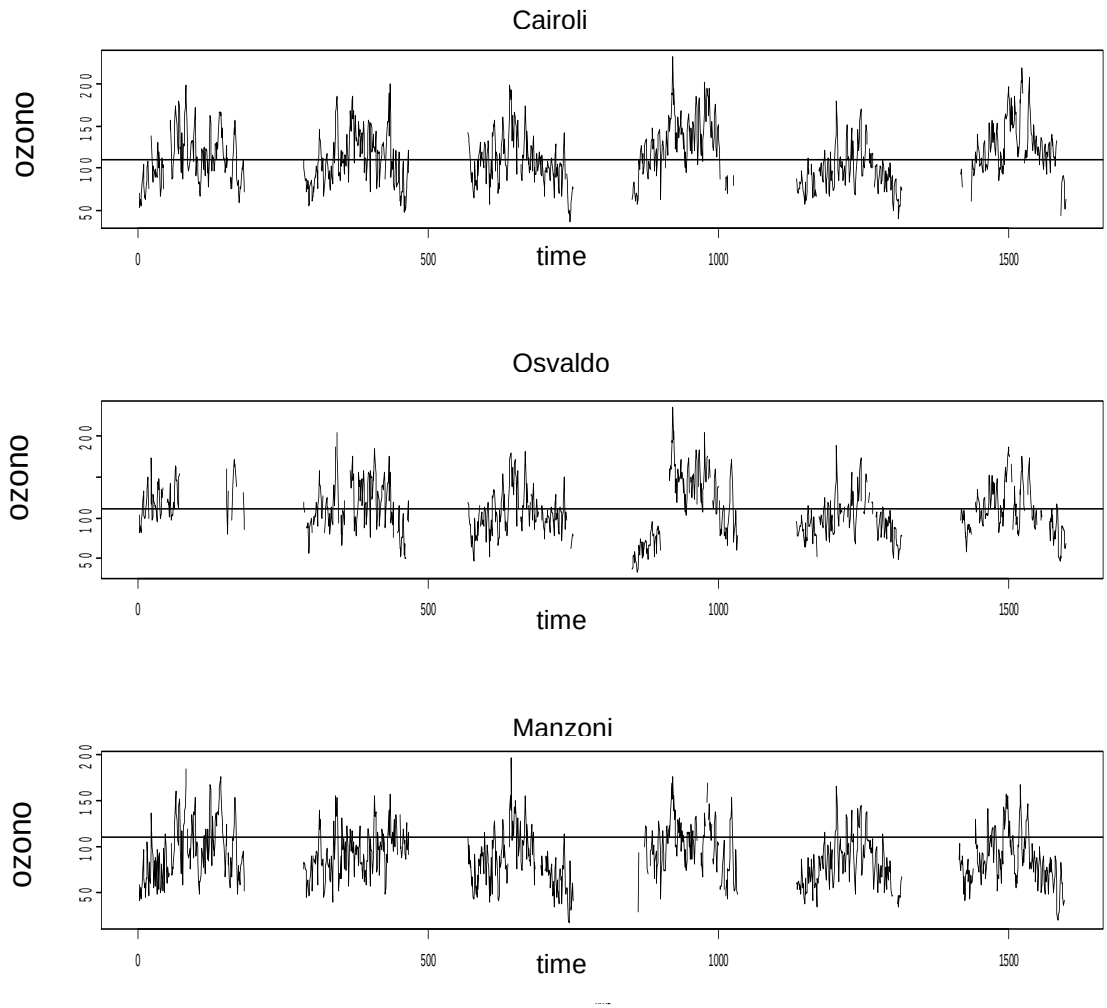
	Manzoni	Osvaldo
Min	17.46	33.95
1st Qu.	68.51	86.91
Median	87.69	104.93
Mean	89.81	108.98
3rd Qu.	107.08	128.54
Max	196.06	233.31
Asimmetria	-0.02	0.01
Curtosi	1.87	1.56

Come possiamo vedere dalle *tabelle 2.2 e 2.4* e dalla *figura 2.7*, il livello di ozono in Cairoli, è più alto rispetto quello presente in Osvaldo e Manzoni. La linea orizzontale rappresenta il livello soglia di 110 µg/m3. Via Manzoni si trova in una zona caratterizzata dalla presenza

Fig. 2.6


Media trascinata
Massimo diurno
Massimo notturno

Fig. 2.7


Cairoli
ozono
time
Osvaldo
ozono
time
Manzoni
ozono
time

di palazzi a molti piani, posta a valle del nucleo urbano rispetto la direzione prevalente del vento; via Cairoli è lontana dalle zone di traffico intenso, in un parco urbano, a Nord Est del centro. Infine, S. Osvaldo è in una zona suburbana, posta a valle della città rispetto alla direzione del vento.

Si osserva un andamento simile dell'ozono nelle stazioni di Cairoli ed Osvaldo, ma nel secondo caso notiamo una più elevata quantità di dati mancanti.

Nella centralina in via Manzoni il livello di O3 è sistematicamente più basso di quello rilevato nelle centraline Cairoli e Osvaldo: per quanto l'ozono sia soggetto a fenomeni di trasporto rilevanti, si suppone qui un effetto dovuto alla diversa presenza degli idrocarburi dovuti al traffico. L'ozono tende infatti a reagire, ossidandoli, con altri inquinanti, quali il monossido d'azoto e gli idrocarburi; in virtù di diverse reazioni chimiche, tuttavia, questi stessi composti partecipano anche alla sua formazione.

	2000	2001	2002	2003	2004	2005
Cairoli	112.28	110.81	108.33	129.41	96.20	126.91
Osvaldo	120.91	114.06	105.91	112.38	98.12	108.80
Manzoni	95.84	93.33	82.52	100.16	82.18	85.92

Sopra sono rappresentati i valori medi dell'ozono nelle tre stazioni, nella stagione estiva e per ciascun anno di riferimento.

Per quanto riguarda i valori massimi orari, questi arrivano a superare ogni anno in tutte tre le stazioni almeno i 150 µg/m3; in Cairoli ed in

Osvaldo, nel 2003 arrivano fino a superare i 230 µg/m3; il massimo in Manzoni è di 196 µg/m3, ed è stato raggiunto nel 2002.

2.4 Variabili usate

Nella seguente tabella sono riportate tutte le variabili utilizzate nello sviluppo dei modelli e i relativi codici identificativi:

Legenda

Variabili usate

cai	stazione di Cairoli
man	stazione di Manzoni
osv	stazione di S' Osvaldo

Inquinanti

run*N*	radice quadrata del massimo giornaliero della media trascinata dell'ozono sulle otto ore
max.O3.N*N*	radice quadrata del massimo notturno dell'ozono
max.O3.D*N*	radice quadrata del massimo diurno dell'ozono
max.NO2.N*N*	massimo diurno del biossido di azoto
max.NO2.D*N*	massimo diurno del biossido di azoto
maxBEN*N*	massimo giornaliero del benzene

Variabili meteorologiche

t.mas*N*	temperatura massima giornaliera

rad.somma*N*	somma giornaliera delle radiazioni solari
pioggia.sommaN	somma giornaliera della quantità di pioggia caduta
uore20.29*N*	media dalle ore 20 alle 5 del giorno successivo della componente est-ovest del vento (positiva da est)
uore.6.9*N*	media dalle ore 6 alle 9 della componente est-ovest del vento (positiva da est)
uore10.17*N*	media dalle ore 10 alle 17 della componente est-ovest del vento (positiva da est)
vore20.29*N*	media dalle ore 20 alle 5 del giorno successivo della componente nord-sud del vento (positiva da nord)
vore.6.9*N*	media dalle ore 6 alle 9 della componente nord-sud del vento (positiva da nord)
vore.10.17*N*	media dalle ore 10 alle 17 della componente nord-sud del vento (positiva da nord)
press.mean*N*	pressione media giornaliera
ur.mean*N*	umidità media giornaliera
ins.somma*N*	tempo di insolazione giornaliero
mese	mese di riferimento (stagione dell'ozono va da aprile a settembre)
giorno	giorno della settimana
anno	anno di riferimento (2000-2005)
fatrun*N*	variabile qualitativa relativa alla suddivisione fatta in base ai quantili della distribuzione e al limite di legge del massimo giornaliero della media trascinata dell'ozono sulle otto ore
fato3D*N*	variabile qualitativa relativa alla suddivisione fatta in

	base ai quantili della distribuzione e al limite di legge del massimo diurno dell'ozono
MAXnotte	variabile qualitativa relativa alla suddivisione fatta in base ai quantili della distribuzione massimo notturno dell'ozono

N è uguale a uno se corrisponde al giorno corrente, , N-1 sono i giorni precedenti

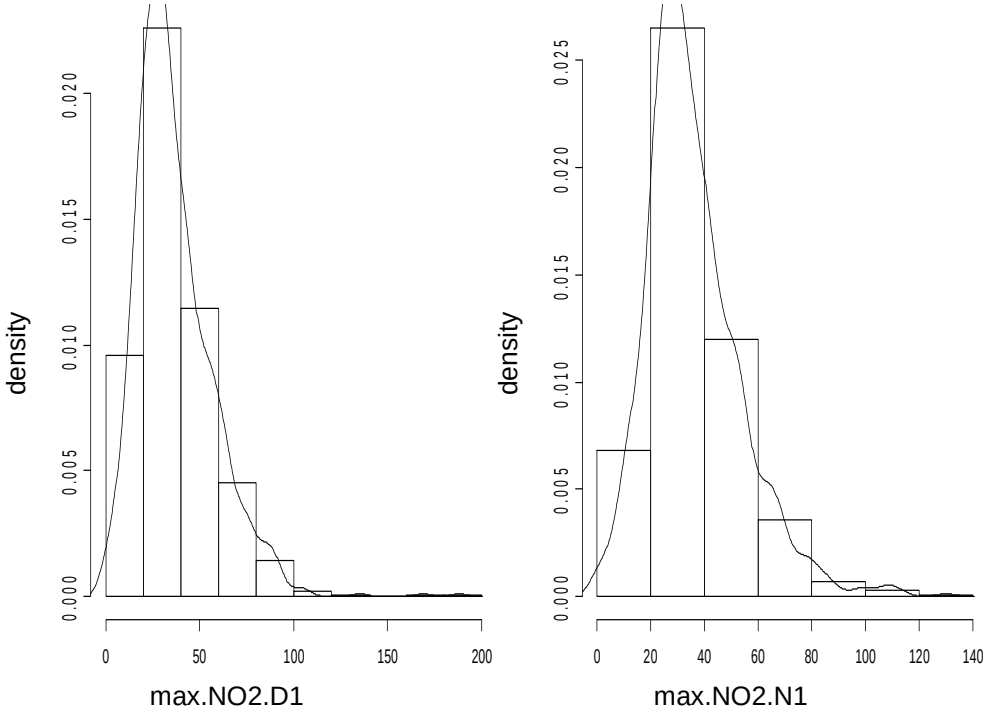
2.5 Precursori

2.5.1 Ossidi di azoto

Anche questa variabile è stata divisa in massimo notturno e diurno, per coerenza con la scelta fatta per l'ozono.

Se osserviamo la tendenza centrale delle distribuzioni di via Cairoli, *figura 2.8 e tabella 2.5*, notiamo che c'è asimmetria positiva, confermata anche dal valore dell'indice di simmetria. Entrambe le distribuzioni hanno un alto grado di curtosi. In tutti e due i casi il livello massimo sta sotto il livello di attenzione di 200 µg/m3.

Fig. 2.8



Tab. 2.5

	max.NO2.D	max.NO2.N
Min	0	0
1st Qu.	22.65	24.28
Median	32.38	32.63
Mean	36.72	35.92
3rd Qu.	46.72	44.55
Max	188.1	130
Asimmetria	0.85	0.67
Curtosi	4.77	3.19
NA	163	170

Per quanto riguarda gli episodi acuti (*tabella 2.6*), nel corso del 2004 si è registrato solamente un superamento del limite orario (a S. Osvaldo), evidenziando una situazione nettamente più tranquillizzante rispetto al 2003.

Anche in via Manzoni troviamo dei livelli più alti di NO2 risalenti ad anni 2000, 2002 e 2003.

Tab. 2.6

	Manzoni max.NO2.D	Osvaldo max.NO2.D
Min	0	0
1st Qu.	37.42	20.14
Median	55.57	32
Mean	61.98	35.18
3rd Qu.	79.29	46
Max	307.7	236.9
Assimmetria	1.04	1.09
Curtosi	5.54	7.43
NA	109	229

Per concludere sono stati esaminati gli scatterplots tra ozono e NO2 (*fig. 2.9 e 2.10*). Dai grafici si può vedere dunque che a bassi livelli di NO2 corrispondano i più alti di ozono.

Fig. 2.9

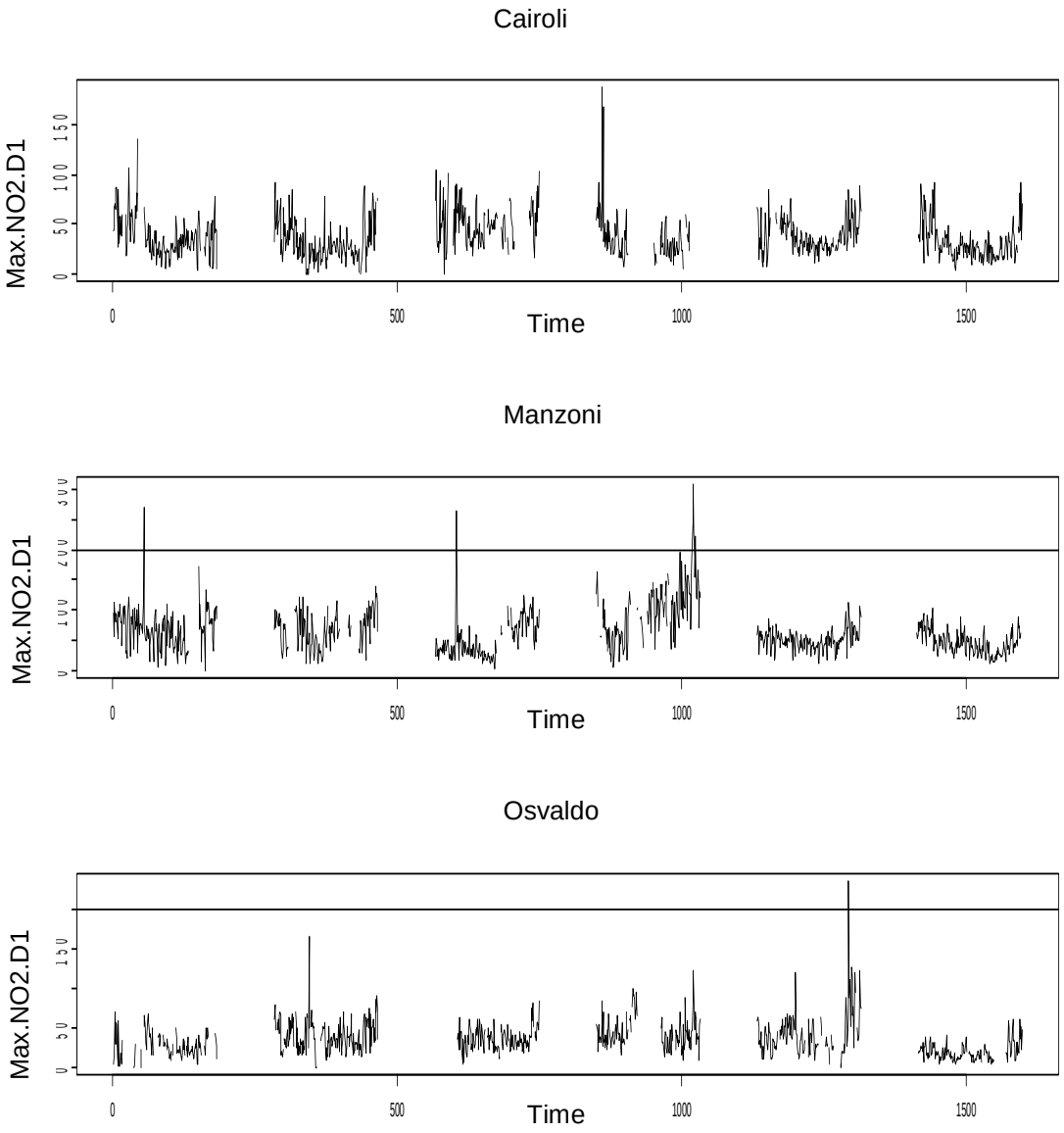
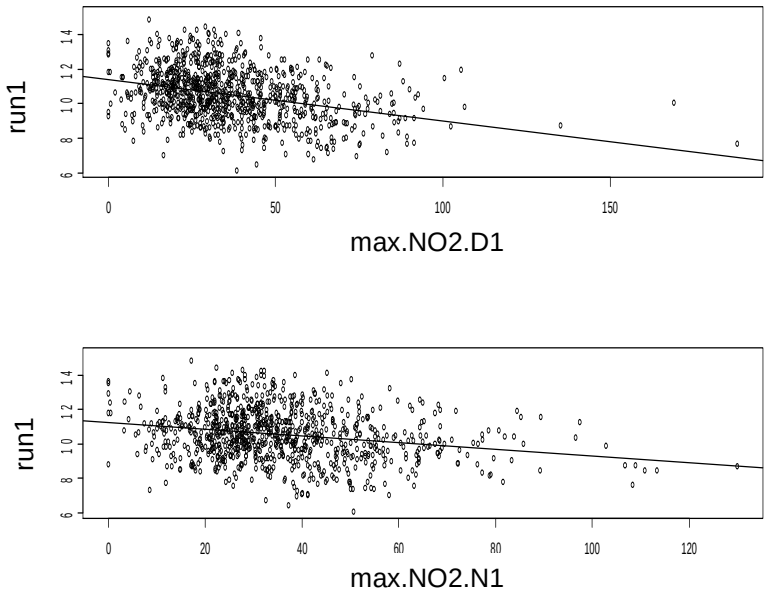

Cairoli
Max.NO2.D1
0
50
100
150
0
500
1000
1500
Time
Manzoni
Max.NO2.D1
0
500
1000
1500
Time
Osvaldo
Max.NO2.D1
0
500
1000
1500
Time

Fig. 2.10

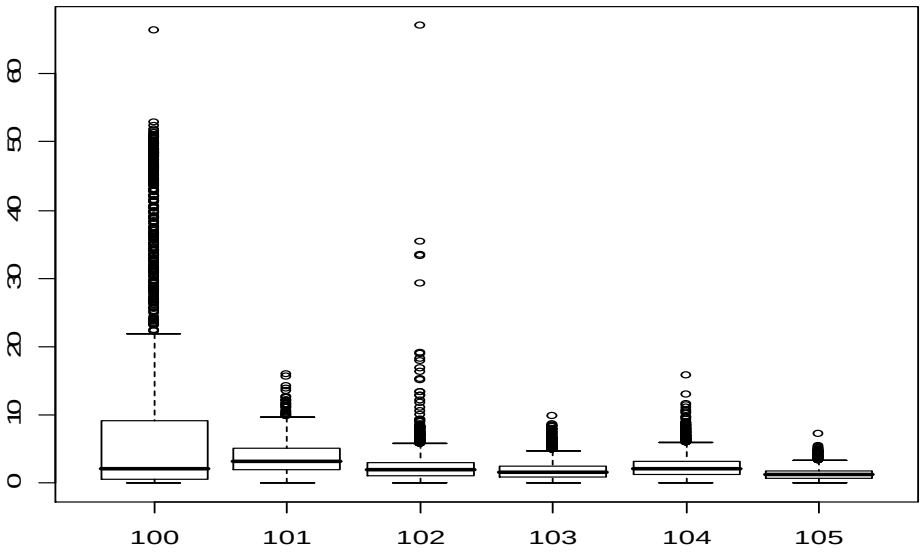


2.5.2 Benzene

I dati che abbiamo sul benzene sono registrati unicamente dalla centralina in via Manzoni; inoltre quelli relativi agli anni dal 2002 al 2005 sono più completi.

La distribuzione dei valori misurati nell'anno 2000 presenta numerosi *outlier (figura 2.11)*; nello stesso anno ci sono molti valori mancanti del benzene e la centralina ha subito diversi cambiamenti: ciò induce ad utilizzare questa variabile con una certa cautela, nello sviluppo dei modelli di dipendenza dell'ozono.

Fig. 2.11



Nella *tabella 2.7* notiamo alcuni valori elevati del massimo giornaliero del benzene; in particolare, il valore massimo (67µg/m3) è stato rilevato 2002-05-09, subito dopo un guasto nell'analizzatore. Oltre a tale outlier, troviamo solo 19 valori che superano i 15 µg/m3, la gran parte appartenenti al 2002 e che sono responsabili degli alti livelli di asimmetria e curtosi.

La linea orizzontale presente nella *figura 2.13* rappresenta il livello di 15 µg/m3.

Vediamo la relazione tra il massimo giornaliero del benzene e la variabile run, riferita agli anni 2000-2005 (*fig. 2.12*). Nella *figura 2.14* si evidenziano invece i dati degli ultimi tre anni (2003-2005).

Vediamo una relazione di anticorrelazione tra benzene e ozono: infatti al diminuire del benzene l'ozono aumenta.

Tab. 2.7

	2000-2005 maxBEN1	2002-2005 maxBEN1
Min	0	0
1st Qu.	2.93	2.96
Median	4.19	4
Mean	6.43	4.77
3rd Qu.	6.65	5.53
Max	67	67
Assimmetria	1.63	3.66
Curtosi	9	53
NA	494	264

Fig. 2.12

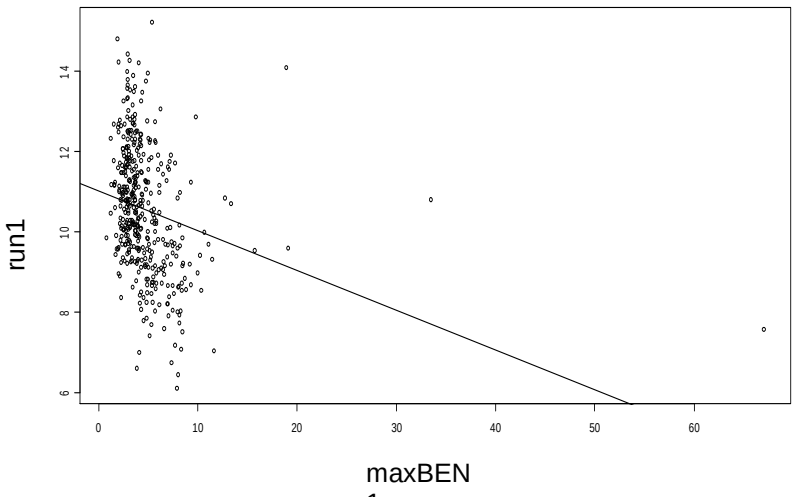


Fig. 2.13

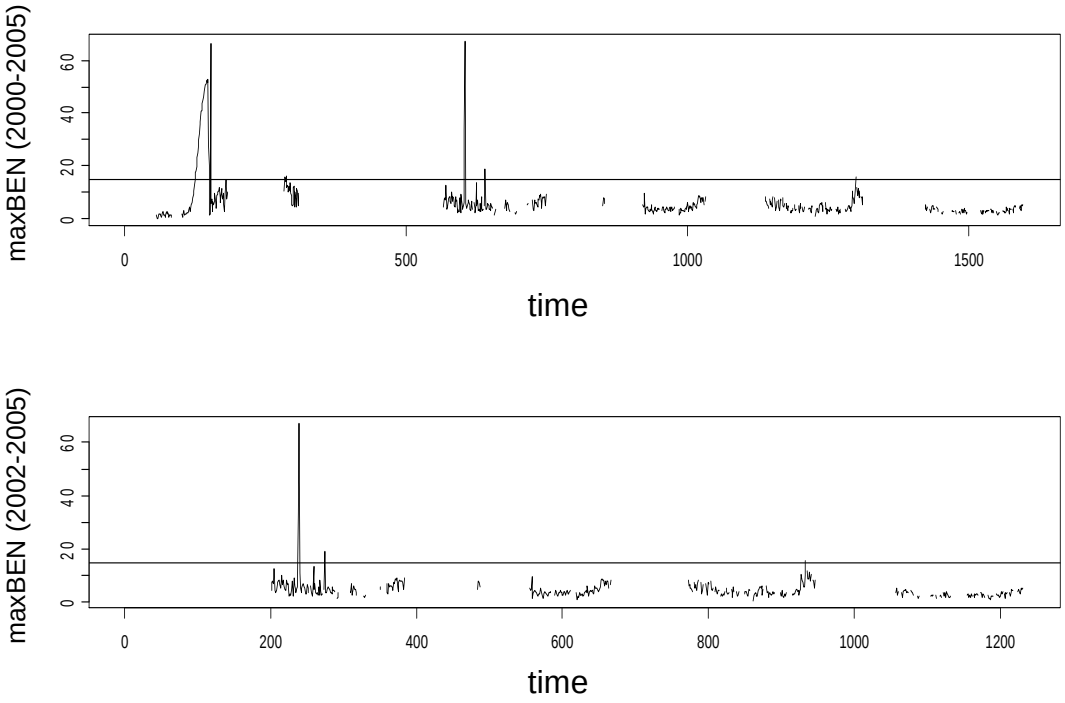
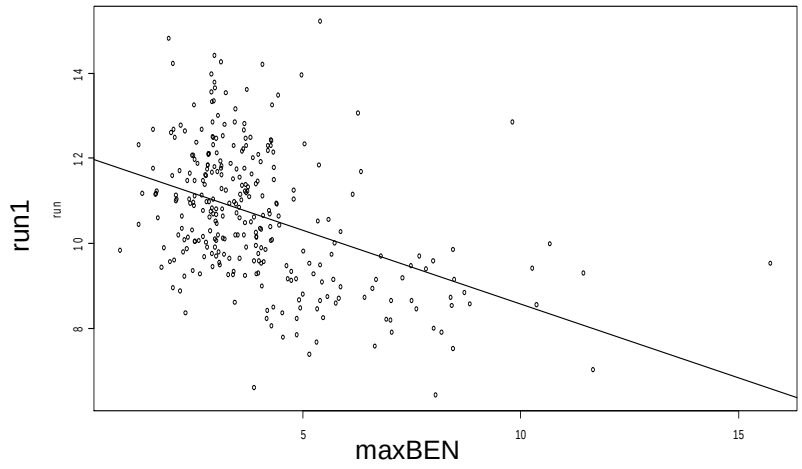


Fig. 2.14



Vediamo quindi come il Benzene influenza l'ozono.

Anche sui dati dei precursori sono state condotte analisi autoregressive: in tutti i casi hanno mostrato una forte persistenza.

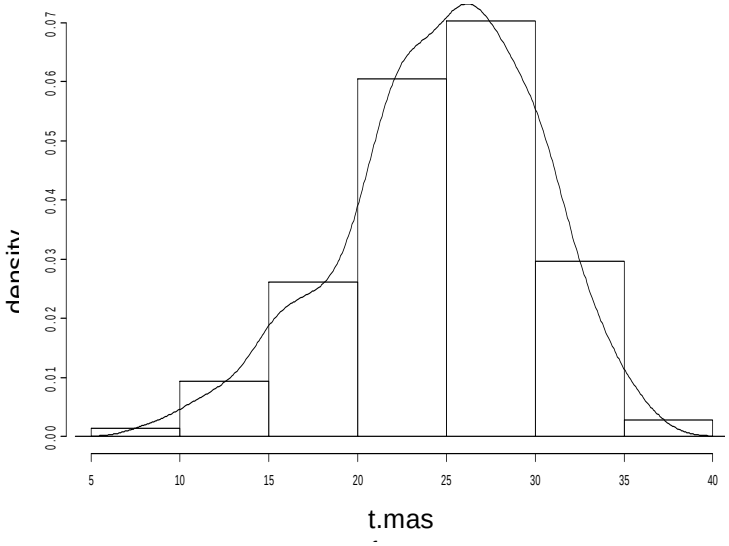
2.6 Variabili meteorologiche:

L'analisi dei dati meteorologici può dare un supporto interpretativo all'interpretazione fornita dalle statistiche descrittive calcolate sui dati chimici.

2.6.1 Temperatura

Poiché la formazione dell'ozono è legata alle condizioni meteo, in particolare all'irraggiamento solare ed alla temperatura, iniziamo ad osservare la distribuzione della temperatura nel periodo considerato (*fig. 2.15*, *tab. 2.8*).

Fig. 2.15



Tab. 2.8

	t.mas
Min	7.7
1st Qu.	21.5
Median	25.3
Mean	24.72
3rd Qu.	28.7
Max	37.4
Assimmetria	-0.3
Curtosi	2
NA	3

Ricordiamo che in questo lavoro sono stati considerati i mesi più caldi dell'anno; uno spiccato aumento della temperatura si ha nei mesi di giugno e luglio di ogni anno.

Il 2003 (*figura 2.16)* è stato un anno molto particolare, in quanto più caldo rispetto agli altri e molto meno piovoso. Nello stesso anno vediamo anche dei picchi di ozono: il clima "torrido" che ha caratterizzato l'estate 2003 ha sicuramente influito in modo significativo sul fenomeno, infatti mediamente il valore massimo della temperatura nel 2003 è il più alto:

2000	2001	2002	2003	2004	2005
24.70	23.99	24.01	**27.14**	24.11	24.35

Fig. 2.16

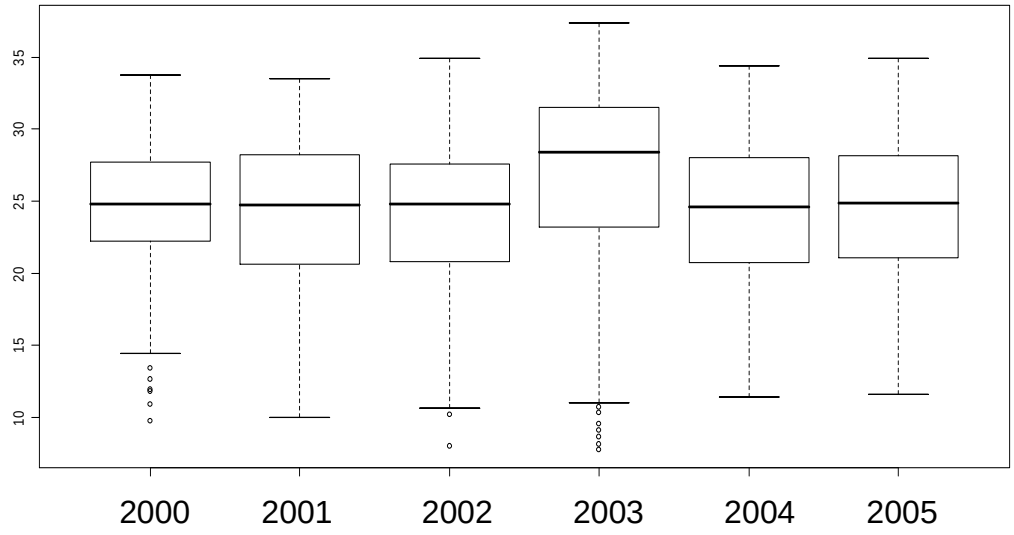
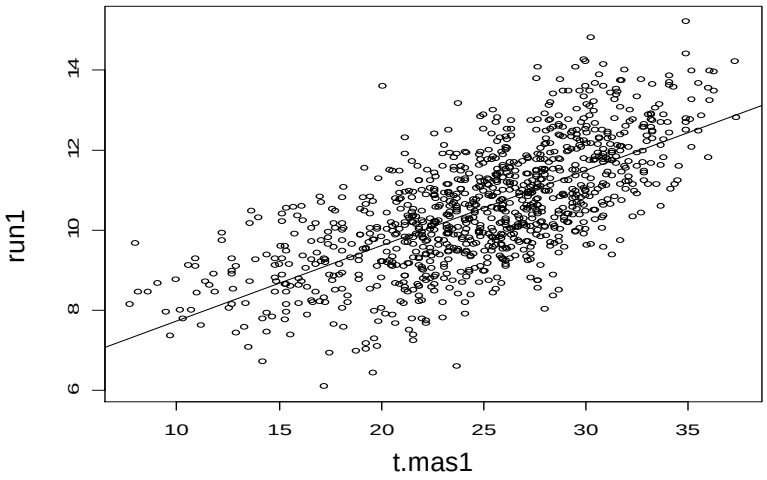


Fig. 2.17



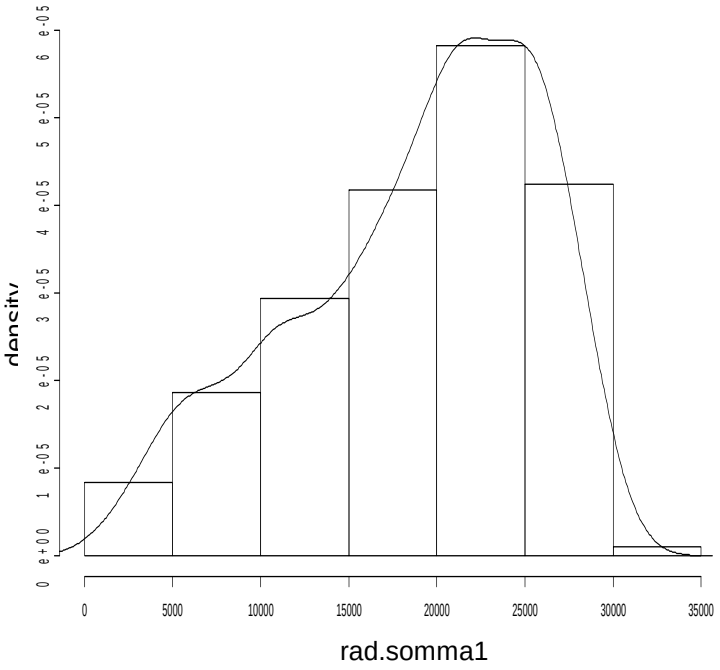
La *figura 2.17* fa vedere la forte correlazione tra ozono e temperratura

2.6.2 Irraggiamento

Per quanto riguarda l'irraggiamento solare, abbiamo due variabili che lo descrivono: il totale giornaliero della radiazione solare incidente sulla superficie (kJoule/m^2) e il tempo totale giornaliero (min) in cui il disco solare è stato visibile (eliofania).

Radiazione solare (*fig. 2.18, 2.19*) e tempo di insolazione *(fig. 2.20, 2.21*), oltre ad essere molto importanti per stabilire il livello di ozono giornaliero, sono molto correlate tra loro (*fig.2. 22*).

Fig. 2.18



Tab. 2.9

	rad.somma
Min	499
1st Qu.	14006
Median	20179
Mean	18792
3rd Qu.	24439
Max	30994
Assimmetria	-0.3
Curtosi	1.65
NA	3

Fig. 2.19

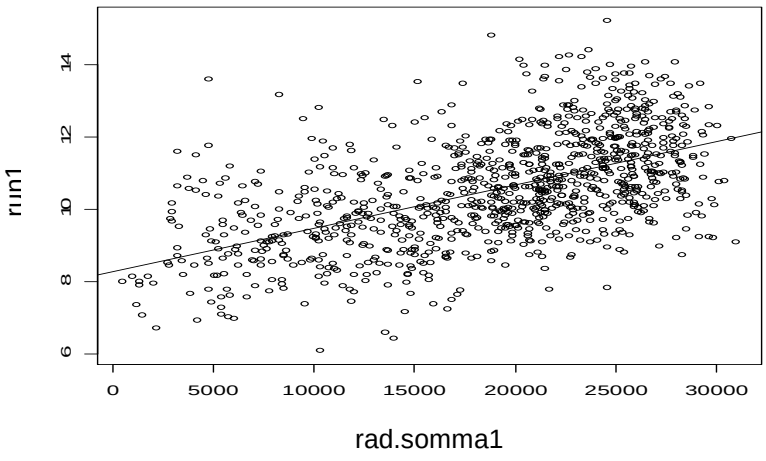
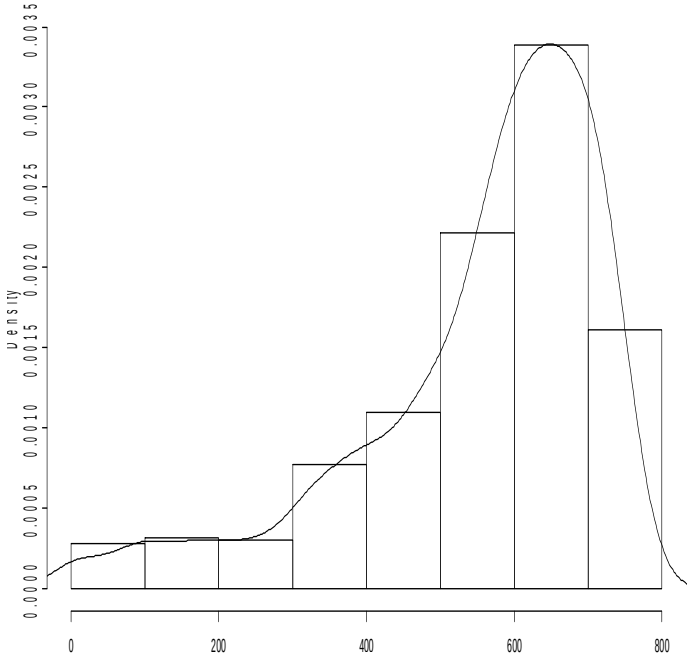


Fig. 2.20



Tab. 2.10

	ins.somma
Min	0
1st Qu.	482
Median	601
Mean	553.4
3rd Qu.	6750
Max	777
Assimmetria	-0.9
Curtosi	2.98
NA	3

Fig. 2.21

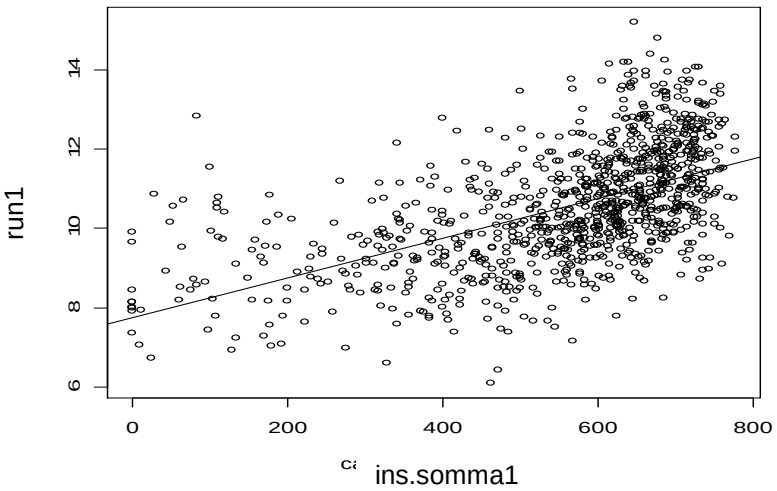
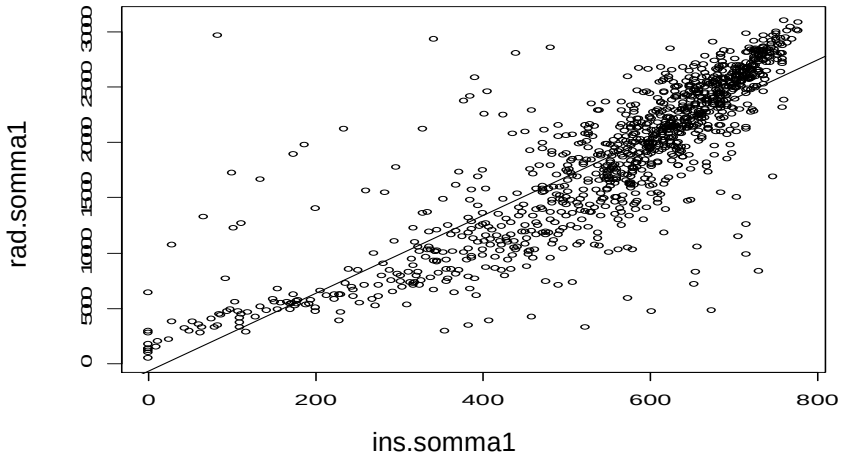


Fig. 2.22



A maggiore tempo di insolazione, corrisponde un più intenso totale giornaliero della radiazione solare; nel 2000, 2001 e 2003 troviamo più alti valori di entrambe le variabili:

	2000	2001	2002	2003	2004	2005
rad.somma:	19377.17	19145.82	18241.62	19097.63	18567.56	18320.99
ins.somma:	575.13	562.41	546.85	569.48	534.06	532.28

2.6.3 Pioggia

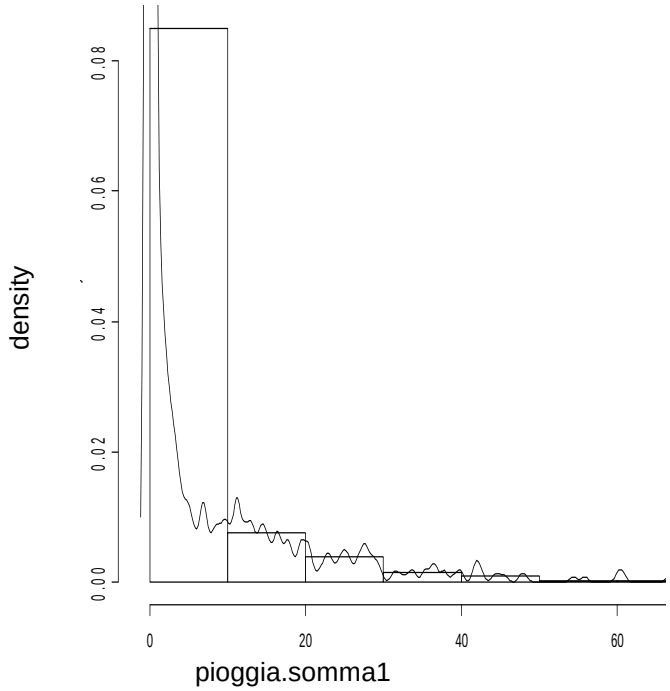
Al contrario, ora consideriamo uno tra i principali meccanismi di rimozione degli ossidi di azoto e dell' ozono: le precipitazioni. La pioggia contribuisce a ripulire l'atmosfera; quindi, anche se il suo effetto riguarda soprattutto la diminuzione di PM10, può abbassare il livello di ozono presente

nell'aria. L'assenza di questo fenomeno favorisce viceversa l'accumulo degli inquinanti e incrementa l'inversione termica[5]. Nella *figura 2.23* e *tabella 2.11* sono rappresentati i dati relativi a questa variabile.

Non si nota, però, una particolare relazione con l'ozono *(fig. 2.24).*

Vediamo che il 2003 è stato l'anno meno piovoso con una somma di pioggia pari a 3.14 mm/giorno (*fig. 2.25*).

Fig. 2.23

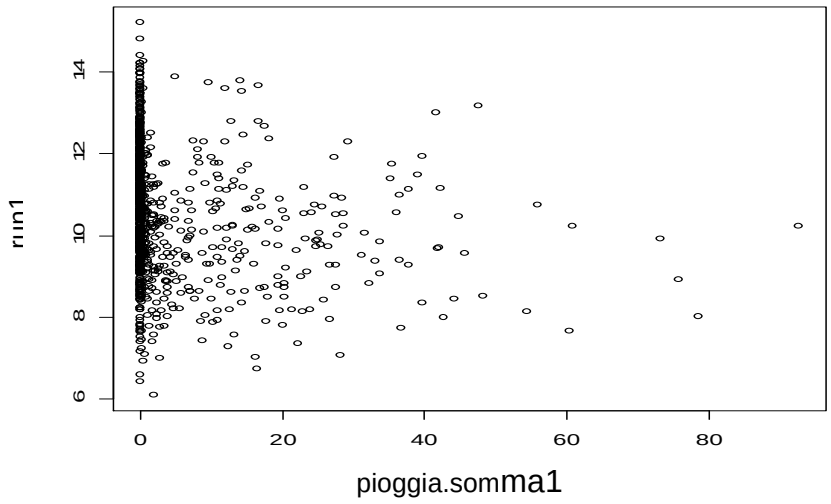


Tab. 2.11

	pioggia.somma
Min	0
1st Qu.	0
Median	0
Mean	4.333
3rd Qu.	2.4
Max	92.6
Assimmetria	2.5
Curtosi	13
NA	0

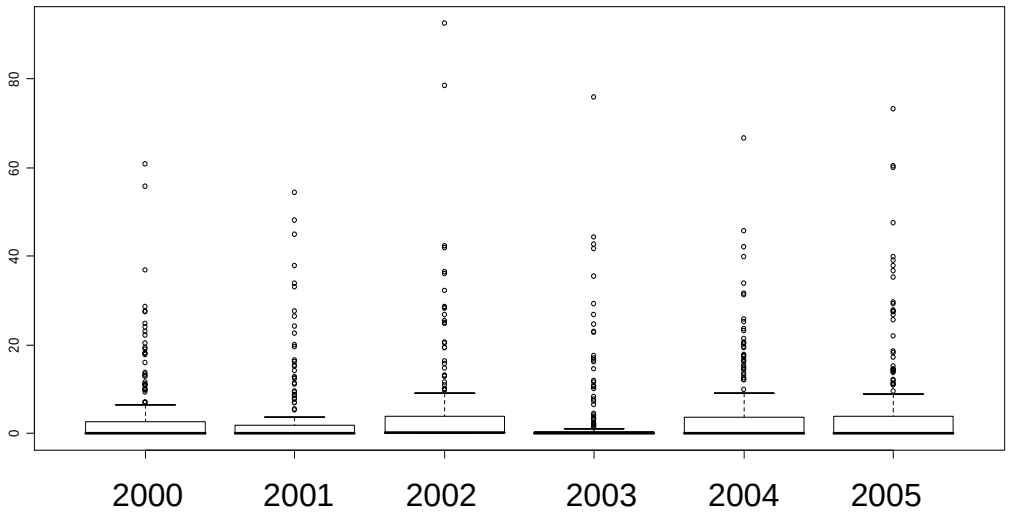
[5] Fenomeno meteorologico per cui la temperatura dell'aria tende ad aumentare al crescere al crescere dell'altitudine anziché diminuire.

Fig. 2.24



2000	2001	2002	2003	2004	2005
3.92	3.69	5.07	**3.14**	4.78	5.38

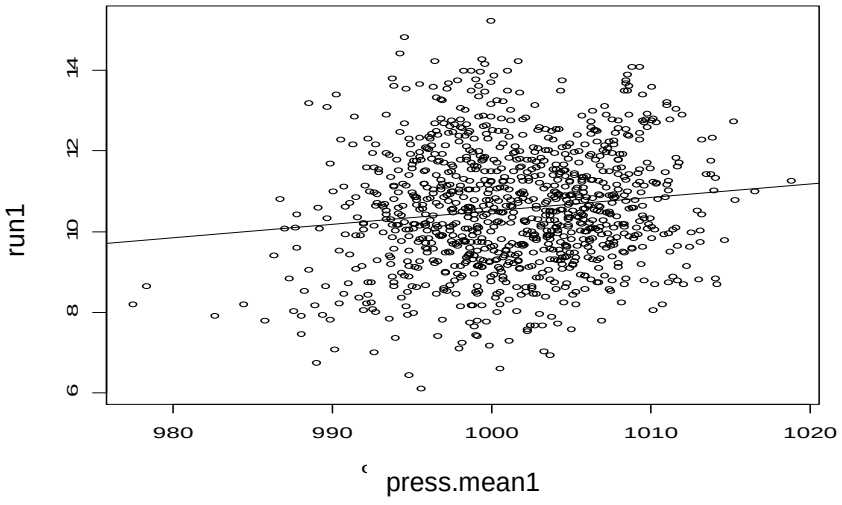
Fig. 2.25



2.6.4 Pressione

A questo punto, effettuiamo un'analisi descrittiva sulle altre variabili di cui disponiamo. La pressione varia durante il giorno perchè dipende dalla variazione della temperatura della colonna d'aria. Infatti un suo riscaldamento rende l'aria meno densa e quindi più leggera. Quindi durante le ore più calde del giorno, in seguito al riscaldamento del suolo e di conseguenza dell'aria negli strati più bassi, si ha un minimo di pressione. Di notte invece la colonna d'aria si raffredda e quindi si ha un aumento del suo peso, ossia della pressione. La pressione può cambiare anche per il movimento delle masse d'aria: nella stessa colonna può entrare o uscire aria per cui se quella che entra è maggiore di quella che esce assisteremo ad un aumento di pressione, viceversa si avrà una sua diminuzione. L'oscillazione diurna descritta sopra è quindi valida in linea di massima, bisogna, cioè, tenere conto anche della variazione del contenuto d'aria nella colonna che stiamo considerando. Di solito alle basse pressioni si associano nubi e piogge.

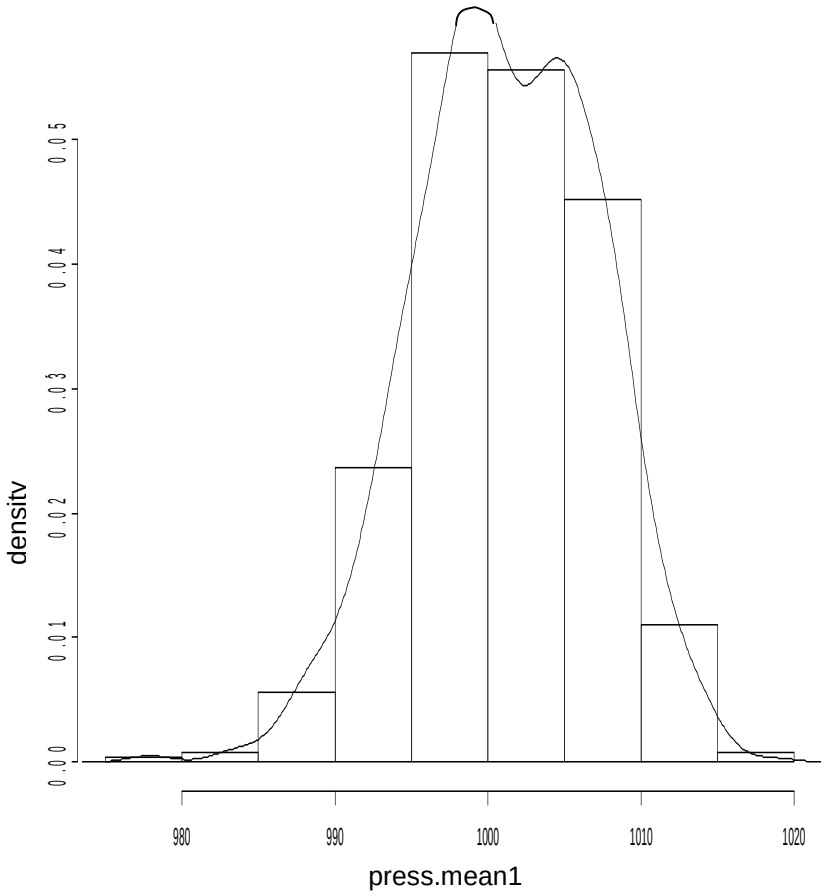
Fig. 2.26



Per le basse pressioni associate alle perturbazioni, la nuvolosità è spiegabile con considerazioni di dinamica atmosferica. Le zone di alta pressione, invece, tengono lontane le nubi: queste zone sono caratterizzate da aria fredda che scende verso il suolo e che va in contro a compressione. Questo scongiura la condensazione e favorisce cielo sereno.

L'ozono aumenta con l'alta pressione (*fig. 2.26, 2,27*): infatti quest'ultima aumenta con l'aumentare delle radiazioni solari e della temperatura.

Fig. 2.27



Tab. 2.12

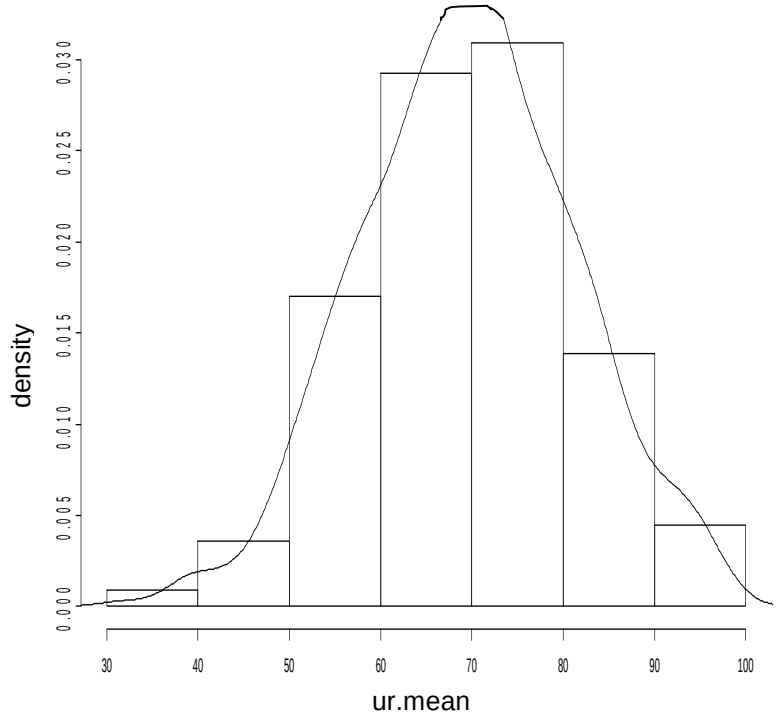
	press.mean
Min	977.5
1st Qu.	997.2
Median	1001.1
Mean	1001.2
3rd Qu.	1005.7
Max	1018.9
Assimmetria	-0.15
Curtosi	2.03
NA	500

2.6.5 Umidità

La misura più utilizzata per dercrivere l'umidità atmosferica è l'umidità relativa. Essa esprime il rapporto percentuale fra la quantità di vapor acqueo presente nell'aria e la quantità che, alla stessa temperatura, sarebbe necessaria perchè il vapore condensi in microscopiche goccioline d'acqua. A questo punto si dice che l'aria è satura di vapore e, aggiungendo altro vapore, questo condensa in ulteriori goccioline.

La saturazione dipende fondamentalmente dalla temperatura dell'aria. Con basse temperature basta poco vapor acqueo perchè si condensi una nuvola, mentre ne occorre molto di più man mano che la temperatura sale.

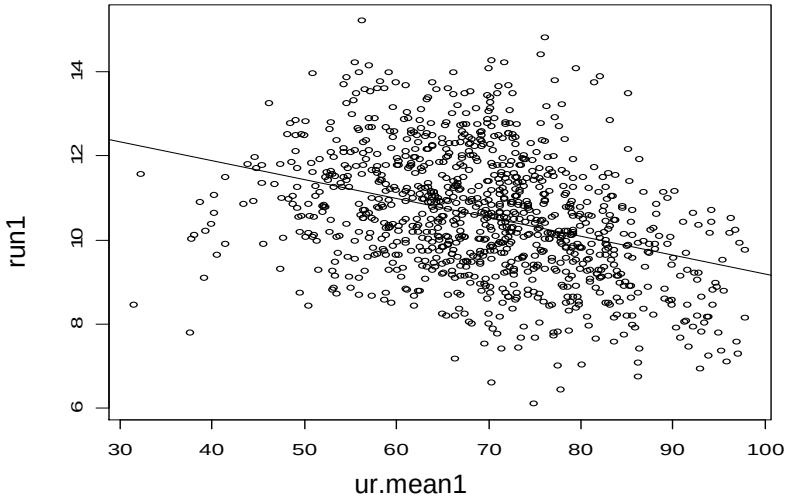
Fig. 2.28



Tab. 2.13

	ur.mean
Min	31.54
1st Qu.	61.75
Median	69.88
Mean	69.54
3rd Qu.	77.6
Max	97.92
Assimmetria	-0.04
Curtosi	1.96
NA	0

Fig. 2.29



Quando l'umidità relativa dell'aria raggiunge il 100%, l'eccesso di vapor acqueo condensa in minuscole goccioline d'acqua e si assiste alla formazione di una nube.

Il tasso di umidità si situa idealmente tra il 40 e 60%. Quindi da questi dati risulta abbastanza elevata, nell'area di interesse *(fig. 2.28, tab. 2.13).*

Dai dati orari la relazione tra pressione ed umidità con l'ozono risulta molto più evidente (*fig. 2.29*), soprattutto la diminuzione dell'ozono con l'aumento dell'umidità. Del resto l'umidità aumenta all'aumentare delle piogge e al diminuire delle radiazioni solari e della temperatura. Anche pressione ed umidità sono correlati negativamente tra loro: alta umidità corrisponde a bassa pressione.

2.6.6 Vento

Consideriamo infine il vento, importante perché funziona come agente di dispersione degli inquinanti e favorisce il rimescolamento dell'aria.

La velocità del vento orizzontale è descritta da due variabili, date dall'angolo e dall'intensità. Nell'analisi sono state create, da quelle originarie, due diverse variabili, u e v: la prima *(fig. 2.30, tab. 2.14)* corrisponde al vento nella direzione est-ovest (segno positivo per il vento proveniente da est, mentre v *(fig. 2.32, tab. 2.15)* corrisponde al vento da nord-sud (positivo da nord).

Sono state calcolate tre velocità del vento medie, per ognuna delle due variabili in esame, relative a tre diverse fasce orarie, una notturna, dalle ore 20 alle 5 del giorno dopo, una relativa alla mattina presto, ore 6 – 9, e una pomeridiana, 10 – 17 (Davis & al., 1999).

I venti dominanti, nel sito esaminato, provengono da Nord-Est.

Si nota innanzitutto che il valore massimo della velocità del vento rilevata è molto bassa (5.2 m/s), che, secondo la scala di Beauford indica *brezza tesa*. La media si attesta sulla *bava di vento*. Quindi notiamo la presenza di venti piuttosto deboli. Questi valori così bassi possono indicare condizioni di ristagno, che però devono essere supportate da informazioni meteorologiche supplementari, quali il gradiente verticale della velocità del vento e della temperatura.

Condizioni di ristagno dell'aria (nebbia), o siccità, scarsa ventilazione ed elevata umidità aggravano in generale i danni provocati dall'inquinamento; nel caso dell'ozono, tuttavia, il forte legame con l'irraggiamento solare contrasta con tali meccanismi.

Tab. 2.14

	uore.20.291	uore.6.91	uore10.171
Min	-1.49	-1.56	-2.46
1st Qu.	0.33	0.38	-0.42
Median	0.58	0.76	0.29
Mean	0.73	1.04	0.56
3rd Qu.	0.95	1.39	1.19
Max	4.64	6.05	7.31
Assimmetria	1.02	1.04	0.82
Curtosi	4.64	4.08	3.36
NA	11	2	0

Fig. 2.30

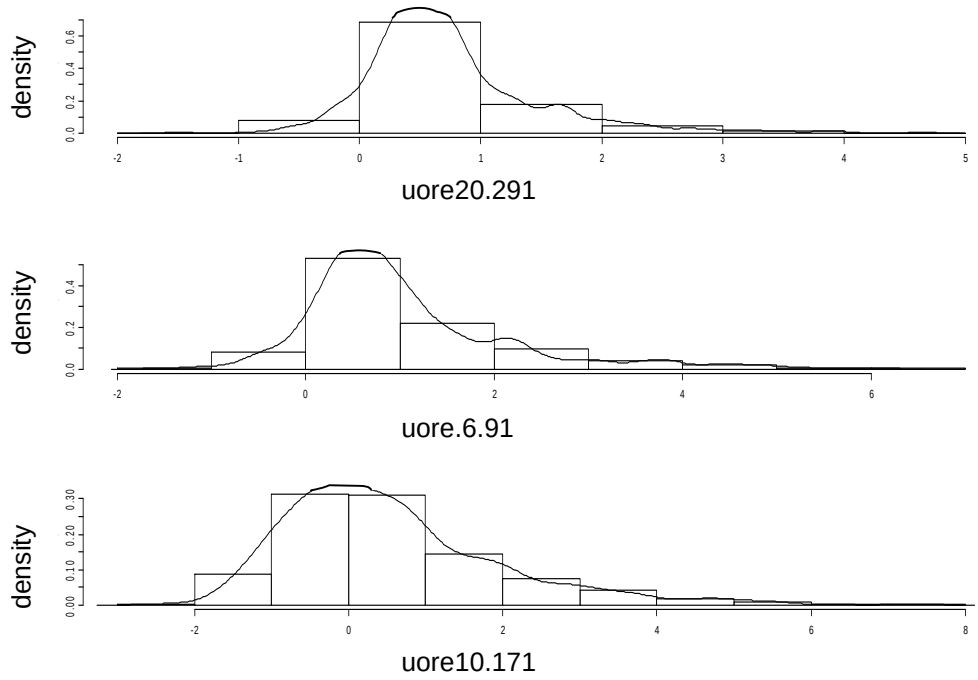
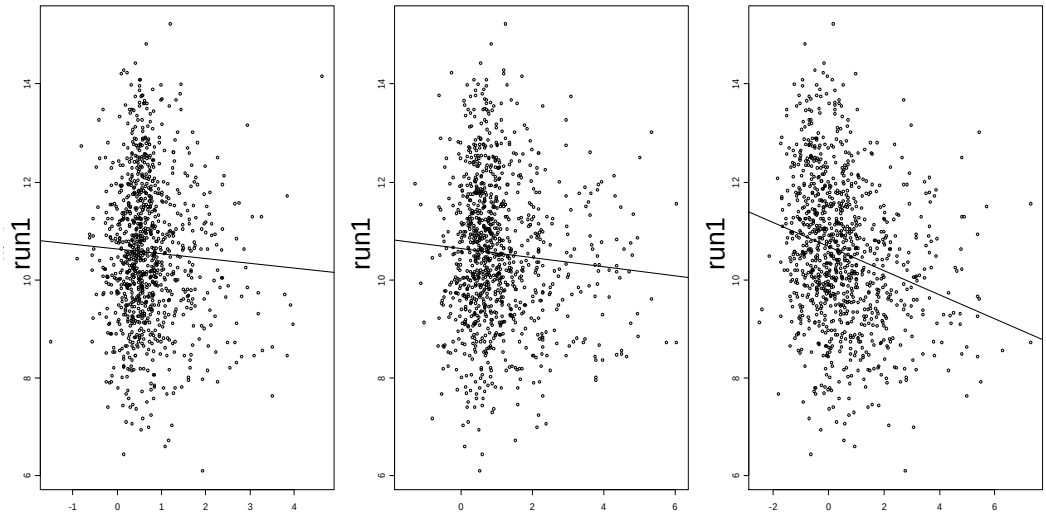


Fig. 2.31



Tab. 2.15

	vore.20.291	vore.6.91	vore.10.171
Min	-2.87	-2.08	-4.68
1st Qu.	0.43	0.32	-1.48
Median	0.81	0.77	-0.79
Mean	0.86	0.84	-0.64
3rd Qu.	1.21	1.30	0.09
Max	3.85	5.22	3.91
Assimmetria	0.26	0.25	0.29
Curtosi	3.6	3.1	2.61
NA	11	2	0

Fig. 2.32

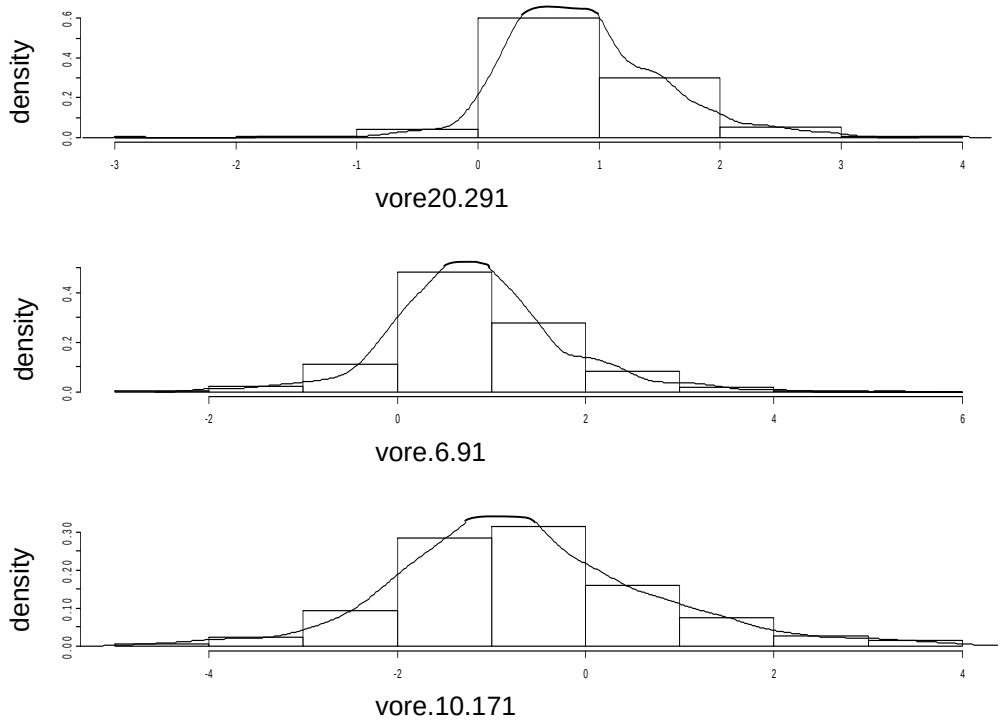
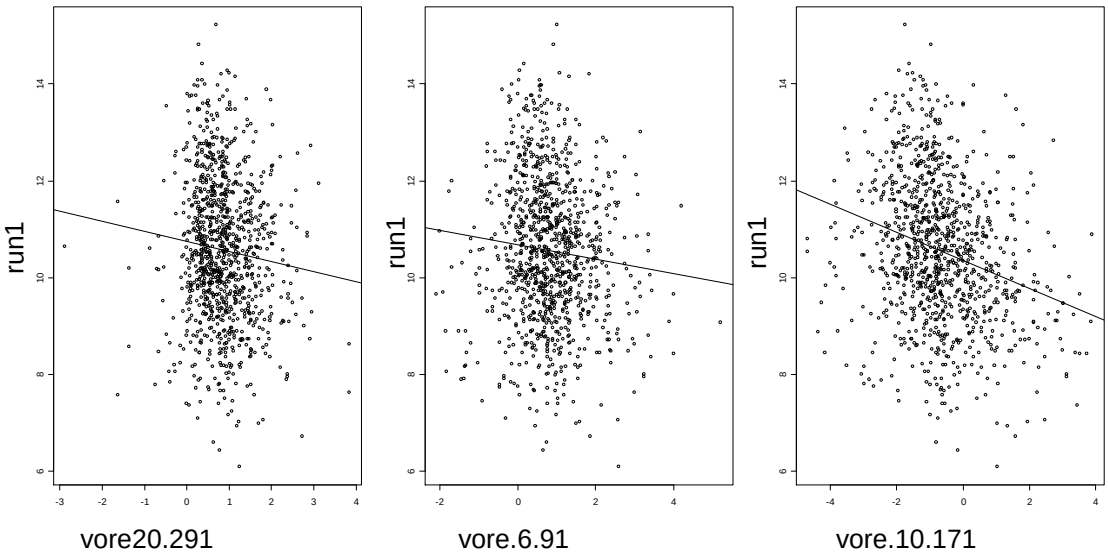


Fig. 2.33



Un’evidente relazione più forte con l’ozono, sia in termini di angolo che intensità, ce l’ha il vento che soffia dalle ore 10 alle 17 *(fig. 2.31, 2.33).*

Si è visto quindi quali sono le variabili che più influenzano l’ozono dagli scatterplot tra quest’ultimo e tutte le variabili che potrebbero essere significative per l’analisi. Sembra che ad alti livelli di NO2 e Benzene, corrispondano bassi valori di ozono.

Per quanto riguarda le variabili meteorologiche, alta temperatura e radiazioni solari, unite a poche piogge, alta pressione, poca umidità, scarso vento, determinino l’aumentare dell’ozono.

3 - I modelli

3.1 Modelli lineari

E’ ampliamente riconosciuto che le relazioni tra ozono, precursori (quali Nox, benzene) e variabili meteorologiche sono complesse e non lineari (Bordignon & al., 2002).

A causa di queste difficoltà, per prevedere le concentrazioni di ozono vengono frequentemente usati modelli stocastici basati sui metodi di regressione multipla non lineare.

E’ stata tuttavia inizialmente tentata una regressione lineare, per la semplicità di interpretazione che caratterizza tale modello.

La regressione ha come scopo principale la previsione (Barrero & al., 2005; Friedman, 1991): si mira, cioè, alla costruzione di un modello attraverso cui prevedere i valori di una variabile dipendente (risposta), a partire dai valori di più (regressione lineare multipla) variabili indipendenti (o esplicative). L’obiettivo sarà identificare quali sono le variabili più significative individuando in tal modo il miglior modello di regressione multipla in grado di spiegare la variabilità dell’ozono.

Attraverso l’analisi dei residui e della varianza, si è ottenuta una prima indicazione sulle variabili principali che spiegano le concentrazioni di ozono. Questa procedura permette il confronto simultaneo tra due o più medie, mantenendo invariata la probabilità alfa complessiva prefissata.

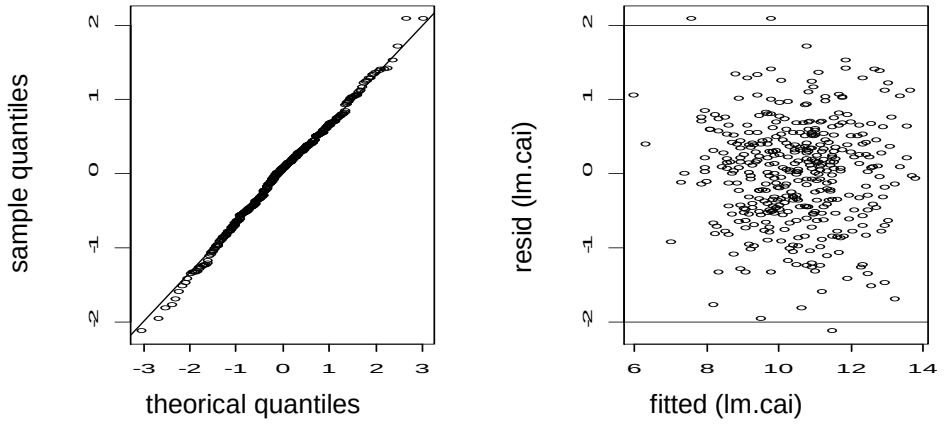
ANOVA utilizza la distribuzione F per verificare la significatività delle differenze tra le medie di vari gruppi.

Le stesse conclusioni sono state confermate usando il metodo un approccio di selezione delle variabili all’indietro. Quest’ ultima consiste in una

procedura di selezione in cui tutte le variabili vengono inserite nell'equazione e in seguito rimosse. Viene considerata in primo luogo la variabile con la più piccola correlazione parziale con la variabile indipendente. Se risponde al test di verifica per l'eliminazione viene rimossa. Dopo di che si considera la variabile con la più piccola correlazione parziale restante nell'equazione. L' algoritmo si arresta quando non ci sono variabili nell'equazione che soddisfano i test di verifica di rimozione. Il modello può considerarsi buono se la differenza tra devianza residua e quella stimata con il novantacinquesimo percentile di una X^2_k (dove k sono i gradi di libertà relativi alla distribuzione in esame) è piccola. Il modello saturo è quello massimo a cui so associa una variabile per ogni osservazione e spiega al 100%.

Il modello finale, riferito all' ozono, indica che le variabili che spiegano meglio il suo livello in Cairoli sono: l'ozono del giorno precedente, il livello di benzene presente nell'aria, la temperatura massima, l'umidità, la pressione media ed il tempo d'irraggiamento solare del giorno stesso, la temperatura, l'umidità e la pressione media del giorno precedente; sono seguite dall' irraggiamento solare, velocità media del vento nelle ore 10-17 della

Fig. 3.1

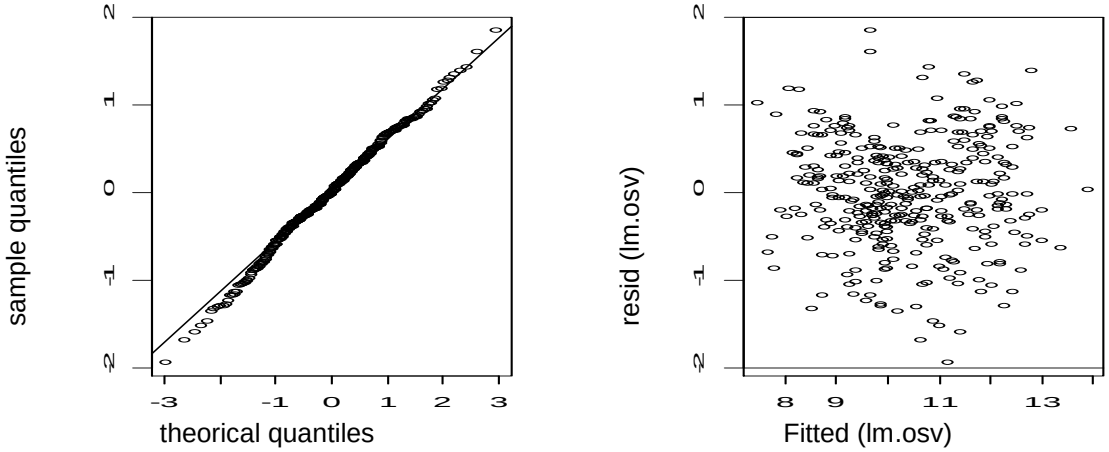


componente nord - sud del giorno corrente oggi,. Il modello ha una devianza residua pari a 185.63 con 369 gradi di libertà (AIC: 882.85)[6].

Dalla *figura 3.1* si osserva che la distribuzione dei residui è quasi normale, le code non sono pesanti, i residui sono raccolti in un intervallo limitato.

Eseguita la stessa analisi per quanto riguarda Osvaldo (*figura 3.2* relativa ai residui del modello), le variabili significative si sono rivelate pressoché le stesse; assumono però importanza anche il mese, l'anno (segno che il set di variabili esplicative scelte potrebbe essere carente di qualche predittore), la quantità di precipitazioni del giorno corrente, mentre la radiazione solare ne perde. Il modello ha una devianza residua pari a 181.44 con 424 gradi di libertà (AIC: 881.08).

Fig. 3.2



[6] Tutte la variabiledei modelli sono riassunte nella *tab.17*

In Manzoni la situazione cambia: il benzene del giorno corrente perde importanza mentre ne assume molta quello del giorno precedente; velocità e direzione del vento importanti sono solo quelle delle ore 10-17 del giorno corrente, in entrambe le direzioni.

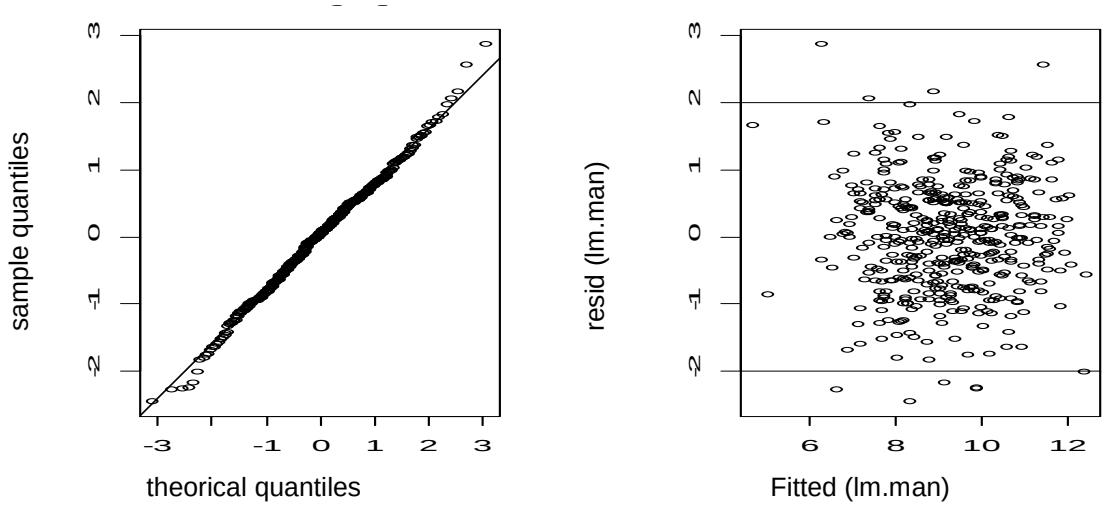
Inoltre da segnalare gli ossidi da azoto sia notturni che diurni, mese e anno. Il modello ha una devianza residua pari a 307.34 con 445 gradi di libertà (AIC: 1165.6).

I valori presenti nella coda destra appartengono all'anno 2000 (*fig. 3.3*).

Il test t, la cui significatività è minore di 0.05, ci permette di accettare l'ipotesi che queste variabili influenzano almeno in parte l'ozono.

Si è provato a rigenerare il modello considerando solo gli anni dal 2002 al 2005, per verificare soprattutto la relazione con il benzene: le concentrazioni di tale inquinante negli anni 2000 e 2001, infatti, risultano essere sensibilmente più alte di quelle degli anni successivi e non si sono

Fig. 3.3



potuti escludere in modo definitivo effetti dovuti alla strumentazione.

Viene confermata la rilevanza del benzene nel modello, mentre l'evidenza più forte è che l'NO_2 assume più importanza in Cairoli. Il modello ha una devianza residua pari a 110.21 con 224 gradi di libertà (AIC: 558.69); in Manzoni il benzene assume molta importanza, sia quello del giorno corrente che di quello precedente. Sono selezionate inoltre: run2, temperatura massima di oggi, radiazioni solari di oggi, quantità di pioggia di ieri, angolo e velocità vento dalle 6,9 e 10,17 di oggi. Tutte le altre variabili perdono importanza. Il modello ha una devianza residua pari a 113.25 con 229 gradi di libertà (AIC: 582.76).

In Osvaldo la situazione rimane pressoché invariata (tranne che per la perdita di importanza della variabile anno).

Nella *tabella 3.1* sono state riassunte le variabili che si sono rilevate più significative per i modelli di regressione lineare multipla stimati. Nessuno di questi ha un p-value superiore a 0.1 (livello di significatività delle regressioni effettuate).

I test chi quadrato sulla devianza residua e rispettivi gradi di libertà e l'analisi dei residui hanno confermato l'adeguatezza dei modelli di regressione individuati. Osservando i grafici precedenti, infatti, non si riconosce alcun andamento regolare dei dati, che al contrario appaiono correttamente disseminati intorno al valore zero.

Si sono quindi sviluppati i modelli lineari sui primi cinque anni di dati, utilizzandoli quindi per prevedere i dati del sesto anno (2005); nelle *figure 3.4, 3.5, 3.6* la linea rossa indica i valori previsti, mentre la linea nera rappresenta i valori originali dei dati dell'ozono rilevati nel 2005.

Tab. 3.1

Cairoli	Cai 2002-2005	**Manzoni**	Man 2002-2005	**Osvaldo**
run2	run2	run2	run2	run2
maxBEN1	maxBen1	maxBen2	maxBen1	maxBen1
press.mean1		press.mean1	maxBen2	
press.mean2		press.mean2		press.mean2
t.mas1	t.mas1	t.mas1	t.mas1	t.mas1
	ins.somma1	ins.somma2		ins.somma1
t.mas2		rad.somma1	rad.somma1	
	mese	mese		mese
rad.somma1	uore6:92	vore10:171	vore10:171	vore10:171
	uore10:171	uore10:171	uore10:171	
vore10:171	uore10:172		vore6.91	
		anno	uore6.91	anno
ur.mean1		ur.mean1		ur.mean1
ur.mean2	manNO2D2,	ur.mean2		ur.mean2
	t.mas2	max.NO2.D1		
		max.NO2.N1	pioggia.somma2	pioggia.somma1
R D: 185.63 on 369 d. f	R D: 110.21 on 224 d.f	R D: 307.34 on 445 d.f	R D:113.25 on 229 d. f	R D: 181.44 on 424 d f
aic: 882.85	aic: 558.69	aic: 1165.6	aic: 582.76	aic: 881.08

Le *tabelle 3.2, 3.3, 3.4*, rappresentano un sintetico paragone tra i valori stimati dal modello e quelli realmente rilevati.

In Manzoni, lo scarto quadratico medio fra valori previsti e misurati è 0.88. Tale valore può essere confrontato con lo scarto quadratico medio calcolato fra i valori di `run1` ed il valore medio dell'ozono degli anni precedenti, allo scopo di verificare quale sia la parte di varianza spiegata dal modello, che ha dato un valore pari a 2.47.

In Cairoli, è stato trovato che l'mse è pari a 0.73, mentre il valore di confronto è pari a 2.24; in Osvaldo: mse 0.63, mentre il test: 1.65.

Al fine di confrontare le prestazioni di tali modelli con quelle degli alberi di classificazione e random forest dei capitoli successivi, si sono trasformate sia `run1` che la previsione del modello in variabili qualitative, computando delle tabelle previsive. Per la suddivisione in 4 categorie (basso, medio, alto, altissimo) si sono utilizzati i quartili della distribuzione dei valori misurati. Sono state eseguite, infine, delle previsioni giorno per giorno dell'ultimo anno. Per questo tipo di previsione, si sono nuovamente usati i primi cinque anni per una stima iniziale del modello; esso veniva quindi ricalcolato inserendo progressivamente ciascun giorno dell'ultimo anno ed utilizzato per la previsione del giorno seguente. Dunque per quanto riguarda i modelli lineari delle tre stazioni, la *tabella 3.5* ne rappresentano il quadro riassuntivo.

Le percentuali di previsioni corrette dei modelli risultano essere inferiori a quelle ottenute dai modelli stimati sui soli anni precedenti. Ciò è dovuto al fatto che, nella stima giorno per giorno, vengono di volta in volta escluse dalla costruzione del modello le variabili non disponibili. Ad esempio, in Cairoli nel 2005, i valori mancanti relativi al benzene sono 80 su 183; il modello stimato sui soli anni precedenti, viceversa, in assenza di tale variabile non restituisce alcuna previsione.

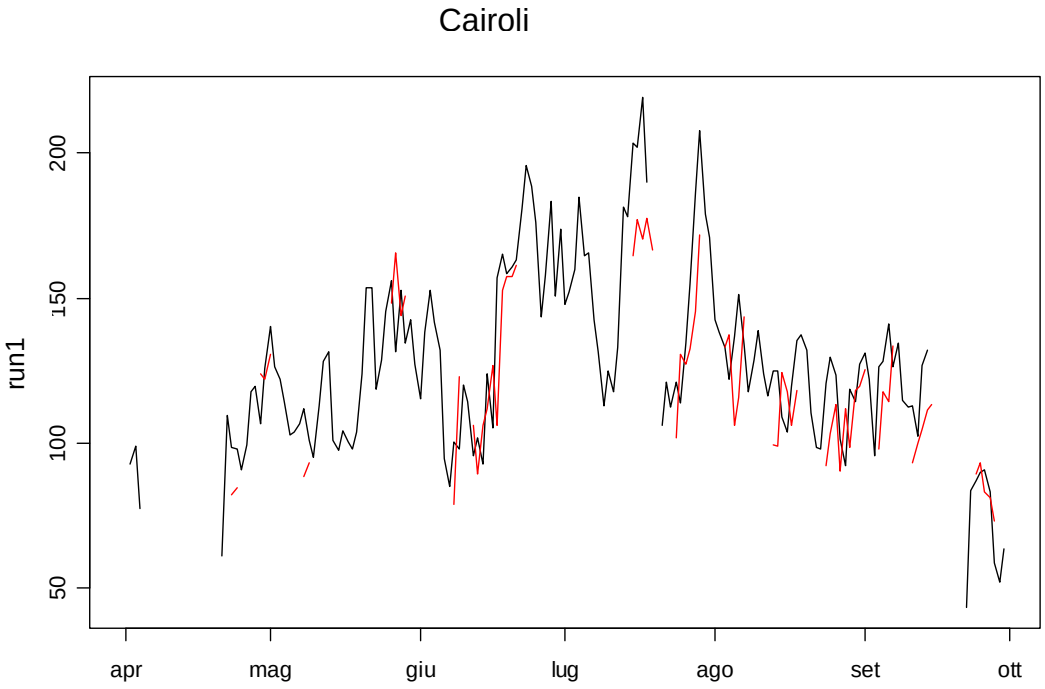
Tab. 3.2

Cairoli
Dati \ previsioni

	altissimo	alto	medio	basso
altissimo	20%	10%	0%	0%
alto	1%	24%	0%	10%
medio	1%	6%	14%	3%
basso	0%	0%	3%	7%

Diag: 65% N.tot= 88

Fig. 3.4



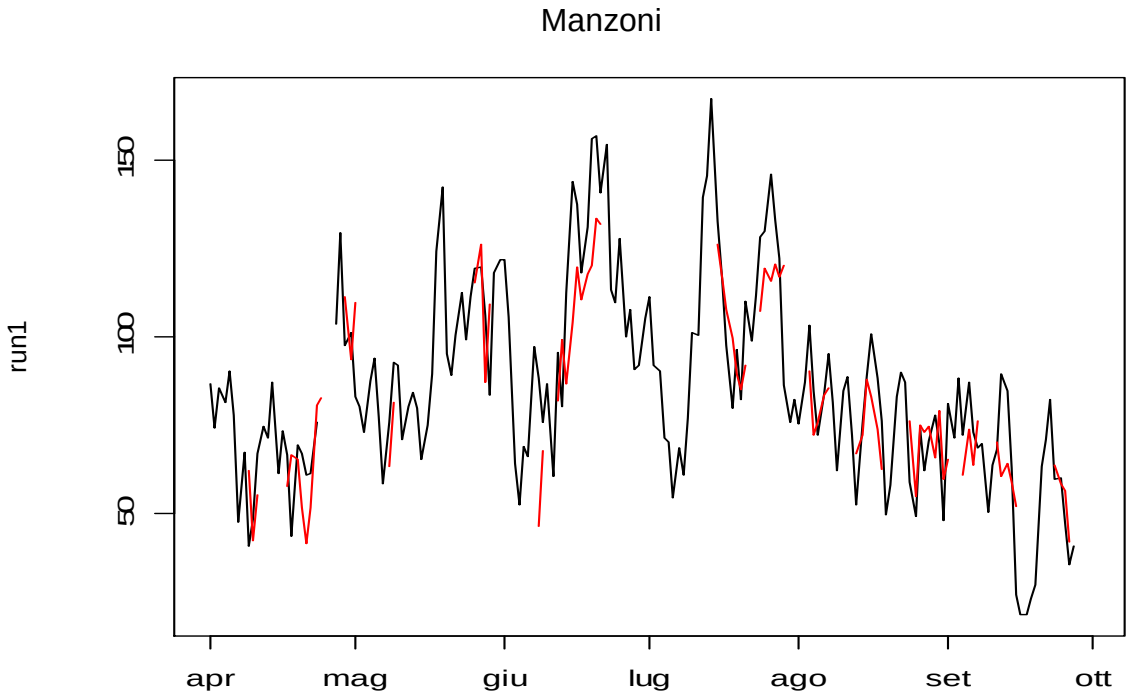
Tab. 3.3

Manzoni
Dati \ previsioni

	altissimo	alto	medio	basso
altissimo	4%	9%	3%	0%
alto	1%	4%	1%	0%
medio	0%	3%	35%	14%
basso	0%	0%	5%	22%

Diag: 65% N.tot=78

Fig. 3.5



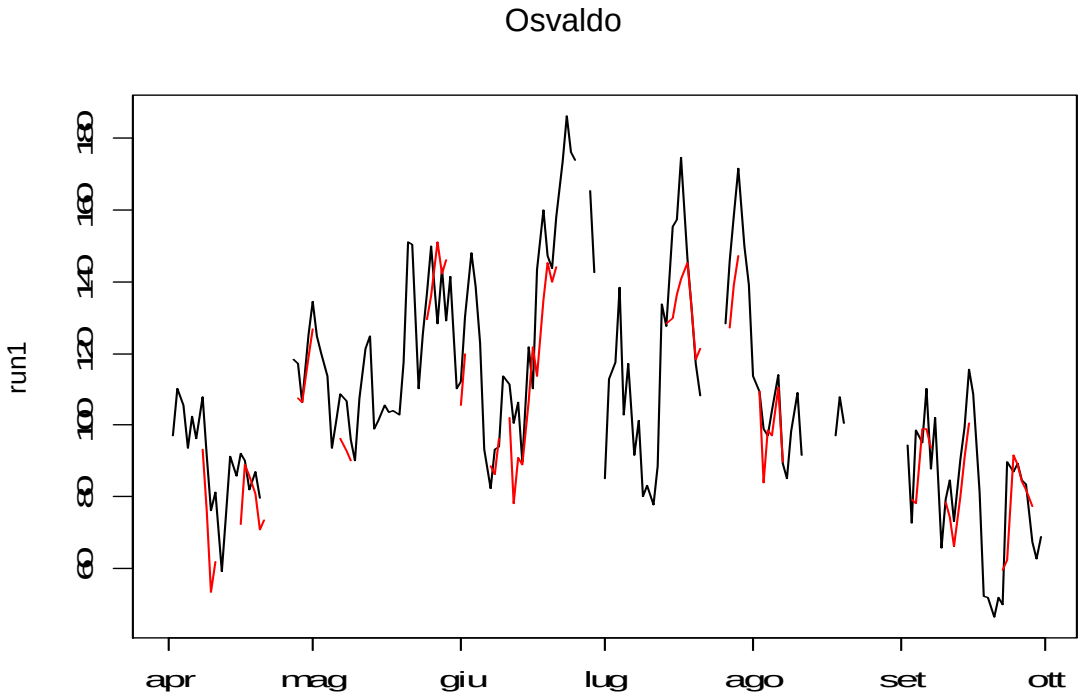
Tab. 3.4

osvaldo
Dati \ previsioni

	altissimo	alto	medio	basso
altissimo	21%	5%	0%	0%
alto	0%	5%	8%	0%
medio	0%	3%	28%	12%
basso	0%	0%	3%	16%

Diag: 70% N.tot=76

Fig. 3.6



Tab. 3.5

Cairoli Dati \	dati.prev altissimo	alto	medio	basso
altissimo	24%	10%	1%	0%
alto	3%	20%	12%	2%
medio	0%	6%	12%	3%
basso	0%	1%	3%	4%

Diag: 60% N.tot=157

Manzoni Dati \	dati.prev altissimo	alto	medio	basso
altissimo	5%	4%	3%	0%
alto	1%	2%	3%	0%
medio	0%	2%	42%	12%
basso	0%	0%	8%	16%

Diag: 65% N.tot=177

Osvaldo Dati \	dati.prev altissimo	alto	medio	basso
altissimo	19%	4%	1%	0%
alto	2%	8%	8%	0%
medio	0%	3%	31%	7%
basso	0%	0%	5%	13%

Diag: 71% N.tot=152

3.2 Alberi di regressione e classificazione[7].

Gli alberi sono una collezione di regole (che permettono di attribuire un'unità statistica ad un certo sottoinsieme sulla base di un valore assunto da una o più variabili osservate o misurate sulla medesima unità statistica), espresse in forma binaria e ottenute attraverso partizionamento ricorsivo (Breiman L. & al., 1984).

Sin dalla loro introduzione, negli anni '80, i metodi di regressione e classificazione ad albero, fondati su un approccio non parametrico, si sono rilevati un utile strumento d'analisi in processi di scoperta della conoscenza e di apprendimento supervisionato dai dati.

Sono stati usati gli alberi perché producono un modello molto facile da interpretare (Kuhnert & al., 2005; Burrows & al., 1995):

- gli alberi decisionali sono non parametrici, quindi non richiedono i presupposti di normalità dei dati;
- possono maneggiare i dati di tipo diverso: continuo, categorico, ordinale, binario, non sono quindi richieste le trasformazioni dei dati e possono cogliere aspetti non lineari;
- gli alberi possono essere utili per la rilevazione delle variabili importanti, interazione ed identificazione dei valori erratici;
- fanno fronte ai dati mancanti identificando le divisioni sostitutive del processo modellante, queste ultime sono altamente connesse con la divisione primaria;

[7] Si è utilizzata la libreria *RPART*, di *R*; la documentazione del software può esser scaricata in pdf da uno dei siti CRAN (ad esempio http://lib.stat.cmu.edu/R/CRAN/)

La metodologia può essere ricapitolata in tre punti principali:

1) Splitting (divisione):

   I dati vengono divisi in gruppi in base ad un algoritmo di minimizzazione dell'indice di Gini (classificazione); quest' ultimo è un indice di eterogeneità intesa come un'equa ripartizione delle frequenze all'interno di una distribuzione di frequenza. L'indice di Gini si basa sulla somma dei quadrati delle frequenze relative.

   $$i(t) = \sum_{i \neq j} p(i/t)\, p(j/t)$$

   dove *p(i|t)* è la probabilità che l'elemento del nodo t sia della classe i e *i(t)* è la misura d''impurità o l'errore di errata classificazione; o della somma dei quadrati (regressione); l'espressione della misura dell'impurità per la regressione è

   $$i(t) = \sum \{y(i/t) \bar{y}(t)\}^2$$

   dove $\bar{y}$ è la media delle osservazioni al nodo t e *y(i/t)* rappresenta l'osservazione i nel nodo t.

   Il processo è ripetuto sui sottogruppi trovati. I dati continuano ad essere divisi a loro volta in sottogruppi fino a creare un albero sufficientemente grande, in cui ad ogni nodo terminale si trova un piccolo gruppo di dati dello stesso tipo.

2) Pruning (potatura):

   Quando l'albero si è sviluppato, si inizia la potatura. L'albero viene progressivamente ridotto ad alberi più semplici, a ciascuno dei quali si

associa una misura di costo complessità. Si considera: $R_\alpha = R + \alpha \times T$, dove T rappresenta il numero di nodi terminali dell'albero, R rappresenta il rischio e α rappresenta il parametro di costo-complessità. α è un termine di penalità che controlla la grandezza dell'albero. Fissato un certo valore di α, l'albero ottimale è quello che minimizza R_α. Per gli alberi di classificazione, R si riferisce all'errore di misclassificazione, mentre per gli alberi di regressione, il rischio corrisponde alla somma dell'errore dei residui.

Una stima dell'errore di classificazione viene attribuito a ciascun grado di complessità dell'albero usando un metodo di validazione incrociata. Vengono creati dieci subset di dati, con procedura casuale. L'albero è creato con nove di questi subset di dati, mentre il decimo è usato per calcolare l'errore di previsione del modello. Questo processo, che viene ripetuto per tutti i sottogruppi, determina complessivamente l'errore di validazione incrociata.

3) Tree Selection:

La selezione dell'albero è fatta con la cosiddetta regola "dell'errore standard" (SE), introdotta da Breiman (1984), con riferimento all'errore di validazione incrociata: viene selezionato il parametro di complessità corrispondente al più piccolo albero tale da rientrare entro uno standard error dall'albero avente il tasso d'errore più basso.

3.2.1 Alberi di classificazione

Per ogni centralina, i dati relativi alle tre variabili che riassumono l'ozono (massimo giornaliero della media trascinata sulle 8 ore (`run`), massimo diurno (`max.O3.D`) e notturno (`max.O3.N`)) sono stati raccolti in quattro gruppi in base ai quantili della relativa distribuzione.

Ove uno dei quantili fosse molto vicino al valore soglia (110 $\mu g/m^3$, la cui radice quadrata vale 10.49), è stato sostituito con tale valore, utile per un rapido riferimento alla normativa.

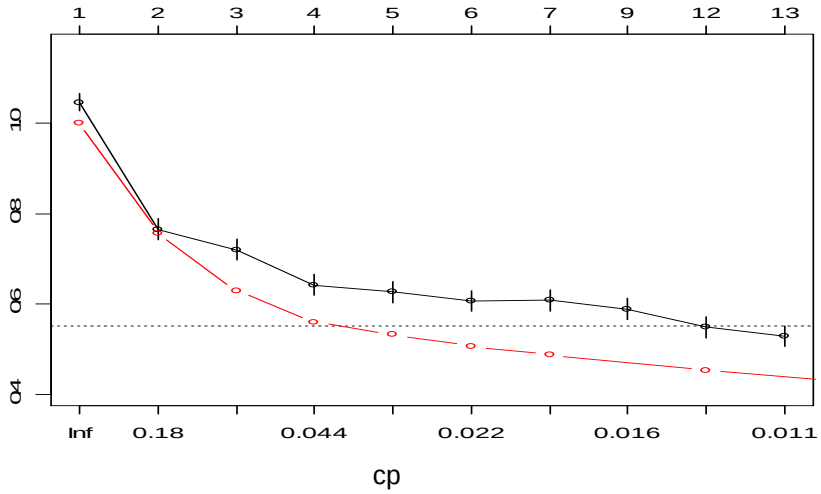
Ad esempio per `run` di Cairoli, la decisione è stata presa nel seguente modo:

```
> quantile(cairit$run1,na.rm=T)
    0%    25%    50%    75%   100%
 6.086  9.547 10.503 11.549 15.223
```

```
basso:      6.086 :  9.547
medio:      9.548 : 10.488
alto:      10.489 : 11.549
altissimo: 11.550 : 15.223
```

Una volta create le variabili qualitative, è stato applicato il metodo degli alberi di classificazione. Le 1598 osservazioni sono state divise casualmente in due gruppi, *analisi*, comprendente 1298 osservazioni su cui è stato stimato

Fig. 3.7



l'albero e *test*, comprendente le restanti 300 osservazioni utilizzate per stimare la capacità previsiva del modello.

Si è iniziato producendo un albero molto grande, fino a raggiungere un indice di costo-complessità corrispondente molto piccolo ($cp=10^{-3}$). Avendo molte variabili, spesso correlate tra loro, è molto utile osservare i primi split che il modello sceglie per dividere i dati e creare l'albero. A questo punto si deve procedere alla potatura ed alla selezione del modello, prendendo un albero piccolo, ma con un errore di validazione incrociata (*cross-validated)* che rispettasse la regola dell'errore standard; il valore di soglia per quest'ultima (SE + min(R_α)) è rappresentato dalla linea tratteggiata della *figura 3.7 e 3.8*.

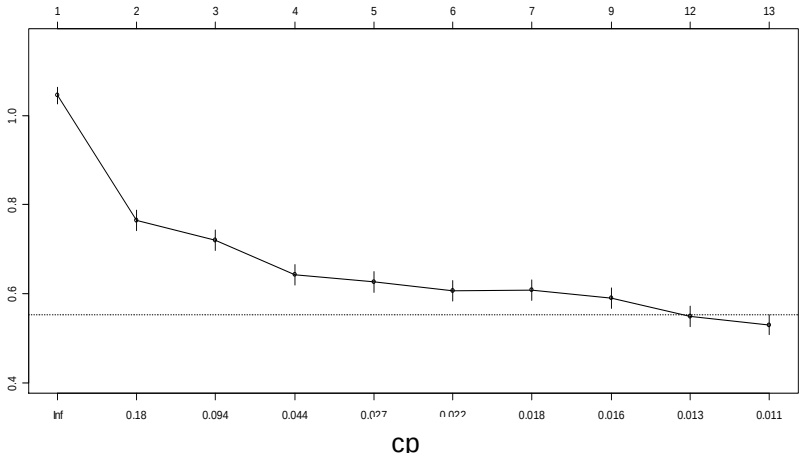
La potatura è necessaria per accertarsi che l'albero sia abbastanza piccolo da evitare la sovrastima nelle previsioni, ma non tanto da introdurre gravi errori di classificazione (*resubstitution*).

La figura 3.7 rappresenta un diagramma dell'errore di classificazione (rosso) e il tasso di errore di validazione incrociata (nero)[8].

In Cairoli vediamo che l'albero scelto ha 9 nodi ed un valore del parametro costo-complessità pari a 0.013 (*figura 3.8*)[9].

Sono stati sostanzialmente ribaditi i risultati del modello di regressione lineare: per spiegare i livelli alti dell'ozono di oggi, le variabili principali sono l'ozono di ieri, la temperatura e il tempo d'insolazione di oggi, la velocità del vento dalle ore 10 alle 17 di oggi (*figura 3.9*). Si sono confrontate le previsioni ed i dati stimati: dalla *tabella 3.6* si può vedere che sulla diagonale sono presenti il 70% delle osservazioni. Gli errori più significativi del modello si evidenziano quando viene previsto un valore medio-basso quando in realtà dai dati risultava alto o altissimo (7%).

Fig. 3.8



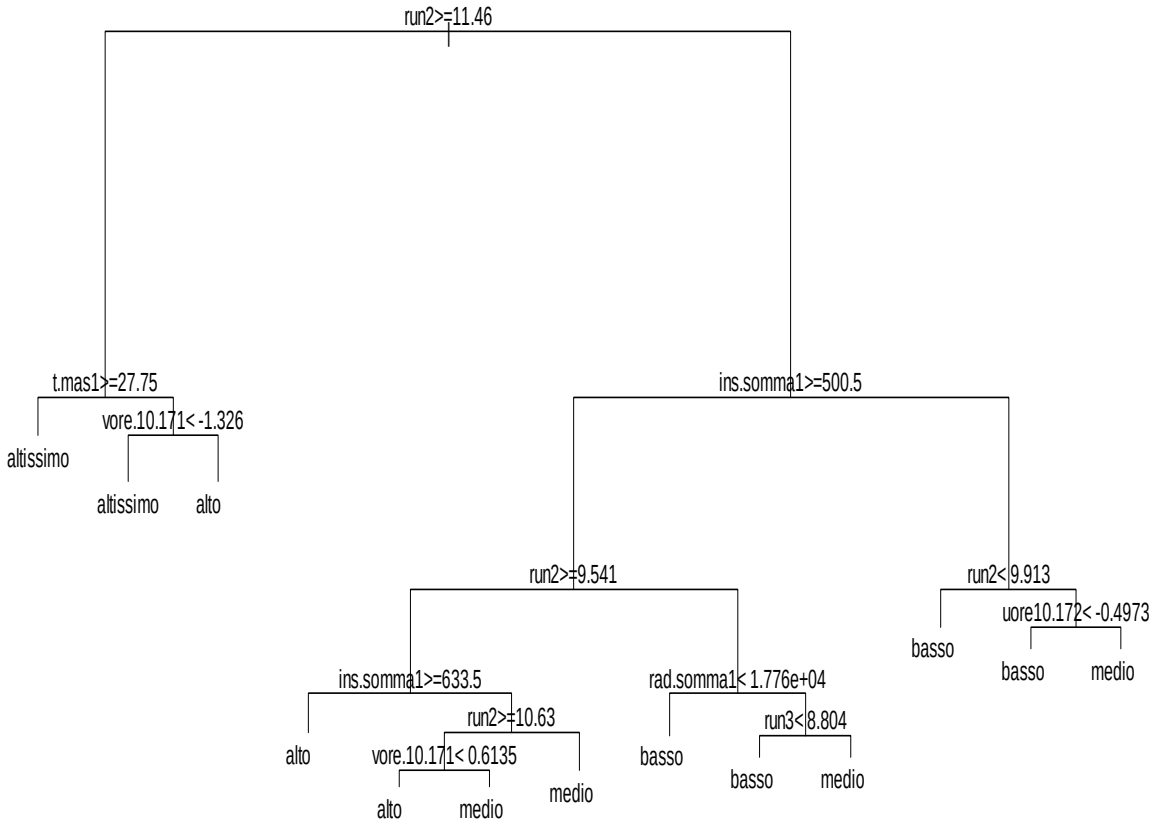
[8] il *resubstitution error* non è così utile nella scelta dell'albero ottimale, in quanto potrebbe condurre alla scelta di un albero *overfitted*; come detto, per selezionare l'albero ottimale, viene usato il cross-validated error.

[9] Sull'asse y c'è il cross validated error (xerror); sull'asse x in alto c'è il numero dei rami dell'albero, mentre in basso si trova un parametro legato all'indice di complessità del modello.

In questa prima tabella ci si riferisce alle previsioni eseguite sugli stessi dati usati per stimare il modello.

Nella *tabella 3.7* vediamo invece il comportamento del modello nel predire i dati *test*, che non erano stati inseriti nel modello: nella diagonale troviamo il 61% delle osservazioni, mentre l'errore principale, definito come per la tabella precedente, è del 9%.

Fig. 3.9



Tab. 3.6

Dati \	cai.pred			
	altissimo	alto	medio	basso
altissimo	17%	2%	1%	0%
alto	7%	17%	5%	1%
medio	0%	4%	17%	5%
basso	0%	1%	3%	19%

Diag: 70% ; N tot=836

Il fatto che la quantità di ozono di oggi risenta soprattutto del proprio livello del giorno precedente significa che quest' ultimo ha un forte potere di sintesi, ma potrebbe anche suggerire che le altre variabili esplicative inserite nel modello sono povere e non riescano a spiegare l'ozono di oggi. A tal scopo, é stato interessante verificare cosa sarebbe successo se nel modello fosse stato inserito come variabile esplicativa solamente l'ozono dei tre giorni precedenti (*tabella 3.8*). Notiamo che la differenza tra inserire o meno le altre variabili esplicative è del 12% (troviamo 470 previsioni esatte su 804, pari al

Tab. 3.7

Dati \	cai.pred.test			
	altissimo	alto	medio	basso
altissimo	18%	8%	2%	1%
alto	7%	14%	5%	1%
medio	1%	6%	13%	3%
basso	1%	1%	5%	15%

Diag: 60% ; N.tot=177

Tab. 3.8

Dati \	cai.pred			
	altissimo	alto	medio	basso
altissimo	18%	4%	1%	0%
alto	7%	12%	6%	2%
medio	0%	5%	9%	3%
basso	0%	3%	8%	19%

Diag: 58% ; N.tot=804

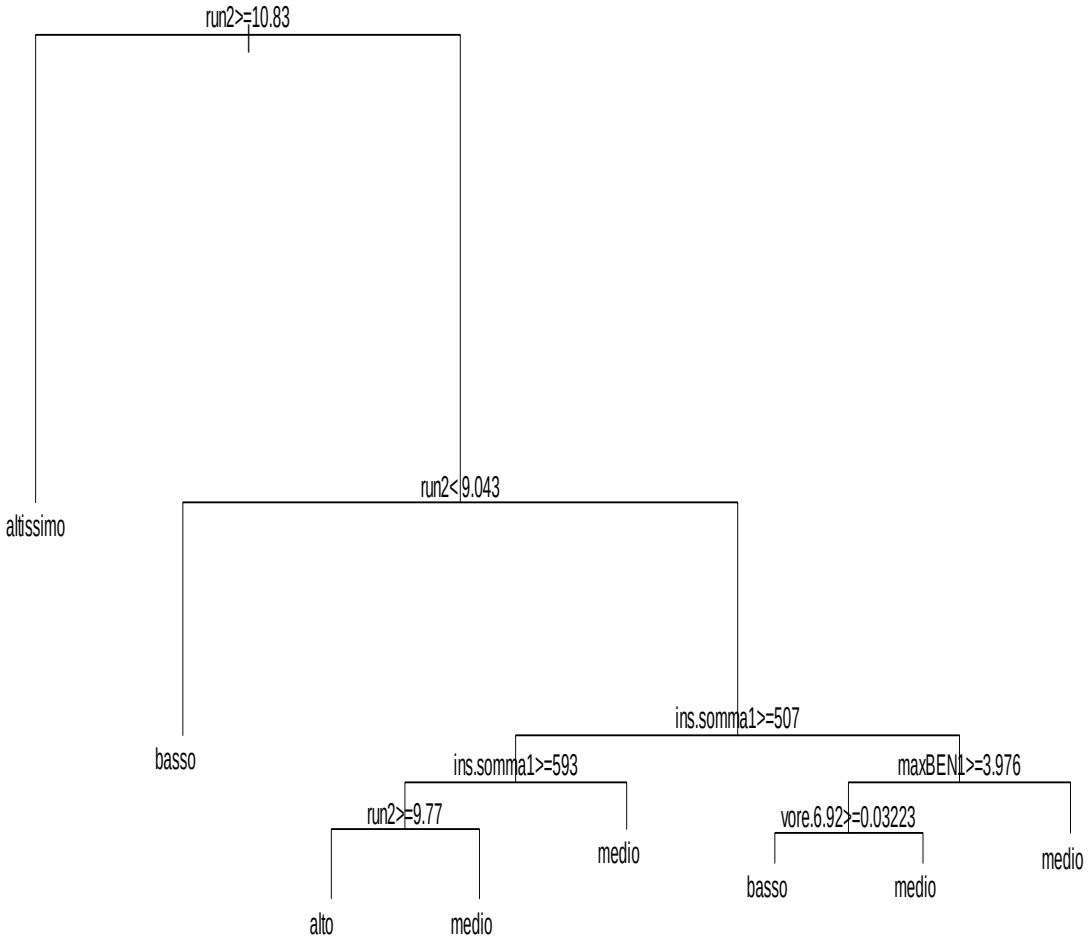
58%, contro il 70% del modello totale).

Per quanto riguarda il massimo notturno e diurno dell'ozono, è interessante notare che per il massimo diurno le variabili d'interesse sono: temperatura, l'eliofania, la quantità di radiazione solare e la velocità del vento dalle ore 10 alle 17 del giorno stesso, oltre che ai massimi di O_3 del giorno e della notte precedenti; per quanto riguarda i massimi notturni, oltre all'ozono presente il giorno e la notte precedenti, è importante il vento, inteso soprattutto come direzione presente nelle ore notturne; assume importanza l'anno di riferimento.

Per gli alberi e le tabelle previsive di Manzoni ed Osvaldo, è stata ripetuta la stessa analisi, iniziando dal fattorizzare le variabili da spiegare sulla base dei loro quantili.

In Osvaldo (*figura 3.10, tabelle 3.9, 3.10*), l'ozono di ieri è senza dubbio la variabile principale per determinare l'alto livello di ozono di oggi, seguita dall'eliofania della stessa giornata. Da evidenziare la presenza di Benzene nel determinare situazioni medie di ozono.

Fig. 3.10


run2>=10.83
altissimo
run2<9.043
basso
ins.somma1>=507
ins.somma1>=593
maxBEN1>=3.976
run2>=9.77
medio
vore.6.92>=0.03223
medio
alto
medio
basso
medio

Tab. 3.9

Dati \ osv.pred	altissimo	alto	medio	basso
altissimo	23%	7%	5%	1%
alto	2%	6%	3%	1%
medio	0%	4%	19%	3%
basso	0%	1%	7%	19%

Diag: 67% N.tot=699

Da segnalare che nel massimo giornaliero in Osvaldo assume importanza la pioggia del giorno stesso (la variabile era segnalata anche dal modello lineare), mentre in quello notturno troviamo NO_2 del giorno stesso.

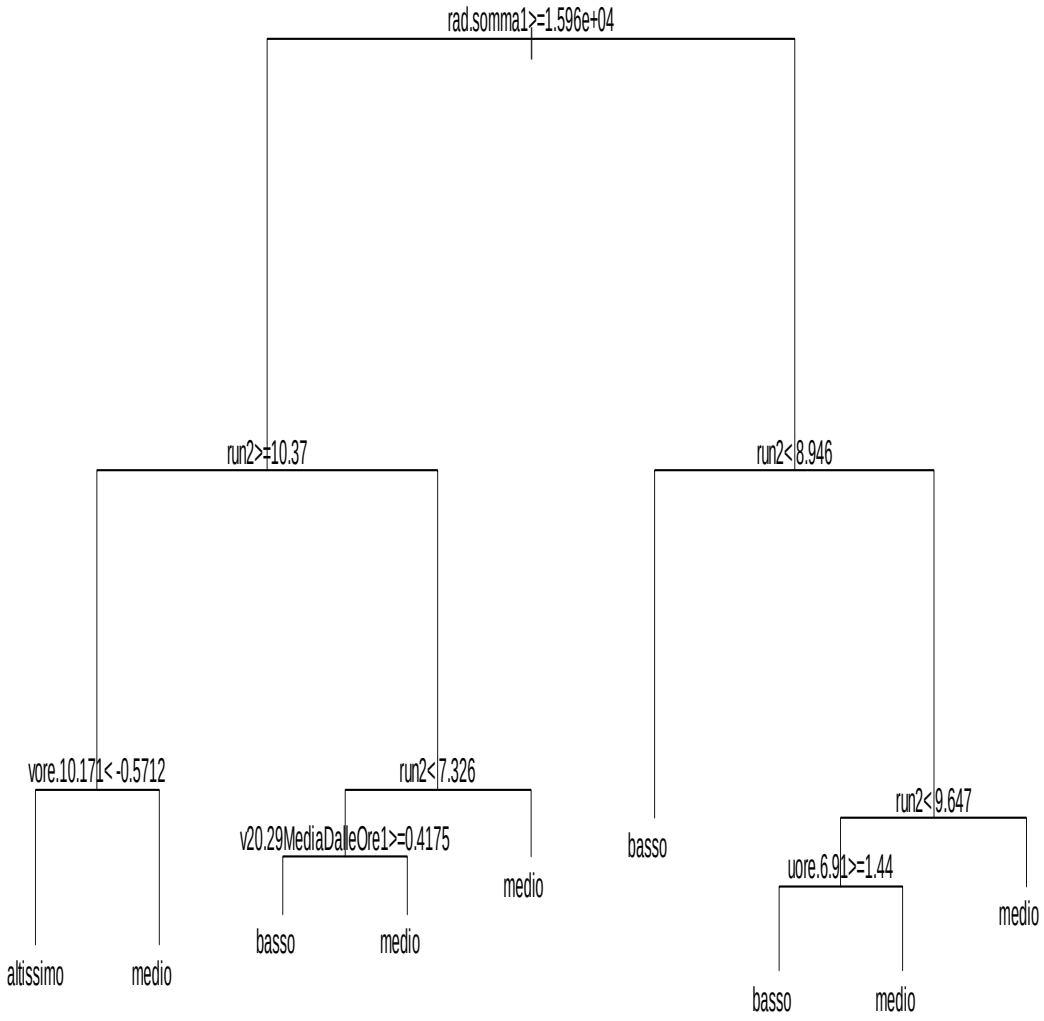
In Manzoni (*figura 3.11, tabelle 3.11, 3.12*), a differenza delle altre, la variabile discriminante più importante risulta essere la quantità di radiazione solare giornaliera, seguita dal livello di ozono di ieri. Da evidenziare l'importanza della velocità del vento dello stesso giorno.

Tab. 3.10

Dati \ osv.pred	altissimo	alto	medio	basso
altissimo	20%	11%	5%	0%
alto	5%	7%	8%	0%
medio	0%	2%	12%	4%
basso	0%	0%	13%	14%

Diag: 53% N.tot=171

Fig. 3.11


rad.somma1>=1.596e+04
run2>=10.37
run2<8.946
vore.10.171< -0.5712
run2<7.326
basso
run2<9.647
v20.29MediaDalleOre1>=0.4175
medio
uore.6.91>=1.44
medio
altissimo
medio
basso
medio
basso
medio

Tab. 3.11

Dati \ man.pred	altissimo	alto	medio	basso
altissimo	10%	3%	2%	0%
alto	0%	0%	0%	0%
medio	4%	5%	47%	8%
basso	0%	0%	4%	17%

Diag: 74% N.tot=836

I valori alti sono in totale, 82 su 1598; di questi, 50 appartengono agli anni dal 2002 in poi.

Ricordiamo che Manzoni presenta dei valori di ozono molto più bassi:

```
> quantile(manrit$run1,na.rm=T)
```

0%	25%	50%	75%	100%
4.18	8.28	9.36	10.35	14.00

Per analogia con le altre stazioni, si è scelto di tenere il valore di 10.49 come soglia fra le classi “medio” ed “alto”; al posto del 75%, per discriminare la classe “alto” da “altissimo”, è stato scelto un valore intermedio (11).

basso	medio	alto	altissimo
4.18	8.28	10.4	11	14.00

Tab. 3.12

Dati \ man.pred.test	altissimo	alto	medio	basso
altissimo	8%	0%	5%	1%
alto	3%	0%	6%	0%
medio	3%	0%	41%	7%
basso	0%	0%	9%	16%

Diag: 65% N.tot=205

Per Manzoni proviamo a considerare unicamente gli anni dal 2002 al 2005 (*figura 3.12),* analogamente a quanto realizzato con il modello lineare.

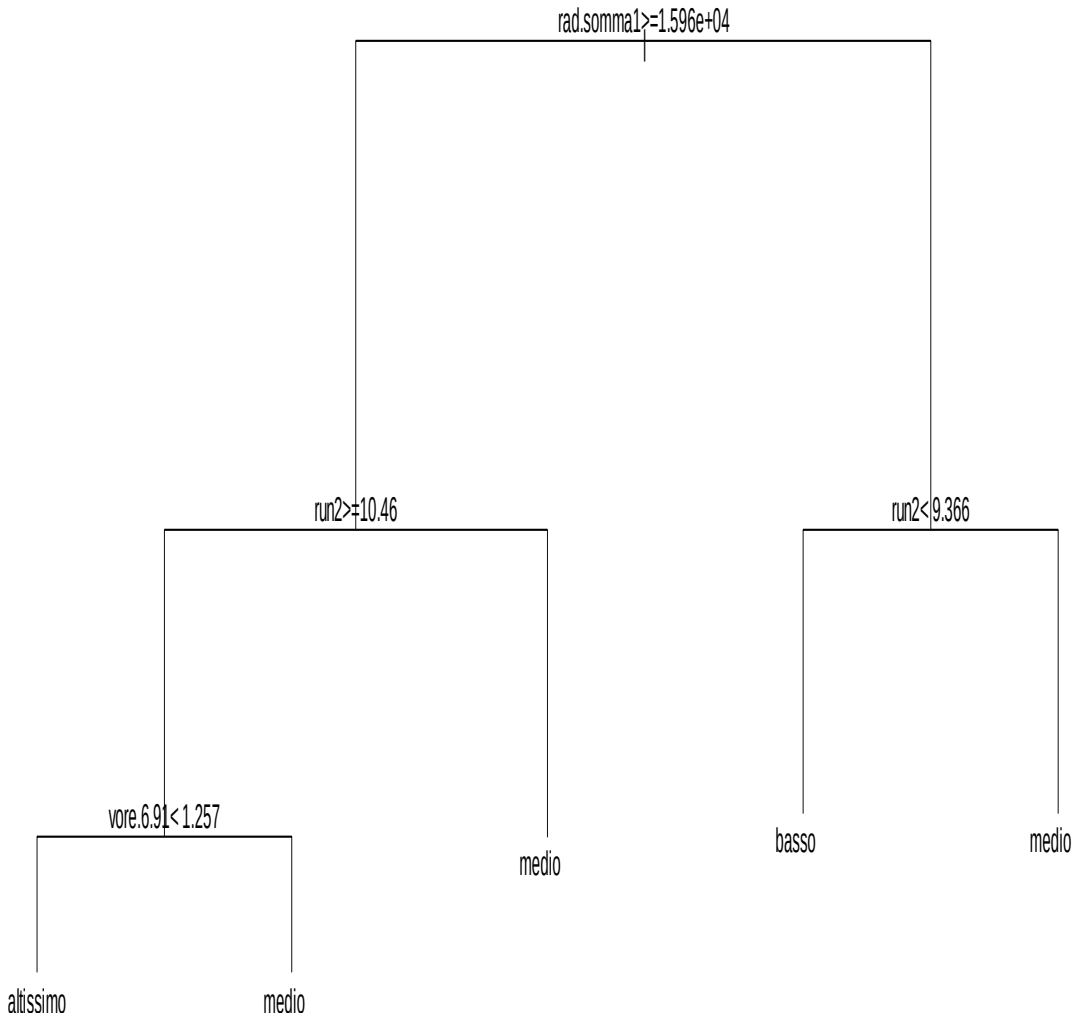
La variabile più importante rimane la quantità di radiazioni solare. Rispetto al modello sviluppato sui sei anni, per determinare alti livelli di ozono, ribadiscono l'importanza la temperatura, la velocità del vento dalle ore 6 alle ore 9, mentre il Benzene, variabile che ci spinge a considerare solamente gli ultimi anni, rimane non significativo.

Nel considerare i sei anni, in *tabella 3.11*, la quantità di valori esatti stimati, situati nella diagonale, sono il 74%, mentre nella *tabella 3.13*, che considera solo gli ultimi quattro anni, troviamo il 75%.

In Manzoni, dal 2002, notiamo la grande importanza della variabile mese assunta per il livello massimo notturno.

Un'applicazione di interesse dei modelli qui sviluppati è la possibilità di prevedere il livello di ozono rilevato da una delle stazioni, noti i valori misurati nelle altre.

Fig. 3.12


rad.somma1>=1.596e+04
run2>=10.46
run2<9.366
vore.6.91<1.257
medio
basso
medio
altissimo
medio

Tab. 3.13

Dati \	Man.pred			
	altissimo	alto	medio	basso
altissimo	10%	2%	2%	0%
alto	1%	3%	1%	0%
medio	2%	4%	46%	12%
basso	0%	0%	2%	14%

Diag. 75%

Cosa succede inserendo l'ozono di tre giorni di Osvaldo e Manzoni nel modello sviluppato per Cairoli?

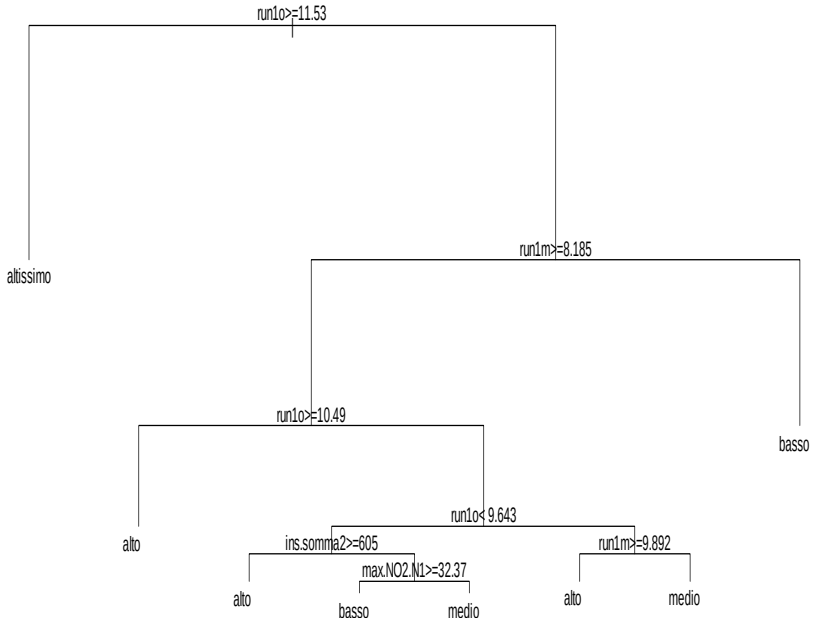
E' evidente il legame tra livelli alti d'ozono rilevati contemporaneamente (*figura 3.13, tabella 3.14*).

Sarebbe stato interessante trovare un valore previsivo elevato; infatti, se le variabili relative all'ozono di ieri in S Osvaldo e Manzoni fossero state sufficienti per conoscere l'ozono di oggi in Cairoli, si sarebbe potuta eliminare la rilevazione dell' O_3 in quest'ultima centralina.

Per quanto riguarda gli alberi, è stato riproposto lo schema delle previsioni giorno-per-giorno nell'ultimo anno, già adottato per i modelli lineari; le tre tabelle seguenti ne rappresentano il quadro riassuntivo *(tabella 3.15).*

Lo stesso esercizio è stato realizzato per la previsione dei valori massimi orari giornalieri: la percentuale di successo dei modelli diminuisce, tranne che nel caso di Osvaldo (*tabella 3.16*).

Fig. 3.13



Tab. 3.14

Dati \	pred			
	altissimo	alto	medio	basso
altissimo	21%	7%	6%	1%
alto	3%	11%	15%	7%
medio	0%	1%	14%	6%
basso	0%	0%	1%	7%

Tab. 3.15

Cairoli Dati \	dati.prev altissimo	alto	medio	basso
altissimo	27%	4%	3%	0%
alto	8%	17%	10%	2%
medio	2%	4%	7%	9%
basso	0%	0%	1%	8%

Diag: 59% N.tot=157

Manzoni Dati \	dati.prev altissimo	alto	medio	basso
altissimo	9%	0%	3%	0%
alto	2%	0%	4%	0%
medio	3%	0%	46%	8%
basso	0%	0%	9%	15%

Diag: 70% N.tot=181

Osvaldo Dati \	dati.prev altissimo	alto	medio	basso
altissimo	18%	4%	1%	0%
alto	7%	5%	7%	0%
medio	3%	4%	26%	7%
basso	0%	0%	9%	10%

Diag: 59% N.tot=152

Tab. 3.16

Cairoli

Dati \	dati.prev altissimo	alto	medio	basso
altissimo	21%	10%	0%	1%
alto	6%	27%	0%	8%
medio	3%	5%	0%	8%
basso	1%	3%	0%	8%

Diag: 56% N. tot=159

Manzoni

Dati \	dati.prev altissimo	alto	medio	basso
altissimo	15%	0%	2%	0%
alto	3%	0%	7%	0%
medio	3%	1%	30%	12%
basso	2%	0%	9%	15%

Diag: 60% N.tot=181

Osvaldo

Dati \	dati.prev altissimo	alto	medio	basso
altissimo	18%	4%	0%	0%
alto	3%	26%	0%	7%
medio	1%	9%	4%	7%
basso	3%	3%	0%	15%

Diag: 63% N.tot=157

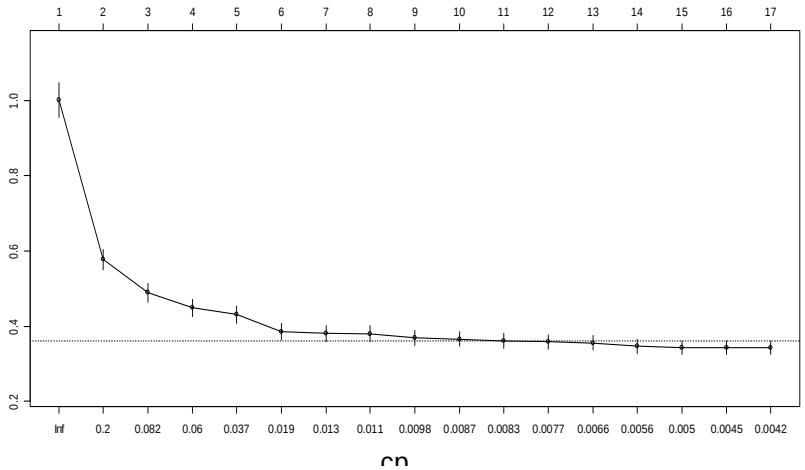
3.2.2 Alberi di regressione

Gli alberi di regressione stimati per le tre centraline, con i rispettivi diagrammi di dispersione tra valori reali e previsti, sono riportati di seguito, per un confronto con i risultati ottenuti con gli alberi di classificazione. Come detto, nei problemi di regressione cambia la definizione dell'impurità attribuita ai nodi ed alle foglie degli alberi, consistendo in tal caso nella somma dei residui al quadrato.

Gli alberi sono stati cresciuti utilizzando il subset di dati *"analisi"*, definito al capitolo precedente.

In Cairoli, l'albero ottimale è stato selezionato attraverso un plot dell'errore di validazione incrociata contro il parametro di complessità dell'albero, con la procedura della regola dell'errore standard già introdotta. È' stato così identificato un parametro di complessità ottimale compreso fra 0.0077 e 0.0066, che ha selezionato un albero di regressione con 12 nodi terminali, o foglie (*figura 3.14*). Il modello mostra una prima discriminazione sul livello di ozono del giorno precedente: ad alti livelli di ozono del giorno

Fig. 3.14



precedente, a certe velocità del vento dalle ore 6-9 e delle ore 10-17 dello stesso giorno e ad un elevato tempo di insolazione, corrispondono i valori più alti di ozono(*figura 3.15*).

Si sono effettuate delle previsioni sulla parte restante dei dataset (subset *"test"*) usando questo modello; per valutarne le prestazioni si è costruito il diagramma di dispersione tra previsioni e risposta reale.

Nel campione *"test"*, lo scarto quadratico medio dei dati è pari a 1.50, quello fra dati misurati e previsti è 0.50 (*figura 3.16*).

Fig. 3.15

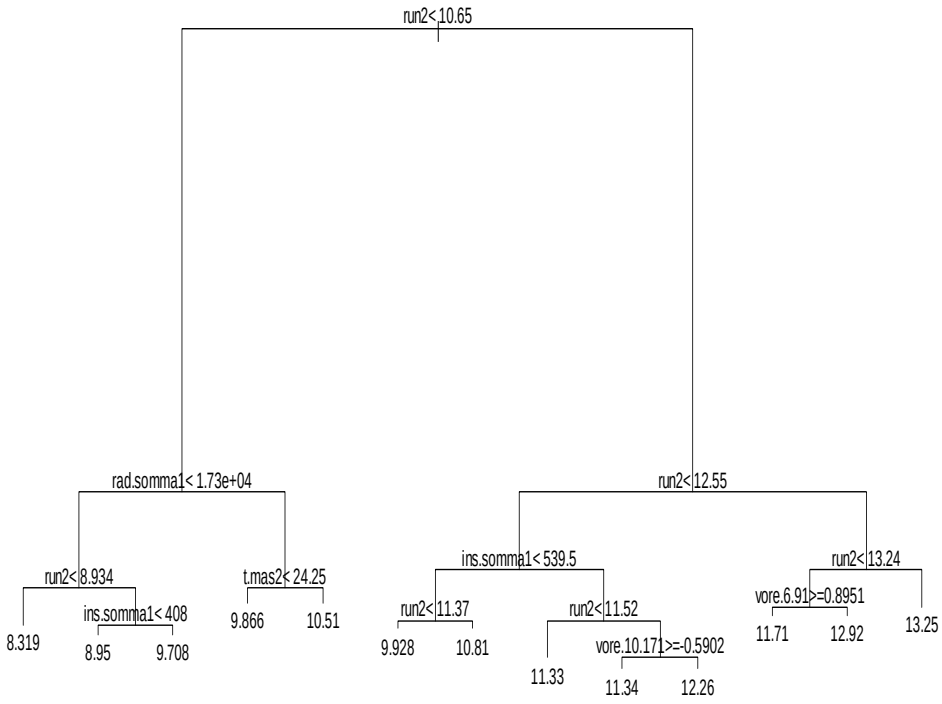
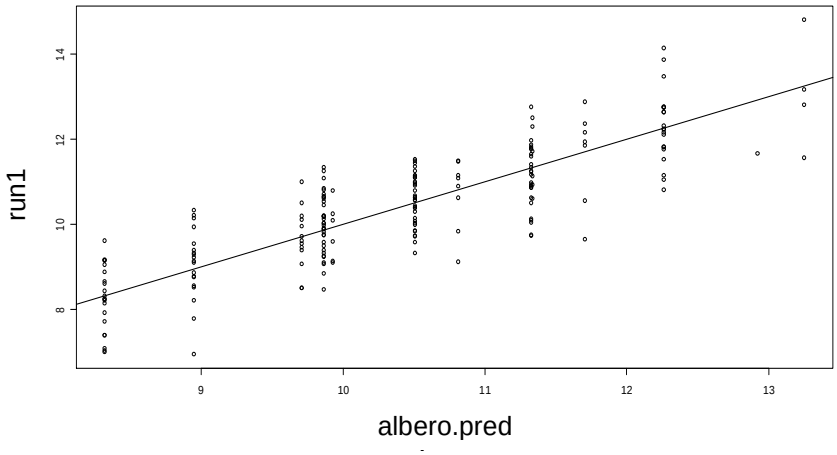


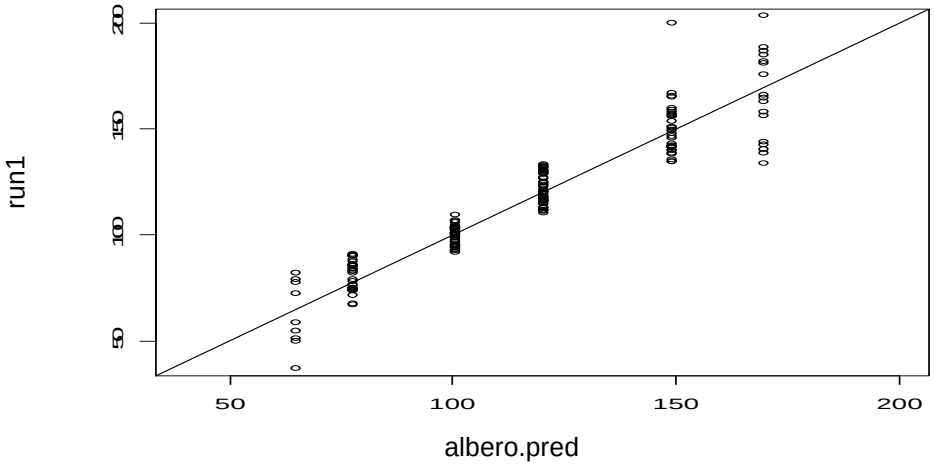
Fig. 3.16



Per chiarezza, viene riportato anche il diagramma di dispersione relativo ai dati non trasformati per mezzo della radice quadrata (*figura 3.17*).

In questo caso nel campione *"test"*, lo scarto quadratico medio dei dati

Fig. 3.17



è pari a 32.21, quello fra dati misurati e previsti è 11.14.

In sant'Osvaldo le foglie dell'albero scelto sono 12, essendo il valore del parametro di complessità scelto tra 0.0086 e 0.0075.

Fig. 3.18

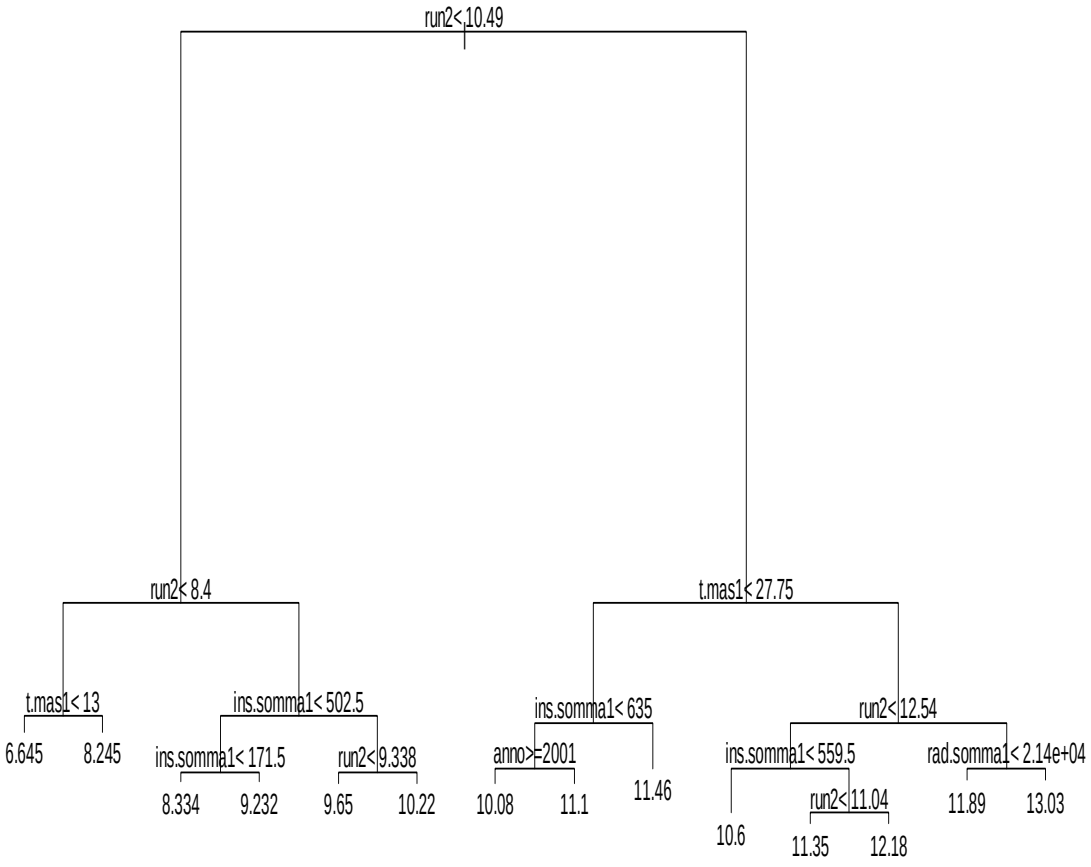
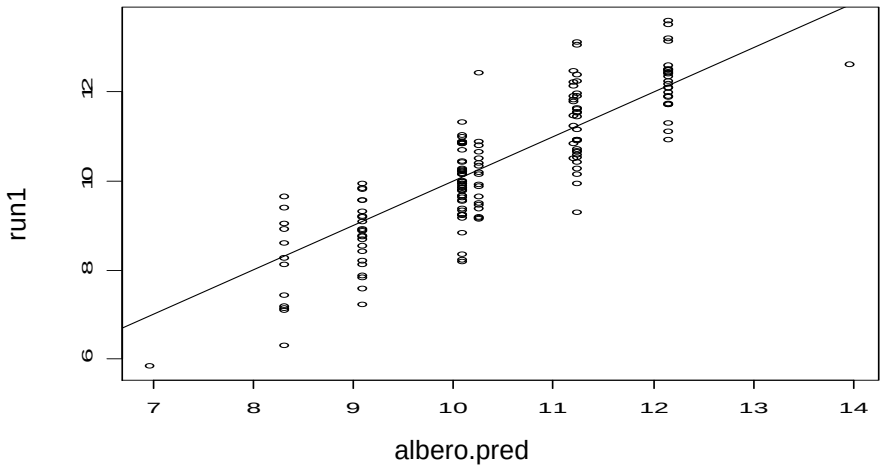


Fig. 3.19



Anche questo modello mostra una prima spaccatura sul livello di ozono del giorno precedente, ma a questo punto, per determinare alti livelli di ozono, assume importanza la temperatura di oggi oltre che l'elevato tempo di insolazione e quantità di radiazioni solari (*figure 3.18, 3.19*).

In Manzoni, troviamo ancora 12 foglie, per un livello del parametro costo-complessità scelto tra 0.0056 e 0.01.

Nel campione *"test"*, lo scarto quadratico medio dei dati è pari ad 1.45, quello fra dati misurati e previsti è 0.50.

Essendoci a Manzoni livelli più bassi di ozono, vediamo meglio ancora l'importanza dell'ozono del giorno precedente, della velocità del vento dalle ore 10 alle 17 dello stesso giorno, delle radiazioni solari, nel determinare livelli alti di ozono. In più viene evidenziata la quantità di pioggia (*figura 3.20*).

Fig. 3.20

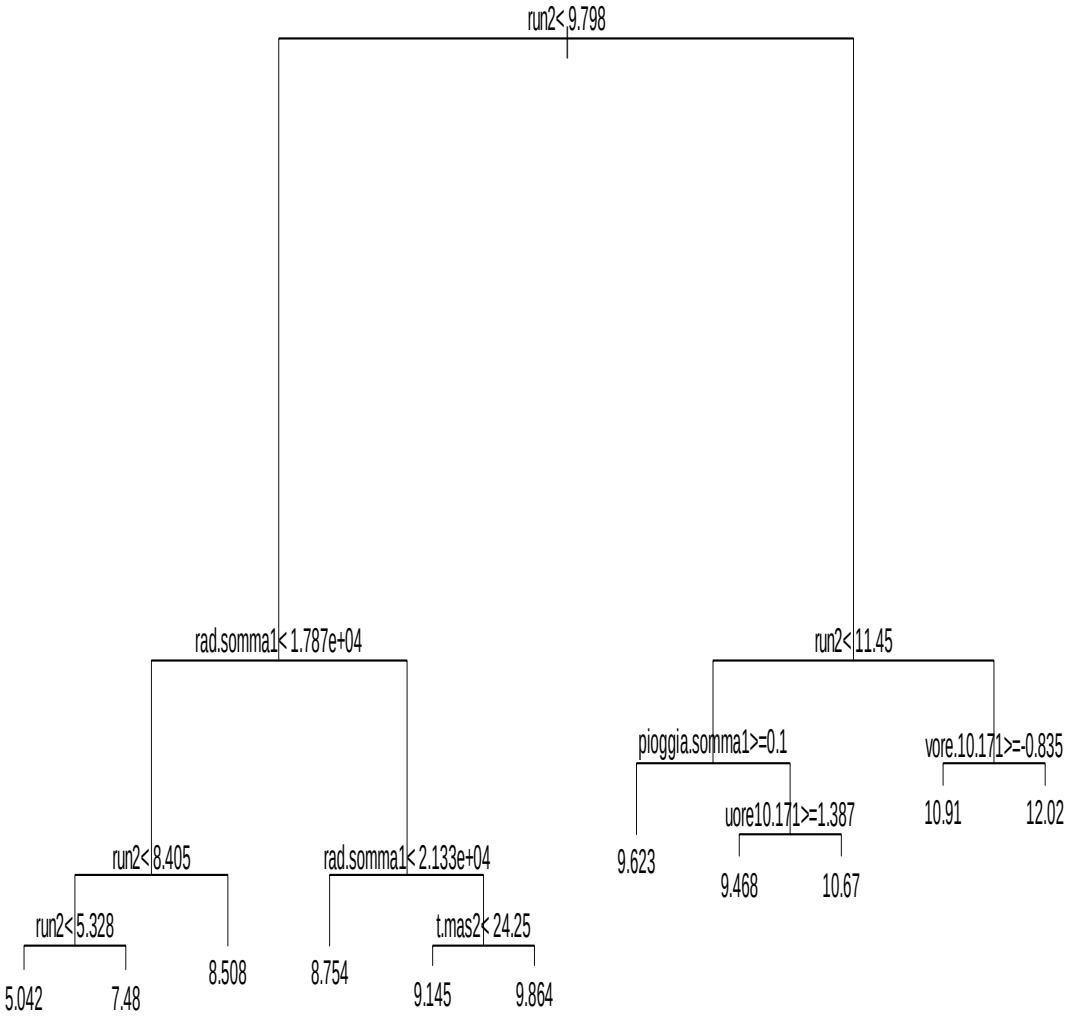
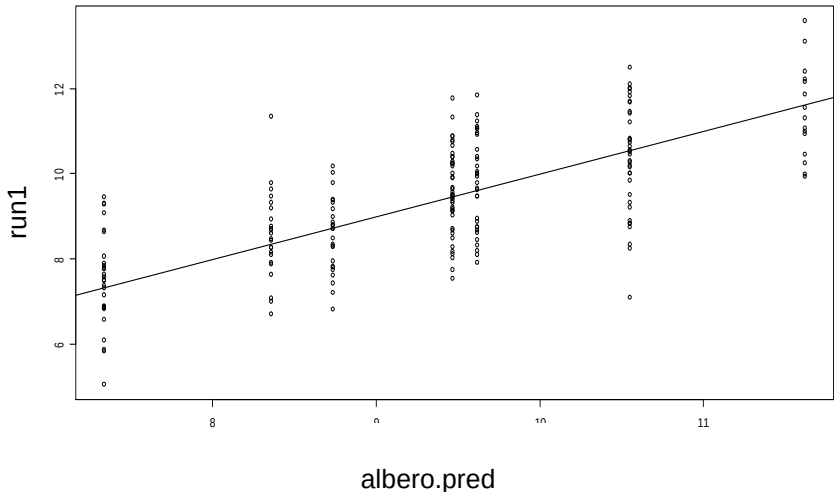

run2< 9.798
rad.somma1< 1.787e+04
run2< 11.45
run2< 8.405
rad.somma1< 2.133e+04
pioggia.somma1>=0.1
vore.10.171>=-0.835
run2< 5.328
t.mas2< 24.25
uore10.171>=1.387
10.91
12.02
9.623
8.508
8.754
9.468
10.67
5.042
7.48
9.145
9.864

Fig. 3.21



Anche per questa classe di modelli proviamo a vedere cosa succede limitandosi agli anni dal 2002 al 2005. L'albero deve essere tagliato già al settimo nodo, con un valore del parametro cp compreso tra 0.019 e 0.013 (*figura 3.21*).

Nel campione *"test"*, lo scarto quadratico medio dei dati è pari ad 1.40, quello fra dati misurati e previsti è 0.55.

L'importanza è data all'ozono di ieri ed alla somma di radiazioni solari.

Gli alberi di regressione e classificazione ci portano per lo più agli stessi risultati, limitando il numero di variabili utilizzate con i modelli lineari: viene infatti data molta meno importanza all'anno e al mese di riferimento, al benzene, alla direzione del vento ed alla pressione.

3.3 Random Forest

Fin'ora abbiamo evidenziato le caratteristiche utili agli alberi decisionali e la loro capacità di maneggiare i dati dando risultati facilmente interpretabili.

Un problema importante però degli alberi è la loro instabilità, in quanto se dovessimo introdurre dati supplementari od ometterne alcuni dal gruppo di dati originale, è probabile che la segmentazione nell'albero decisionale risulterebbe diversa (Hamza & al., 2005).

Sono stati studiati vari metodi per superare questo problema, uno di questi è il *boosting* (Freud & al., 1997; Friedman & al., 2000; Breiman & al., 1996) degli alberi di classificazione: il risultato ottenuto non è un singolo albero ma una foresta di alberi in successione, in cui ogni albero è costruito tenendo conto degli errori di previsione commessi dai precedenti. Gli alberi vengono quindi aggregati con quella che è sostanzialmente un'operazione di media aritmetica pesata, creando un unico classificatore.

Ciò può rendere difficile l'interpretazione del modello. Il criterio ottimale per la scelta del numero di alberi da utilizzare nel modello è ancora oggetto di dibattito.

Altri metodi simili (es: *Boosting Aggregation*, conosciuto anche come Bagging[10]) hanno come elemento in comune il fatto che per l'albero k-esimo viene generato un vettore random O_k, indipendente dai passati vettori casuali $O_1 \ldots O_{k-1}$, ma con la stessa distribuzione: un albero cresce sia usando il training set sia O_k; il risultato viene quindi inserito in un classificatore $h(\mathbf{x}, O_k)$ dove **x** è un vettore di input.

Random Forest, introdotto nel 2001 da Breiman, cerca di sviluppare ulteriormente quest'idea (Breiman, 2001).

[10] Tecnica computazionale basata sul ricampionamento che consente di ridurre la sensibilità di una regola di previsione, sia in termini di varianza che in termini di distorsione delle stime.

Esso può essere utilizzato perché maneggia bene diversi tipi di variabili, è invariante per trasformazioni monotone dei dati di input, è robusto rispetto gli *outlier*, dà una stima dell'errore, della correlazione e dell'importanza delle variabili, e consente di adottare diverse strategie per trattare i dati mancanti. E' stato inoltre dimostrato che, per la legge dei grandi numeri, i modelli di tipo *Random Forest* convergono sempre, cosicché non presentano problemi di *overfitting*. L'accuratezza di *Random Forest* dipende dalla forza della classificazione dei singoli alberi e dalla misura della loro dipendenza reciproca (OOB). I risultati risultano essere insensibili al numero di elementi selezionati per ogni nodo.

Random Forest non necessita della messa in atto di tecniche di validazione incrociata o della verifica su un set separato di variabili per avere una valutazione imparziale dell'errore, in quanto ciò è garantito intrinsecamente dal metodo (Liaw & al., 2002; Gislason & al., 2005).

Ogni albero, infatti, è costruito usando un diverso campione *bootstrap* dei dati originali (composto, cioè, di N elementi estratti, con sostituzione, in un insieme di N); i casi non selezionati per la costruzione dell' albero sono utilizzati per la stima degli errori di classificazione.

Negli alberi standard, il criterio di decisione associato a ciascun nodo è scelto considerando tutte le variabili esplicative disponibili; in Random Forest, ogni nodo è deciso sulla base di un diverso sottoinsieme dei predittori, scelti casualmente. E' questa strategia a garantire la robustezza rispetto l'*overfitting* (Breiman, 2001). Inoltre, il modello è di semplice utilizzo, in quanto prevede l'inserimento di soltanto due parametri (il numero di variabili nel sottoinsieme di variabili casuali usate in ogni nodo ed il numero di alberi nella foresta) e non è molto sensibile ai loro valori.

Dopo che è stato generato un gran numero di alberi, essi “votano” e, per ciascun vettore di input x, viene selezionata la classe più popolare.

Più in dettaglio, dunque, l’algoritmo può essere descritto come segue:

1) estrae un certo numero n di campioni *bootstrap* dai dati originali;
2) per ciascuno degli n campioni *bootstrap*, sviluppa un albero di classificazione o regressione fino alla massima estensione (*unpruned*), con la seguente modifica: per ogni nodo, invece di scegliere il criterio di classificazione migliore su tutti gli M predittori, campiona casualmente m predittori (m<M) e limita la scelta a queste variabili[11];
3) esegue la previsione sui nuovi dati usando la moda delle previsioni degli n alberi, nei problemi di classificazione, o la media, per la regressione.

Una stima degli errori può essere ottenuta come segue:

1) usando l’albero cresciuto con il campione *bootstrap*, si predicono i dati non presenti nel campione *bootstrap* (OOB: dati *“out of bag”*);
2) calcolando il tasso di errore nelle previsioni degli OOB (in media, ogni dato sarebbe *out of bag* attorno al 36% delle volte).

Random Forest fornisce ulteriori informazioni: la possibilità di misurare l’importanza della variabile predittore ed una misura della struttura interna dei dati (prossimità).

Importanza della variabile: è difficile da determinare, perché può essere dovuta alla possibile interazione con le altre variabili. Breiman ha proposto quattro diverse misure che quantificano la rilevanza di ogni variabile.

[11] *bagging* può essere pensato come un caso speciale di *Random Forest* ottenuto quando m = M, numero dei predittori disponibili

Nella *misura 1*, *Random Forest* valuta l'importanza guardando a quanto l'errore di previsione aumenta quando, nei casi OOB, i valori della variabile considerata vengono permutati, mentre tutte le altre restano invariate.

Nella *misura 2* e nella *misura 3* si considera, alla fine della simulazione, il margine dell'*n*-esima unità statistica. Il *margine* è dato dalla proporzione dei voti per la sua vera classe di appartenenza (nota) meno il massimo tra le proporzioni di voti per ognuna delle rimanenti classi. La seconda misura per la *m*-esima variabile si ottiene come media dei margini che si sono abbassati per ogni caso quando la variabile *m*-esima è permutata come per la *misura 1*. La *misura 3* rappresenta il conteggio di quanti margini si sono abbassati diminuita del numero di margini che si sono alzati.

Nella *misura 4,* ad ogni suddivisione, una delle variabili è usata per formare la suddivisione, evento che comporta una riduzione dell'indice del Gini. La somma di tutti i decrementi nella foresta dovuti ad una certa variabile, normalizzato per il numero di alberi, costituisce la *misura 4*.

Prossimità: l'elemento (i,j) della matrice di prossimità è la frazione degli alberi in cui gli elementi i e j cadono nello stesso nodo terminale. L'intuizione suggerisce che osservazioni "simili" dovrebbero trovarsi negli stessi nodi terminali più spesso di quelle fra loro dissimili. La tabella di prossimità può essere usata per identificare la struttura dei dati o per una classificazione non supervisionata.

È stato implementato il metodo sugli stessi dataframe usati per gli alberi di classificazione ed i modelli lineari; in particolare, sul sottinsieme di dati denominato *"analisi"*.

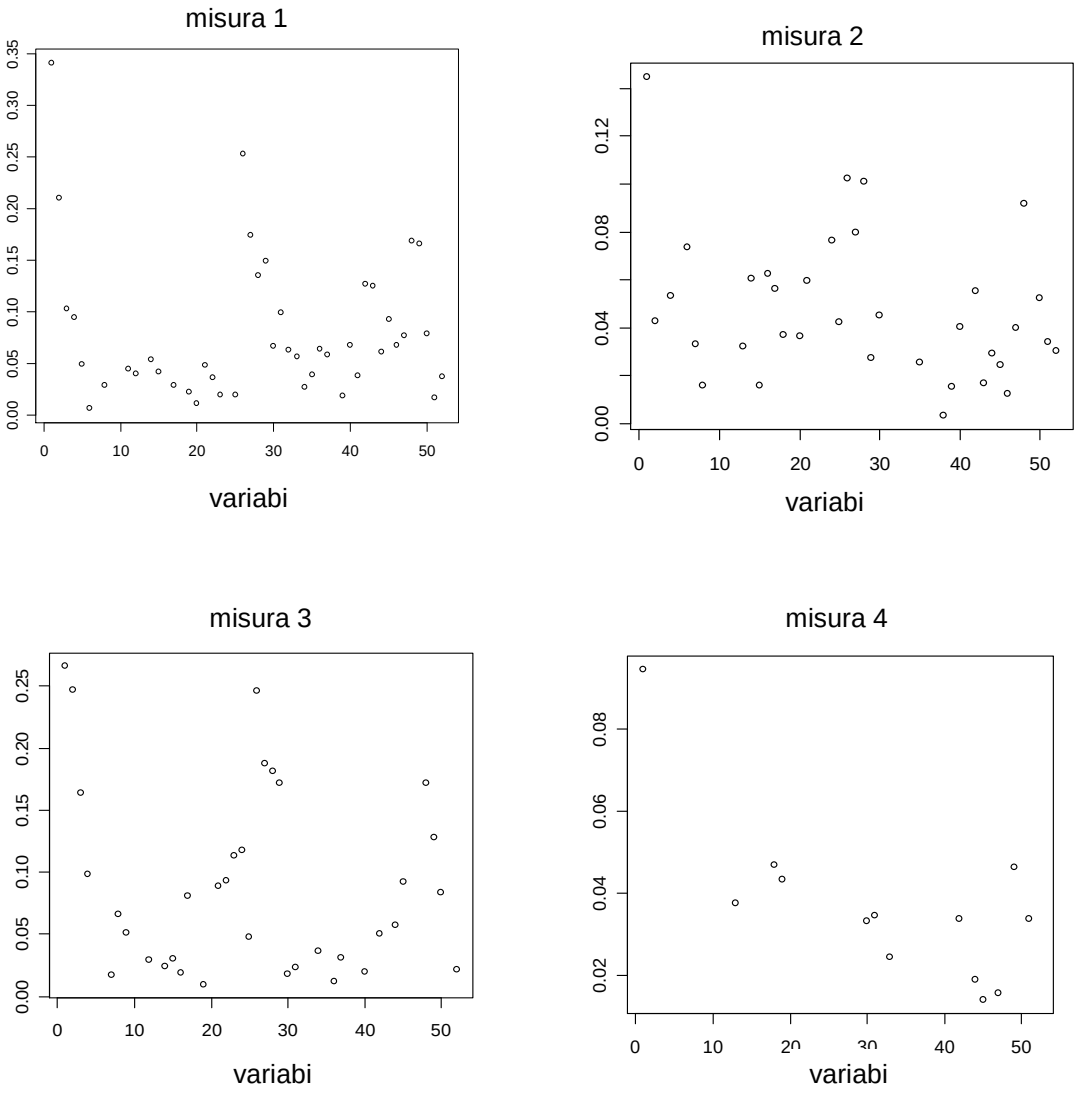
Per Cairoli, il numero dei casi su cui il modello riesce ad effettuare le previsioni, utilizzando tutte le variabili disponibili, è relativamente basso:

157 previsioni corrette su 260 (60.38%).

La stima degli errori di classificazione OOB è stata di 39.77%.

Le variabili più importanti sono evidenziate dai quattro grafici della *figura 3.22*, raffiguranti le stime dell'importanza delle variabili secondo Breiman; il grafico in alto a sinistra, rappresenta la *misura 1* dell'importanza delle variabili.

Fig. 3.22



Essa è stata assunta come riferimento principale, fissando una soglia del 10%; le variabili che superano tale soglia di importanza sono:

`run2, run3, t.mas1, t.mas2, rad.somma1, rad.somma2, uore10.171, pioggia.somma1, pioggia.somma2, vore.10.171, vore.10.172, press.mean1, press.mean2, ur.mean1, ur.mean2, ins.somma1, ins.somma2.`

Tab. 3.17

Dati \	cai.pred			
	altissimo	alto	medio	basso
altissimo	21%	7%	0%	0%
alto	7%	16%	2%	1%
medio	1%	7%	2%	10%
basso	0%	1%	4%	22%

Diag: 61% N.tot=259

Sono stati ribaditi i risultati ottenuti con i metodi precedenti, con l'eccezione dell'aumento di importanza dato all'ozono di 2 giorni precedenti (anche se appena sopra soglia) e dell'importanza aggiunta alla temperatura, al tempo e all'intensità delle radiazioni solari del giorno precedente e della pioggia, cosa che negli altri modelli non era stata evidenziata.

Da evidenziare l'importanza che il modello non sottostimi il livello di ozono quando è "alto" o "altissimo"; per questi casi vediamo che questo modello commette un errore del 3% (*tabella 3.17*), mentre negli alberi di classificazione l'errore dello stesso tipo è stato del 7%.

È stato ristimato il modello usando solo le variabili d'interesse:

sono stati predetti correttamente 619 su 961 (64%). L'errore sulle classi "alto" ed "altissimo" è ora del 7% (*tabella 3.18*); l'errore OOB del 36%.

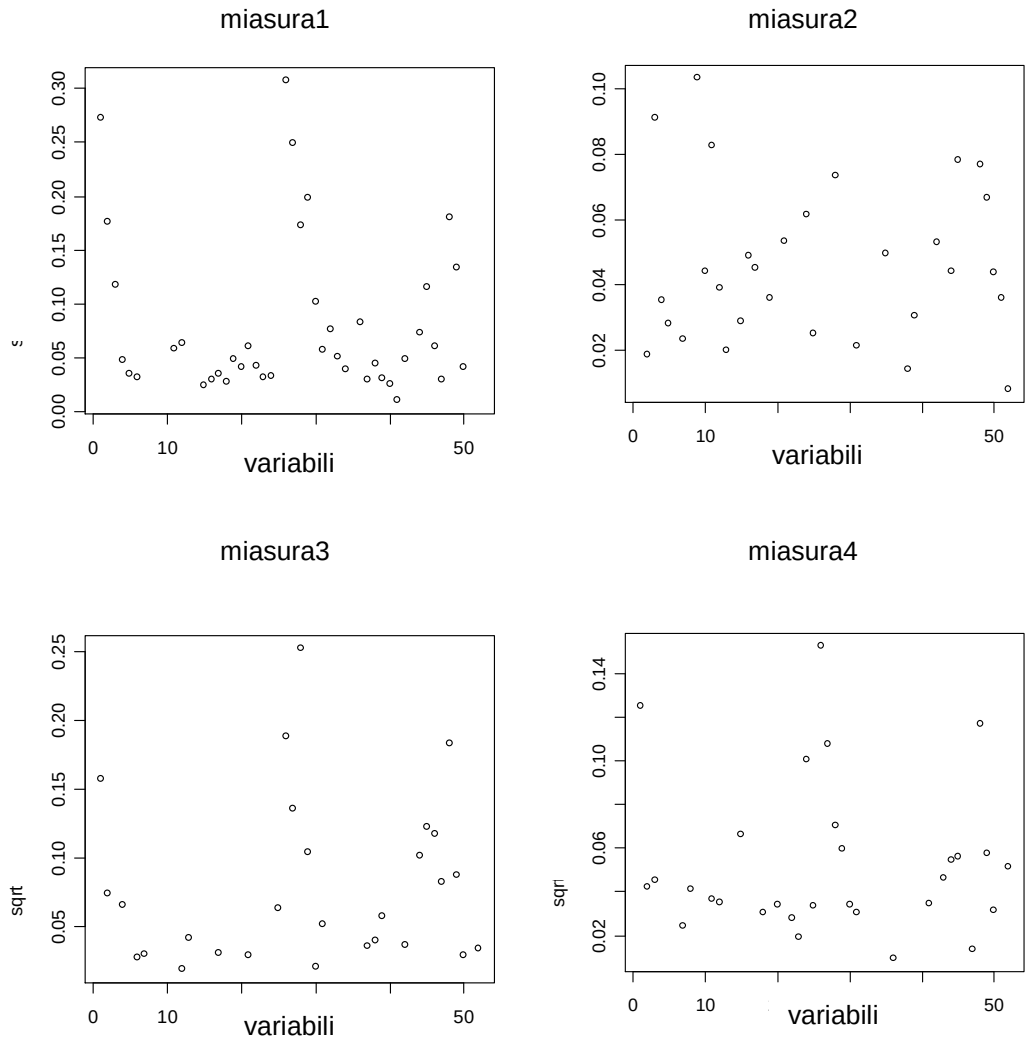
Per i massimi diurni, le variabili significative sono sostanzialmente le stesse.

Tab. 3.18

Dati \	cai.pred			
	altissimo	alto	medio	basso
altissimo	20%	5%	0%	0%
alto	5%	14%	6%	1%
medio	1%	7%	12%	4%
basso	0%	1%	5%	18%

Diag: 64% N.tot=961

Fig. 3.23



Per quanto riguarda i massimi notturni, le variabili significative restano solamente `max.O3.N1`, `max.O3.D1`, `max.O3.D2` e la temperatura di oggi e di ieri, mentre l'anno e il vento, che risultavano importanti per gli alberi, perdono importanza. Al limite dell'importanza troviamo `max.NO2.D1`.

Le stesse analisi vengono rifatte anche per Osvaldo e Manzoni. Le variabili più importanti per Osvaldo, vengono riassunte nella *figura 3.23*.

Questi grafici permettono di capire che le variabili importanti selezionate restano le stesse che per Cairoli.

Tab. 3.19

Dati \	osv.pred			
	altissimo	alto	medio	basso
altissimo	25%	0%	4%	0%
alto	4%	2%	8%	0%
medio	5%	2%	27%	3%
basso	0%	0%	12%	8%

Diag: 62% N.tot=129

A differenza delle conclusioni tratte dagli alberi, il Benzene perde importanza; la pioggia, che era significativa solo per spiegare il massimo giornaliero di ozono, acquista importanza anche per il massimo trascinato sulle 8 ore. Per quanto riguarda le previsioni, la *tabella 3.19* mostra la presenza del 62% di osservazioni predette correttamente, a differenza del 67% trovate con gli alberi. La stima degli errori di classificazione OOB è del 40%. Ricostruendo il modello con le sole le variabili ritenute più importanti (*tabella 3.20*), si ottiene un errore OOB pari al 32%

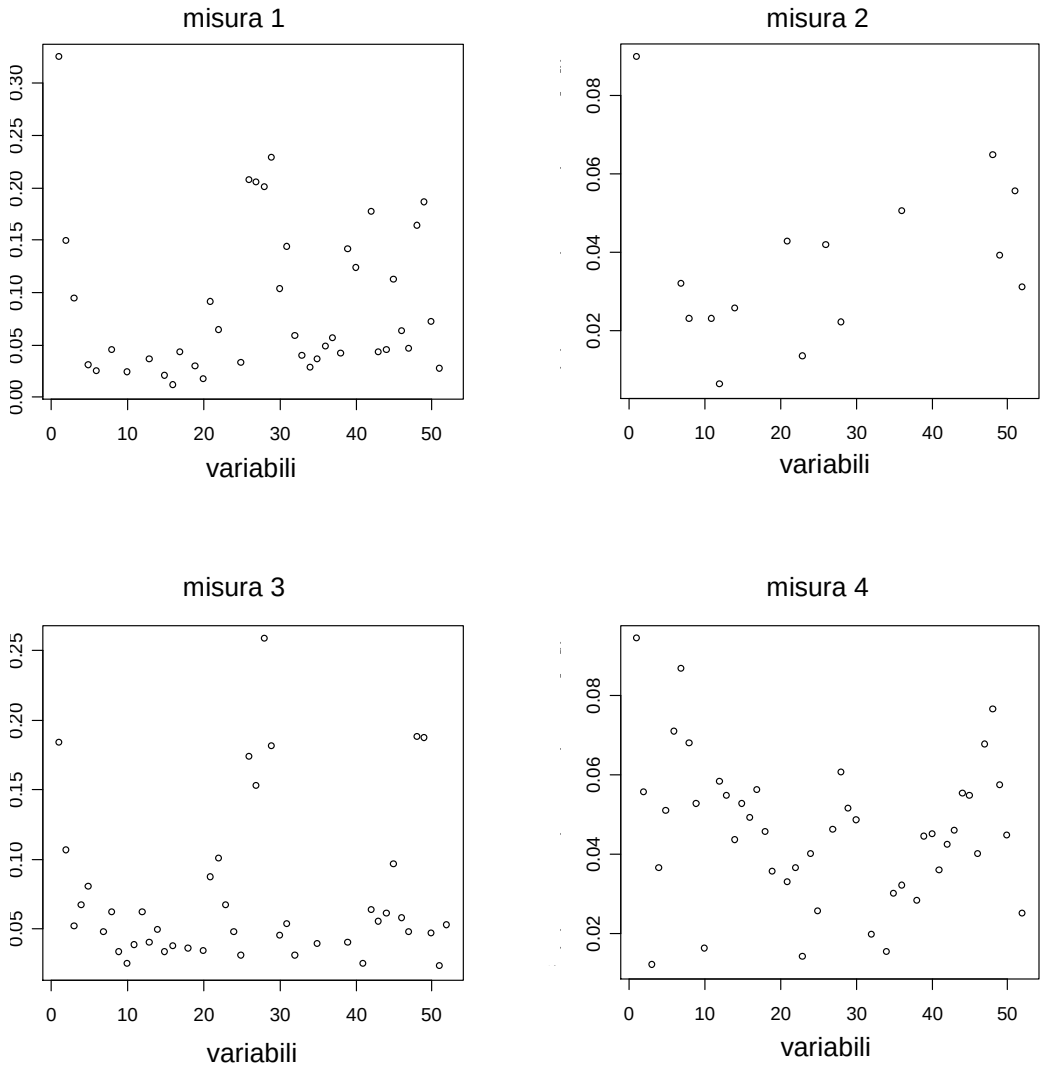
Le variabili più significative sono:

`run2, run3, t.mas1, t.mas2, rad.somma1, uore.6.91, uore10.171, vore.6.91 , vore.10.171, press.mean1, press.mean2, ur.mean1, ur.mean2.`

Tab. 3.20

Dati \	osv.pred			
	altissimo	alto	medio	basso
altissimo	21%	2%	2%	0%
alto	5%	4%	8%	0%
medio	2%	2%	27%	4%
basso	0%	0%	7%	16%

Fig. 3.24



Passando a considerare i massimi diurni e notturni, le variabili significative rimangono le stesse con l'eccezione che nel primo caso pioggia, umidità e pressione perdono importanza, mentre nel massimo notturno ritornano, al limite dell'importanza, pressione, umidità, NO2 della notte stessa presa in considerazione e l'intensità di radiazioni solari del giorno

Tab. 3.21

Dati \	man.pred			
	altissimo	alto	medio	basso
altissimo	10%	0%	6%	0%
alto	1%	0%	5%	0%
medio	1%	0%	43%	6%
basso	0%	0%	10%	18%

Diag: 71% N.tot=292

precedente: le variabili significative in questo caso sono in particolar modo l'ozono dei giorni precedenti e la temperatura. Il benzene non risulta importante.Per quanto riguarda Manzoni, le tabelle d'interesse sono indicate dalla *figura 3.24.* Se si fissa una soglia di importanza del 10%, le variabili che la superano, sono le stesse di Cairoli e Osvaldo. In Manzoni, il benzene sembra essere più importante e il vento assume importanza anche in altre fasce temporali (velocità media del vento del giorno stesso, in tutte e tre le fasce orarie considerate).

Le previsioni fatte da *Random Forest* sono rappresentate nella *tabella 3.21*; l'errore OOB corrispondente è del 29%, sulla diagonale troviamo il 71.6% di previsioni contro il 73.8% degli alberi.

Per quanto riguarda i massimi notturni, se guardiamo alla significatività al 10% troviamo `run2, run3, t.mas1` e `t.mas2` e la somma della radiazione solare del giorno precedente. I massimi diurni sono invece spiegati dalle stesse variabili importanti di `run`.

Se si ristima il modello tenendo le variabili significative:

`run2, run3, maxBEN1, maxBEN2, t.mas1, t.mas2, rad.somma1, rad.somma2, u20.29MediaDalleOre2, uore.6.92,`

uore10.171 v20.29MediaDalleOre2, vore.6.91, vore.10.171, press.mean1, press.mean2, ins.somma1, ins.somma2.

La stima OOB dell'errore è del 27% (*tabella 3.22*).

Tab. 3.22

Dati \	dati.pred			
	altissimo	alto	medio	basso
altissimo	12%	0%	4%	0%
alto	3%	0%	5%	0%
medio	2%	0%	42%	4%
basso	0%	0%	8%	18%

Diag: 72% N.tot= 499

Per quanto riguarda Random Forest, è stato rifatto il modello delle previsioni giorno per giorno per le previsioni dell'ultimo anno paragonate ai valori reali, le tre tabelle seguenti ne rappresentano il quadro riassuntivo:

Tab. 3.23

Cairoli

Dati \	dati.prev			
	altissimo	alto	medio	basso
altissimo	24%	10%	1%	0%
alto	1%	17%	18%	1%
medio	1%	2%	13%	6%
basso	0%	0%	1%	8%

Diag: 62% N.tot=157

Manzoni

Dati \	dati.pred			
	altissimo	alto	medio	basso
altissimo	8%	0%	5%	0%
alto	3%	0%	3%	0%
medio	1%	1%	46%	9%
basso	0%	0%	9%	16%

Diag: 70% N.tot:=177

Osvaldo

Dati \	dati.pred			
	altissimo	alto	medio	basso
altissimo	20%	3%	1%	0%
alto	3%	5%	10%	0%
medio	2%	2%	30%	7%
basso	0%	0%	5%	13%

Diag: 68% N.tot=152

Per quanto riguarda il massimo diurno dell'ozono per RF, è stato rifatto il modello delle previsioni giorno per giorno per le previsioni dell'ultimo anno paragonate ai valori reali, le tre tabelle seguenti ne rappresentano il quadro riassuntivo:

Tab. 3.24

Cairoli				
Dati \	dati.prev			
	altissimo	alto	medio	basso
altissimo	20%	11%	1%	0%
alto	3%	30%	1%	6%
medio	0%	7%	1%	8%
basso	0%	4%	1%	6%

Diag: 57% N.tot=158

Manzoni				
Dati \	dati.pred			
	altissimo	alto	medio	basso
altissimo	15%	1%	2%	0%
alto	3%	1%	6%	0%
medio	2%	1%	30%	12%
basso	0%	0%	11%	16%

Diag: 62% N.tot:=177

Osvaldo				
Dati \	dati.pred			
	altissimo	alto	medio	basso
altissimo	20%	3%	0%	0%
alto	4%	26%	1%	5%
medio	1%	8%	3%	8%
basso	0%	6%	1%	13%

Diag: 62% N.tot=156

4 - Considerazioni conclusive

4.1 I modelli sviluppati

In questo Studio è stato adottato, inizialmente, un approccio basato su un modello lineare, sperimentando quindi modelli ritenuti capaci di interpretare la non linearità del problema; in particolare, alberi di regressione, di classificazione e *Random Forest*.

I risultati sono stati confrontati fra loro e con il modello *naive*, basato sulla semplice ipotesi di persistenza del livello di ozono del giorno precedente.

L'esplorazione dell'efficacia dei modelli è stata limitata, per alcuni aspetti, dalla disponibilità delle variabili. Per esempio, non c'è ancora la disponibilità dei dati relativi agli indicatori della stabilità atmosferica (altezza di rimescolamento, classi di stabilità di Pasquill-Gifford, etc.); né è stato possibile reperire, presso alcun Ente, indicatori relativi all'intensità del traffico veicolare nell'area urbana di Udine.

Per poter confrontare tutti i modelli sviluppati, la *tabella 4.1* rappresenta uno schema riassuntivo delle percentuali delle previsioni corrette effettuate

Tab. 4.1

Previsioni media trascinata "giorno per giorno"

	via Cairoli	S. Osvaldo	via Manzoni
modello naive	55%	59%	59%
modello lineare	60%	71%	65%
alberi di classificazione	59%	59%	70%
Random Forest	62%	68%	70%

Previsioni massimo giornaliero "giorno per giorno"

	via Cairoli	S. Osvaldo	via Manzoni
modello naive	54%	54%	50%
alberi di classificazione	56%	63%	60%
Random Forest	57%	62%	62%

giorno per giorno per ogni modello e in tutte tre le stazioni.

4.1.1 Massimo giornaliero della media nelle 8 ore

Cairoli: il vento risulta essere una variabile molto significativa, soprattutto nelle fasce tra le ore 10-17 e 6-9 (nella direzione Nord-Sud discrimina soprattutto fra i valori alti/altissimi di ozono, mentre Est-Ovest tra quelli medio/bassi); ciò si nota soprattutto grazie agli alberi e a Random Forest. I valori altissimi dell'ozono sono determinati da alti valori di ozono del giorno precedente, dalle alte temperature, dall'insolazione del giorno stesso e dalla direzione Nord-Sud del vento nelle due fasce orarie scritte sopra.

Osvaldo: vento (10-17 e 6-9), pressione e umidità sono variabili significative per Random Forest e per i modelli lineari, mentre gli alberi tendono ad evidenziare soprattutto l'ozono del giorno precedente, le alte temperature, radiazione e insolazione del giorno stesso; è presente e significativa la variabile anno: indice che persiste qualche componente del fenomeno di cui le altre variabili esplicative utilizzate non riescono a render conto. Il benzene è significativo per tutti i modelli, ma è richiamato soprattutto per determinare valori medio bassi di ozono.

Manzoni: ozono del giorno precedente, vento, radiazione solare sono le variabili più importanti; interessante la rilevanza del vento e della variabile pioggia, soprattutto negli alberi di regressione.

Un risultato migliore viene raggiunto dagli alberi e da *Random Forest* in Manzoni, mentre il modello lineare trova un maggior riscontro in Osvaldo.

Com'è lecito attendersi, i modelli *naive* danno una percentuale più bassa rispetto gli altri. Ciò non accade in Osvaldo, dove si può notare la coincidenza con il risultato trovato con gli alberi nella previsione del massimo

della media trascinata nelle otto ore. Da evidenziare, a riguardo, il fatto che le previsioni nei tre modelli studiati sono sviluppati dall'analisi dei primi cinque anni e poi implementati sull'ultimo anno (da 150 a 170 osservazioni), mentre quello *naive* prende in considerazione tutti sei gli anni (da 990 a 1030 dati).

Complessivamente:

- i modelli di tutti e tre i tipi segnalano l'ozono del giorno precedente come variabile essenziale. Radiazione solare globale e tempo di insolazione (eliofania) sono altamente correlate fra loro e vengono scelte alternativamente da tutti i modelli;
- *Random Forest* dà più peso rispetto agli altri anche all'ozono di due giorni precedenti, così come, più in generale, ad altre variabili ritardate;
- variabili come benzene, pressione ed umidità, che erano state segnalate dalla regressione lineare, tornano ad essere importanti in *Random Forest*;
- il benzene manifesta la sua importanza in tutti tre i modelli relativi ad Osvaldo. In Manzoni, stazione in cui viene rilevato, assume significatività solo nei modelli lineari e in *Random Forest*;
- per quanto riguarda il vento, la fascia oraria predominante, selezionata per tutte le stazioni, è quella 10-17. La sua importanza nel determinare valori alti di ozono si nota soprattutto negli alberi;
- anno e mese di riferimento erano importanti per il modello lineare in Osvaldo e Manzoni; non vengono utilizzati negli altri modelli, tranne che nel modello ad albero di Osvaldo.

4.1.2 Massimo orario giornaliero (diurno)

Le considerazioni appena svolte valgono, sostanzialmente, anche per la previsione di questa variabile; in generale, le percentuali di successo dei

modelli, compreso il modello *naive*, sono inferiori, tranne che per l'albero di classificazione di Osvaldo.

4.1.3 Massimo orario notturno

Questa variabile si è rivelata di difficile previsione. Soprattutto, appare necessario riuscire a distinguere fra fenomeni di *persistenza* di alti valori di ozono nelle ore serali da episodi di veri e propri *picchi notturni*, che vanno ricondotti a fenomeni chimico-fisici diversi (come l'adduzione di masse d'aria provenienti dallo strato d'accumulo posto al margine dello Strato Limite Planetario (Sozzi R., 2003).

4.2 Evidenze legate alla natura fisico-chimica dei fenomeni

Le concentrazioni di ozono registrate in via Manzoni, area maggiormente interessata dal traffico veicolare, sono sensibilmente più basse di quelle misurate nelle altre postazioni: ciò è coerente con l'effetto atteso in presenza di intense emissioni di NO.

Confermato dai modelli anche il forte legame con l'irraggiamento solare.

La presenza di venti, collegati a fenomeni di trasporto, nelle ore diurne, viene interpretata in modo non univoco dai modelli, e richiede approfondimenti.

4.3 Utilizzo operativo dei modelli per fini predittivi

La previsione dei livelli di ozono è di estrema utilità per garantire la tempestiva dell'informazione al pubblico, da parte delle autorità competenti,

al fine di suggerire comportamenti che riducano l'esposizione dei soggetti a rischio.

Tipicamente le previsioni dovrebbero essere disponibili con almeno un giorno d'anticipo, ma preferenzialmente con più giorni, tempo richiesto per preparare le misure di prevenzione.

I modelli per le previsioni dell'ozono troposferico possono essere basati su equazioni deterministiche derivate dalle teorie relative ai processi fisici e chimici nell'atmosfera; tuttavia tali modelli sono inadatti in molti contesti operativi, perché richiedono molti input fisici complessi e l'impiego di molto personale. Inoltre, causa l'influenza delle variabili meteo sul livello dell'ozono e la grande incertezza associata all'utilizzo di questi dati di input, è molto difficile ottenere un buon accordo tra dati osservati e previsti.

Al giorno d'oggi, è ampliamente riconosciuto che le relazioni tra ozono, precursori e variabili meteorologiche sono complesse e non lineari.

A causa di queste difficoltà, come alternativa ai modelli deterministici, per prevedere le concentrazioni di ozono sono stati usati modelli stocastici basati sui metodi di regressione multipla non lineare, reti neurali, modelli addittivi, autoregressivi, che includono oltre ai valori di ozono e dei precursori quali NO_x, benzene e variabili meteorologiche.

È frequente l'utilizzo congiunto di più modelli *deterministici* e *statistici*, i cui risultati vengono confrontati e valutati criticamente dagli operatori ai fini della formulazione definitiva della previsione (ARPA Emilia Romagna, 2006).

In generale, l'utilizzo operativo dei modelli ai fini predittivi richiede di disporre delle previsioni per alcune delle variabili esplicative. Dato il carattere preliminare di questo Studio, si è ritenuto importante avviare il lavoro su variabili *misurate*, e non *previste*, per stabilire la massima efficacia previsiva ottenibile dai modelli.

Tab. 4.2

Dati \	pred			
	altissimo	alto	medio	basso
altissimo	21%	7%	6%	1%
alto	3%	11%	15%	7%
medio	0%	1%	14%	6%
basso	0%	0%	1%	7%

Diag: 53%

È necessario attendersi dunque che, in un utilizzo operativo dei modelli qui sviluppati, le percentuali di successo siano destinate ad abbassarsi, a causa dell'incertezza aggiuntiva sulle previsioni delle variabili esplicative.

AI fini del successivo passaggio all'operatività, i risultati sono stati tuttavia confortanti: i modelli hanno identificato, come variabili di primaria importanza, la temperatura massima, la somma della radiazione solare incidente, il tempo di irraggiamento solare diretto, che possono tutte essere previste con alti livelli di confidenza, oltre all'ozono del giorno precedente, garantito dalle misure.

4.4 Analisi orientate alla razionalizzazione della rete

La *tabella 4.2,* che riprende la *tabella 3.14* riportata nel testo, riassume un'analisi fatta con gli alberi di classificazione allo scopo di riuscire a prevedere l'ozono in v.Cairoli, avendo a disposizione esclusivamente le variabili relative alle altre centraline: ciò per indicare un'eventuale possibilità di smettere la rilevazione dell'inquinante nella centralina in esame.

I risultati ottenuti non sono sufficienti per riuscire a garantire una tale possibilità.

4.5 Indicazioni per successivi sviluppi

Specificamente per fini predittivi:

- utilizzo di variabili esplicative *previste*, anziché *misurate*, per il giorno di interesse;
- creazione e previsione di un unico indice di concentrazione dell'ozono per l'area urbana di Udine.

Più in generale:

- esplorazione dei meccanismi NO/NO_2, riconsiderando le serie dei dati orari;
- introduzione di una variabile indicante il livello di ozono di background, da ottenere mediando opportunamente rilevazioni provenienti da un'area più vasta o acquisendo i risultati di modelli fotochimici fatti operare da atri Enti su scala di bacino padano, nazionale od europea;
- introduzione di variabili meteorologiche misurate in quota, per mezzo del radiosondaggio di Campoformido (UD);
- introduzione di parametri di stabilità atmosferica (altezza dello strato di rimescolamento, ecc.), stimate da opportuni *processori* meteorologici;
- nei modelli di classificazione, introduzione di una matrice di pesi per i vari tipi di errore, in modo da ottimizzare l'efficacia nella previsione dei valori alti ed altissimi;
- sulla base della conoscenza dei fenomeni fisico-chimici:
  - analisi mirate sui dati orari;

- sviluppo dei modelli su sottinsiemi dei dati originali, corrispondenti a particolari condizioni meteorologiche;
- introduzione di termini di interazione fra le variabili;
- esplorazione dei Generalized Additive Models (GAM).

Ai fini dell'utilizzo dei modelli per un supporto agli operatori nella *validazione* dei dati, infine, è opportuno condurre l'analisi delle serie storiche dei dati *orari*.

L'utilizzo di modelli del tipo qui implementato richiede, in tal caso, di affrontare le complessità legate alla *ciclostazionarietà* delle serie storiche.

5 - Bibliografia

ARPA Emilia Romagna (2006), *Previsione dei livelli di ozono.* www.arpa.emr.it/sim/archivio/downloads/ambiente/ozononuovo.pdf

ARPA FVG (2005), *Rapporto annuale sulla qualità dell'aria nel comune di Udine*, a cura del Servizo Tematico Analitico del Dipartimento provinciale di Udine – anno 2004.

ARPA FVG (2002), *Rapporto sullo Stato dell'Ambiente del Friuli Venezia Giulia anno 2002*.

Baird C. (a) (1997), *Chimica Ambientale*, Zanichelli, cap.3, pagg.45-73.

Baird (b) (1997), *Chimica Ambientale*, Zanichelli, Appendice II, pagg.235-237.

Barrero M.A., Grimalt J.O, Canto´n, L. (2005*), Prediction of daily ozone concentration maxima in the urban atmosphere,* Chemical Engineering Group, Department of Applied Chemistry, Faculty of Chemistry, University of the Basque Country, P. Manuel de Lardizabal, 3, 20018 San Sebastia´n, Spain.

Benvenuto, F. and Marani, A. (2000), *Nowcasting of urban air pollutants by neural networks,* Nuovo Cimento, 23C, 567-586.

Bernard G.J., Massart Olav M. Kvalheim (1998), Ozone forecasting from meteorological variables Part I. *Predictive models by moving window and partial least squares regression, Department of Chemistry, Uníersity of Bergen, N-5007 Bergen, Norway.*

Bolzacchini E.(2005-2006), Corso di Chimica dell'Atmosfera, Università di Milano Bicocca - Dipartimento di Scienze Ambientali, Corso di Laurea in Scienze dell'Ambiente e del Territorio, a.a. http://www.disat.unimib.it/chimamb/CHI_ATM_05/OZONO%20TROPOSFERICO.ppt

Bordignon S., Gaetan C. and Lisi F. (2002), *Nonlinear models for ground-level ozone forecasting,* Journal of the italian statistical society, 11-2, 227-245.

Breiman L., Friedman J., Olshen R. and Stone C. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Breiman L. (2001), *Random Forest,* Statistics Department, University of California, machine learning, 45, 5-32.
Breiman L. (1996), *Bagging Predictors*, Machine Learning, 24(2):123-140.

Burrows W., Benjamin M., Beauchamp S., Lord E., McCollor D. and Thomson B. (1995), *CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada*, *Journal of Applied Meteorology*, 34, 1848–1862.

Davis J. and Speckman P. (1999), *A model for predicting maximum and 8h average ozone in Houston*, *Atmospheric Environment*, 33, 2487–2500.

Damon J. and Guillas S. (2001),*The inclusion of exogenous variables in functional autoregressive ozone forecasting,* Institut de Statistique de l'Universit_e de Paris Laboratoire de Statistique Th_eorique et Appliquée.

Freund Y. and Schapire R. (1997), *A decision-theoretic generalization of ondine learning and an application to boosting*, *Journal of Computer and System Sciences*, 55, 119–139.

Friedman J. (1991), *Multivariate adaptive regression splines*, *The Annals of Statistics*, 19, 1–50.

Friedman J. Hastie T. Tibshirani R. (2000), *Special invited paper additive logistic regression: a statistical view of boosting, The Annals of Statistics*, 28, 337-374.

Giaiotti D. e Stel F. (2006), comunicazione privata, Visco (Ud), aprile 2006.

Gislason P. G. Benediktsson J. A. Sveinsson J. R. (2005), *Random Forests for land cover classification,* Department of Electrical and Computer Engineering, University of Iceland, Hjardarhaga 2-6, IS 107 Reykjavik, Iceland.

Graf-Jacottet M. and Jaunin M. (1998*), Predictive models for ground ozone and nitrogen dioxide time series*, *Environmetrics*, 9, 393–406.

Hamza M. and Larocque L. (2005), *An empirical comparison of ensemble methods based on classification trees,* Department of Management

Sciences, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine,Montréal (Québec), Canada H3T 2A7.

Kumar A., Vedula S. and Sud A., (2000), *Development of an Ozone Forecasting Model for Non-Attainment Areas in the State of Ohio,* Environmental Monitoring and Assessment, 62: 91-111.

Kuhnert P. and Venables B. (2005), *An Introduction to R: Software for StatisticalModelling & Computing,* CSIRO Mathematical and Information Sciences, Cleveland, Australia.

Liaw A. and Wiener M. (2002), *Classification and Regression by Random Forest*, R News , 2\3, 18-22.

Liguori F. (1996), *Urban pollution data interpretation and nowcasting by neural networks*, laurea in scienze ambientali.

Moimas F. (2006), comunicazione privata, Udine, marzo 2006.

Pastore A. (1997), *Modelli neuronali per lo smog fotochimica,* laurea in scienze ambientali.

Rindone B. (2005-2006), Materiale di Chimica dell'Ambiente per STAT, Università di Milano Bicocca.
http://www.disat.unimib.it/ita/corso/news_chimamb.htm

Sozzi R. (2003), *La micrometeorologia e la dispersione degli inquinanti aria*, ARPAT CTN – ACE.

Therneau Terry M and Beth Atkinson. R port by Brian Ripley, <ripley@stats.ox.ac.uk>. (2005), rpart: Recursive Partitioning. R, package version 3.1-23. S-PLUS 6.x original at http://www.mayo.edu/hsr/Sfunc.html

US EPA, (1999), *Guideline for developing an ozone forecasting program* EPA-454/R-99-009.

6 - Legenda

Variabili usate	
cai	stazione di Cairoli
man	stazione di Manzoni
osv	stazione di S. Osvaldo
cairit	matrice dati Cairoli
osvrit	matrice dati Osvaldo
manrit	matrice dati Manzoni
Inquinanti	
run*N*	radice quadrata del massimo giornaliero della media trascinata dell'ozono sulle otto ore
max.O3.N*N*	radice quadrata del massimo notturno dell'ozono
max.O3.D*N*	radice quadrata del massimo diurno dell'ozono
max.NO2.N*N*	massimo diurno del biossido di azoto
max.NO2.D*N*	massimo diurno del biossido di azoto
maxBEN*N*	massimo giornaliero del benzene
Variabili meteorologiche	
t.mas*N*	temperatura massima giornaliera
rad.somma*N*	somma giornaliera delle radiazioni solari
pioggia.sommaN	somma giornaliera della quantità di pioggia caduta
uore20.29*N*	media dalle ore 20 alle 5 del giorno successivo della componente est-ovest del vento (positiva da est)
uore.6.9*N*	media dalle ore 6 alle 9 della componente est-ovest del vento (positiva da est)

uore10.17*N*	media dalle ore 10 alle 17 della componente est-ovest del vento (positiva da est)
vore20.29*N*	media dalle ore 20 alle 5 del giorno successivo della componente nord-sud del vento (positiva da nord)
vore.6.9*N*	media dalle ore 6 alle 9 della componente nord-sud del vento (positiva da nord)
vore.10.17*N*	media dalle ore 10 alle 17 della componente nord-sud del vento (positiva da nord)
press.mean*N*	pressione media giornaliera
ur.mean*N*	umidità media giornaliera
ins.somma*N*	tempo di insolazione giornaliero
mese	mese di riferimento (stagione dell'ozono va da aprile a settembre)
giorno	giorno della settimana
anno	anno di riferimento (2000-2005)
fatrun*N*	variabile qualitativa relativa alla suddivisione fatta in base ai quantili della distribuzione e al limite di legge del massimo giornaliero della media trascinata dell'ozono sulle otto ore
fato3D*N*	variabile qualitativa relativa alla suddivisione fatta in base ai quantili della distribuzione e al limite di legge del massimo diurno dell'ozono
MAXnotte	variabile qualitativa relativa alla suddivisione fatta in base ai quantili della distribuzione massimo notturno dell'ozono

N è uguale a uno se corrisponde al giorno corrente, , N-1 sono i giorni precedenti

7 - Allegati

Elenco delle procedure:

Di seguito viene riportato un elenco delle procedure utilizzate in queste analisi.

In una prima fase sono state utilizzate delle procedure specifiche relative ai dati orari disponibili, allo scopo di trasformarli in dati giornalieri; vengono elencato al solo scopo di essere un punto di riferimento:

- ozono: media trascinata, massimo giornaliero diurno e notturno;
- ossidi di azoto: massimo giornaliero diurno e notturno;
- benzene: massimo giornaliero;
- temperatura : massimo giornaliero;
- radiazione solare, pioggia, insolazione: somma giornaliera;
- vento diviso nelle due direzioni (u, v) : media giornaliera in tre intervalli di tempo (6-9, 10-17, 20-5);
- pressione, umidità: media giornaliera.

Per quanto riguarda le procedure utilizzate per fare le previsioni giorno per giorno, vengono di seguito riportati i codici per i due metodi principali utilizzati:

- previsioni giorno per giorno per gli alberi;
- previsioni giorno per giorno per Random Forest.

Procedure analoghe sono state fatte non solo per la media trascinata (riportate di seguito), ma anche per i massimi notturni e diurni dell'ozono.

Alberi:

Creazione: 20.07.2006

File: alberi

Finalità: calcola la previsione giorno per giorno dell'ultimo anno di osservazione: 2005, crea delle tabelle previsive ogni 15 giorni, da la percentuale dei dati che si trovano sulla diagonale

Input: MAT.DATI = MATRICE DI DATI (cairit,osvrit,manrit) la variabile y : fatrun1, fato3D1, fato3N1

Returns: stampa a video delle percentuali di successo delle previsioni ogni 15 giorni, stampa in C i valori effettivamente rilevati nel 2005 tratti dal dataste e le previsioni effettuate giorno per giorno di quest'ultimo anno: salva i dati in "C:\\mat.txt"

Esempio:

```
# inserire la matrice di dati
mat.dati = manrit
```

```
#dire il numero minimo di osservazioni per ogni nodo finale
 minsplit = 5

#numero di righe totale
 n = nrow ( mat.dati )

# predice l'ultimo anno in base agli altri 5 e alle osservazioni
# mancanti inserite dentro
step0 <- round((5/6)*n)

# crea un dataframe: nella prima colonna troviamo il valore
# osservato, nella seconda la previsione, nella terza il numero di
# nodi dell'albero selezionato
 valorevero<- mat.dati$fatrun1[(step0+1):n]
 valoreprevisto<-valorevero
 valoreprevisto[1:(n-step0)]<-NA
 nodi<- as.integer(rep(NA,(n- step0)))
  sintesi <- data.frame(valorevero,valoreprevisto,nodi)

 # stima il modello usando inizialmente il modello con “step0” dati,
 # prevedo step0+1; stima di nuovo il modello con step0+1 e prevede #
step0+2
 for (i in step0:(n-1))  {
 set.seed(123456)
```

```
x.rp<-
rpart(fatrun1~.,data=mat.dati[1:i,],method="class",control=rpart.control(minsplit=5,cp=
0.0001))

# queste righe sono prese pari pari dal listato di cpplot()
p.rpart <- x.rp$cptable
xstd <- p.rpart[, 5]
xerror <- p.rpart[, 4]
cp0 <- p.rpart[, 1]
cp <- sqrt(cp0 * c(Inf, cp0[-length(cp0)]))

# minpos è l'albero col miglior cross validate error
minpos <- min(seq(along = xerror)[xerror == min(xerror)])

# trova l'albero più piccolo, tale che  xerror < min(xerror+xstd)
minpos.first <- min (seq (along = xerror) [xerror<((xerror+xstd)[minpos])])
x.rp2<-prune(x.rp,cp=cp0[minpos.first])

# calcola le previsioni e le inserisce nel dataframe
pre<- predict(x.rp2,mat.dati[i+1,] )
valore<-colnames(pre)
sintesi[i-step0+1,2]<-valore[pre==max(pre)][1]
sintesi[i-step0+1,3]<- p.rpart[minpos.first, 2]

livelli<- x.rp2$frame$var[x.rp2$frame$var!="<leaf>"]
```

```
 sintesi[i-step0+1,(4:8)]<-as.character(livelli[1:5])


}

# stampa in C il dataframe ottenuto

 write.table(data.frame(sintesi),file="C:\\mat.txt",row.names=F,quote=F)


# trova le percentuali dei valori esatti della diagonale di sintesi

# ogni 15 gg


nstep<- round(( n-1-step0)/15)

percentualesuccesso<-numeric(length=nstep)

for (v in  0: (nstep-1))  {

 j <-  v*15

diago <- sum(sintesi$valorevero[j:(j+15)]==sintesi$valoreprevisto[j:(j+15)],na.rm = T)

s<- sum(table(sintesi$valorevero[j:(j+15)],sintesi$valoreprevisto[j:(j+15)]))

percentualesuccesso[v+1] <- diago/s

  }


# stampa le la percentuale delle previsioni dei dati delle matrici che si # trovano sulla diagonale

 percentualesuccesso
```

Random Forest:

Creazione: 25.07.2006

File: RF

Finalità: è stata fatta una funzione che permette a RF di trovare la previsioni giorno per giorno dell'ultimo anno anche nel caso in cui non sia disponibile nel dataframe una variabile significativa nel allo scopo di ottenere delle tabelle previsive confrontabili con gli alberi; questi ultimi infatti calcolano automaticamente le previsioni anche in assenza delle variabili significative.

Input: MAT.DATI = MATRICE DI DATI (cairit,osvrit,manrit)
la variabile y : fatrun1, fato3D1, fato3N1

Returns: stampa a video delle percentuali delle previsioni corrette ogni 15 giorni, stampa in C i valori effettivamente rilevati nel 2005 tratti dal dataste e le previsioni effettuate giorno per giorno di quest'ultimo anno: salva i dati in "C:\\mat.txt"

Esempio:

```
# inserire la matrice di dati
mat.dati = manrit

#numero di righe totale
n = nrow ( mat.dati )

# predice l'ultimo anno in base agli altri 5 e alle osservazioni
# mancanti inserite dentro
step0 <- round((5/6)*n)

#crea un dataframe : nella prima colonna si trova il valore
# osservato, nella seconda la previsione
valorevero<- mat.dati$fatrun1[(step0+1):n]
valoreprevisto<-valorevero
valoreprevisto[1:(n-step0)]<-NA
sintesi <- data.frame(valorevero,valoreprevisto)

# stima il modello usando inizialmente step0 dati, prevede step0+1;
# ristima il modello con step0+1 e prevede step0+2
for (i in step0:(n-1))  {
set.seed(123456)
```

```
 # esclude dal calcolo dell'albero le variabili che hanno NA
 # e calcola lo stesso la previsione
 #crea un vettore da 1 a “numero colonne”
 nagiorno <-  (1:45)[is.na(mat.dati[1+i,(1:45)])]
 x.rf<-randomForest(mat.dati$fatrun1[1:i]~.,data=mat.dati[1:i,-c(nagiorno)],
importance=T,do.trace=100,na.action=na.omit)

# usa la matrice di importanza per selezionare le variabili più
# importanti per il modello
 c = (x.rf$importance[,1]) > 0.01

 # ricostruisce il modello usando solo le variabili più importanti
  x.rf2<-randomForest(mat.dati$fatrun1[1:i]~.,data=mat.dati[1:i,names(c[c==T])],
importance=T,do.trace=100,na.action=na.omit)

 # fa le previsioni e le inserisce nel dataframe
 pre<- predict(x.rf2,mat.dati[i+1,] )
 sintesi[i-step0+1,2]<-pre
 sintesi[i-step0+1,(3:7)]<-names(sort(x.rf2$importance[,1],decreasing = T))[1:5]

 }

# stampa in C il dataframe ottenuto
 write.table(data.frame(sintesi),file="C:\\mat.txt",row.names=F,quote=F)
```

```
# trova le percentuali dei valori esatti della diagonale di sintesi ogni
# 15 giorni
nstep<- round(( n-1-step0)/15)
percentualesuccesso<-numeric(length=nstep)
for (v in  0: (nstep-1))  {
j <-  v*15
diago <- sum(sintesi$valorevero[j:(j+15)]==sintesi$valoreprevisto[j:(j+15)],na.rm = T)
s<- sum(table(sintesi$valorevero[j:(j+15)],sintesi$valoreprevisto[j:(j+15)]))
percentualesuccesso[v+1] <- diago/s
}

# stampa le la percentuale delle previsioni dei dati delle matrici che si #
trovano sulla diagonale
percentualesuccesso
```