

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



**Modelli additivi per lo studio della dispersione del PM_{2.5}
nel territorio di Venezia-Mestre e Porto Marghera**

Relatore:

Prof. Guido Masarotto
Dipartimento di Scienze Statistiche

Correlatori:

Prof. Carlo Gaetan
Dott.ssa Eliana Pecorari
Dipartimento di Scienze Ambientali,
Informatica e Statistica

Laureando: Andrea Longo

Matricola N. 1039959

Anno Accademico 2013/2014

Sommario

INDICE DELLE FIGURE	V
INDICE DELLE TABELLE	VII
CAPITOLO 1	1
INTRODUZIONE	1
CAPITOLO 2	5
2.1 I PROBLEMI DELL'INQUINAMENTO	5
2.2 MODELLI DI QUALITÀ DELL'ARIA	7
<i>2.2.1 Modelli fotochimici di dispersione per lo studio del particolato atmosferico (PM)</i>	8
<i>2.2.2 Performance</i>	9
CAPITOLO 3	11
3.1 LA REGRESSIONE LINEARE	11
3.2 LA MODELLAZIONE SEMI-PARAMETRICA	18
<i>3.2.1 Il problema del confondimento</i>	18
<i>3.2.2 I Modelli Additivi</i>	19
<i>3.2.3 Il metodo del lisciamiento o Smoothing</i>	23
<i>3.2.4 Alcune estensioni dei modelli additivi</i>	25
3.3 LA VALUTAZIONE DEL MODELLO	27
<i>3.3.1 Gli indicatori quantitativi</i>	28
<i>3.3.2 Gli indicatori qualitativi</i>	30

3.4 BIAS	32
3.5 IL SOFTWARE R	33
CAPITOLO 4	35
4.1 AREA DI STUDIO	35
4.2 LA STRUTTURA DEI DATI	37
4.3 ANALISI PRELIMINARE	39
4.4 REGRESSIONE LINEARE	53
4.5 MODELLI ADDITIVI	63
4.6 IL CONFRONTO TRA I MODELLI	72
4.7 BIAS	81
CONCLUSIONI	95
ALLEGATI	99
ALLEGATO 1: INDICI PER IL CONFRONTO DEI MODELLI CALCOLATI SULL'INTERO <i>DATASET</i>	99
ALLEGATO 2: INDICI PER IL CONFRONTO DEI MODELLI CALCOLATI SUL TEST SET	101
BIBLIOGRAFIA/SITOGRAFIA	103
SITI CONSULTATI	106
RINGRAZIAMENTI	107

Indice delle Figure

Figura 1: Siti di campionamento e principali stazioni meteorologiche.....	38
Figura 2: Serie storica del PM _{2.5} per i siti di Via Lissa e Via Malcontenta, anno 2009	40
Figura 3: Rosa dei venti rappresentante la velocità del vento in relazione alla direzione, anno 2009, mediana e moda giornaliera	41
Figura 4: Rosa dei venti rappresentante la velocità del vento in relazione alla direzione, rilevazione in suolo ed in quota, anno 2009, mediana e moda giornaliera	42
Figura 5: Serie storiche della velocità del vento, rilevazione in suolo ed in quota, anno 2009, mediana giornaliera	42
Figura 6: Esempio di perturbazione delle correnti d'aria ed il trasporto degli inquinanti in aree urbane ed industriali	43
Figura 7: Temperatura rilevata nelle stazioni di Via Lissa e Via Malcontenta, anno 2009, mediana giornaliera	44
Figura 8: Umidità relativa rilevata per i siti di Via Lissa e Via Malcontenta, anno 2009, mediana giornaliera	45
Figura 9: Pressione atmosferica a terra nei due siti di rilevazione, anno 2009, mediana giornaliera.....	46
Figura 10: Precipitazioni giornaliere nell'area di Venezia-Mestre, anno 2009	47
Figura 11: Radiazione solare nell'area di Venezia-Mestre, anno 2009, massimo giornaliero.....	48
Figura 12: Frequenza annua delle classi di stabilità per i due siti, anno 2009, moda giornaliera	49
Figura 13: Effetti della stabilità atmosferica sulla dispersione verticale di una sorgente in quota	50
Figura 14: Serie storica relativa al gas NO ₂ per il sito di Via Lissa, anno 2009, mediana giornaliera.....	51
Figura 15: Serie storica relativa al gas NO ₂ per il sito di Via Malcontenta, anno 2009, mediana giornaliera	51
Figura 16: Serie storica relativa al gas SO ₂ per il sito di Via Lissa, anno 2009, media giornaliera	52
Figura 17: Grafici diagnostici dei residui dei minimi quadrati ordinari, Via Lissa	55
Figura 18: Grafici diagnostici dei residui dei minimi quadrati ordinari, Via Malcontenta	56
Figura 19: Autocorrelazione dei residui globale e parziale, Via Lissa	58
Figura 20: Autocorrelazione dei residui globale e parziale, Via Malcontenta	58

Figura 21: Grafici dei residui GLS e del termine incorrelato, Via Lissa	61
Figura 22: Grafici dei residui GLS e del termine incorrelato, Via Malcontenta	62
Figura 23: Confronto tra i residui e le variabili esplicative, Via Lissa	64
Figura 24: Confronto tra i residui e le variabili esplicative, Via Malcontenta.....	65
Figura 25: Funzione di lisciamiento stimata per la variabile Precipitazioni giornaliere, Via Lissa	68
Figura 26: Funzione di lisciamiento stimata per la variabile Precipitazioni giornaliere, Via Malcontenta	69
Figura 27: Funzioni di lisciamiento stimate per variabili <i>Time</i> e Temperatura, Via Lissa	70
Figura 28: Funzioni di lisciamiento stimate per variabili <i>Time</i> e Temperatura, Via Malcontenta	71
Figura 29: Diagramma di Taylor per i modelli costruiti sull'intero database, Via Lissa	74
Figura 30: Diagramma di Taylor per i modelli costruiti sull'intero database, Via Malcontenta	75
Figura 31: Diagramma di Taylor per i modelli costruiti sul <i>training test</i> , Via Lissa	79
Figura 32: Diagramma di Taylor per i modelli costruiti sul <i>training test</i> , Via Malcontenta	80
Figura 33: Relazione tra <i>bias</i> e le variabili esplicative, Via Lissa	83
Figura 34: Relazione tra <i>bias</i> e le variabili esplicative, Via Malcontenta	84
Figura 35: Residui del modello AM con outlier, Via Lissa.....	86
Figura 36: Residui del modello AM con outlier, Via Malcontenta.....	87
Figura 37: Distanza di Cook, Via Lissa	88
Figura 38: Distanza di Cook, Via Malcontenta	89
Figura 39: Grafici dei residui del modello GAM per <i>bias</i> , Via Lissa.....	90
Figura 40: Grafici dei residui del modello GAM per <i>bias</i> , Via Malcontenta.....	91
Figura 41: Stima delle funzioni di lisciamiento, <i>bias</i> Via Lissa.....	93
Figura 42: Stima della funzione di lisciamiento per l' SO_2 , <i>bias</i> Via Malcontenta	94

Indice delle Tabelle

Tabella 1: Stima dei minimi quadrati ordinari della log dispersione del $PM_{2.5}$	53
Tabella 2: Test F e coefficiente di determinazione lineare semplice e corretto	54
Tabella 3: Test D-W per l'autocorrelazione dei residui del modello lineare, Via Lissa.....	57
Tabella 4: Test D-W per l'autocorrelazione dei residui del modello lineare, Via Malcontenta.....	57
Tabella 5: Schema riassuntivo delle stime GLS per i siti di Via Lissa e Via Malcontenta	59
Tabella 6: Stime dei parametri autoregressivi per Via Lissa e Via Malcontenta	60
Tabella 7: Stime del modello additivo e additivo misto per Via Lissa.....	67
Tabella 8: Stime del modello additivo e additivo misto per Via Malcontenta.....	67
Tabella 9: Indici di performance per i modelli, Via Lissa	72
Tabella 10: Indici di performance per i modelli, Via Malcontenta	72
Tabella 11: Confronto MFB per Via Lissa	73
Tabella 12: Confronto MFB per Via Malcontenta	73
Tabella 13: Indici di performance per i modelli, Via Lissa	77
Tabella 14: Indici di performance per i modelli, Via Malcontenta	77
Tabella 15: Confronto MFB per Via Lissa	78
Tabella 16: Confronto MFB per Via Malcontenta	78
Tabella 17: Stime modello GAM per <i>bias</i> , Via Lissa e Via Malcontenta	92
Tabella 18: Coefficienti di correlazione di Spearman per Via Lissa e Via Malcontenta	99
Tabella 19: <i>Mean fractional bias</i> per Via Lissa e Via Malcontenta	99
Tabella 20: <i>Root mean square error</i> per Via Lissa e Via Malcontenta.....	100
Tabella 21: Normalized mean square error per Via Lissa e Via Malcontenta ...	100
Tabella 22: Coefficienti di correlazione di Spearman per Via Lissa e Via Malcontenta	101
Tabella 23: <i>Mean fractional bias</i> per Via Lissa e Via Malcontenta	101
Tabella 24: <i>Root mean square error</i> per Via Lissa e Via Malcontenta.....	102
Tabella 25: Normalized mean square error per Via Lissa e Via Malcontenta ...	102

Capitolo 1

Introduzione

Negli ultimi anni, l'attenzione della comunità scientifica si è sempre più rivolta verso il problema dell'inquinamento a causa degli effetti sulla salute umana. Molte sono le discipline che si adoperano all'interno di questo ambito, dove la complessità del problema auspica al coinvolgimento di diverse professionalità, in un contesto multidisciplinare nel quale convivono analisi chimico-fisiche, modelli statistico-matematici e valutazioni clinico-epidemiologiche.

L'utilizzo della statistica per lo studio dell'inquinamento atmosferico è prevalentemente legato alla valutazione di modelli di regressione e di analisi delle variabili misurate. In questo ambito è utilizzata la statistica più avanzata. Al contrario una statistica base è ancora in uso per quanto concerne la valutazione di performance di modelli matematici atti alla simulazione della dispersione di inquinanti.

Il più delle volte si utilizzano degli indicatori sintetici di distorsione illustrati in letteratura, come quelli suggeriti dall'EPA (2007), senza però analizzare le cause di una cattiva stima mediante modelli statistici complessi. Infatti, questi indicatori si rivolgono prevalentemente allo studio di una sola variabile alla volta, quando la natura generatrice dei dati coinvolge invece un'infinità di

processi chimico-fisici strettamente correlati tra loro. La difficoltà di approfondire questa tematica, come per altri aspetti connessi allo studio delle scienze ambientali, nasce dalla difficoltà di integrare competenze complementari. Nonostante gli studi fatti che ne attestano la reale necessità, risulta ancora difficile, specialmente nel nostro contesto geografico, l'applicazione di una visione più ampia che abbraccia più discipline.

Questo lavoro nasce con lo scopo di supportare l'utilizzo di un modello fotochimico per la simulazione della formazione, del trasporto e della trasformazione del particolato atmosferico nella zona orientale della pianura padano-veneta. Saranno due i siti scelti per il rilevamento ed il confronto dei risultati: il primo urbano (situato in Via Lissa), il secondo industriale (in Via Malcontenta).

L'obiettivo della tesi è quello di utilizzare la famiglia dei modelli additivi per individuare quei fattori che risultano essere correlati alla concentrazione del particolato rilevato nei due siti. Si vuole inoltre esaminare la capacità prognostica degli stessi modelli, comparandola con la simulazione del modello fotochimico, ed apprendere inoltre quali possano essere le cause della distorsione di queste ultime.

Dopo un primo *excursus* atto a presentare il fenomeno dell'inquinamento dell'aria ed i principali modelli matematici impiegati in letteratura (Capitolo 2), si presentano, nel Capitolo 3, gli strumenti utilizzati in questo lavoro.

Nel Capitolo 4 vengono riportati i risultati e lo sviluppo dell'analisi. Nello specifico, si adopereranno i modelli di regressione lineare ed i modelli additivi, per evidenziare le possibili relazioni tra alcuni fattori meteorologici e la dispersione del $PM_{2.5}$. In seguito si andranno a valutare la capacità previsiva di questi, confrontandoli con il modello fotochimico attraverso alcuni indicatori statistici. Infine si implementerà un'analisi sulla distorsione tra i valori osservati e quelli previsti dal modello matematico, per far emergere

quei fattori che determinano una differenza tra i valori predetti e quelli osservati.

Capitolo 2

2.1 I problemi dell'inquinamento

Negli ultimi anni si è diffusa sempre più la consapevolezza del problema dell'inquinamento atmosferico, definito come "l'alterazione della normale composizione chimica dell'aria, dovuta alla presenza di sostanze in quantità e con caratteristiche tali da alterare le normali condizioni di salubrità" (ARPAV 2013). Le più diffuse in atmosfera risultano essere: il biossido di zolfo (SO_2), gli ossidi di azoto (NO_x), il monossido di carbonio (CO), l'ozono, il benzene, gli idrocarburi policiclici aromatici (IPA), i composti organici volatili (COV) e le polveri, soprattutto il particolato di diametro aerodinamico inferiore a 10 milionesimi di metro (PM_{10}).

La recente attenzione della comunità scientifica per i livelli dell'inquinamento da particolato atmosferico è oltretutto motivata dal fatto che esiste una dipendenza significativa tra i livelli di concentrazione delle particelle fini (PM_{10} e $\text{PM}_{2.5}$) e l'incremento del rischio di morte (Schwartz, 2002; Englert, 2004). Tale problema, quindi, si concentra soprattutto nelle aree metropolitane, dove il traffico, gli impianti industriali e il riscaldamento degli edifici hanno un forte impatto sulla qualità dell'aria (Agostini, 2012).

L'inquinamento atmosferico non è la causa di una malattia specifica, tuttavia può contribuire ad una vasta gamma di processi multi-causali. Numerosi studi scientifici, tra cui si ricordano Kunzli (2000) ed Englert (2004), hanno infatti correlato

l'esposizione all'inquinamento da particolato ad una varietà di problemi, tra cui:

- morte prematura nelle persone con malattie cardiache o polmonari;
- attacchi cardiaci non fatali;
- battito cardiaco irregolare;
- asma aggravata;
- riduzione della funzione polmonare;
- aumento dei sintomi respiratori, come irritazione alle vie respiratorie, tosse o difficoltà respiratorie.

Il particolato può essere distinto sia in funzione delle particelle che lo compongono, sia in base ai processi che lo hanno generato; è quindi possibile distinguere un particolato di origine primaria e un particolato di origine secondaria.

Il particolato primario è costituito da particelle originatesi direttamente da processi meccanici di erosione, dilavamento e rottura di particelle più grandi, da processi di evaporazione dello spray marino in prossimità delle coste e da processi di combustione. Viene emesso in atmosfera direttamente nella sua forma finale da sorgenti identificabili ed è generalmente più concentrato nell'aria immediatamente circostante il suo punto di emissione.

Al contrario, il particolato secondario è costituito dagli aerosol, contenenti quasi esclusivamente particelle fini che si generano dalla mutazione dei gas in particelle solide anche attraverso reazioni chimiche tra gli inquinanti primari presenti in atmosfera (Molinaroli e Masiol, 2006). Un ruolo importante è pertanto assunto dalle condizioni meteorologiche di contesto, che influiscono più o meno fortemente sulla formazione degli inquinanti secondari (per un approfondimento sintetico ma esaustivo dei processi di formazione e la composizione chimica del particolato atmosferico si veda: Agostini C., 2012).

2.2 Modelli di qualità dell'aria

Lo studio dell'inquinamento atmosferico, da come esso si sia formato alla sua dispersione, necessita di molte informazioni riguardanti la descrizione meteorologica e geografica dell'area di studio, le fonti di emissione e gli aspetti chimico-fisici coinvolti. Lo studio della qualità dell'aria è un fenomeno complesso e l'uso della modellizzazione matematica rappresenta uno strumento fondamentale sia per la ricerca che per il supporto delle politiche decisionali.

Questi modelli vengono ampiamente utilizzati sia per controllare l'inquinamento dell'aria, ma anche per identificare i contributi delle varie sorgenti emmissive e contribuire al disegno di strategie efficaci per la riduzione degli inquinanti.

Un tale sistema di modellizzazione è, di fatto, complesso e si compone di modelli secondari, necessari per creare l'input del modello chimico-fisico principale. Si devono avere a disposizione: un inventario delle emissioni adeguato; la simulazione degli eventi meteorologici; la simulazione della dispersione e del trasporto delle sostanze considerate.

I modelli per la qualità dell'aria utilizzano tecniche matematiche e algoritmi numerici per simulare i processi chimico-fisico che influenzano gli inquinanti, come questi si disperdono e le loro reazioni in atmosfera. Sulla base di input meteorologici ed informazioni sulle sorgenti di emissione, questi modelli sono progettati anche per identificare gli inquinanti primari che vengono emessi direttamente in atmosfera e, in alcuni casi, gli inquinanti secondari che si formano a seguito di reazioni chimiche complesse all'interno della stessa.

Generalmente, i modelli più utilizzati sono:

- Modelli di dispersione: tipicamente utilizzati nel processo di autorizzazione, per stimare la concentrazione di inquinanti al

suolo attorno ad una specifica sorgente di emissioni.

- Modelli foto-chimici: utilizzati nelle valutazioni di norme o di politiche per simulare gli impatti di tutte le sorgenti attraverso la stima delle concentrazioni di inquinanti e della deposizione su grandi scale spaziali.
- Modelli recettori: tecniche di osservazione che utilizzano le caratteristiche chimico-fisiche dei gas e delle particelle misurate alla sorgente, per identificare sia la presenza che misurare i contributi delle varie sorgenti nelle concentrazioni di inquinamento.

Le scale spaziali di questi modelli variano molto, si passa da un minimo di pochi chilometri (per le fonti puntuali industriali), a 100km (per le singole aree urbane), a qualche migliaia di chilometri (per le più grandi aree regionali). Quando si definisce un dominio, la scala spaziale degli importanti fenomeni atmosferici che influenzano il problema della qualità dell'aria deve essere analizzata accuratamente.

2.2.1 Modelli fotochimici di dispersione per lo studio del particolato atmosferico (PM)

Nella modellizzazione della qualità dell'aria, la simulazione del trasporto e delle reazioni chimico-fisiche del PM, assume un ruolo cruciale. La criticità principale nasce dalla reale capacità di collegare le emissioni del PM, ed i suoi precursori, alle concentrazioni osservate in atmosfera e ad altre proprietà importanti.

I principali modelli per la qualità dell'aria, utilizzati nello studio della distribuzione del particolato, sono generalmente chiamati *Chemical-Transport Models* (CTM). Questi modelli descrivono, attraverso rappresentazioni matematiche, entrambi i processi, fisici

e chimici, che si sviluppano in atmosfera. La risoluzione mediante algoritmi numerici permette di ottenere concentrazioni di inquinanti in funzione dello spazio e del tempo per un dato insieme di emissioni e condizioni meteorologiche. Sono pertanto modelli prognostici.

Sebbene la maggior parte dei modelli attuali tenda a trattare gli stessi principali processi, si trovano significative differenze nella caratterizzazione della composizione chimica del PM e della distribuzione granulometrica.

I *Chemical-Transport Models* (CTM) sviluppati per il PM, sono più complessi rispetto ai modelli di dispersione semplici, generalmente utilizzati per studiare l'andamento dei gas, e più completi essendo in grado di prevedere anche questi ultimi. La complessità matematica, tuttavia, richiede costi computazionali non indifferenti. Possono essere indicati, per questo motivo, anche come “*one atmosphere*”, “multi-inquinante”, o modelli di qualità dell'aria “unificata” (Seigneur e Moran, 2004).

Nello specifico, il modello fotochimico considerato è il modello FARM (*Flexible Air quality Regional Model*).

2.2.2 Performance

Generalmente, le prestazioni di un sistema di modellizzazione comprendono metriche statistiche calcolate rispetto alle concentrazioni medie (orarie o giornaliere a seconda del dato osservato disponibile). Nell'ambito della qualità dell'aria, l'EPA (2007) suggerisce parametri semplici di cui Boylan e Russell (2006) definiscono degli intervalli di confidenza empiricamente testati. Una descrizione più approfondita di tali indicatori verrà, tuttavia, presentata nel paragrafo 3.3.

Il limite comune di questi indicatori risiede nel fatto che si rivolgono prevalentemente all'esame di una sola variabile, lasciando all'analista il compito, non privo di rischi, di estendere i risultati a relazioni che debbono invece comprendere più variabili. Inoltre, nel caso di cattiva *performance*, non aiutano a prendere in considerazione una particolare evoluzione dell'analisi, né a cogliere eventuali fattori che possono causare l'aumento della distorsione.

Capitolo 3

La reale complessità del problema, assieme alle innumerevoli variabili che possono giocare ruoli decisivi nei fenomeni di dispersione degli inquinanti in atmosfera, suggerisce l'uso di metodi e tecniche flessibili, trattandosi di fenomeni dove o non è possibile conoscerne il vero processo generatore o non si è in grado di stimarlo correttamente.

Il proseguo di questo capitolo vuole presentare alcuni modelli statistici, utilizzati per evidenziare quelle relazioni esistenti tra le variabili esplicative e la dispersione del $PM_{2.5}$ (in due siti di campionamento), aventi assunzioni più o meno forti. In seguito si presenteranno degli strumenti per supportare il confronto della performance del "rigido" modello matematico con quella di un modello statistico, più elastico. Infine verrà introdotto il software utilizzato per il lavoro di analisi, R.

3.1 La regressione lineare

Il primo modello presentato, in grado di esprimere la relazione esistente tra la variabile di risposta (concentrazione del $PM_{2.5}$) e le variabili esplicative (meteo e gas), è il modello di regressione lineare multipla, la cui formulazione risulta:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_p + \varepsilon$$

con y la variabile dipendente, x_i la i -esima variabile esplicativa, β_i il suo coefficiente ed ε il termine d'errore.

Questo modello, grazie alla sua facilità di interpretazione e di stima dei parametri incogniti β_i , viene comunemente applicato in molti studi in cui si ricerca una relazione funzionale tra variabili, rendendolo tra gli strumenti statistici più diffusi. Il suo utilizzo porta tuttavia a dover fare delle assunzioni, più o meno forti, ed in primis il fatto che la variabile risposta y derivi dalla somma di una componente sistematica che è una funzione lineare nei parametri:

$$r(x_1, \dots, x_p) = \alpha + \beta_1 x_1 + \dots + \beta_p x_p$$

ed una componente accidentale, rappresentata dal termine d'errore.

Supponendo di avere n osservazioni prodotte dal modello (y_1, \dots, y_n) , viste ognuna come determinazione di n variabili casuali (Y_1, \dots, Y_n) , utilizzando la formulazione matriciale si può scrivere:

$$Y = X\underline{\beta} + \underline{\varepsilon}$$

dove: $Y = (Y_1, \dots, Y_n)^T$ è il vettore contenente le n variabili risposta; $X = [x_{ij}]$ è la matrice di regressione, di dimensione $n \times p$, e contiene i valori delle variabile esplicative; $\underline{\beta} = (\beta_1, \dots, \beta_p)^T$ è il vettore dei parametri di regressione; infine $\underline{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ è il vettore delle n variabili casuali errore.

I coefficienti, in questo contesto, possono essere stimati tramite il criterio dei minimi quadrati ordinari, assumendo che:

- i regressori siano non stocastici e la matrice X sia di rango pieno;
- il termine d'errore non sia osservabile;
- il termine d'errore abbia media nulla e varianza costante (omoschedasticità ed incorrelazione).

Si tratterà di scegliere $\underline{\beta}$ in modo da minimizzare il quadrato della distanza euclidea tra il vettore osservato ed il suo valore atteso previsto:

$$\|y - X\underline{\beta}\|^2$$

Nel caso di corretta specificazione del modello e delle assunzioni fatte, lo stimatore dei minimi quadrati ordinari risulta essere corretto ed efficiente. Per l'analisi inferenziale, basata sulla teoria della verosimiglianza, è necessario introdurre un'ulteriore ipotesi riferita alla distribuzione di $\underline{\varepsilon}$. Nei modelli classici di regressione lineare si assume pertanto che l'errore sia distribuito normalmente, ossia:

$$\underline{\varepsilon} \sim N_n(0, \sigma^2 I_n)$$

con σ^2 positivo non noto, rappresentante la varianza di ogni variabile casuale ε_i , e stimabile in maniera non distorta dalla varianza residua corretta. Vista la semplicità interpretativa e la diffusione del modello, non si ritiene necessario approfondire la base teorica su cui esso è stato sviluppato, di seguito vengono tuttavia presentate alcune caratteristiche utili nelle analisi successive. La variabile dipendente verrà, nell'analisi, trasformata nel suo logaritmo a causa di una forma di eteroschedasticità presente in entrambe le realtà, urbana ed industriale. Tale trasformazione risulterà sufficiente per soddisfare le ipotesi di specificazione precedentemente descritte e non comporterà l'utilizzo di ulteriori strumenti per risolvere tale problema. Le variabili esplicative, sulle quali verrà regredito il logaritmo della concentrazione del PM_{2.5}, saranno:

- Direzione del vento
- Velocità del vento
- Temperatura
- Umidità
- Pressione
- Precipitazioni giornaliere
- Radiazione Solare Massima
- Classe di stabilità
- Diossido di azoto
- Diossido di zolfo

Quest'ultima osservata per il solo sito urbano di Via Lissa.

In questo contesto, la stima del parametro fornisce, a parità di tutti gli altri fattori, la variazione nel logaritmo del $PM_{2.5}$ quando il valore dell'esplicativa aumenta di una unità; $\Delta \log PM_{2.5} | \Delta x_i = 1$. Tuttavia per un'interpretazione più agevole è comune, per piccoli cambiamenti in x_i , l'approssimazione:

$$\Delta \% PM_{2.5} \cong (100 * \beta_i) \Delta x_i$$

dimostrabile dal fatto che $\log(1 + x_i) \cong x_i$ con $x_i \cong 0$.

In ogni modo, se si volesse invece ricavare la variazione esatta, è possibile provare che:

$$\Delta \% PM_{2.5} = 100 * (\exp \{ \beta_i \Delta x_i \} - 1)$$

La regressione, così sviluppata, offre in R un output completo. Per ogni variabile esplicativa (e intercetta) è presente la stima del parametro ad essa associato, una stima della deviazione standard dello stimatore, il valore della statistica test nell'ipotesi di nullità di tale parametro con il suo livello di significatività osservato. Nella parte superiore si trovano delle informazioni per la simmetria ed il *range* dei residui e, sotto alle stime, degli indicatori sintetici per la valutazione della bontà di adattamento del modello.

La diagnostica dei residui risulta, di conseguenza, necessaria per la verifica delle assunzioni fatte e abbraccia diverse possibilità, alcune di tipo esplorativo basate sulla costruzione di opportuni grafici, altre affidate all'uso di test specifici.

Per supportare tale diagnostica, vengono eseguiti dei test per la convalida delle ipotesi di linearità del modello, attraverso il test RESET (*REgression Specification Error Test*) e le valutazioni disgiunte di asimmetria, curtosi ed eteroschedasticità.

Con dati seriali può risultare ulteriormente critica la situazione di indipendenza degli errori. Per questo motivo, assieme ai grafici delle autocorrelazioni globali e parziali, si utilizza nell'analisi il test

di *Durbin Watson*, specifico per saggiare l'ipotesi di autocorrelazione.

Nello specifico, il test DW si usa nei modelli di regressione dei minimi quadrati per verificare l'ipotesi nulla di indipendenza tra i residui del modello contro l'ipotesi alternativa di autocorrelazione del primo ordine.

$$H_0 : \rho_1 = 0$$

$$H_1 : \rho_1 > 0$$

con $\rho_1 = \text{Corr}[\varepsilon_t, \varepsilon_{t-1}]$ dove ε_t è l'errore al tempo t . Il test assume di conseguenza la seguente formulazione:

$$DW = \frac{\sum_{t=2}^n (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^n \varepsilon_t^2}$$

ed è compreso tra 0 e 4. Con perfetta correlazione positiva, DW sarebbe pari a 0; con correlazione negativa, DW sarebbe pari a 4; se non ci fosse correlazione, DW sarebbe pari a 2.

Nello specifico, valori piccoli della statistica indicano che i residui successivi sono, in media, vicini in valore l'uno all'altro (ossia correlati positivamente), viceversa valori grandi denotano che i residui successivi sono, in media, differenti l'uno dall'altro, quindi correlati negativamente. La distribuzione della statistica di Durbin-Watson non è nota, ma gli stessi autori hanno tabulato i valori critici attraverso delle simulazioni condotte con il metodo Monte Carlo. Per verificare la presenza di autocorrelazione al livello di significatività α , la statistica test DW viene pertanto confrontata con dei valori critici inferiori e superiori ($d_{L\alpha}$ e $d_{U\alpha}$), nel modo seguente:

- $DW < d_{L\alpha} \rightarrow$ rifiuta H_0 (autocorrelazione positiva)
- $DW > d_{U\alpha} \rightarrow$ non si rifiuta H_0 (non autocorrelazione positiva)
- $d_{L\alpha} < DW < d_{U\alpha} \rightarrow$ test non è conclusivo

Per la verifica dell'ipotesi di autocorrelazione negativa, vengono confrontati gli stessi valori critici con la trasformata $4 - DW$.

Al fine di questa analisi, si suppone che l'errore del modello lineare sia generato da un processo autoregressivo di primo ordine, AR(1), di formulazione:

$$\varepsilon_t = \varphi \varepsilon_{t-1} + u_t$$

con $u_t \sim IIN(0, \sigma_u^2)$ rumore bianco e φ parametro autoregressivo, che per la stazionarietà del modello viene ipotizzato essere: $|\varphi| < 1$ (per ulteriori approfondimenti si veda: Hamilton, 1994).

L'autocovarianza del processo AR(1) è data da:

$$\gamma_k = \begin{cases} \frac{\sigma_u^2}{1 - \varphi^2} & k = 0 \\ \varphi^k \gamma_0 & k > 0 \end{cases}$$

con k il ritardo, e si ricava moltiplicando il modello degli errori, ambo i lati, per ε_{t-k} e calcolandone successivamente il valore atteso.

L'autocorrelazione, sempre a ritardo k , deriva direttamente dalla funzione di autocovarianza:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \varphi^k$$

Infine la funzione di autocorrelazione parziale, per completezza, è:

$$P_k = \begin{cases} \varphi^k & k = 1 \\ 0 & k > 1 \end{cases}$$

Il coefficiente di autocorrezione ad un ritardo dei residui è pertanto la stima del parametro autoregressivo:

$$\hat{\rho}_1 = \text{Corr}[e_t, e_{t-1}] = \hat{\varphi}$$

Per verificare infine se il vero parametro ρ_k della serie storica degli errori sia effettivamente diverso da zero, valore di un'eventuale processo *white noise*, si può utilizzare la sua stima $\hat{\rho}_k$.

La varianza di un rumore bianco è infatti circa $1/n$ per ogni ritardo k , sotto le ipotesi di normalità si ricava che la regione di accettazione di livello di significatività al 5% è data dall'intervallo:

$$[(-1.96) / \sqrt{n} , 1.96 / \sqrt{n}]$$

un valore di $\hat{\rho}_k$ fuori dall'intervallo, porta a ritenere ρ_k significativamente diverso da zero (Cryer J. D., Chan K. S., 2008).

Nel modello lineare per la dispersione del $PM_{2.5}$, l'autocorrelazione dei residui porta, come diretta conseguenza, al fatto che la matrice di varianza – covarianza degli stessi non assuma più la struttura diagonale, bensì sia del tipo:

$$V[\varepsilon|X] = \frac{\sigma_u^2}{1 - \varphi^2} \begin{bmatrix} 1 & \varphi & \dots & \varphi^{T-2} & \varphi^{T-1} \\ \varphi & 1 & \dots & \varphi^{T-3} & \varphi^{T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \varphi^{T-2} & \varphi^{T-3} & \dots & 1 & \varphi \\ \varphi^{T-1} & \varphi^{T-2} & \dots & \varphi & 1 \end{bmatrix} = \sigma_u^2 \Psi$$

Lo stimatore dei minimi quadrati ordinari risulta ancora non distorto, ma non efficiente; le procedure standard di verifica d'ipotesi basate sui test t ed F non sono più valide.

È quindi necessario, in questo contesto, rivedere le ipotesi di specificazione del modello e presentare la stima dei minimi quadrati generalizzati o, più correttamente, la loro versione calcolabile (in quanto la matrice Ψ non è nota). Lo stimatore FGLS (*Feasible Generalized Least Squares*) è asintoticamente più efficiente dello stimatore OLS:

$$\hat{\beta}_{FGLS} = (X' \hat{\Psi}^{-1} X)^{-1} X' \hat{\Psi}^{-1} y$$

con $\hat{\Psi}$ stima consistente di Ψ .

3.2 La modellazione semi-parametrica

Negli ultimi anni, ai modelli di regressione standard, usati nella modellazione statistica delle serie temporali relative all'inquinamento atmosferico (Smith et al., 2000), sono stati affiancati con enfasi sempre più crescente, approcci e formulazioni semi – parametriche.

In questi ultimi modelli, è consuetudine trovare dei termini lineari corrispondenti agli effetti dei singoli inquinanti, che sono i reali componenti di interesse dell'analisi, ed altri termini, detti “di lisciamiento”, rappresentati solitamente da *smoother* (lisciatori) del tempo e della temperatura, variabili considerate “rumore” (si vedano i lavori di: Speckman, 1988; Hastie and Tibshirani, 1990).

Rispetto all'usuale specificazione, totalmente parametrica, questi modelli dimostrano una maggiore flessibilità nel controllo dei fattori confondenti non lineari, quali stagionalità e trend, e nel controllo delle variabili meteorologiche.

Per meglio comprendere la metodologia applicata in questo lavoro, si ritiene opportuno introdurre dapprima il concetto di *confondimento*, necessario per giustificare e comprendere l'uso dei modelli additivi, per poi passare alla descrizione degli stessi, il metodo del lisciamiento e la successiva estensione dei modelli additivi generalizzati misti.

3.2.1 Il problema del confondimento

Nello studio della dispersione del particolato atmosferico, così come in molti altri casi soprattutto di tipo epidemiologico, è necessario controllare la presenza di fattori confondenti che possono danneggiare lo studio.

Con il termine *confondimento* si fa riferimento al caso in cui la misura di associazione tra un'esposizione e l'esito è “confusa”

dall'effetto di un'altro fattore. Si dice pertanto che la stima di effetto è affetta da distorsione (Agabiti et al., 2011).

La maggior parte delle variabili che si riferiscono al tempo atmosferico, all'inquinamento e alla salute, presentano una variazione sistematica naturale durante l'anno. Questo impegna il modellista a rimuovere l'effetto confondente dovuto a questi andamenti per poter evidenziare l'effettivo rapporto tra le variabili. Risulta quindi necessario, nel nostro caso, separare gli andamenti giornalieri e stagionali naturali, come l'altezza dello strato di rimescolamento (*Planetary Boundary Layer*), la pressione atmosferica, l'umidità relativa, e le altre variabili meteorologiche che possono considerarsi di disturbo, dall'effetto degli inquinanti.

Il metodo più usato per affrontare questo problema, come si è visto, è quello di introdurre nel modello additivo funzioni di liscio (*smoothing*) che tengano conto di tali situazioni confondenti e ne riducano la variabilità; lo scarto viene utilizzato per stimare l'effetto distinto dell'inquinamento.

Uno dei maggiori problemi di questo approccio nasce dal riconoscere e conseguentemente inserire correttamente le funzioni di *smoothing*; capire di conseguenza quale grado di liscio è più coerente per poi modellare nel modo migliore i residui, un problema che verrà affrontato nel seguito. Altri elementi, che possono giocare un ruolo di confondente, sono quelli associati a specifici giorni di calendario e coincidenti con feste o vacanze. Alcune strutture produttive e di servizio hanno infatti abitudini differenziate durante i giorni o le festività.

3.2.2 I Modelli Additivi

I modelli additivi possono essere considerati come una generalizzazione dei modelli lineari, in cui la funzione di

regressione lineare lascia il posto alla somma degli effetti che delle specifiche funzioni di lisciamento, diverse per ciascuna esplicativa, apportano alla variabile risposta.

Un primo esempio chiarificatore nasce dal problema dell'andamento di una qualsivoglia struttura di dati (in questo caso la dispersione del PM_{2.5}) in relazione a due fattori, che a titolo esemplificativo sono denominati x e z . Per semplicità si suppone non esserci altre variabili in gioco. Un possibile modello, atto a focalizzare la struttura del nuovo approccio, è ragionevolmente del tipo:

$$y = f_1(x) + f_2(z) + \varepsilon$$

dove y è la variabile risposta della dispersione, $f_1(x)$ rappresenta il contributo dato dalla prima variabile esplicativa e $f_2(z)$ il contributo dato dalla seconda variabile esplicativa. Il termine d'errore, che solitamente si considera indipendente e identicamente distribuito con media zero e varianza σ^2 , è indicato con ε . Le f_j sono funzioni arbitrarie, solitamente una per ogni variabile esplicativa, ma possono dipendere anche da due o più variabili; in questo caso le componenti sono funzioni multidimensionali.

In queste tipologie di modello, utili per studiare casi come la velocità del vento, il monitoraggio del livello di inquinamento e le sue conseguenze, l'ottimizzazione della produzione di alcune realtà, ecc., è necessario specificare tali funzioni arbitrarie poiché il vero processo generatore dei dati non è conosciuto; se questo meccanismo fosse noto, si potrebbe istituire un modello parametrico ben definito, il modello sarebbe correttamente specificato e la distorsione praticamente nulla.

Il contesto in cui tali modelli operano, incentiva a procedere sostanzialmente secondo un approccio non parametrico, tuttavia la scelta dei modelli additivi cerca di evitare una serie di problemi propri dell'utilizzo di questi stimatori non parametrici come la

maledizione della dimensionalità (Hastie, Tibshirani e Friedman 2001). Ipotizzando infatti tra la variabile risposta e le variabili esplicative la singola relazione:

$$Y = f(\underline{x}) + \varepsilon$$

con f funzione non nota definita in $D \subset \mathbb{R}^d$ ed ε il consueto termine d'errore, le stime non parametriche si basano su medie locali d-dimensionali; preso un valore $\underline{x}_0 \in D$, viene comunemente stimato $f(\underline{x}_0)$ attraverso una media ponderata delle y osservate in un intorno di \underline{x}_0 . A causa della dimensionalità, se nel caso unidimensionale un intervallo di variazione viene suddiviso in m intervalli richiedendo la stima di m parametri, nel caso bidimensionale il campo di variazione della coppia di variabili viene diviso in m^2 rettangoli (si necessita quindi la stima di m^2 parametri), nel caso generale d-dimensionale, il campo di variazione delle d variabili viene suddiviso in m^d rettangoloidi costringendo ad una stima di m^d parametri (Bellman, 1961). Inoltre, all'aumentare della dimensione del numero di predittori, diventa esponenzialmente più difficile trovare i punti di ottimo globale per la stima dei parametri. Ad esempio una numerosità di 100 osservazioni copre relativamente bene l'intervallo unidimensionale $[0,1]$ sulla retta reale, permettendo di fare dell'inferenza su tali dati. Se si considera invece un ipercubo di 10 dimensioni, le 100 osservazioni si presenteranno come punti isolati in uno spazio vasto e vuoto. Infatti per ottenere una copertura simile a quella osservata sullo spazio unidimensionale sarebbero necessarie 10^{20} osservazioni (Bellman, 1961). Infine nel caso multidimensionale si ha la necessità di sintetizzare gli elementi in metriche e indici, spesso di difficile interpretazione, in particolare quando le variabili sono misurate in diverse unità di misura o sono altamente correlate.

Dopo questa premessa, la definizione generale di modello additivo viene definita, da Hastie e Tibshirani (1990), nel modo seguente.

Sia \underline{Y} il vettore casuale delle risposte (Y_1, \dots, Y_T) , α un termine costante che rappresenta l'intercetta del modello, f_j funzioni non note (stimate attraverso una procedura di lisciamento), X la matrice contenente le p variabili esplicative osservate ed ε il vettore dei termini d'errore. La formulazione di un modello additivo (AM) è del tipo:

$$Y_i = \alpha + \sum_{j=1}^p f_j(x_{ij}) + \varepsilon_i$$

dove, come nei modelli di regressione lineare, si assume che:

1. gli errori siano indipendenti dalle variabile esplicative;
2. $E[\varepsilon_i] = 0$;
3. $Var[\varepsilon_i] = \sigma^2$;
4. f_j sia funzione della j -sima variabile esplicativa.

Il modello additivo presentato pone tuttavia un vincolo a priori molto importante per l'analisi dei dati: ogni variabile viene rappresentata separatamente. Questa peculiarità mantiene i vantaggi della modellizzazione tramite la regressione lineare senza però dover sottostare a vincoli eccessivamente restrittivi sulla forma di dipendenza funzionale della risposta dai predittori. Una volta adattato il modello additivo ai dati, si possono in questo modo disegnare separatamente le p funzioni relative alle diverse variabili per esaminare separatamente l'apporto delle stesse nel descrivere la risposta.

Anche le funzioni stimate possono essere considerate l'analogo dei parametri in un modello di regressione lineare; ciò nonostante, come si è già accennato, per stimare tali funzioni vengono usati metodi non parametrici. Tali stimatori si determinano nel caso semplice di regressione con una sola variabile esplicativa e cercano di stimare al meglio la funzione di regressione senza dover effettuare scelte pesanti sulla forma della stessa.

Dopo aver stimato queste funzioni, può essere utile conoscere quanti sono i *gradi di libertà* utilizzati, in una nozione che deriva alla regressione lineare parametrica (Hastie e Tibshirani, 1990). Supponendo di scrivere lo stimatore come applicazione lineare definita da $\hat{Y} = SY$, la matrice lisciante S dipende dalle variabili esplicative e dalla scelta dello stimatore, ma non da Y (Hastie e Tibshirani, 1990).

In letteratura si trovano almeno tre possibili definizioni di *gradi di libertà*, che dipendono dal contesto in cui vengono usate, derivate per analogia dal modello di regressione lineare:

- $tr(SS^T)$
- $tr(2S - S^T S)$
- $tr(S)$

Per stimatori simmetrici con autovalori della matrice di lisciamiento non negativi i tre modi portano a risultati asintoticamente equivalenti. Inoltre, che se S è una matrice di proiezione simmetrica, i tre metodi coincidono.

3.2.3 Il metodo del lisciamiento o *Smoothing*

Lo *smoother*, o “lisciatore”, è un oggetto matematico utile al fine di riassumere l’andamento delle misure di una variabile risposta Y attraverso una funzione delle misure di uno o più predittori X_1, \dots, X_p . Esso produce una stima di tale andamento che è meno variabile di quanto lo sia Y stessa, di qui il nome di *smoother*.

Un’importante caratteristica è la sua natura non parametrica: esso non assume, infatti, nessuna forma rigida di dipendenza tra Y e X_1, \dots, X_p , e per questo motivo ci si riferisce spesso ai modelli che produce con il termine di regressione non parametrica. L’esito della procedura di *smoothing* nel caso in cui vi sia un singolo predittore è detto *scatterplot smoothing*.

Molte le proposte per stimatori di questo tipo come le *medie mobili*, *kernel smoothers*, *smoothing spline* e così via. In questo studio si ritiene esaustivo presentare quest'ultima tipologia di stimatori in quanto quella utilizzata nell'analisi.

Spline di regressione. Il termine *spline* deriva dalle aste in legno utilizzate nella costruzione degli scafi delle navi. Fissati alcuni punti sulla sezione trasversale, il resto della curva si otteneva forzando le aste a passare per tali punti e lasciandola libera di disporsi per il resto del profilo secondo la sua naturale tendenza. Questo approccio alternativo, derivante dalla logica del meccanismo appena spiegato, è quello di approssimare delle funzioni in cui si conosce solo il valore in alcuni punti che si andranno ad interpolare. Si creano così delle funzioni polinomiali a tratti (solitamente di grado 3) in cui le regioni sono separate da sequenze di nodi ξ_1, \dots, ξ_K , in corrispondenza dei quali si forzano i polinomi stessi ad una continuità che di norma arriva fino alle derivate seconde. La difficoltà principale che sorge quando si ha a che fare con le *spline* di regressione, è determinare il numero e la posizione ottimale dei nodi. Da questo concetto sono state sviluppate una serie di estensioni e generalizzazioni. È il caso delle *spline* di lisciamiento, in cui si penalizza l'irregolarità della funzione, e quindi il *trade-off* tra variabilità e distorsione, tramite un parametro, λ , detto di lisciamiento. Si rimanda, per ulteriori sviluppi ed approfondimenti, ad Azzalini e Scarpa (2004), Hastie e Tibshirani (1990) e all'estensione pratica in R di Wood (2006).

Il software usato nell'analisi, utilizza di default le *spline* di lisciamiento cubiche e la loro generalizzazione a più dimensioni detta *thin plate regression splines*. Essa gode di buone proprietà e, nonostante sia molto costosa, a livello computazionale, per grandi *dataset*, non si ritiene necessario l'utilizzo di altre tipologie di *smoothers*.

3.2.4 Alcune estensioni dei modelli additivi

Come nel caso dei modelli di regressione lineare, anche la gamma dei modelli additivi offre una serie di estensioni e generalizzazioni che permettono all'analista di adattarli ad una miriade di circostanze.

I modelli additivi generalizzati (GAM), ad esempio, estendono i modelli additivi nello stesso modo in cui i modelli lineari generalizzati (GLM) estendono quelli di regressione lineare, e permettono di modellare gli effetti non lineari facendo uso di funzioni di lisciamiento. Il modello additivo generalizzato è infatti definito dalla relazione:

$$g(E[Y_i|x_1, \dots, x_p]) = \alpha + \sum_{j=1}^p f_j(x_{ij})$$

che coinvolge stimatori non parametrici al posto dei coefficienti di regressione del caso GLM. La funzione g rappresenta il legame tra il valore atteso della variabile risposta e la parte additiva, ed è chiamata *link*.

In questo studio, però, verrà utilizzata la formulazione più semplice di modello additivo, in quanto risulta quella che più si adatta alle osservazioni, tuttavia, come nella regressione lineare, si andrà prima a valutare il caso in cui gli errori assumano distribuzione Normale con l'espressione del modello (avente *link* identità) data da:

$$Y_i = \alpha + \sum_{j=1}^p f_j(x_{ij}) + \varepsilon_i$$

successivamente si dovrà rilassare questa ipotesi e si stimerà un modello additivo con errori autocorrelati, che rientra nella famiglia dei modelli additivi generalizzati misti. La nuova specificazione deriva da quella classica in cui si valuta una struttura autoregressiva di primo ordine per i residui:

$$Y_t = \alpha + \sum_{j=1}^p f_j(x_{tj}) + \varepsilon_t$$

$$\varepsilon_t = \varphi \varepsilon_{t-1} + u_t$$

con $u_t \sim IIN(0, \sigma_u^2)$ componente di rumore bianco e φ parametro autoregressivo come nel caso lineare. Per ulteriori approfondimenti si rimanda a Hastie e Tibshirani (1990) e Wood (2006).

La variabile di interesse continua ad essere il logaritmo del $PM_{2.5}$ e vengono valutati come predittori gli stessi fattori del modello lineare precedente:

- Direzione del vento
- Velocità del vento
- Temperatura
- Umidità
- Pressione
- Precipitazioni giornaliere
- Radiazione Solare Massima
- Classe di stabilità
- Diossido di azoto
- Diossido di zolfo

Con l'intento di catturare possibili effetti dovuti alla stagionalità, si inserisce nel modello un'ulteriore variabile esplicativa, impiegata spesso in letteratura e negli esempi pratici in Wood (2006), costruita sui giorni di campionamento e avente valori crescenti da 1 a 365, denominata *time*.

Se alcune variabili, tuttavia, entrano nel modello in forma lineare, la funzione di lisciamiento viene stimata con un solo grado di libertà e sarà per l'appunto lineare; si può pertanto reinserire tali variabili come fossero semplici predittori lineari, esattamente come nei modelli precedenti, alleggerendone la formulazione. Le variabili qualitative (direzione del vento e classe di stabilità), vista la loro natura nominale, per la quale presentano già un coefficiente per ogni loro possibile realizzazione, ritornano anch'esse nel modello nel modo classico.

3.3 La valutazione del modello

I modelli statistici, introdotti per simulare la dispersione del PM_{2.5} per i siti di Via Lissa e Via Malcontenta, necessitano di un metro di valutazione e confronto che gli permettano di essere comparati facilmente tra loro. Tali strumenti devono anche consentire il raffronto con lo stesso modello matematico di dispersione, trasformazione e trasporto introdotto precedentemente, al fine di poter valutare l'effettiva capacità di quest'ultimo di simulare il processo generatore e la diffusione degli inquinanti, rispetto ad un più semplice modello statistico.

Alcuni dei più comuni indicatori di qualità sono riportati nell'Appendice III "Criteri per l'utilizzo dei metodi di valutazione diversi dalle misurazioni in siti fissi" del DLgs 155/10. Essi hanno una natura o quantitativa o qualitativa, e ciascuno svolge un ruolo particolare nella valutazione del modello. La scelta di quale indicatore sia più opportuno usare dipende pertanto dallo scopo dell'applicazione modellistica e dalla disponibilità dei dati ottenuti dalle stazioni di misurazione per il confronto.

Il DLgs 155/10 individua:

- indicatori quantitativi:
 - Coefficiente di correlazione di Spearman;
 - *Fractional bias*;
 - *Root mean square error*;
 - *Normalize mean square error*.

- indicatori qualitativi
 - Diagramma di dispersione;
 - Grafici quantile-quantile;
 - Grafico dei residui;
 - Diagramma di Taylor.

La valutazione di un modello, o una serie di modelli, mediante questi stessi indicatori è utile, ma non sufficiente a comprendere le ragioni per le quali i risultati delle simulazioni possano essere vicini o lontani dai dati ottenuti dalle stazioni di campionamento. Per tale motivo, lo stesso decreto afferma come la valutazione del modello debba essere accompagnata anche da uno studio specifico dei processi riprodotti dallo stesso.

Nei paragrafi successivi si presentano gli indicatori, suddivisi in quantitativi e qualitativi, che verranno utilizzati successivamente nel confronto.

3.3.1 Gli indicatori quantitativi

Coefficiente di correlazione di Spearman o, più correttamente, indice di correlazione per ranghi di Spearman. Diversamente del coefficiente di correlazione lineare di Pearson, esso misura il grado di correlazione tra due variabili per le quali la relazione può essere descritta utilizzando una funzione monotona. Se non ci sono valori di dati ripetuti, si ha perfetta correlazione di Spearman quando l'indice assume valore +1 (se positivo) o -1 (se negativo), e si verifica quando ciascuna delle variabili è funzione monotona dell'altra, al valore zero corrisponde la non correlazione. La formulazione risulta:

$$\rho = \frac{\sum_{i=0}^n (c_i - \bar{c})(m_i - \bar{m})}{\sqrt{\sum_{i=0}^n (c_i - \bar{c})^2 (m_i - \bar{m})^2}}$$

dove c_i ed m_i sono rispettivamente il rango dell' i -esimo valore previsto e quello osservato, \bar{c} ed \bar{m} i valor medi ed n la numerosità campionaria.

Fractional bias, o mean fractional bias. A differenza della distorsione media, quest'indice normalizza il *bias* rispetto alla media algebrica calcolata sulla coppia delle osservazioni, presumendo che possa esserci distorsione rispetto al valor vero

anche nelle osservazioni a causa di misurazioni e campionamenti non sempre corretti. Un ulteriore vantaggio sta nel fatto che le sovrastime e le sottostime hanno un peso uguale, rendendo più riconoscibili eventuali fenomeni di stagionalità. Infine, poiché il MFB varia da -2 a +2, ha il vantaggio di delimitare il massimo errore di polarizzazione, impedendo che pochi *outlier* dominano la metrica.

Si ha la migliore prestazione del modello quando l'indice assume valore prossimo allo zero, viceversa un valore vicino a ± 200 indica una pessima adattabilità alle osservazioni. Boylan e Russel (2006) suggeriscono, per le applicazioni scientifiche, tre intervalli di confronto. La prima fascia è il *goal range*, con $-0.3 \leq \text{MFB} \leq +0.3$; poi si trova una seconda zona, con $-0.6 \leq \text{MFB} \leq +0.6$, che corrisponde ad un rendimento mediocre del modello; infine, la zona più esterna, indica un modello che poco si adatta ai dati misurati.

L'espressione dell'MFB può essere scritta come:

$$\text{MFB} = \frac{1}{n} \sum_{i=1}^n \frac{(C_i - M_i)}{(C_i + M_i) / 2}$$

con C_i ed M_i l'i-esimo valore previsto e osservato ed n sempre la numerosità del campione.

Root mean square error. Calcolato dalla radice dell'errore quadratico medio, è la deviazione standard degli scarti fra i valori osservati e quelli stimati. L'RMSE è una buona misura di precisione ma può essere usato solo per confrontare stime riferite ad una stessa variabile, in quanto dipende dalla metrica della stessa, non essendo un indice adimensionale.

L'RMSE varia tra zero e $+\infty$ ed indica un peggioramento della performance al crescere del valore. Utilizzando la stessa notazione degli indici precedenti, la sua espressione risulta:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (C_i - M_i)^2}$$

Normalize mean square error. È una versione normalizzata dell'errore quadratico medio ed è implementabile nei casi in cui il fenomeno osservato e previsto assume valori sempre positivi, in caso contrario perde di significato. Esso esamina il caso in cui la distorsione venga determinata principalmente da delle fluttuazioni casuali ma che è tuttavia influenzata anche dalla presenza di coppie anomale che si discostano in modo atipico. Un basso NMSE indica la buona performance del modello, d'altra parte un valore elevato non è segno di cattivissima stima, in quanto si tratta di un indice molto sensibile ai picchi e risente della scala di misura delle osservazioni.

Definendo \bar{C} ed \bar{M} le medie aritmetiche dei dati previsti ed osservati, otteniamo:

$$NMSE = \frac{1}{n} \sum_{i=1}^n \frac{(C_i - M_i)^2}{\bar{C} * \bar{M}}$$

3.3.2 Gli indicatori qualitativi

Diagramma di dispersione o scatter plot. È il grafico che proietta in uno spazio cartesiano un set di dati riferiti a due variabili.

Il grafico di dispersione risulta molto utile quando si vuole valutare la relazione di due insiemi di osservazioni. Più i *dataset* corrispondono, più i punti tendono a concentrarsi esattamente sulla linea d'identità; esso ha inoltre la capacità di mostrare relazioni non lineari tra variabili e, in questo contesto, evidenziare rapporti di sotto o sopra stima. Verrà disegnato introducendo, per ogni coppia, i valori calcolati nell'asse delle ascisse ed i valori

previsti in quello delle ordinate, determinando così il punto di intersezione.

Grafico quantile-quantile o Q-Q plot. Contrappone i quantili delle due distribuzioni poste a confronto rappresentando, in un piano cartesiano, i punti che hanno per coordinate i valori dei quantili dello stesso ordine delle due distribuzioni. Più il grafico dei quantili si avvicina alla bisettrice, maggiore è la somiglianza tra i due gruppi di osservazioni.

Grafici dei residui. Come per l'analisi diagnostica, la distribuzione dei residui può essere cruciale anche nella valutazione della performance dei modelli. Difatti, ipotizzando che la distorsione sia casuale, grazie a questi grafici si possono individuare particolari andamenti deterministici che non si sono ancora riusciti a modellare.

Diagramma di Taylor.

Il diagramma di Taylor (Taylor, 2001) fornisce un riassunto grafico di quanto un modello (o un insieme di modelli) rappresenti bene le osservazioni di riferimento. La somiglianza tra i due *dataset* viene valutata in termini di correlazione, di *root mean square error* e dalle loro deviazioni standard. Questo schema risulta, infatti, particolarmente utile sia per valutare congiuntamente diversi aspetti di modelli complessi, sia per confrontare la relativa capacità di adattamento di modelli differenti.

La correlazione viene mostrata nella partizione dell'angolo tra ascisse ed ordinate, mentre l'RMSE è proporzionale alla distanza del punto riferito al modello rispetto a quello rappresentato nell'asse come riferimento. Le deviazioni standard di entrambi i modelli, calcolato e simulato, corrispondono invece alla distanza radiale dall'origine.

Nel caso di confronti fra più modelli simulati, saranno migliori quelli rappresentati più vicini al punto marcato nell'asse x. Questi

modelli avranno sostanzialmente alta correlazione e basso RMSE. Quei modelli che giaceranno alla sua stessa distanza dall'origine, avranno invece un'uguale deviazione standard, data dalla capacità di questi di modellare la reale variabilità del processo osservato.

3.4 Bias

La formulazione di modelli statistici capaci di simulare al meglio il fenomeno studiato, non è il solo scopo di questa analisi. La ricerca di quei fattori che causano una maggiore distorsione nelle stime, possiede la prospettiva di un futuro miglioramento della qualità delle stesse, in un'ottica di continuo perfezionamento. La messa a punto del modello fotochimico, precedentemente descritto, può difatti venire agevolata e difesa da una serie di conclusioni nate da un utilizzo differente degli strumenti da poco trattati.

Nello specifico, si vuole utilizzare la formulazione dei modelli additivi per osservare quali variabili causano un aumento nella differenza tra le concentrazioni del $PM_{2.5}$ calcolate dal modello e quelle osservate realmente nei due siti. Il fine sarà quello di caratterizzare quali elementi risultano correlati all'aumento della distorsione, congiuntamente alle altre variabili presenti, e la stima della funzione di liscio o del parametro ad essi associata.

Si sceglie di valutare lo scarto tra il valore predetto e quello reale, formulazione utilizzata prettamente nell'ambito modellistico, nell'ottica di valutare la sovra o sotto stima.

Il problema, così sviluppato, assume questa struttura:

$$PM_{2.5}^{Calc} - PM_{2.5}^{Oss} = \alpha + f_1(x_1) + f_2(x_2) + f_3(x_3) + \dots + \varepsilon$$

con x_i la i -esima variabile esplicativa, f_i la funzione stimata ed ε il termine d'errore. Tale formulazione risulta, pertanto, avere le

stesse caratteristiche del modello additivo sviluppato precedentemente.

Tra le esplicative troviamo ancora:

- Direzione del vento
- Velocità del vento
- Temperatura
- Umidità
- Pressione
- Precipitazioni giornaliere
- Radiazione Solare Massima
- Classe di stabilità
- Diossido di azoto
- Diossido di zolfo

Tuttavia i valori ad esse associato non sarà lo stesso attribuito nelle analisi precedenti. Come si chiarirà successivamente, si andrà ad introdurre l'*input* del modello fotochimico, per porre l'attenzione agli elementi di distorsione generati dallo stesso. Proprio per questo motivo, si ha a disposizione la variabile relativa alla dispersione del diossido di zolfo anche per il sito industriale.

3.5 Il software R

R, il software utilizzato per l'analisi, nacque da un progetto di Ross Ihaka e Robert Gentleman che prevedeva lo sviluppo di una piattaforma per l'analisi statistica utilizzabile liberamente e derivante dal pacchetto S non *open source*. La prima versione, 0.90.0, venne pubblicata il 22 novembre 1999.

Attualmente R è un potente ambiente di analisi computazionale avente licenza gratuita e dotato di un linguaggio completo, con cui si può interagire. È scritto in C ed è disponibile per tutti i principali sistemi operativi (dal sito *r-project*, ultimo accesso ottobre 2013).

Il software gode inoltre di una vasta comunità di specialisti che ne contribuiscono la crescita attraverso la condivisione delle proprie

soluzioni. Con l'obiettivo di organizzare e supportare questa realtà rendendola facilmente accessibile, è stato costituito il CRAN (*Comprehensive R Archive Network*).

La particolarità di R consiste nel poter ampliare il programma attraverso numerose estensioni, disponibili sotto forma di *package*, che vengono sviluppate principalmente dagli stessi utilizzatori. Un'altra peculiarità è l'ambiente grafico, capace di supportare l'analisi computazionale ed aiutare ad intuire informazioni particolari, come nel caso della distribuzione dei dati. A parte i classici istogrammi e *boxplot*, è possibile tracciare funzioni, disegnare punti, inserire notazioni di testo o legende: tutto questo con grande precisione e qualità.

La versione utilizzata in questo lavoro è la 3.0.1 ed è stata estesa dalle principali library di *Rmetrics* e da altri *package* utili nell'analisi come *openair*, *mgcv* e *gamair*.

Capitolo 4

La tesi svolta si colloca all'interno di un progetto più ampio finalizzato allo studio del particolato atmosferico. Tale indagine fu svolta durante gli anni 2009-2010 da un gruppo di ricerca dell'Università Ca' Foscari, Venezia, in collaborazione con l'Ente Zona Industriale di Porto Marghera ed alcune delle maggiori aziende della zona industriale (Enel, Eni, Edison, Polimeri Europa).

Dopo un'introduzione all'area di studio e ai dati, verranno presentati e discussi i principali risultati ottenuti durante lo studio dei *dataset*. Ad una prima analisi grafica, utile a comprendere il quadro generale del problema, seguiranno gli esiti della modellazione lineare ed il loro sviluppo nei modelli additivi. Successivamente si andranno a comparare tali formulazioni principali con le stime del modello matematico di dispersione, utilizzando alcuni indicatori di performance già anticipati. Infine, caratterizzando il *bias* come la differenza tra la concentrazione di $PM_{2.5}$ calcolata dal modello e quella osservata nei campionamenti, si mostreranno i risultati di una modellazione lineare parametrica rispetto alle variabili usate, con lo scopo di correlare il fenomeno di distorsione ai principali fattori che la determinano.

4.1 Area di studio

La pianura padano-veneta è considerata una delle aree più inquinate d'Europa (EEA, 2012; Benassi et al., 2011; Carnevale et al., 2008, 2010; Lonati et al., 2010) e la sua morfologia,

caratterizzata dai gruppi montuosi (Alpi ed Appennini) che la delimitano per la quasi totalità e dallo sbocco sul Mare Adriatico, contraddistingue un ecosistema complesso e fortemente influenzato dai diversi ambienti (montuoso, lagunare e marino). La discontinuità spaziale genera quindi un ambiente di transizione che influenza la meteorologia. L'indagine, precedentemente introdotta, esaminava il territorio, di quasi 2'500 km², attorno alla città di Venezia che si trova sulla costa al margine della pianura.

La parte orientale della pianura padana è storicamente conosciuta per i pesanti livelli di inquinamento atmosferico ed è inoltre caratterizzata dalle diverse sorgenti di emissione antropiche e naturali che riversano sopra le zone principali, tra cui: il centro storico, l'insediamento urbano di Mestre (270.000 abitanti) e il distretto industriale di Porto Marghera.

Le principali fonti di emissione sono: i siti chimici e metallurgici, le centrali elettriche a carbone e le raffinerie di petrolio; il traffico stradale e autostradale, il trasporto e il traffico delle imbarcazioni, le operazioni portuali commerciali e da crociera; il traffico aereo nazionale e internazionale; le vetriere artistiche di Murano (Rampazzo 2008 a, b). A causa della vicinanza del mare Adriatico, anche lo spray marino rappresenta una componente importante del PM, così come i materiali cristallini e biologici (Masiol et al., 2012 a).

La zona attorno alla città di Venezia è caratterizzata da stabilità atmosferica e inversioni termiche che inducono le masse fredde verso il livello del suolo durante il periodo invernale. L'umidità è alta, causando eventi di nebbia in inverno e nelle stagioni intermedie, con una bassa dispersione di inquinanti. Un aumento di temperatura è accoppiato ad un aumento del *Planetary Boundary Layer*, ossia quella parte di atmosfera che viene direttamente influenzata dalla presenza della superficie terrestre. La sua altezza condiziona la dinamica atmosferica e, di

conseguenza, la densità delle polveri e degli inquinanti (ARPAV, 2013); per questo sia il mescolamento che la dispersione delle sostanze inquinanti aumentano nel periodo estivo (Agostini, 2012).

4.2 La struttura dei dati

I dati utilizzati in questa tesi sono di tipo meteorologico e chimico, e si riferiscono all'anno 2009.

La diversa tipologia dei dati porta questi stessi a derivare da molteplici fonti ed avere tra loro campionamenti differenti. Le centraline chimiche osservano dati giornalieri, mentre la meteorologia viene presa con cadenza diversa, solitamente oraria, perché misurata da enti ed istituzioni che hanno altre finalità. Per uniformare il database, è pertanto necessaria una mediazione dei dati meteorologici.

Per conciliare le esigenze logistiche ed economiche del progetto con le reali possibilità dei dati disponibili, nella qualità dell'aria vengono fatti compromessi descrittivi; se da un lato la numerosità di questi enti permette di avere un buon numero di stazioni per descrivere la meteorologia, non è possibile un'altrettanto adeguata descrizione dei dati chimici.

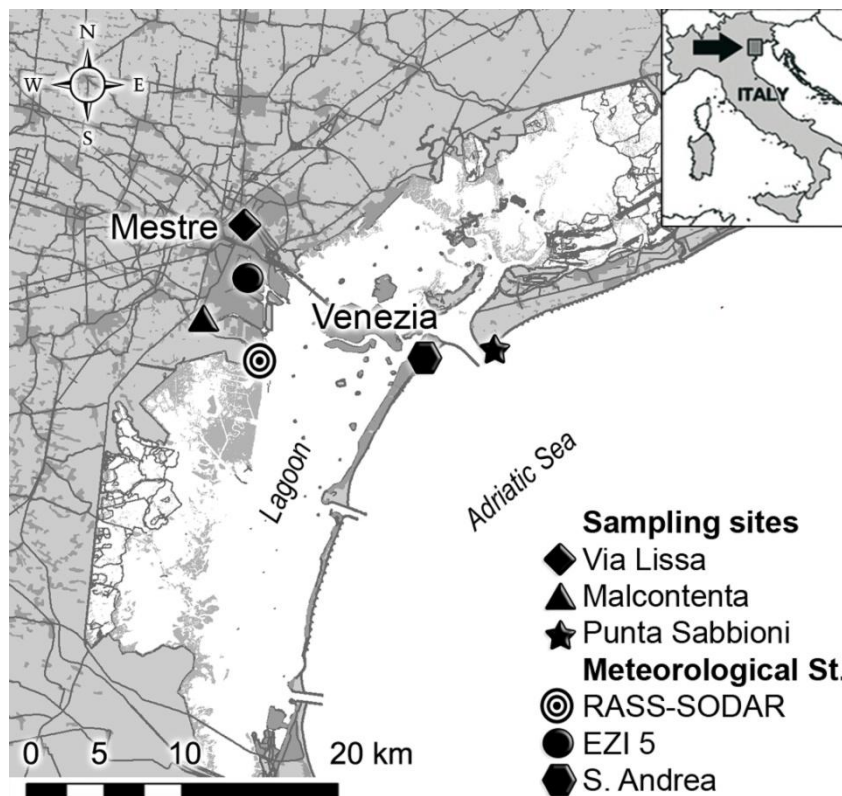
Nello specifico i dati meteorologici sono stati forniti, per l'anno 2009, da: ARPAV, Centro Maree del comune di Venezia, Ente Zona Industriale di Porto Marghera, ISPRA e Magistrato delle Acque di Venezia (per un totale di 27 stazioni); campionati mediante centraline fisse aventi sensori meteo-climatici e idrologici. I dati relativi ai gas, NO₂ e SO₂, e al PM_{2.5}, sempre per l'anno 2009, sono ricavati dal progetto citato precedentemente e forniti da ARPAV.

Le centraline del PM_{2.5} si trovano in Via Lissa e nell'adiacente Via Tagliamento a Mestre, area urbana continentale di Venezia, e in

Via Malcontenta a Marghera, zona industriale e portuale a circa 5 km di distanza.

In Figura 1 è raffigurata la posizione dei siti di campionamento del progetto di ricerca ed alcune stazioni meteorologiche. Il terzo sito di Punta Sabbioni non è stato considerato in questa tesi per mancanza di dati relativi ai gas.

Figura 1: Siti di campionamento e principali stazioni meteorologiche



Dal momento che si stimeranno dei modelli per tali concentrazioni di PM nelle due aree, per ogni variabile considerata si utilizzano le osservazioni provenienti dalla stazione meteo più vicina al punto di rilevamento del particolato. Nel caso di valore mancante si prende il dato “offerto” della stazione successivamente più vicina, in un’idea che ricorda il metodo del donatore o *hot deck* (Trimarchi F., 1990).

Sia per il sito di Via Lissa che per quello di Via Malcontenta, nei 365 giorni di campionamento dell'anno 2009, la variabile risposta manca di 13 valori.

L'*output* del modello fotochimico, tuttavia, non copre l'intero anno di studio. Per la sua generazione sono strettamente necessari dei dati in quota, osservati solo per alcuni periodi dell'anno:

- primavera: 26/02/2009 - 16/03/2009
- estate: 11/06/2009 - 16/07/2009
- autunno: 05/10/2009 - 31/10/2009
- inverno: 22/12/2009 - 31/12/2009

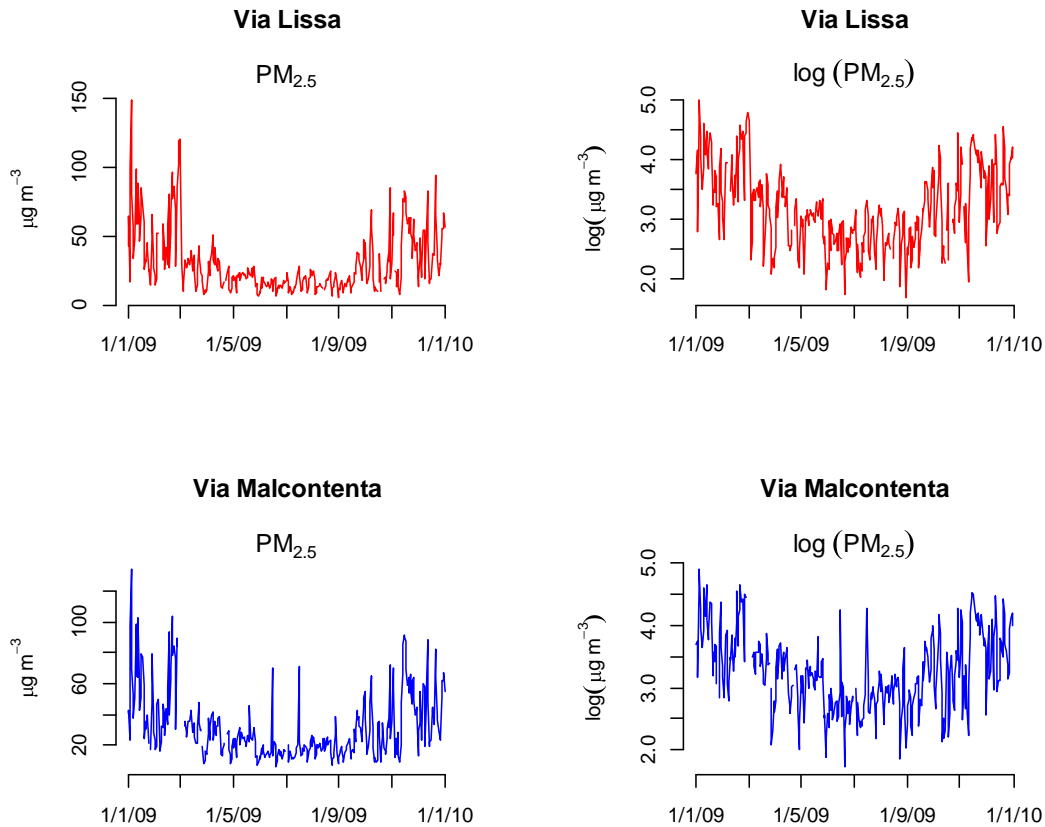
Per coerenza, il confronto sulla capacità predittiva tra i modelli e, necessariamente, lo studio della distorsione del CTM, riguarderà questi stessi intervalli.

4.3 Analisi preliminare

L'andamento temporale della serie del PM_{2.5} mostra la presenza di una componente stagionale molto marcata in entrambi i siti di rilevazione. Durante il periodo invernale, infatti, si osservano valori notevolmente più alti rispetto al resto dell'anno ed un più ampio campo di variazione.

Per ridurre questa differenza di variabilità tra le stagioni, è stata applicata fin da subito la trasformata logaritmica alle osservazioni (strettamente positive) con buoni risultati. Il grafico delle serie storiche originarie e la loro trasformata, è presente in Figura 2.

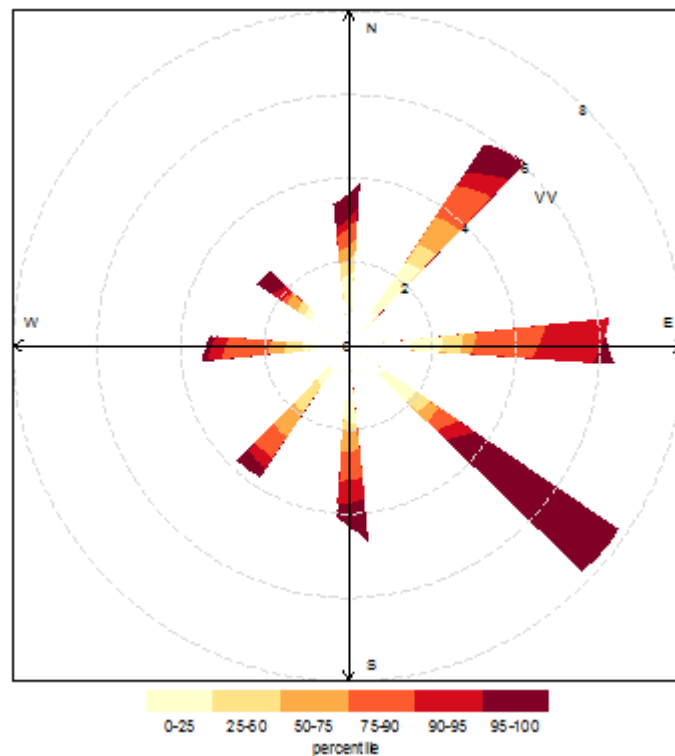
Figura 2: Serie storica del PM_{2.5} per i siti di Via Lissa e Via Malcontenta, anno 2009



Le variabili riferite al vento offrono l'opportunità di relazionare la direzione dello stesso, con la velocità, osservate nel corso dell'intero anno nella zona esaminata. Nei due siti ci si aspetta la quasi uniformità delle osservazioni, dato che distano solamente pochi km, e la centralina a terra utilizzata per il campionamento di questi fattori risulta essere la n° 5 dell'Ente Zona Industriale, in comune per entrambi i siti.

Dal grafico riportato in Figura 3 si osserva come i venti più veloci e ricorrenti risultano essere quelli provenienti dall'arco Nord-Est (relativi alla Bora) e Sud-Est. Verso tutte le altre direzioni si riscontrano venti deboli.

Figura 3: Rosa dei venti rappresentante la velocità del vento in relazione alla direzione, anno 2009, mediana e moda giornaliera



Può risultare tuttavia interessante valutare come si comportino i venti in quota, dove si riscontrano alcune importanti differenze. È presente infatti, in zona industriale vicino al sito di Via Malcontenta, una stazione meteorologica in quota (n° 22 dell'EZI).

In Figura 4 sono riportate le tendenze difformi per la direzione dei venti e la velocità, nelle due rilevazioni. Mentre, per quanto riguarda l'andamento della sola velocità visibile in Figura 5, i valori a terra risultano più bassi di circa un 1 ms^{-1} rispetto a quelli campionati in quota. Nonostante questo, l'andamento nel corso dell'anno è simile, e non si rilevano particolari effetti dovuti alla componente stagionale in alcun caso.

Nelle successive analisi, per una maggior coerenza rispetto agli altri fattori in esame, si utilizzano comunque i soli dati osservati al suolo.

Figura 4: Rosa dei venti rappresentante la velocità del vento in relazione alla direzione, rilevazione in suolo ed in quota, anno 2009, mediana e moda giornaliera

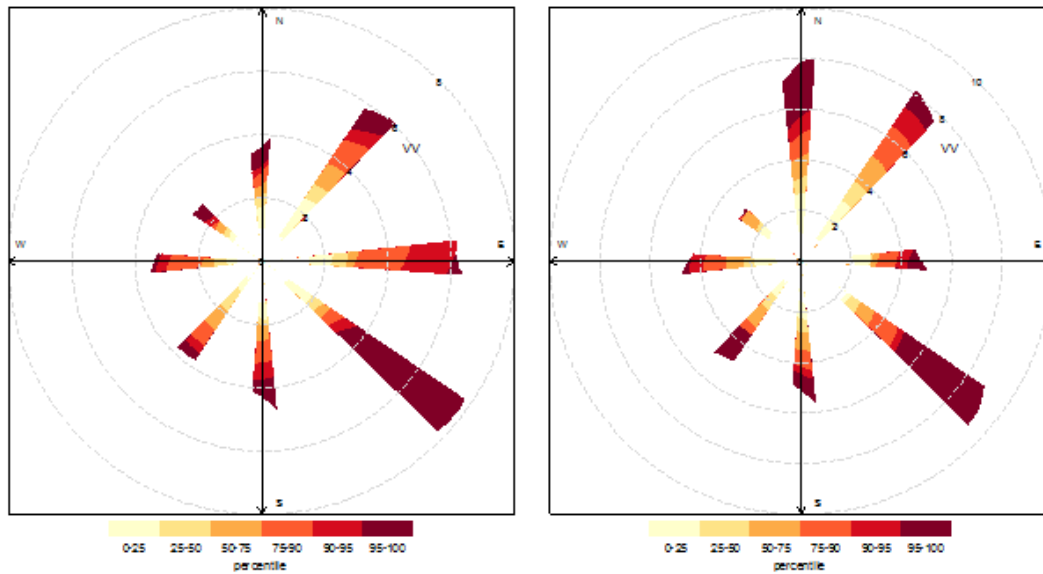
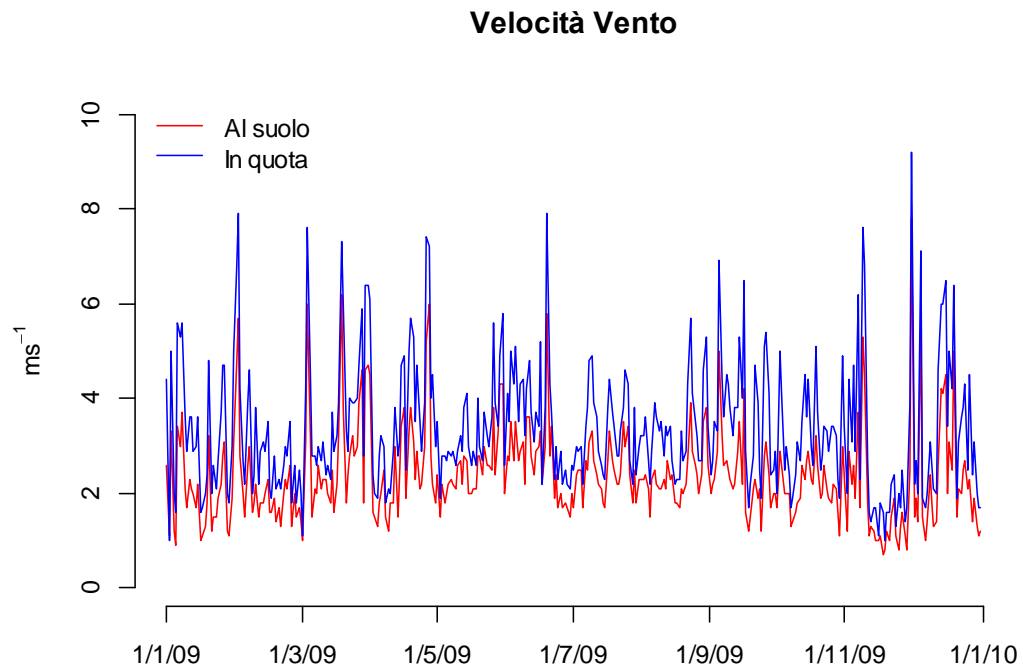
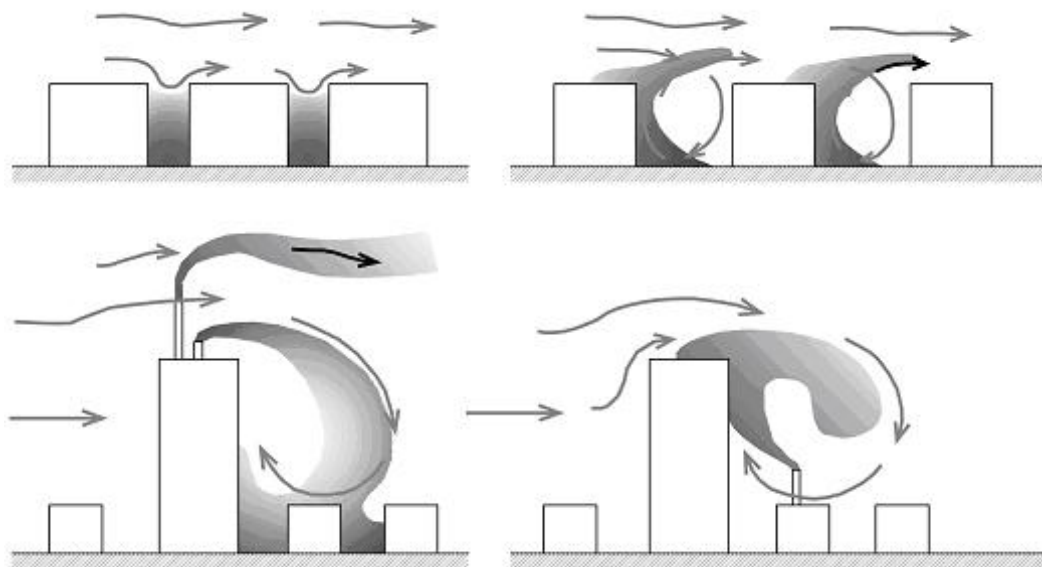


Figura 5: Serie storiche della velocità del vento, rilevazione in suolo ed in quota, anno 2009, mediana giornaliera



Tale discontinuità in realtà è stata confermata da alcuni studi (Barnaba, 2007; Camuffo, 1979) che sembrerebbero evidenziare andamenti disaccoppiati sia tra gli strati più prossimi al suolo e quelli più alti, che tra l'entroterra e l'area costiera. Alcune differenze tra siti così vicini possono inoltre essere causate dall'incanalamento delle correnti d'aria per la presenza di edifici (*street canyon*) (Figura 6).

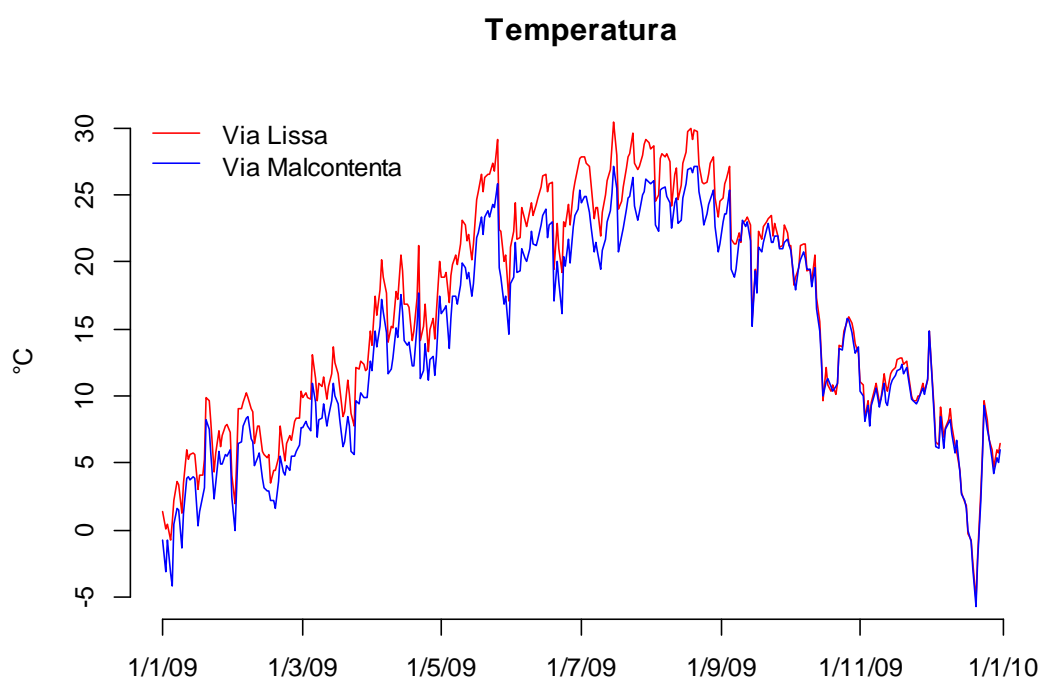
Figura 6: Esempio di perturbazione delle correnti d'aria ed il trasporto degli inquinanti in aree urbane ed industriali (dal sito di EcoEdility)



La temperatura, misurata in gradi Celsius ($^{\circ}\text{C}$), è tra le variabili che subiscono la maggiore influenza stagionale, in quanto causata prevalentemente dall'irraggiamento solare. Essa è fondamentale per la comprensione della cinetica dei gas e quindi della diffusione in atmosfera di polveri ed inquinanti. Semplificando, a temperature più basse i gas sono meno dinamici e tendono a disperdersi più lentamente, viceversa per valori più alti si ha una maggiore energia cinetica tra le particelle con conseguenza opposta. Un aumento di temperatura è anche abbinato, come già trattato, ad un aumento del *Planetary Boundary Layer* e, di conseguenza, il rimescolamento

e la dispersione delle sostanze inquinanti aumentano nel periodo estivo.

Figura 7: Temperatura rilevata nelle stazioni di Via Lissa e Via Malcontenta, anno 2009, mediana giornaliera



I dati relativi all'umidità relativa provengono, per entrambi i siti, dalla centralina n° 23 dell'Ente Zona Industriale, situata a Porto Marghera tra il canale industriale ovest ed il canale industriale sud.

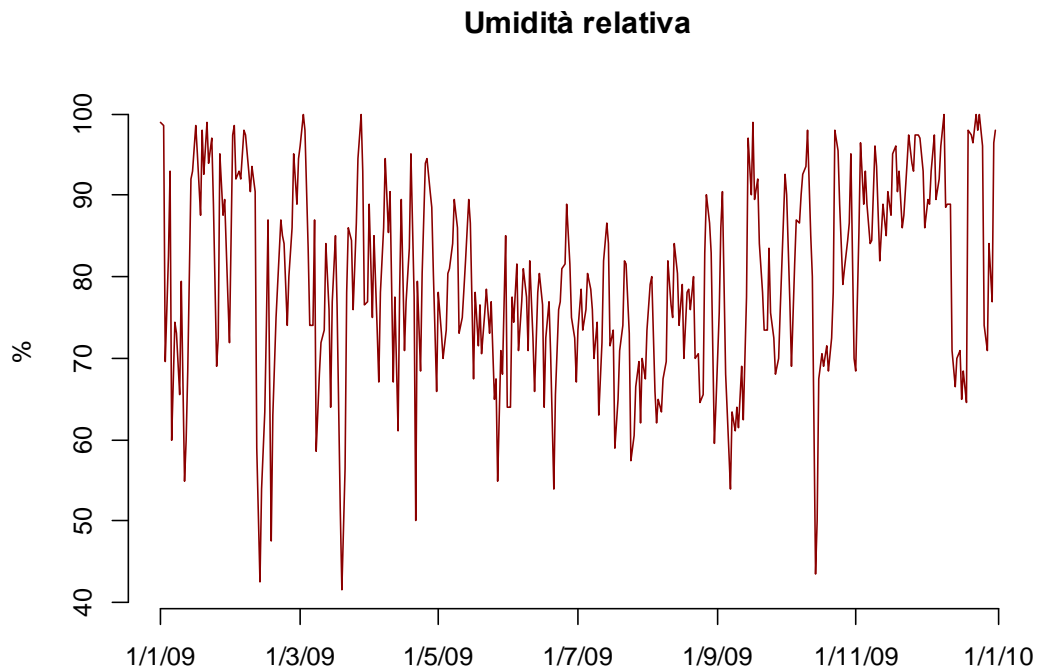
Essa è definita come il rapporto della densità del vapore acqueo nell'aria e la densità del vapore saturo alla temperatura della miscela: un'umidità relativa unitaria (100%) indica che il miscuglio gassoso contiene la massima quantità di umidità possibile per le date condizioni di temperatura e pressione, e non esistono valori superiori (a meno di sovrasaturazioni).

Questa centralina risulta essere, sia per Via Lissa che per Via Malcontenta, la più vicina avente tali dati. Ciò, tuttavia, non

appare un problema, in quanto l'umidità relativa è costante in quasi tutte le centraline di rilevamento, anche quelle più distanti.

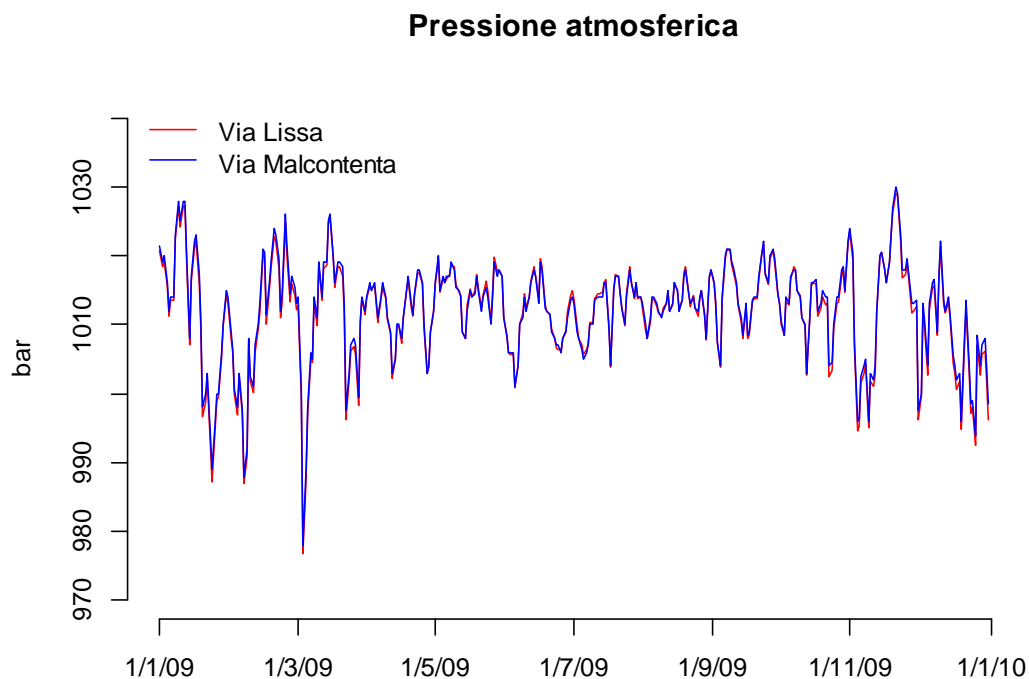
Nel corso dell'anno la sua distribuzione, in Figura 8, appare più regolare e con minore variabilità durante il periodo primaverile; diversamente, durante le altre fasi dell'anno, si osservano mutamenti di alta e bassa umidità molto dinamici.

Figura 8: Umidità relativa rilevata per i siti di Via Lissa e Via Malcontenta, anno 2009, mediana giornaliera



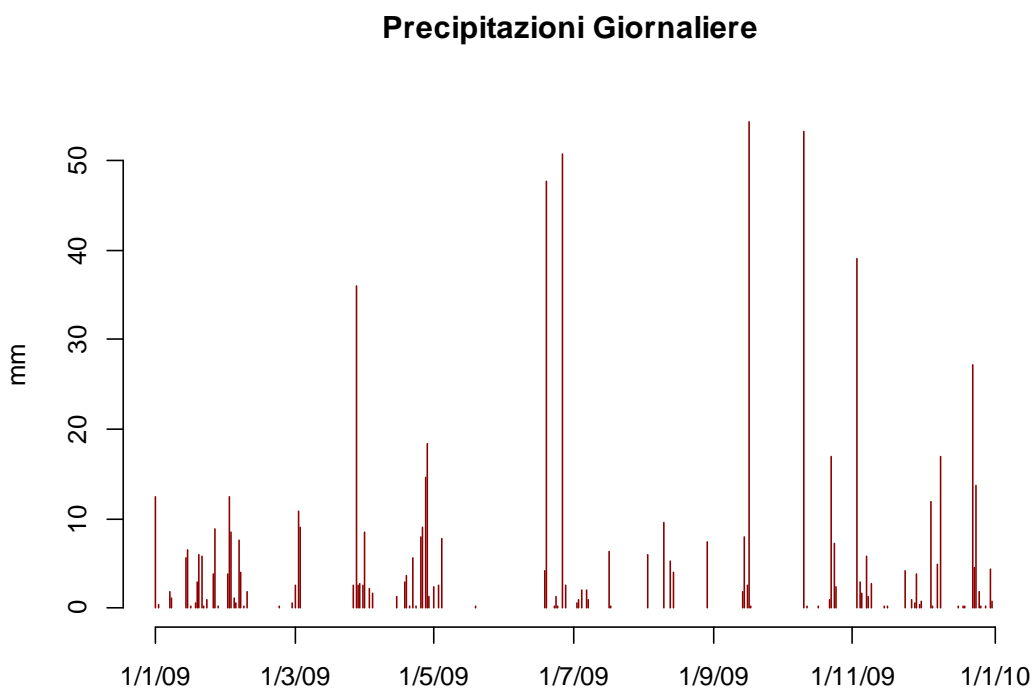
La pressione, misurata in bar ($1 \text{ bar} = 10^5 \text{ Pa}$), si mostra anch'essa quasi totalmente invariabile tra i due luoghi di rilevamento del $\text{PM}_{2.5}$. Come l'umidità, mostra un comportamento stabile durante il periodo primaverile - estivo, ed una fase più perturbata durante i mesi freddi, in cui c'è una forte alternanza di alta e bassa pressione.

Figura 9: Pressione atmosferica a terra nei due siti di rilevazione, anno 2009, mediana giornaliera



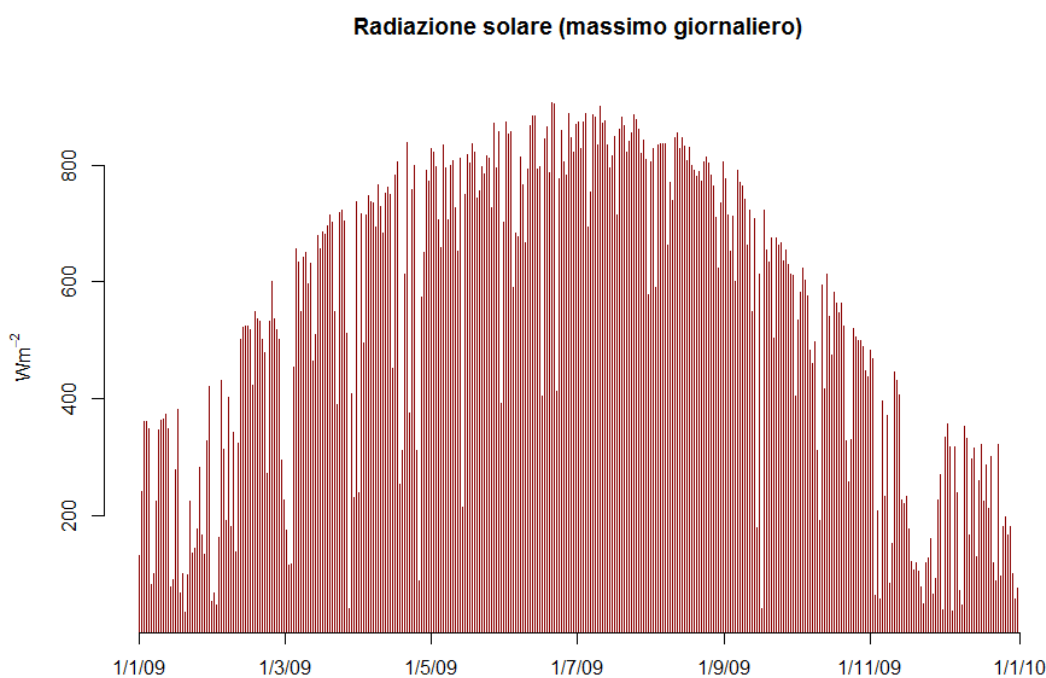
Misurate in mm, le precipitazioni assumono un ruolo fondamentale per l'azione di lavaggio dell'atmosfera da aerosol e particolato atmosferico; esercitano così un'azione naturale di mitigazione in situazioni di forte inquinamento come nel caso di grandi aree metropolitane o industriali. In Figura 10 si riportano le precipitazioni giornaliere dell'anno 2009. La stagione più piovosa risulta essere la primavera, tuttavia è in autunno che si hanno i maggiori picchi, con ben tre delle sette precipitazioni eccezionali sopra i 25mm di pioggia.

Figura 10: Precipitazioni giornaliere nell'area di Venezia-Mestre, anno 2009



La radiazione solare presenta una stagionalità simile alla temperatura, vista in precedenza, in quanto anch'essa derivante principalmente dall'irraggiamento solare. Pure in questo caso entrambi i siti adoperano le stesse rilevazioni, presentate in Figura 11. Per questo fattore, si è scelto di adoperare il valore massimo registrato nell'arco della giornata in quanto alcune reazioni chimico-fisiche, che svolgono un ruolo rilevante nella formazione del particolato, vengono attivate in natura al superamento di una specifica soglia di irraggiamento.

Figura 11: Radiazione solare nell'area di Venezia-Mestre, anno 2009, massimo giornaliero

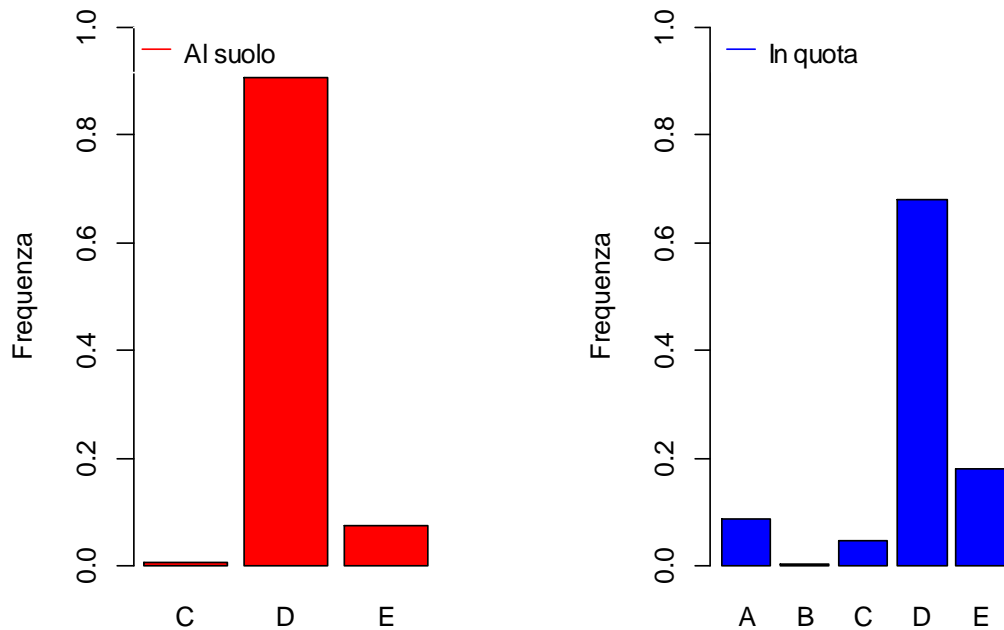


La stabilità atmosferica, classificata da Pasquill nel 1961, è l'indicatore per eccellenza della turbolenza dell'atmosfera. Essa può favorire o meno il rimescolamento dell'aria e quindi il processo di diluizione del particolato, così come la dispersione dell'inquinamento delle sorgenti ad alta quota. La scala comprende le lettere dalla A alla F, alle quali corrispondono, progressivamente, dei sistemi molto instabili fino a dei sistemi stabili. La categoria D si riferisce ad una stabilità atmosferica neutra.

Anche in questo caso risulta interessante il confronto tra i valori al suolo e quelli in quota campionati dalle stesse centraline meteorologiche viste per il vento.

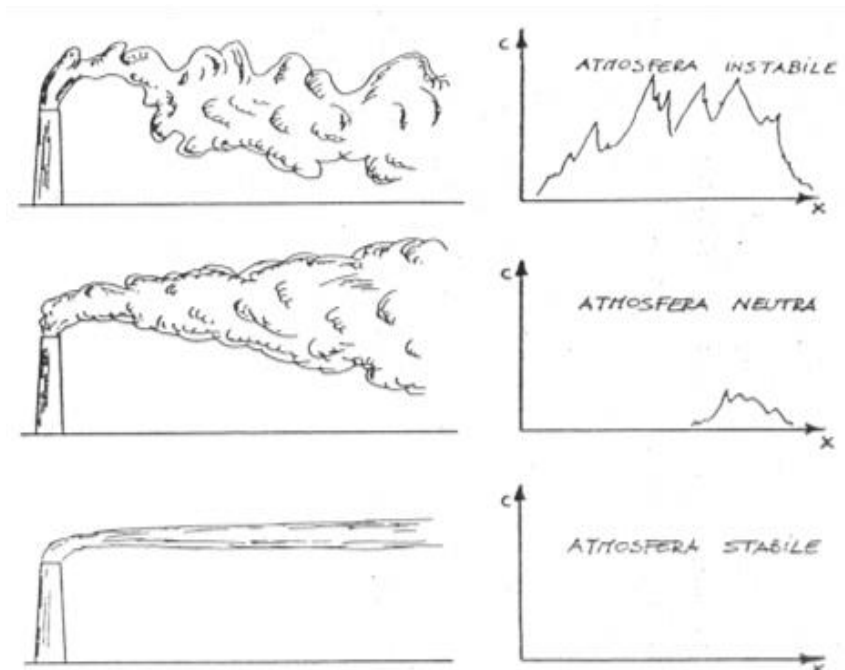
In Figura 12 si nota come la quasi totalità delle classi di stabilità misurate al suolo siano di categoria D, segno di una più bassa turbolenza atmosferica rispetto a ciò che succede in quota. In quest'altro caso infatti si osservano frequenze anche per le altre classi di stabilità, nonostante l'atmosfera comunque sia più frequentemente neutra o stabile.

Figura 12: Frequenza annua delle classi di stabilità per i due siti, anno 2009, moda giornaliera



Questo aspetto di difformità, come nel caso delle correnti d'aria, ha un forte impatto nella diffusione del particolato delle sorgenti in quota. Il sito industriale di Porto Marghera presenta infatti numerose ciminiere che emettono i loro fumi ad un'altezza di circa 80-100 metri che risentono degli effetti delle diverse condizioni di stabilità (Figura 13).

Figura 13: Effetti della stabilità atmosferica sulla dispersione verticale di una sorgente in quota



Nello studio sono inseriti anche due gas tra le variabili esplicative, il diossido di azoto, NO_2 , ed il diossido di zolfo (o anidride solforosa), SO_2 . La loro presenza in atmosfera può causare gravi danni all'ambiente e alla salute stessa dell'uomo, ed assumono un ruolo importante nella formazione del particolato secondario. L' NO_2 , per entrambi i siti, appare piuttosto stabile e presenta una lieve componente stagionale percepibile nei mesi invernali.

Figura 14: Serie storica relativa al gas NO₂ per il sito di Via Lissa, anno 2009, mediana giornaliera

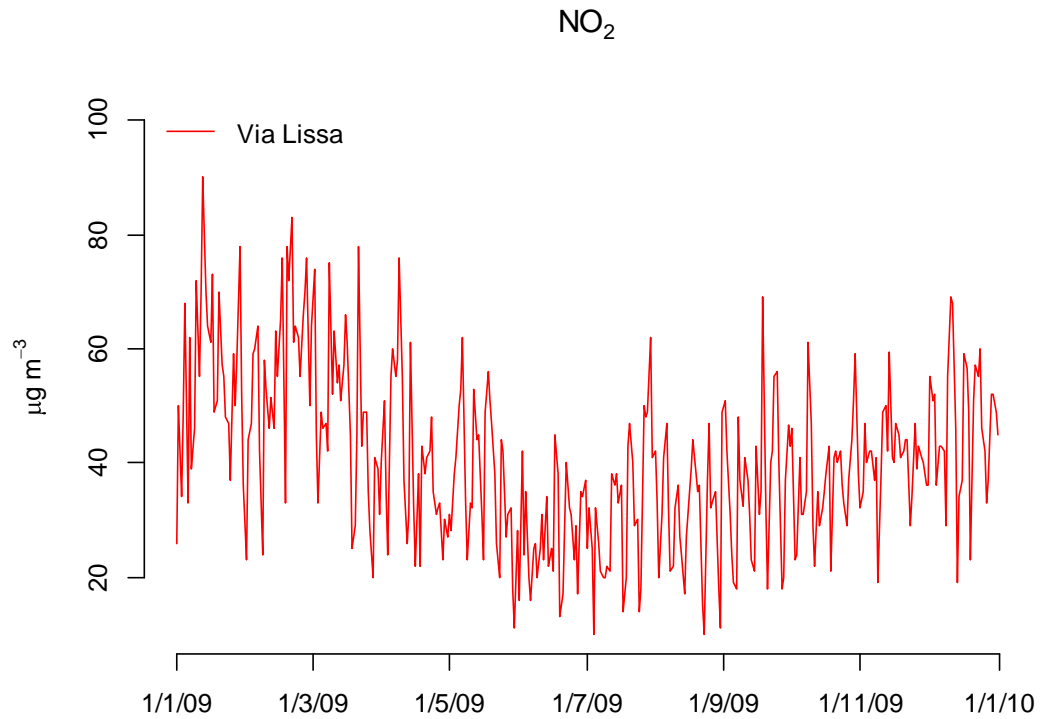
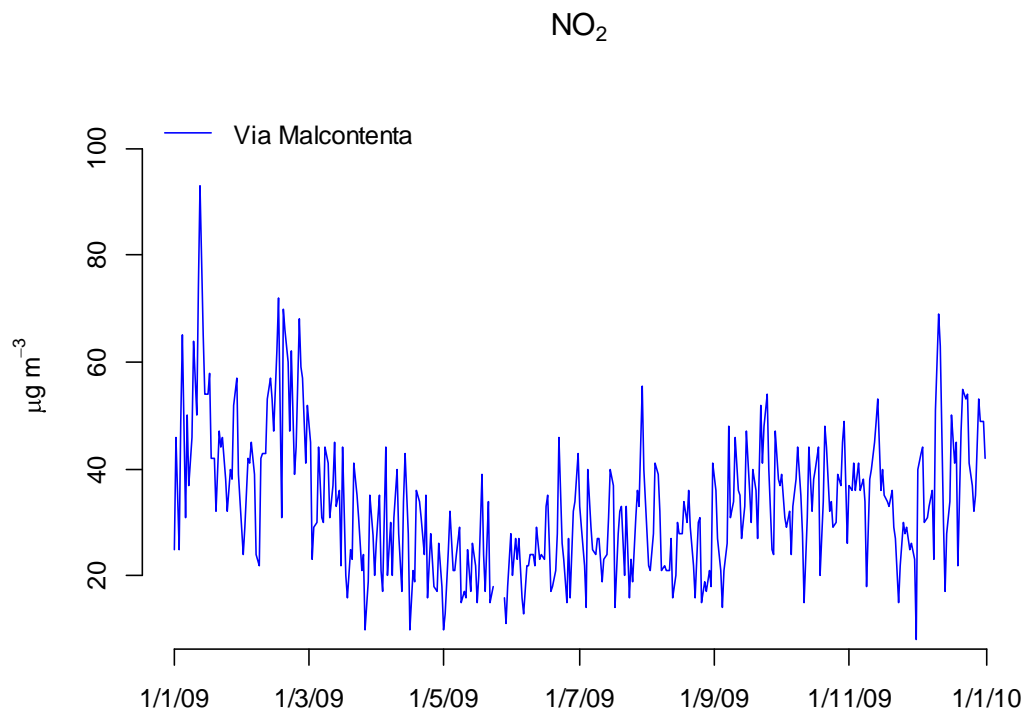


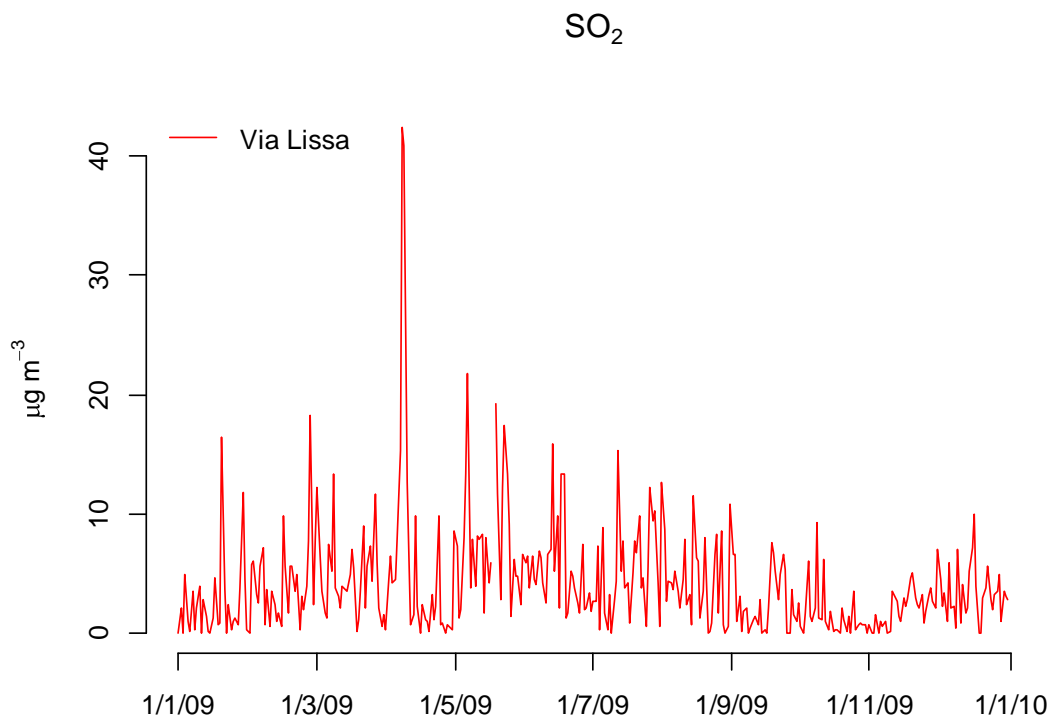
Figura 15: Serie storica relativa al gas NO₂ per il sito di Via Malcontenta, anno 2009, mediana giornaliera



Il diossido di zolfo, SO₂, campionato solo per il sito di Via Lissa, risente di forti picchi stagionali anch'essi in prevalenza durante il periodo freddo (Figura 16). Ad inizio primavera si riscontra, il valore più alto, oltre 40 $\mu\text{g m}^{-3}$ che, come vedremo successivamente nell'analisi, è verosimilmente causato da una sorgente eccezionale.

Il valore di concentrazione dell'SO₂, tuttavia, potrebbe risentire della taratura degli strumenti effettuata riguardo al limite imposto dalla direttiva europea 2008/50/CE, che comporta una poca sensibilità rispetto ai reali valori bassi misurabili.

Figura 16: Serie storica relativa al gas SO₂ per il sito di Via Lissa, anno 2009, media giornaliera



4.4 Regressione lineare

L'analisi grafica precedente pone in evidenza soprattutto il comportamento annuale delle variabili prese in esame.

Con il primo modello presentato, la regressione lineare, si è invece capaci di cogliere la relazione lineare esistente tra la concentrazione del PM_{2.5} osservata in atmosfera e gli altri fattori considerati.

La variabile dipendente, come anticipato, è in forma logaritmica, ed il coefficiente cattura la variazione relativa della concentrazione del PM_{2.5} rispetto ad una variazione dell'esplicativa.

Nella Tabella 1 vengono riportate le stime dei minimi quadrati ordinari per la formulazione più ampia del modello, contenente tutte le variabili esplicative.

Tabella 1: Stima dei minimi quadrati ordinari della log dispersione del PM_{2.5}

Variabile	Via Lissa			Via Malcontenta		
	Coefficienti	Deviazione Standard	Significatività 5%	Coefficienti	Deviazione Standard	Significatività 5%
Intercetta	-10.6499	3.0479	Si	-17.6085	3.1446	Si
DV-Nord	0.0603	0.0940	No	0.0415	0.1176	No
DV-NordEst	-0.0318	0.0958	No	0.0807	0.1223	No
DV-NordOvest	0.1671	0.1298	No	0.1574	0.1457	No
DV-Ovest	-0.0507	0.1349	No	-0.0748	0.1415	No
DV-Sud	0.0417	0.1363	No	0.1928	0.1396	No
DV-SudEst	0.0505	0.0982	No	0.1502	0.1361	No
DV-SudOvest	0.0912	0.1171	No	0.0572	0.1389	No
Velocità Vento	-0.2010	0.0262	Si	-0.1439	0.0197	Si
Temperatura	-0.0202	0.0042	Si	-0.0193	0.0040	Si
Umidità relativa	0.0079	0.0022	Si	0.0071	0.0023	Si
Pressione	0.0135	0.0029	Si	0.0205	0.0030	Si
Precipitazioni	-0.0102	0.0032	Si	-0.0118	0.0031	Si
Radiazione Solare	-0.0005	0.0001	Si	-0.0005	0.0001	Si
Classe Stabilità B	Non osservata			0.4443	0.2637	No
Classe Stabilità C	Non osservata			0.1358	0.1137	No
Classe Stabilità D	0.0378	0.2143	No	0.0648	0.0743	No
Classe Stabilità E	0.2030	0.2260	No	0.1614	0.0868	No
NO ₂	0.0127	0.0018	Si	0.0121	0.0020	Si
SO ₂	-0.0013	0.0050	No	Non osservata		

Tabella 2: Test F e coefficiente di determinazione lineare semplice e corretto

Via Lissa		Via Malcontenta	
Test F =	54.300	Test F =	39.700
R ² =	0.737	R ² =	0.685
R ² corretto =	0.7230	R ² corretto =	0.6680

In Tabella 2 sono riportati i valori dei test F ed i coefficienti di determinazione semplici e corretti. Entrambi i modelli stimati presentano un test di significatività che porta al rifiuto dell'ipotesi di nullità congiunta dei coefficienti. Tuttavia, presi singolarmente, alcuni parametri risultano non significativi (visibili in Tabella 1) e possono suggerire l'eliminazione di queste variabili dal modello. Nello specifico, si accetta la nullità dei coefficienti al 5% nelle variabili riferite a: direzione del vento, classi di stabilità e, per via Lissa, concentrazione del diossido di zolfo. Infine si osserva che questa specificazione di modello lineare sembra adattarsi meglio ai dati per Via Lissa, spiegando maggior variabilità rispetto all'altro sito, con un R² pari a 0.737.

Il comportamento dei residui soddisfa sufficientemente le assunzioni di simmetria, omoschedasticità e Normalità dei modelli, suggerendo che non ci siano grossi fattori sistematici omessi o di dover cambiare la forma funzionale del modello.

In Figura 17 e Figura 18 sono presenti alcuni grafici diagnostici. I residui riferiti a Via Malcontenta hanno un campo di variazione più ampio rispetto agli altri e, come si evince anche dal grafico quantile-quantile, presentano code che più si discostano dalla linea di perfetta Normalità.

Figura 17: Grafici diagnostici dei residui dei minimi quadrati ordinari, Via Lissa

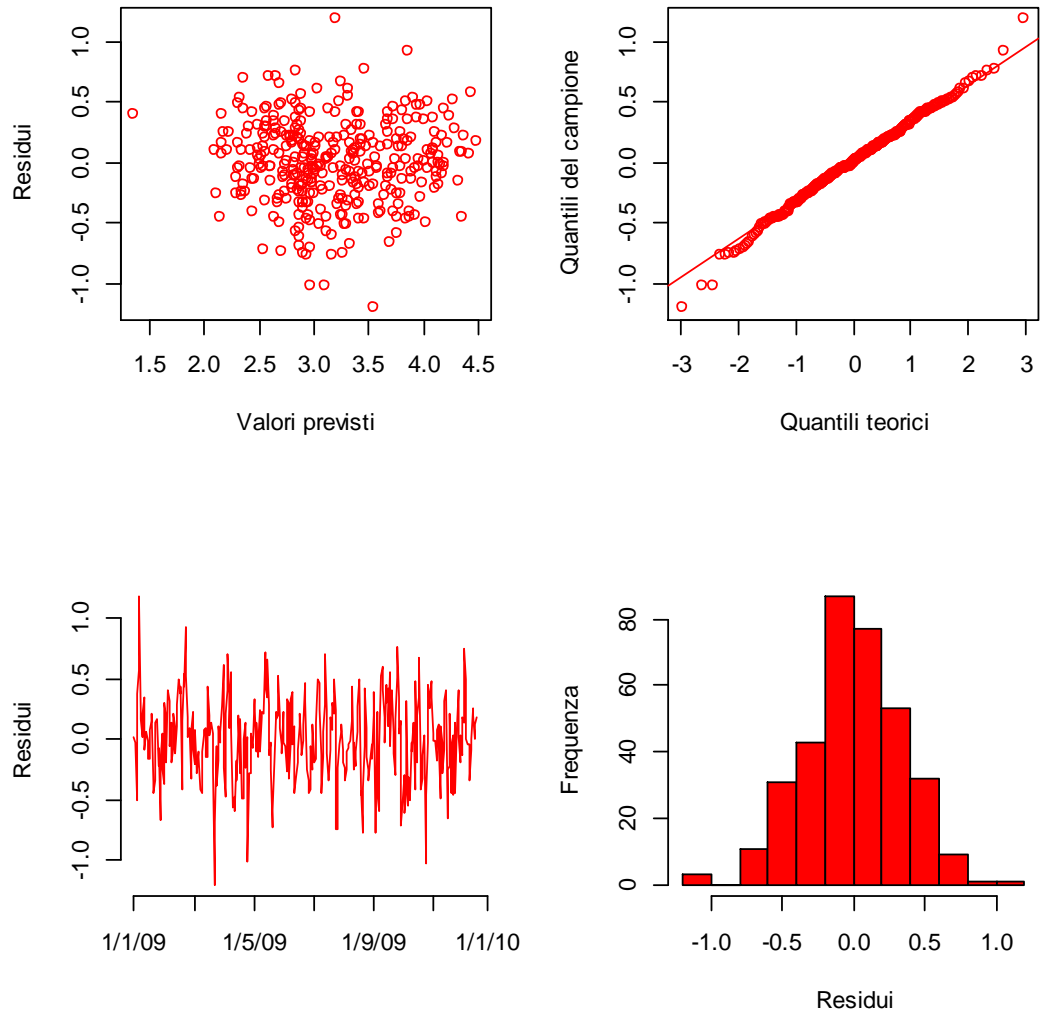
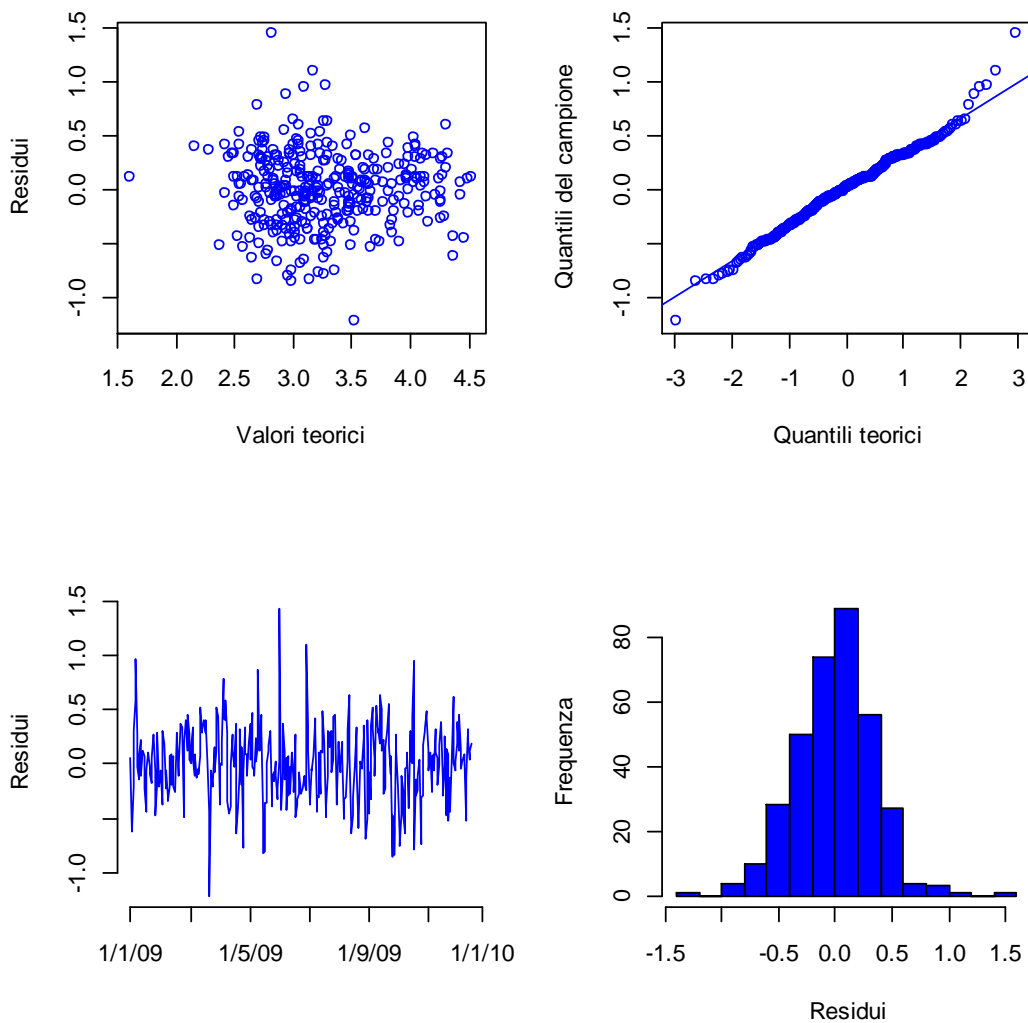


Figura 18: Grafici diagnostici dei residui dei minimi quadrati ordinari, Via Malcontenta



L'assunzione di indipendenza degli errori risulta essere invece più critica. Trattandosi infatti di dati giornalieri, è sensato pensare che una concentrazione alta di particolato osservata in un dato giorno, può causare ripercussioni sulla concentrazione dello stesso inquinante nel giorno successivo, e al contrario per valori piccoli. Questo effetto di trascinamento può venire inoltre amplificato dalle caratteristiche morfologiche e meteorologiche della pianura

padano-veneta spiegate precedentemente. Diversi autori hanno infatti stimato gli effetti dell'inquinamento diffuso che caratterizzano l'area della pianura padana e che sono i più alti di tutta Europa (EEA, 2012; Benassi et al., 2011; Carnevale et al., 2008, 2010).

Un primo test di Durbin Watson, riferito ai residui dei due modelli, evidenzia la possibilità di una componente autocorrelata di primo ordine; il test rifiuta, in entrambe le serie, l'ipotesi di incorrelazione.

Tabella 3: Test D-W per l'autocorrelazione dei residui del modello lineare, Via Lissa

Test Durbin Watson	Via Lissa
$DW = 1.152$	$p\text{-value} < 2.2e-16$
Si rifiuta l'ipotesi nulla a favore dell'alternativa: autocorrelazione positiva	

Tabella 4: Test D-W per l'autocorrelazione dei residui del modello lineare, Via Malcontenta

Test Durbin Watson	Via Malcontenta
$DW = 1.339$	$p\text{-value} = 5.799e-11$
Si rifiuta l'ipotesi nulla a favore dell'alternativa: autocorrelazione positiva	

Una più approfondita analisi grafica sulle autocorrelazioni globali e parziali, mostra come esse risultino significative: fino al secondo ritardo, per il sito urbano; ad un ritardo, per quello industriale.

Grafici di questo tipo sono causati, credibilmente, da processi autoregressivi di primo ordine, AR(1), aventi un basso coefficiente

di autoregressione. In Figura 19 e Figura 20 vengono presentati i grafici delle funzioni di autocorrelazione globale e parziale. Le linee tratteggiate indicano la regione di accettazione al 5%; un valore di $\hat{\rho}_k$ fuori dall'intervallo, porta a ritenere ρ_k significativamente diverso da zero.

Figura 19: Autocorrelazione dei residui globale e parziale, Via Lissa

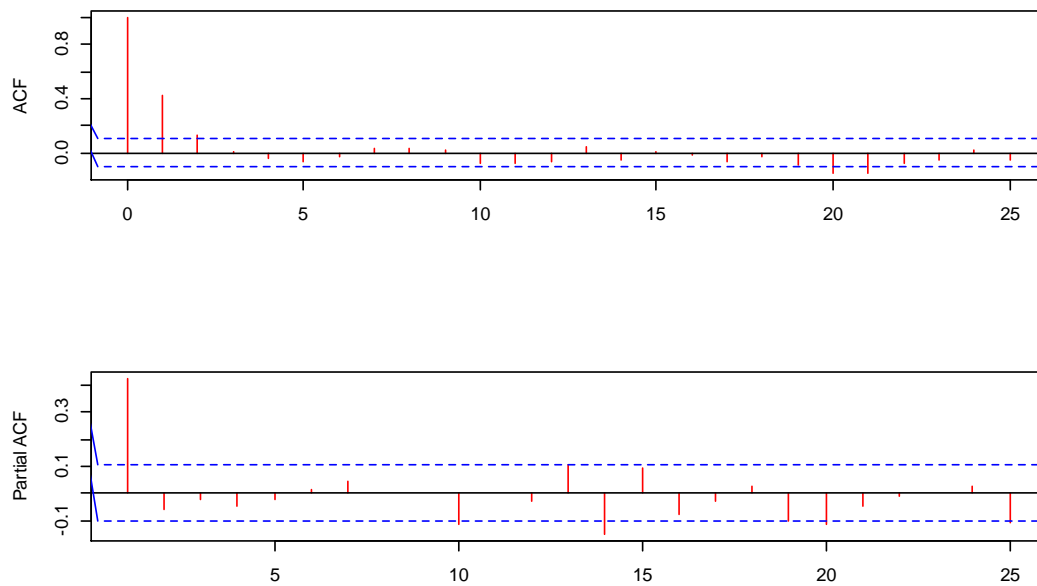
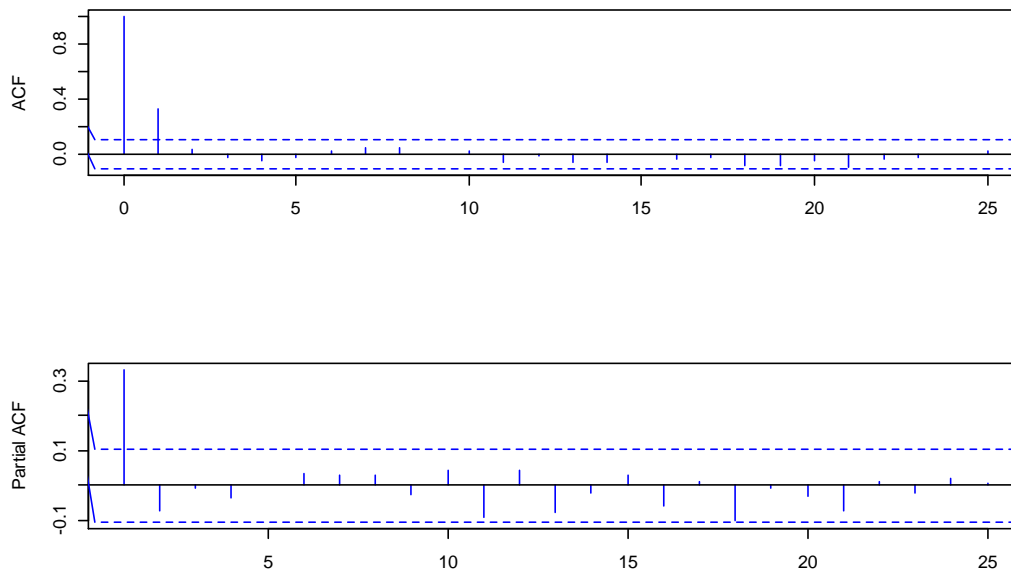


Figura 20: Autocorrelazione dei residui globale e parziale, Via Malcontenta



Alla luce di questi risultati, è necessario rivedere il metodo di stima dei parametri, considerando la struttura autoregressiva di primo ordine negli errori. In Tabella 5 vengono quindi riportate le stime eseguite attraverso lo stimatore dei minimi quadrati generalizzati in cui si ipotizza tale modello per gli errori.

Tra le variabili esplicative di Via Malcontenta manca, come già anticipato, il diossido di zolfo. Questo tuttavia non è visto come un problema; la reale concentrazione del gas risente, per Via Lissa, della poca sensibilità degli strumenti che offrono campionamenti di discutibile bontà. La criticità viene infatti catturata dal processo di regressione lineare che ne fa risaltare un coefficiente non significativo, nonostante l'importante ruolo chimico svolto dall' SO_2 nella formazione delle polveri.

Tabella 5: Schema riassuntivo delle stime GLS per i siti di Via Lissa e Via Malcontenta

Variabile	Via Lissa			Via Malcontenta		
	Coefficienti	Deviazione Standard	Significatività 5%	Coefficienti	Deviazione Standard	Significatività 5%
Intercetta	-5.1142	3.9338	No	-12.2612	3.9099	Si
DV-Nord	0.0938	0.0734	No	0.0177	0.1049	No
DV-NordEst	-0.0496	0.0742	No	0.0367	0.1079	No
DV-NordOvest	0.2363	0.0996	No	0.1198	0.1291	No
DV-Ovest	0.0491	0.1081	No	-0.0200	0.1306	No
DV-Sud	0.0735	0.1021	No	0.0983	0.1267	No
DV-SudEst	0.0717	0.0740	No	0.0709	0.1257	No
DV-SudOvest	0.1451	0.0917	No	0.0654	0.1231	No
Velocità Vento	-0.1612	0.0224	Si	-0.1210	0.0187	Si
Temperatura	-0.0201	0.0054	Si	-0.0167	0.0051	Si
Umidità relativa	0.0092	0.0021	Si	0.0070	0.0022	Si
Pressione	0.0077	0.0038	Si	0.0152	0.0038	Si
Precipitazioni	-0.0050	0.0025	Si	-0.0081	0.0027	Si
Radiazione Solare	-0.0003	0.0001	Si	-0.0004	0.0001	Si
Classe Stabilità B	Non osservata			0.4285	0.2276	No
Classe Stabilità C	Non osservata			0.0374	0.0960	No
Classe Stabilità D	0.1703	0.1628	No	0.0102	0.0675	No
Classe Stabilità E	0.1854	0.1729	No	0.0867	0.0767	No
NO ₂	0.0117	0.0017	Si	0.0115	0.0020	Si
SO ₂	-0.0069	0.0047	No	Non osservata		



Per entrambi i siti si può notare l'assenza di due categorie (una per ogni variabile categoriale), in quanto presa come riferimento dal

modello a causa del problema della multicollinearità. In Via Lissa mancano anche le stime per le classi di stabilità A e B; queste non vengono mai osservate nel periodo di riferimento.

Lo stimatore dei minimi quadrati generalizzati, per il sito di Via Malcontenta, riconosce significativi i coefficienti riferiti alle stesse variabili rilevanti dell'analisi per Via Lissa. In più esso conserva la significatività dell'intercetta, cosa che si va a perdere aggiungendo la specificazione AR(1) per gli errori dell'altro sito.

È possibile, inoltre, stimare il parametro autoregressivo $\hat{\phi}$ nei due nuovi modelli sviluppati. Di seguito sono presentate le stime e i test di significatività di tali coefficienti.

Tabella 6: Stime dei parametri autoregressivi per Via Lissa e Via Malcontenta

	Stima	Std error	t value	p value
 Via Lissa	0.5583	0.0445	12.6	< 2e-16
 Via Malcontenta	0.4206	0.0487	8.64	< 2e-16

Il sito urbano, presentando un coefficiente più alto, si dimostra più influenzato dal proprio passato rispetto al sito industriale. Questo comportamento deriva principalmente dalla minor capacità di ricambio dell'aria della zona di Via Lissa, che si trova in una posizione centrale del contesto cittadino e, quindi, è più confinata. Diversamente, la zona industriale assume una posizione più favorevole al ricircolo.

Nei grafici presenti in Figura 21 e Figura 22, si nota come la dipendenza temporale viene totalmente catturata dal processo autoregressivo dei residui. Il termine \hat{u}_t risulta non auto correlato, presenta un campo di variazione più piccolo e un andamento simile alla distribuzione Normale.

Figura 21: Grafici dei residui GLS e del termine incorrelato, Via Lissa

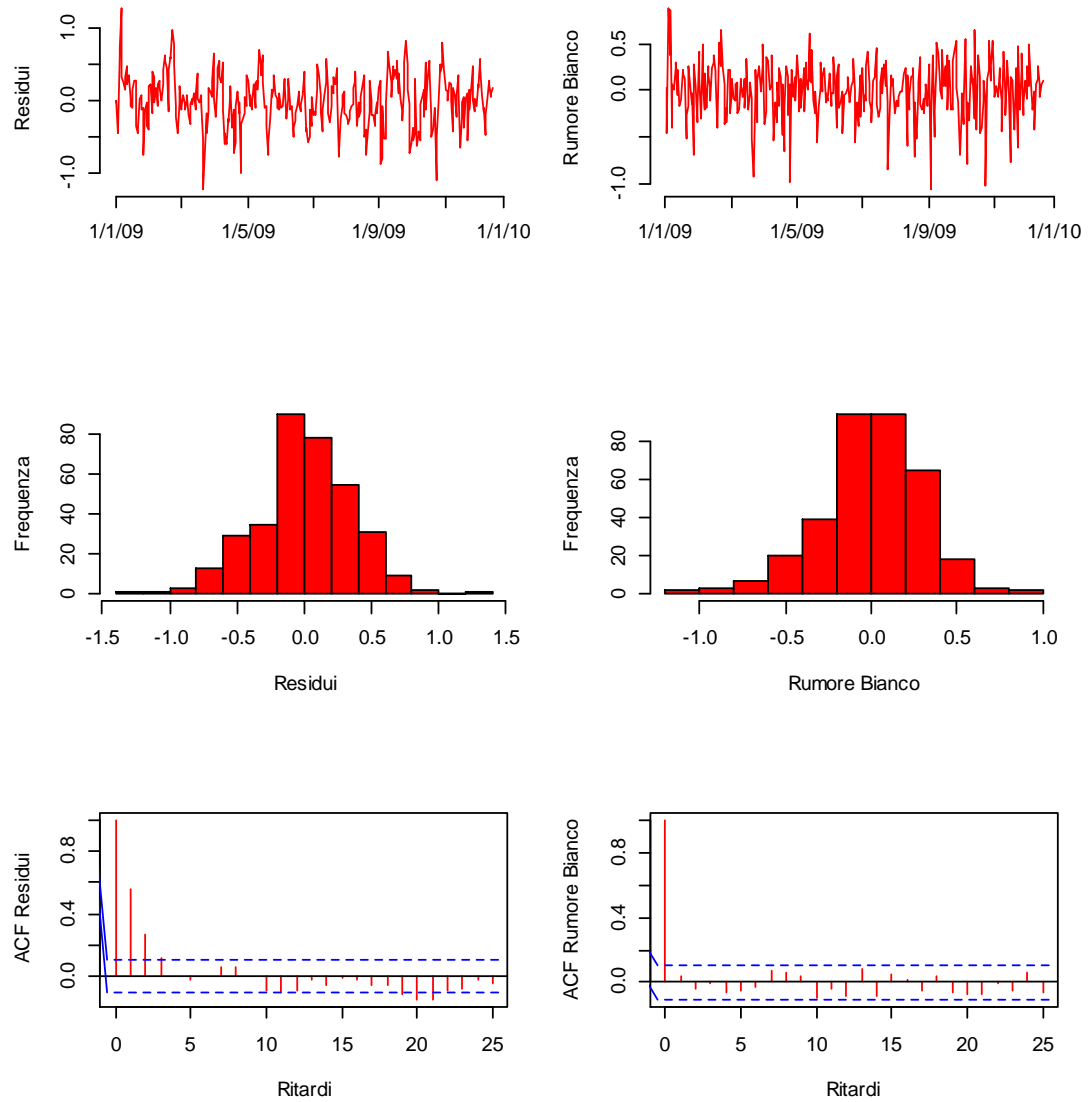
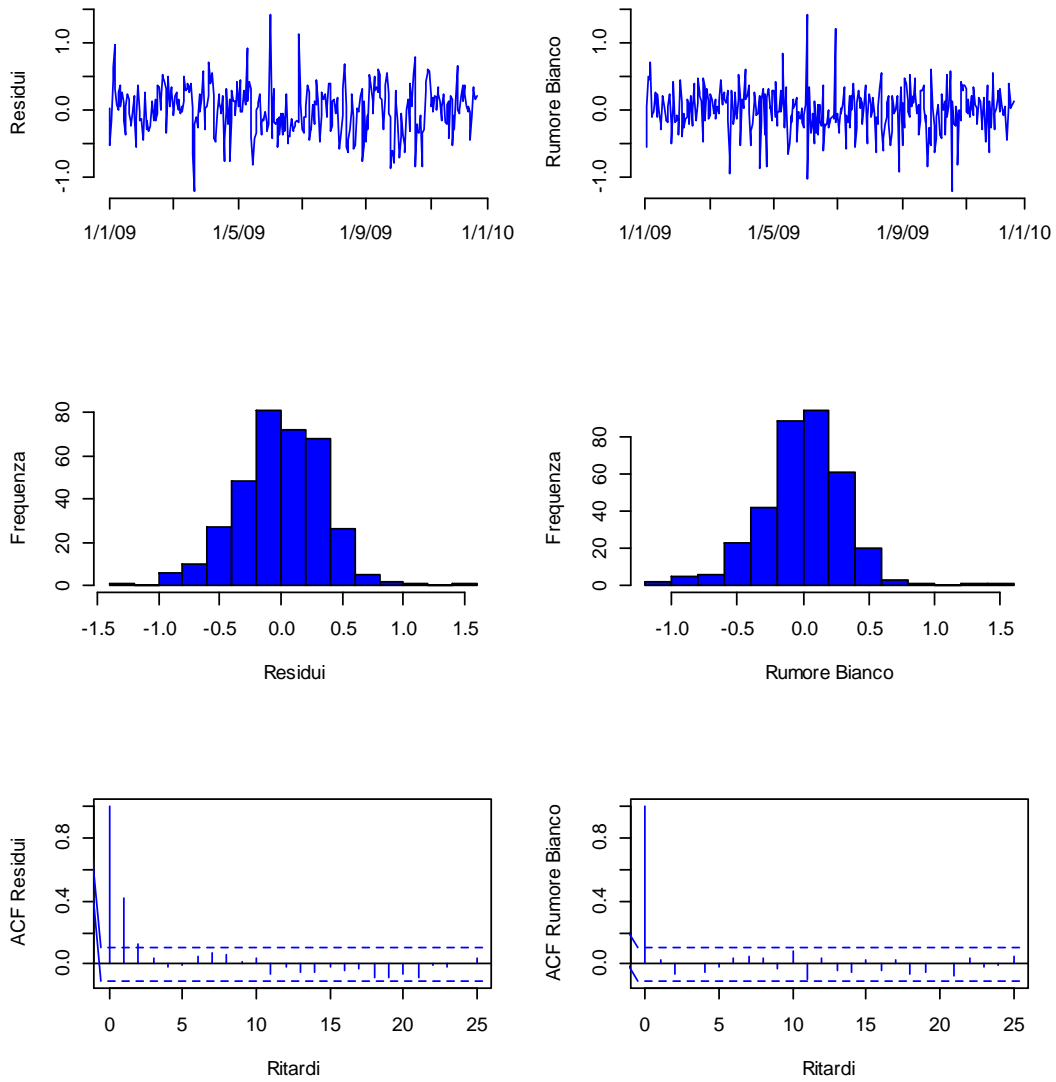


Figura 22: Grafici dei residui GLS e del termine incorrelato, Via Malcontenta



4.5 Modelli additivi

Rispetto alle specificazioni precedenti, totalmente parametriche, è possibile sviluppare una nuova classe di modelli semi-parametrici adatti alla stima della dispersione del $PM_{2.5}$ in atmosfera. Infatti, come già introdotto, i modelli additivi dimostrano una maggiore flessibilità nel controllo dei fattori confondenti non lineari, propri delle variabili meteorologiche, adoperando funzioni di lisciamento affinché ne riducano la variabilità.

Osservando il confronto tra i residui ottenuti dai modelli lineari precedenti e le variabili esplicative (Figura 23 e Figura 24), si riescono ad individuare degli andamenti di natura non esclusivamente causale.

Si osserva un possibile trend quadratico nella temperatura del sito urbano, tendenza che si ritrova anche per la variabile *time* in entrambe le realtà. Nelle variabili riferite alla velocità del vento e all'umidità, invece, sembra venir meno l'ipotesi di omoschedasticità, anche in questo caso per entrambi i siti di campionamento. Infine, per la pressione atmosferica e la concentrazione del diossido di azoto, si nota la presenza di un trend lineare crescente.

Questi andamenti indicano, verosimilmente, una componente sistematica, che la natura lineare dei modelli precedenti non è in grado di cogliere.

Figura 23: Confronto tra i residui e le variabili esplicative, Via Lissa

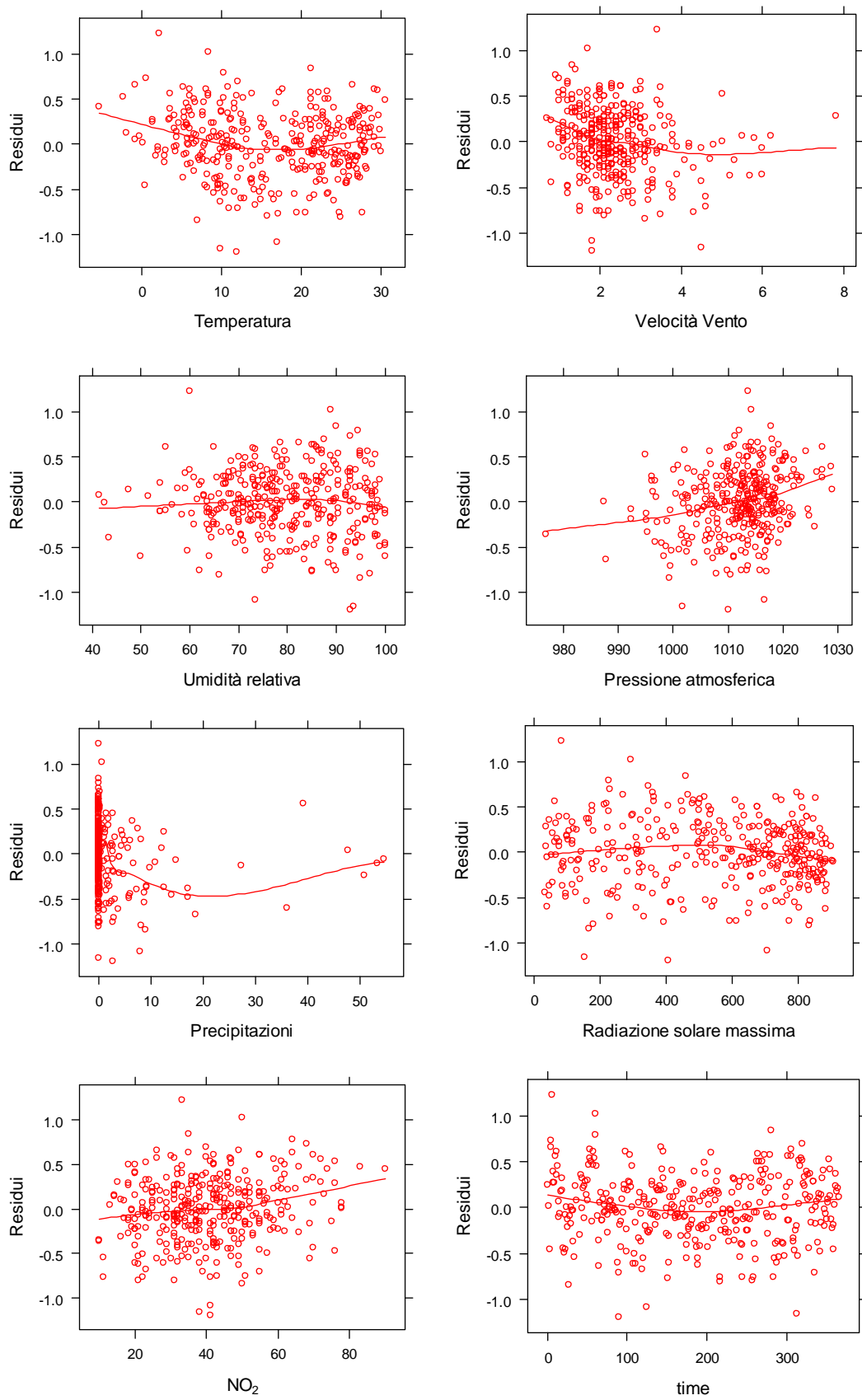
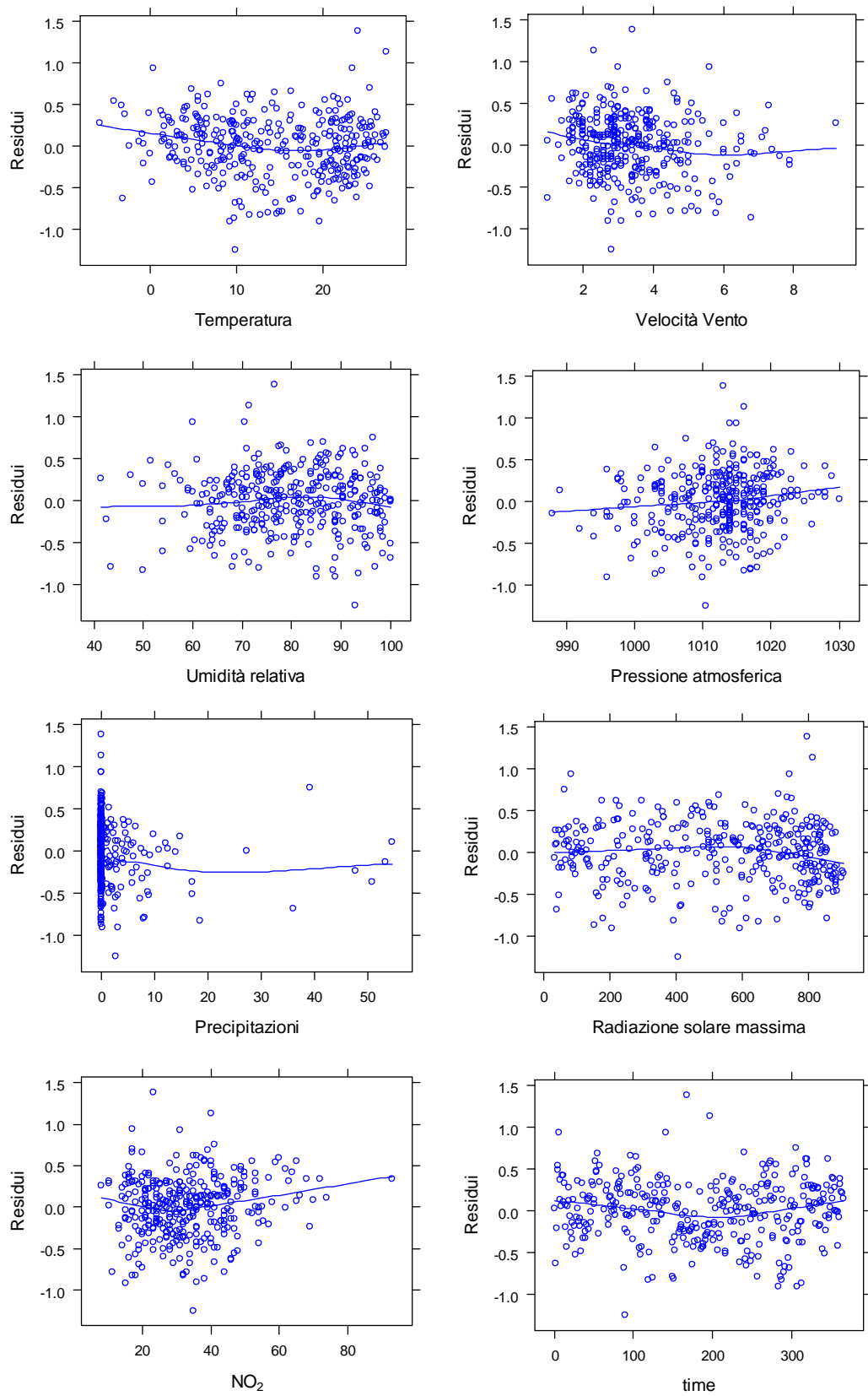


Figura 24: Confronto tra i residui e le variabili esplicative, Via Malcontenta



Assieme alle esplicative considerate nei modelli precedenti, si inserisce tra i regressori anche la variabile *time*, costruita, come già introdotto, sui giorni di campionamento. Ciò permette la stima di un possibile trend non lineare, e di catturare gli eventuali effetti dovuti alla stagionalità.

Il modello additivo che meglio si adatta alla distribuzione degli errori risulta essere nuovamente quello semplice, gaussiano. Si sceglie di mantenere il *link* di identità, continuando a lavorare sul logaritmo delle dispersioni del particolato, per una maggiore semplicità concettuale, e uniformità rispetto allo sviluppo dei precedenti modelli lineari. Sono state comunque valutate, durante l'analisi, differenti famiglie e *link* per entrambi i modelli; il risultato migliore resta, per l'appunto, quello descritto sopra.

Rimane presente, nei residui del modello additivo così formulato, una correlazione seriale positiva in entrambi i siti di rilevamento. Nuovamente risulta poter essere generata da un processo AR(1). Anche in questo caso, è possibile sviluppare una formulazione che tenga conto di questo fatto e rientra nella famiglia dei modelli additivi generalizzati misti, GAMM, presentati nel Capitolo 3.

In Tabella 7 e Tabella 8 sono riportati i gradi di libertà e le significatività delle stime delle funzioni di lisciamiento, nelle due ipotesi di presenza o meno dell'autocorrelazione dei residui. Laddove le variabili entrino nel modello in forma lineare, vengono date le stime dei parametri significativi.

Dopo la selezione di tali esplicative, la formulazione additiva mista riduce il numero delle funzioni di lisciamiento presenti nei modelli a tre, che, sia per il sito urbano che industriale, riguardano le variabili: *time*, temperatura e precipitazioni giornaliere.

Non vengono considerate tra i fattori la direzione del vento, le classi di stabilità e il diossido di zolfo in quanto, anche in questo contesto, non risultano significative al 5% in alcuna formulazione.

Tabella 7: Stime del modello additivo e additivo misto per Via Lissa

Via Lissa					
GAM			GAMM		
Parametri	stima	Signif 5%	Parametri	stima	Signif 5%
Intercetta	2.3732	Si	Intercetta	-4.6035	No
Umidità	0.0103	Si	VV	-0.1729	Si
Smoothers	stima gdl	Signif 5%	Umidità	0.0112	Si
s(time)	7.78	Si	Press	0.0070	Si
s(VV)	8.38	Si	Rsm	-0.0003	Si
s(Temp)	6.24	Si	NO2	0.0094	Si
s(Press)	1.16	Si	Smoothers	stima gdl	Signif 5%
s(Precip)	2.39	Si	s(time)	5.73	Si
s(Rsm)	1.94	Si	s(Temp)	3.89	Si
s(NO2)	1.67	Si	s(Precip)	2.38	Si

Tabella 8: Stime del modello additivo e additivo misto per Via Malcontenta

Via Malcontenta					
GAM			GAMM		
Parametri	stima	Signif 5%	Parametri	stima	Signif 5%
Intercetta	-15.4282	Si	Intercetta	-13.8101	Si
Press	0.0185	Si	VV	-0.1018	Si
Smoothers	stima gdl	Signif 5%	Umidità	0.0086	Si
s(time)	7.73	Si	Press	0.0164	Si
s(VV)	6.46	Si	Rsm	-0.0003	Si
s(Temp)	4.11	Si	NO2	0.0103	Si
s(Umidità)	2.91	Si	Smoothers	stima gdl	Signif 5%
s(Precip)	1.99	No	s(time)	6.03	Si
s(Rsm)	2.80	Si	s(Temp)	3.81	Si
s(NO2)	1.43	Si	s(Precip)	2.15	Si

In Figura 25 e Figura 26 sono riportati i grafici delle stime delle funzioni per la variabile riferita alle precipitazioni, simile in entrambi i siti, che assume una forma inaspettata. Tale andamento non risulta propriamente in linea con la fenomenologia fisica del processo di deposizione umida; è ragionevole che all'aumentare delle precipitazioni aumenti l'azione pulente delle stesse verso il

particolato atmosferico e tutte quelle altre sostanze disciolte. Si nota invece che in entrambi i siti, al crescere delle precipitazioni, si ottiene una prima diminuzione della concentrazione del $PM_{2.5}$ seguita da una successiva stabilizzazione al superamento dei 25mm di pioggia.

Verosimilmente, al protrarsi del fenomeno, diminuisce la quantità di particolato presente nell'aria e, pertanto, l'effetto marginale sarà progressivamente più blando. Inoltre, la rarità di queste eccezionali precipitazioni (solo 7 nel corso del 2009), porta le stime della funzione di liscio ad avere bande di confidenza più ampie.

Figura 25: Funzione di liscio stimata per la variabile Precipitazioni giornaliere, Via Lissa

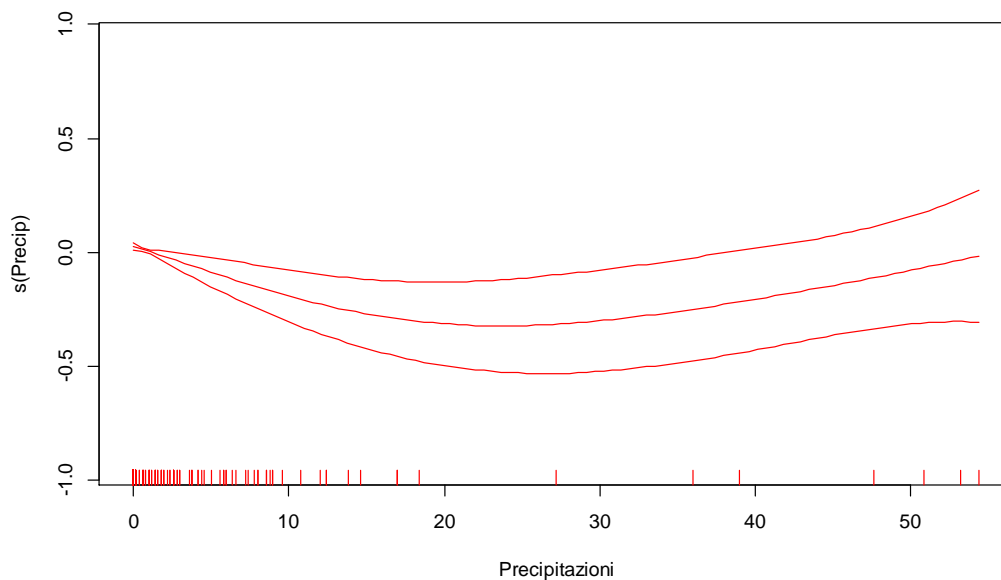
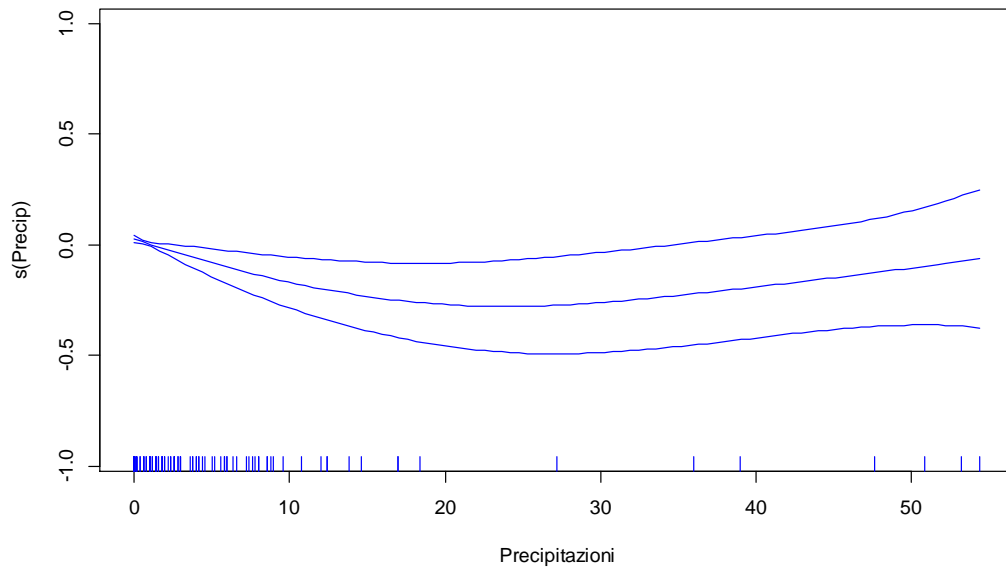


Figura 26: Funzione di lisciamo stimata per la variabile Precipitazioni giornaliere, Via Malcontenta



Le stime delle rimanenti due funzioni di lisciamo, vengono espone, congiuntamente, in Figura 27 per Via Lissa e Figura 28 per Via Malcontenta.

Anche per queste funzioni, l'andamento risulta essere simile tra i due siti, urbano e industriale, in un grafico a forma di paraboloide ellittico.

Nello specifico, il mese di giugno e una temperatura di circa 10°C, apportano il minor aumento della dispersione del PM_{2.5} nell'aria; la concentrazione risulta maggiore durante i mesi invernali, ed a temperature estreme.

Figura 27: Funzioni di lisciamento stimate per variabili *Time* e *Temperatura*, Via Lissa

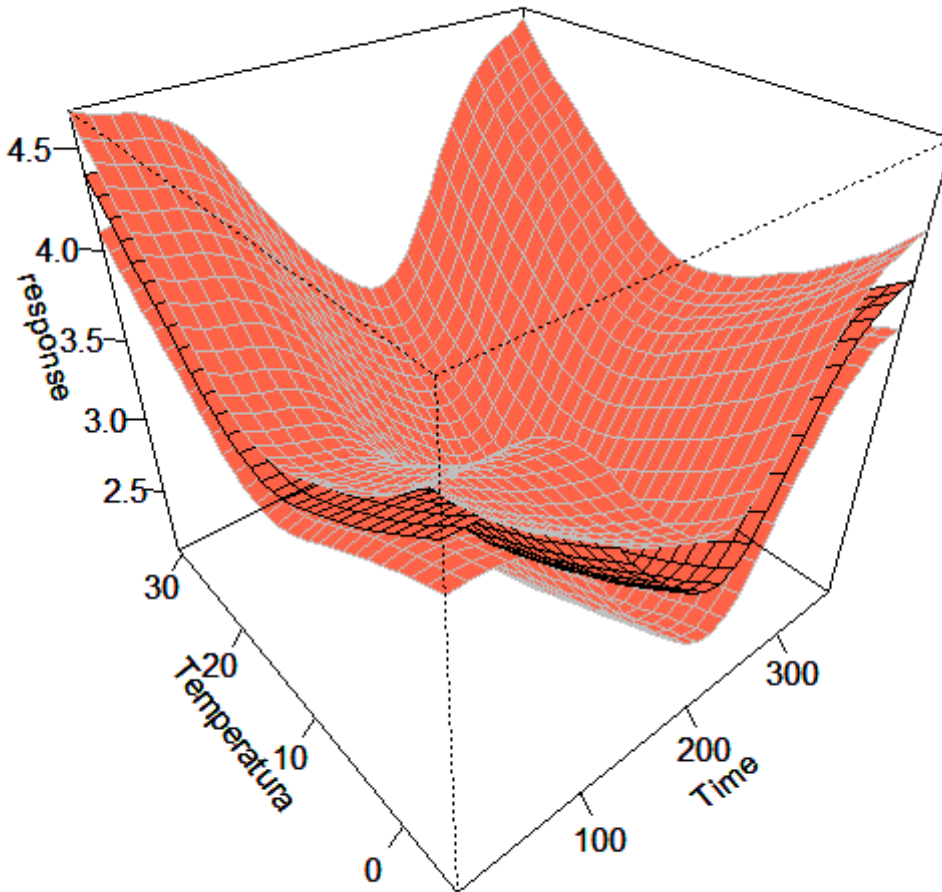
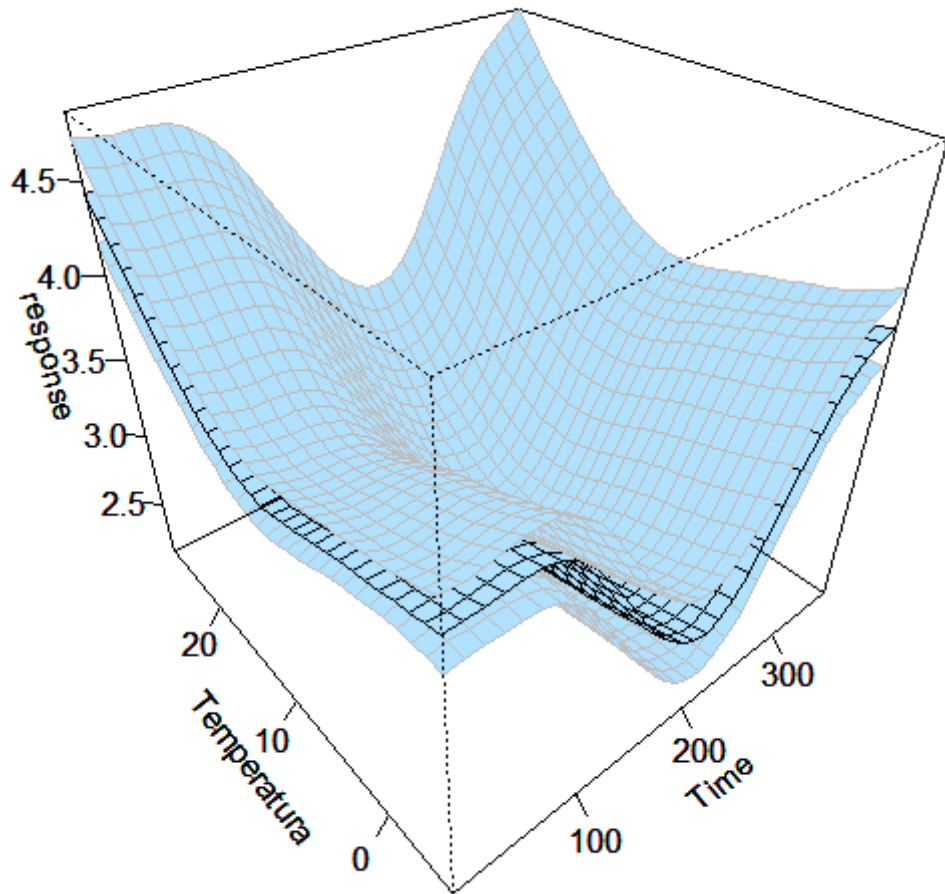


Figura 28: Funzioni di lisciamento stimate per variabili *Time* e *Temperatura*, Via Malcontenta



4.6 Il confronto tra i modelli

Per un confronto sulla capacità previsiva delle varie specificazioni, in Tabella 9 e Tabella 10 sono presenti gli indici di performance riferiti ai modelli statistici stimati, paragonati al modello matematico. Il modello fotochimico considerato è il modello FARM (*Flexible Air quality Regional Model*). Tale modello è stato applicato al dominio considerato. L'output del modello FARM è trasformato nel suo logaritmo per uniformare il confronto.

Tabella 9: Indici di performance per i modelli, Via Lissa

Via Lissa					
	Fotochimico	LM	GLM	GAM	GAMM
ρ	0.830	0.863	0.840	0.894	0.890
MFB	8.912	0.701	1.315	-0.471	-0.809
RMSE	0.463	0.382	0.415	0.322	0.348
NMSE	0.0201	0.0146	0.0172	0.0105	0.0123

Tabella 10: Indici di performance per i modelli, Via Malcontenta

Via Malcontenta					
	Fotochimico	LM	GLM	GAM	GAMM
ρ	0.708	0.879	0.847	0.904	0.889
MFB	11.295	0.668	0.597	-0.523	-0.940
RMSE	0.544	0.340	0.362	0.287	0.309
NMSE	0.0273	0.0117	0.0133	0.0084	0.0098

In entrambi i siti, il modello fotochimico predice la reale dispersione del PM_{2.5} peggio dei modelli statistici più semplici, come ci si poteva aspettare vista la natura dei modelli considerati. Interessante come l'informazione relativa all'autocorrelazione dell'errore porti ad avere, nel complesso, modelli dove la distorsione risulta, anche se di poco, più grande rispetto a quelli costruiti sotto l'ipotesi di indipendenza.

Tabella 11: Confronto MFB per Via Lissa

Via Lissa					
	Fotochimico	LM	GLM	GAM	GAMM
Primavera	-0.539	0.443	-0.051	0.895	-0.494
Estate	15.675	0.474	3.387	-0.461	-0.867
Autunno	10.355	7.255	7.818	2.479	3.682
Inverno	0.167	-11.779	-16.996	-9.302	-10.639
Totale	8.912	0.701	1.315	-0.471	-0.809

Tabella 12: Confronto MFB per Via Malcontenta

Via Malcontenta					
	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.643	-3.956	-5.072	-1.948	-2.567
Estate	16.304	1.924	3.557	0.277	0.629
Autunno	14.233	1.320	0.966	-1.251	-2.206
Inverno	0.206	0.442	-3.289	0.369	-1.005
Totale	11.295	0.668	0.597	-0.523	-0.940

Le formulazioni lineari parametriche, così come il modello matematico, tendono a sovrastimare la concentrazioni di particolato, tendenza invece ribaltata nei modelli additivi dove l'MFB (Tabella 11 e Tabella 12) ne evidenzia una leggera sottostima. Ad ogni modo, in tutti i modelli, si ha maggiore *bias* durante il periodo estivo, al quale corrisponde tuttavia la minor concentrazione di particolato (si vedano le tabelle presenti nell'Allegato 1).

I diagrammi di Taylor, costruiti per i modelli principali visti precedentemente, e suddivisi anche in questo caso per Via Lissa e Via Malcontenta, mettono in luce il reale limite che il modello fotochimico mostra rispetto ai modelli statistici.

Il CTM infatti sembra non riuscire a cogliere la variabilità del processo che genera la dispersione del particolato nel corso

dell'anno, tendendo ad appiattare la concentrazione calcolata. In Figura 29 e Figura 30 si può notare come le previsioni del modello matematico per i due siti si discostano dal *cluster* dei modelli statistici che hanno una deviazione standard molto più simile ai dati osservati.

Figura 29: Diagramma di Taylor per i modelli costruiti sull'intero database, Via Lissa

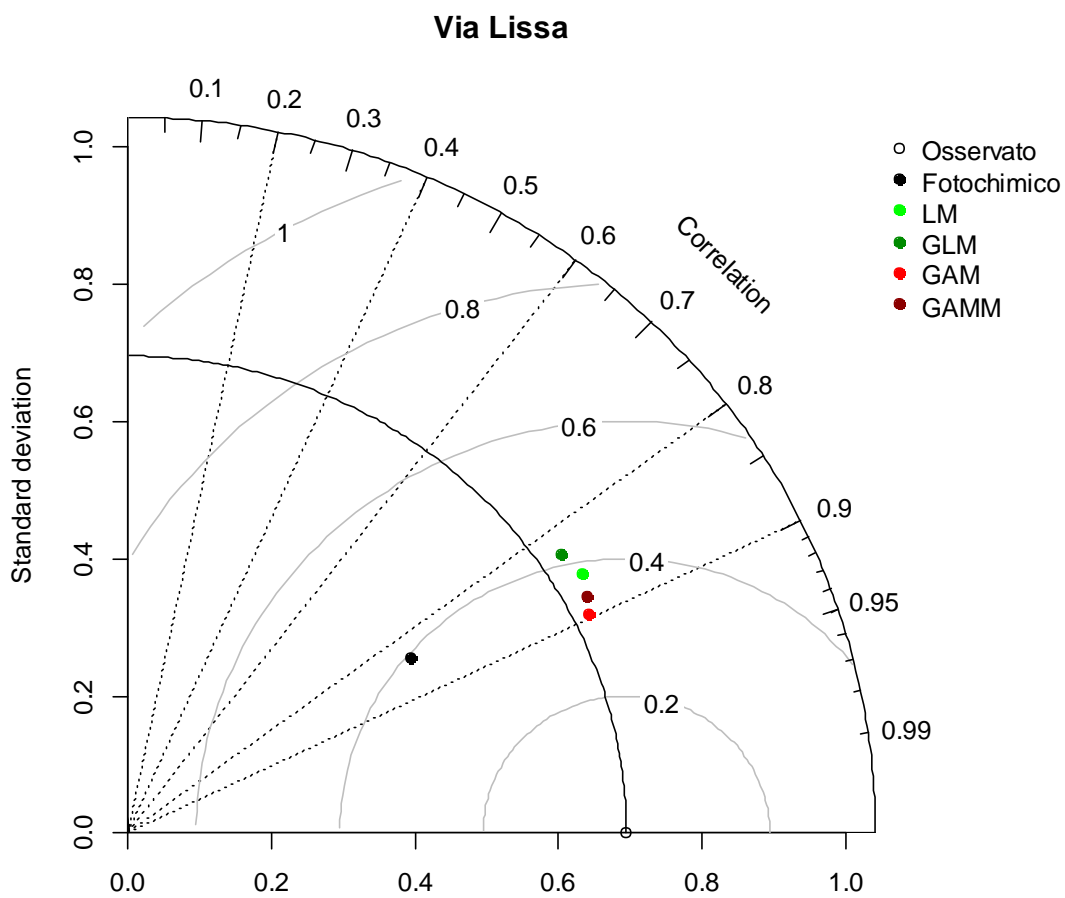
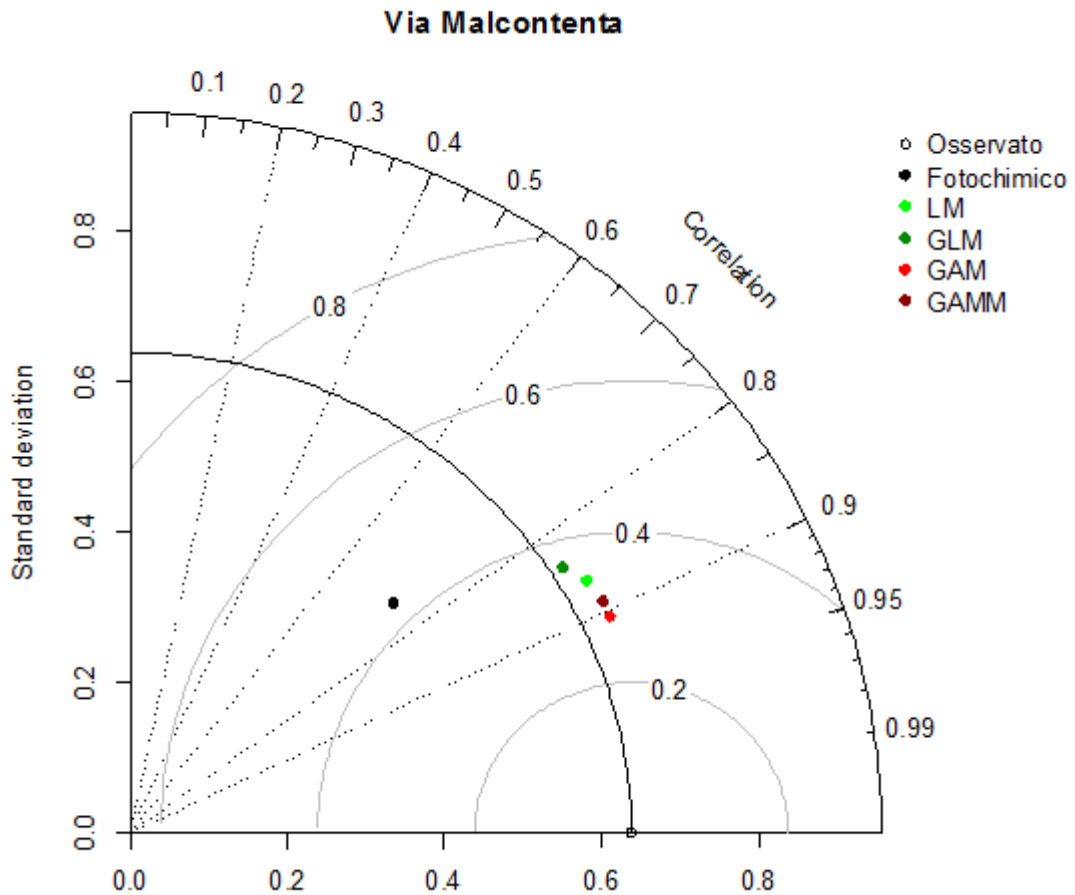


Figura 30: Diagramma di Taylor per i modelli costruiti sull'intero database, Via Malcontenta



Sempre grazie all'immediata diagnostica grafica, possibile attraverso i diagrammi di Taylor presentati, si evince come per Via Lissa il modello matematico ed i modelli statistici stimati abbiano comunque correlazione pressoché simile (tra lo 0.8 e lo 0.9), e le migliori derivanti da questi ultimi derivano dalla loro capacità di cogliere la variabilità del fenomeno. Viceversa, il modello fotochimico di Via Malcontenta, presenta una correlazione molto minore (≈ 0.7) rispetto agli altri modelli statistici, che si dimostrano invece performanti come gli analoghi dell'altro sito (ad ogni modo questo valore rappresenta una buona prestazione se confrontata con la correlazione abituale dei modelli matematici).

Verosimilmente, in questa seconda area sono presenti dei fattori che rendono più critiche le stime del modello matematico ma che diversamente non è così decisivo nella distorsione dei modelli statistici.

Questi risultati possono tuttavia presentarsi normali ed aspettati, trattandosi di un confronto tra un modello prognostico e dei modelli di regressione, a maggior ragione se questi ultimi vengono confrontati sugli stessi dati utilizzati per la loro stima.

Per collaudare più concretamente la capacità previsiva dei modelli statistici, si ritiene opportuno analizzare la loro previsione in un *dataset* di dati svincolato da quello utilizzato per la loro stima.

Nel proseguo dell'analisi si vuole, pertanto, utilizzare lo strumento del *training set-test set*, proprio dell'apprendimento automatico (Mitchell T. M., 1997). Il campione originario viene suddiviso in due *dataset*, ponendo l'attenzione nel garantire un confronto coerente con il modello fotochimico e cercando di evitare particolari problemi legati alla natura stessa del processo. Ad esempio, una suddivisione dell'anno nelle due metà non avrebbe colto il trend stagionale (stimato nei modelli additivi), viceversa un campionamento casuale sarebbe andato a perdere l'informazione riguardante l'autocorrelazione dei residui.

Con l'idea di confrontare le previsioni dei modelli statistici negli stessi periodi in cui si possiedono anche le stime del modello matematico, il *test set* viene composto dai seguenti intervalli:

- primavera: 26/02/2009 - 16/03/2009
- estate: 11/06/2009 - 16/07/2009
- autunno: 05/10/2009 - 31/10/2009
- inverno: 22/12/2009 - 31/12/2009

Di conseguenza il *training test* racchiude le restanti osservazioni dell'anno.

Operativamente, dopo la stima dei modelli statistici sul nuovo insieme di dati (*training set*), si vanno a calcolare i valori previsti di concentrazione del particolato nel *test set*. Il risultato viene infine confrontato con il modello fotochimico.

In Tabella 13 e Tabella 14 sono riportati i nuovi valori per gli indici di performance.

Tabella 13: Indici di performance per i modelli, Via Lissa

Via Lissa					
	Fotochimico	LM	GLM	GAM	GAMM
ρ	0.830	0.847	0.847	0.839	0.873
MFB	8.912	1.020	2.502	1.355	0.399
RMSE	0.463	0.354	0.384	0.387	0.361
NMSE	0.0201	0.0127	0.0148	0.0151	0.0133

Tabella 14: Indici di performance per i modelli, Via Malcontenta

Via Malcontenta					
	Fotochimico	LM	GLM	GAM	GAMM
ρ	0.708	0.792	0.792	0.760	0.819
MFB	11.295	4.497	5.366	1.962	3.028
RMSE	0.544	0.399	0.424	0.409	0.374
NMSE	0.0273	0.0157	0.0176	0.0168	0.0140

I modelli statistici, anche in questo caso, riescono a prevedere meglio il particolato atmosferico, tuttavia, il precedente divario riscontrato con il modello fotochimico viene colmato. Tra tutti, i migliori risultano essere i modelli additivi misti, la cui concentrazione stimata si discosta in minor volume da quella osservata. Questo risultato è frutto della capacità di questi modelli di considerare relazioni non unicamente lineari tra la dispersione del PM_{2.5} e le esplicative, e la capacità di servirsi dell'informazione data dalla natura autocorrelata degli errori.

Per il sito urbano si nota, in Tabella 15, come i modelli statistici tendano a sottostimare nel periodo freddo e sovrastimare durante i periodi estivi e soprattutto autunnali. La formulazione additiva mista è quella che riesce a controllare meglio tale andamento, lasciandolo maggiormente. All'opposto, il modello matematico, come già descritto, sovrastima eccetto che in primavera.

Tabella 15: Confronto MFB per Via Lissa

Via Lissa					
	Fotochimico	LM	GLM	GAM	GAMM
Primavera	-0.539	-3.186	-2.750	-5.582	-4.930
Estate	15.675	2.022	5.053	3.164	0.769
Autunno	10.355	4.479	4.598	8.329	4.818
Inverno	0.167	-1.757	-0.848	-6.443	-0.049
Totale	8.912	1.020	2.502	1.355	0.399

Nel sito industriale, i modelli statistici si comportano pressoché come in Via Lissa, con una sola differenza per il modello additivo misto. Quest'ultimo, infatti, sovrastima anche durante il periodo invernale (Tabella 16). Gli altri indici di performance dimostrano come, nel complesso, esso rimanga il più adeguato nella previsione del PM_{2.5} (si vedano le tabelle di *performance* presenti nell'Allegato 2).

Tabella 16: Confronto MFB per Via Malcontenta

Via Malcontenta					
	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.643	-1.545	-2.912	-4.149	-2.500
Estate	16.304	5.383	8.046	0.976	2.121
Autunno	14.233	8.622	8.477	8.686	7.223
Inverno	0.206	-1.491	-1.318	-4.260	2.172
Totale	11.295	4.497	5.366	1.962	3.028

I diagrammi di Taylor in Figura 31 e Figura 32 riportano, graficamente, l'avvicinamento delle performance dei modelli statistici verso il modello fotochimico.

Figura 31: Diagramma di Taylor per i modelli costruiti sul *training test*, Via Lissa

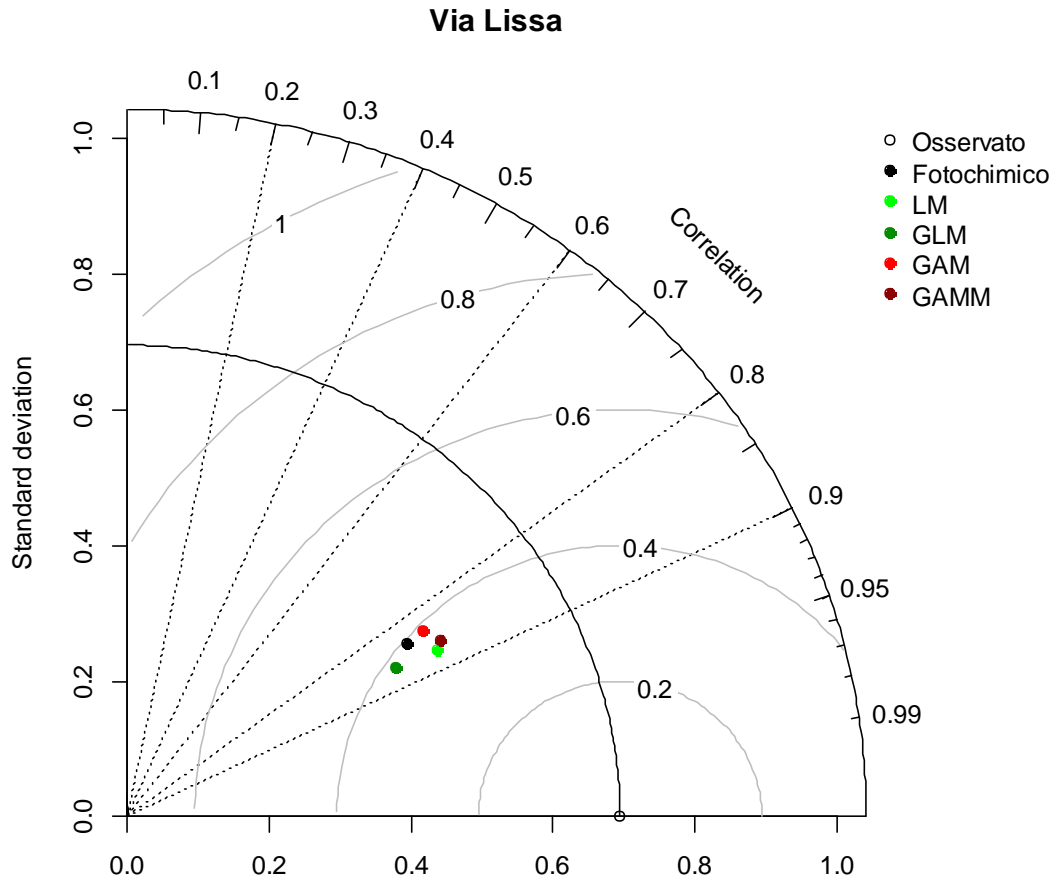
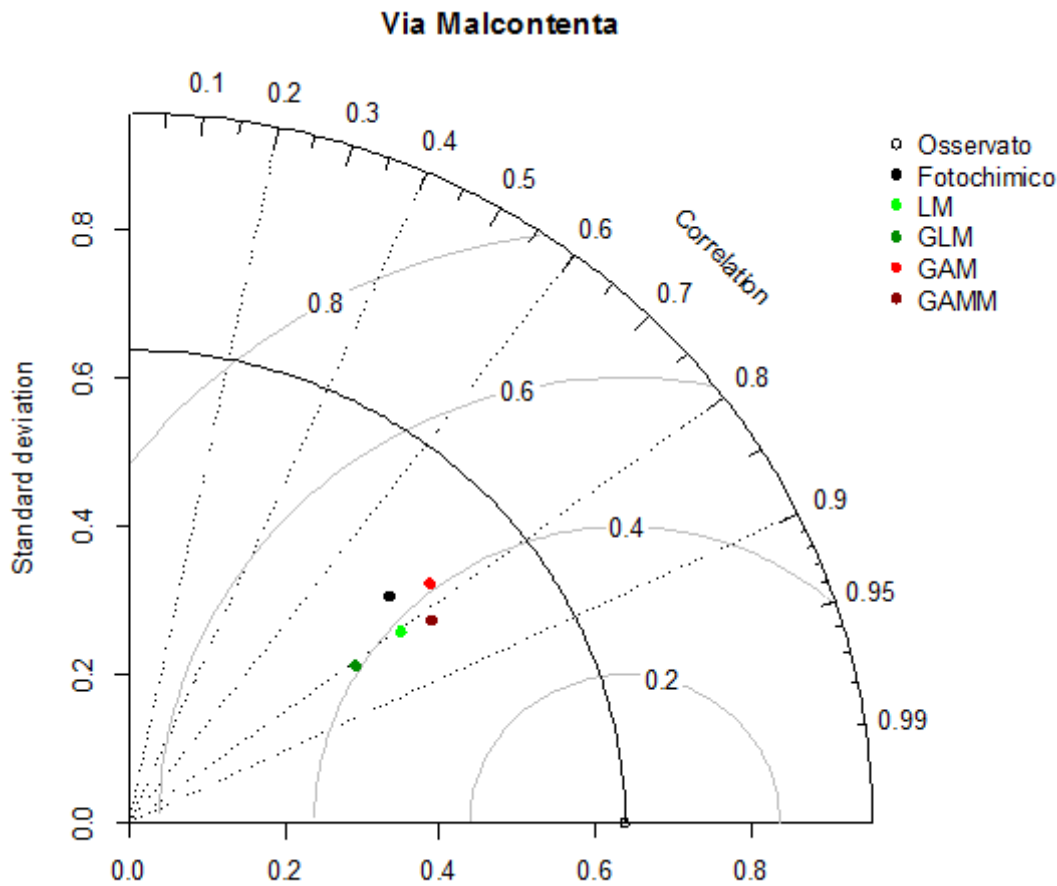


Figura 32: Diagramma di Taylor per i modelli costruiti sul *training test*, Via Malcontenta



In Via Lissa, tutti i modelli risultano vicini tra loro e si osserva quanto commentato precedentemente. È in Via Malcontenta che la situazione si dimostra più dinamica e dissimile. Come già osservato, il GAMM è più vicino ai dati osservati sia per quanto riguarda la correlazione, che per la variabilità rappresentata.

In entrambi i siti si osserva, per tutti i modelli statistici, una deviazione standard minore rispetto ai diagrammi precedenti; la capacità di cogliere la variabilità del fenomeno si riduce alla stregua del modello fotochimico. Peggiori, nella simulazione di questo fattore, sono i modelli lineari generalizzati, che uniformano maggiormente la dispersione.

A conclusione di questa analisi, si osserva una performance minore nella previsione della concentrazione del particolato nel sito industriale, sia del CTM che dei modelli statistici.

Come già anticipato, in questo contesto sono presenti degli elementi non conosciuti che distorcono le stime. I modelli statistici precedenti, costruiti in questo insieme di dati, erano capaci di coglierli. Censurando invece il campione e testando le stime solo successivamente, perdono questa abilità, pur mantenendo, nel complesso, una buona qualità.

4.7 Bias

Entra tra gli scopi di questa analisi la ricerca di quei fattori che causano distorsione (*bias*) nelle previsioni della concentrazione del $PM_{2.5}$ per i due siti del modello fotochimico.

Il *bias*, così definito, assume nel modello di analisi questa struttura, già anticipata precedentemente:

$$PM_{2.5}^{Calc} - PM_{2.5}^{Oss} = \alpha + f_1(x_1) + f_2(x_2) + f_3(x_3) + \dots + \varepsilon$$

Si pone una maggiore attenzione ai gas (NO_2 e SO_2), fortemente legati ai processi di formazione del particolato, che il modello matematico tenta di simulare al meglio.

In questa valutazione non si utilizzano gli stessi valori delle variabili considerati negli studi precedenti. Il modello fotochimico, infatti, impiega una serie di dati in *input* provenienti da un pre-processore meteorologico appositamente calibrato. Quest'ultimo, dai dati osservati nelle 27 stazioni meteorologiche, calcola una previsione in tutto il dominio, simulando l'evoluzione nel tempo e nello spazio di queste variabili e creando così l'*input* del CTM.

Inserendo, tra i regressori, i valori associati a tale *input*, si vuole delineare l'analisi affinché esamini la distorsione provocata dalla parte del modello fotochimico che ha il compito di simulare la formazione, il trasporto e la dispersione del PM_{2.5}, e non la parte antecedente, legata alla previsione della meteorologia.

Viene tolta la variabile relativa alla pressione atmosferica, poiché il processore meteorologico simula, per essa, variazioni non significative nell'anno.

È stata inoltre analizzata la correlazione tra i dati meteorologici osservati e quelli calcolati prima di intraprendere questa analisi. La simulazione si adatta bene ai dati reali e si osserva un coefficiente di Spearman sempre compreso tra 0.9 e 1.

In questo caso, tuttavia, il modello di regressione lineare non riesce a cogliere l'effettivo apporto delle variabili esplicative al *bias* e risulta, in entrambi i casi, non significativo. In Figura 33 e Figura 34 vengono riportati i grafici della distorsione, rispetto ad ogni regressore, nei quali si individuano comunque delle relazioni, nonostante ci sia il disturbo di alcuni *outlier* (specie per il sito urbano).

Figura 33: Relazione tra *bias* e le variabili esplicative, Via Lissa

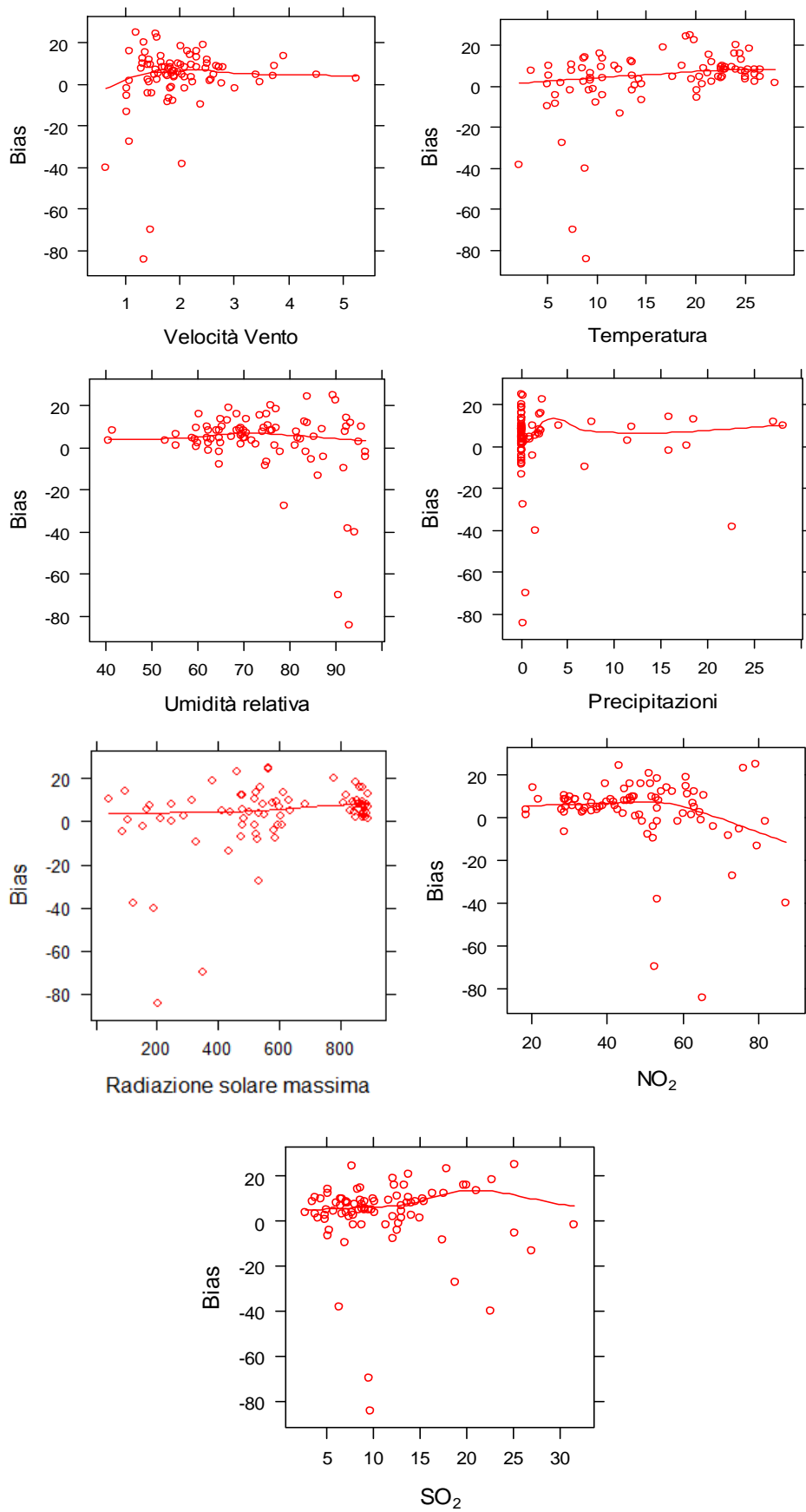
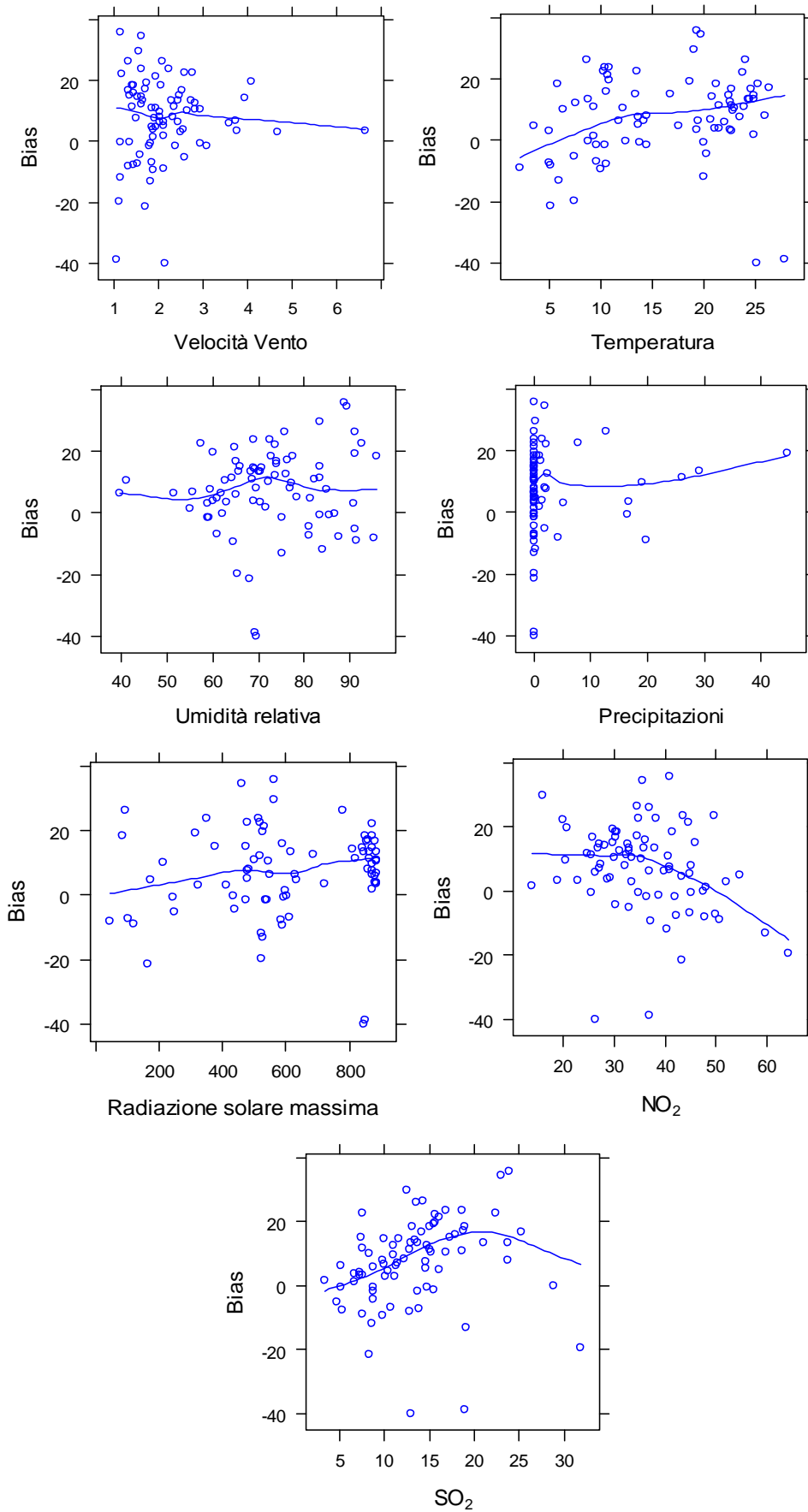


Figura 34: Relazione tra *bias* e le variabili esplicative, Via Malcontenta



Attraverso i modelli additivi, è tuttavia possibile individuare quelle variabili che significativamente causano una maggior differenza tra il modello fotochimico e l'osservato. Anche in questo caso, l'ipotesi di normalità degli errori è, tra le formulazioni possibili, quella che meglio interpreta la loro distribuzione. Nonostante quanto appena definito, le stime di entrambi i modelli tuttavia producono degli errori le cui code si discostano dall'assunzione di normalità, minando la qualità stessa delle stime.

Non si ritiene necessario, per quanto si andrà ad esporre successivamente, presentare le stime derivanti da questi primi modelli. Ciononostante, in Figura 35 e Figura 36, vengono presentati alcuni grafici diagnostici relativi ai residui di entrambe le distorsioni valutate, urbana ed industriale.

Si evince, come la presenza degli *outlier*, individuati anche nei grafici precedenti, possa essere la ragione della criticità riscontrata. Questo a maggior ragione per il sito di Via Lissa, dove i valori anomali sono, in numero, maggiori.

Figura 35: Residui del modello AM con outlier, Via Lissa

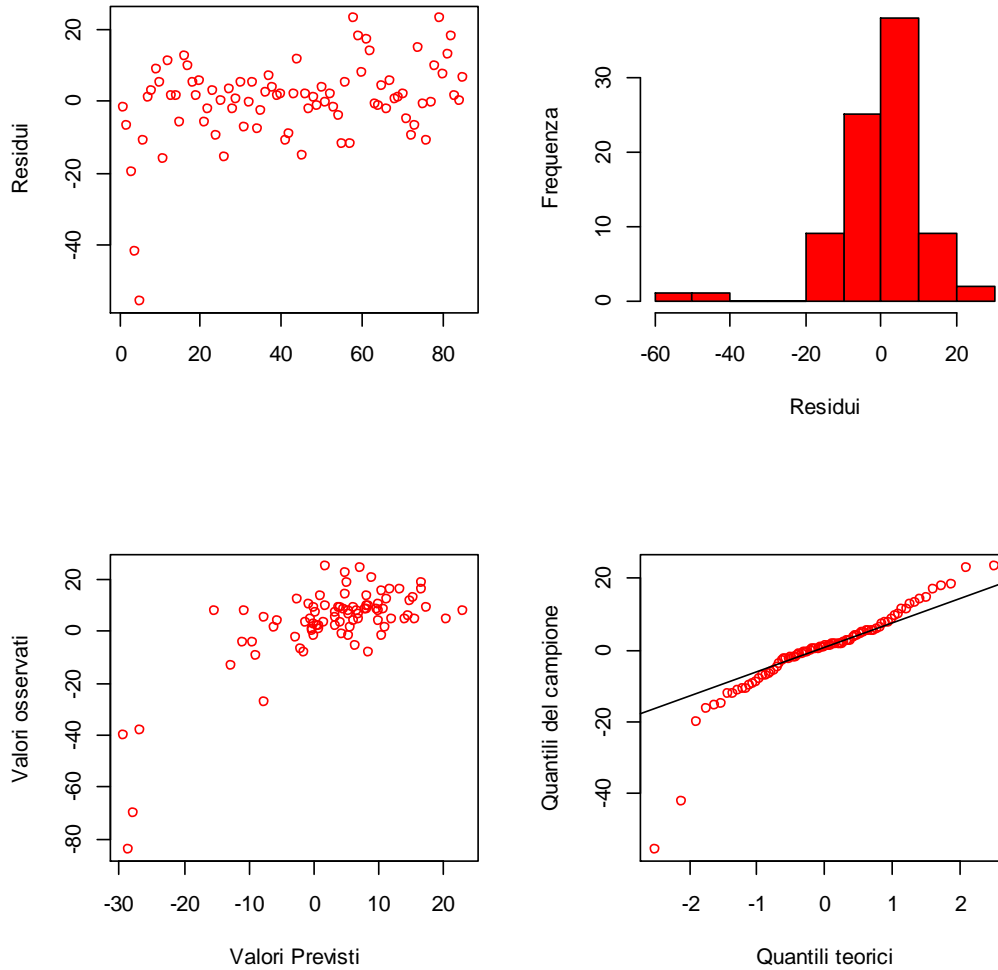
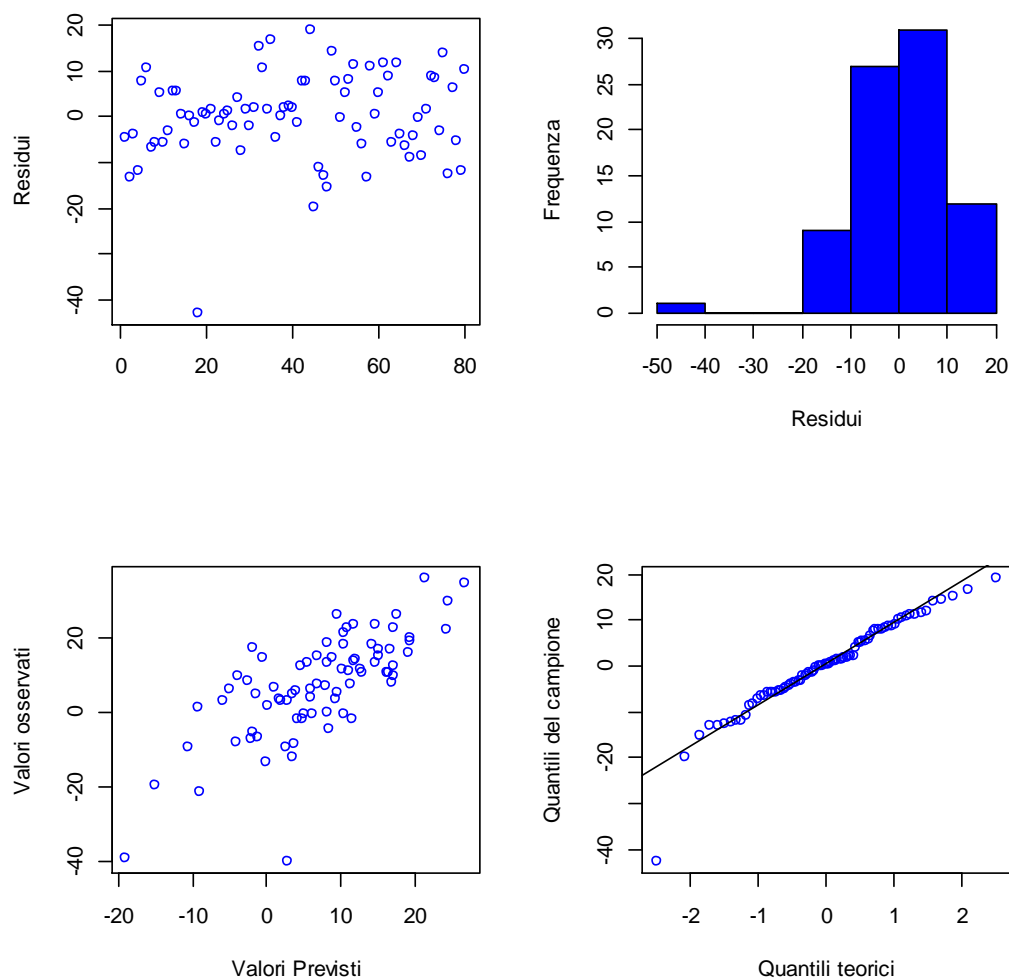


Figura 36: Residui del modello AM con outlier, Via Malcontenta



Per poter comprendere meglio il contributo dei fattori studiati al *bias* nella previsione del modello fotochimico, si decide di togliere i valori anomali riscontrati. Questi non sono correlati, in alcun caso, a particolari valori nell'*input*, né al divario con i fattori meteorologici osservati.

Nello specifico, si sceglie di togliere le osservazioni per le quali la distanza di Cook supera il valore $4/n$, suggerito da Bollen e Jackman (1985), in cui n è il numero di osservazioni a

disposizione. I valori tolti sono 7 nel sito urbano e 3 nel sito industriale (Figura 37 e Figura 38).

In alcune di queste giornate si sono riscontrate sorgenti di emissione atipiche, che non erano state caratterizzate nel modello fotochimico per la loro natura eccezionale.

Figura 37: Distanza di Cook, Via Lissa

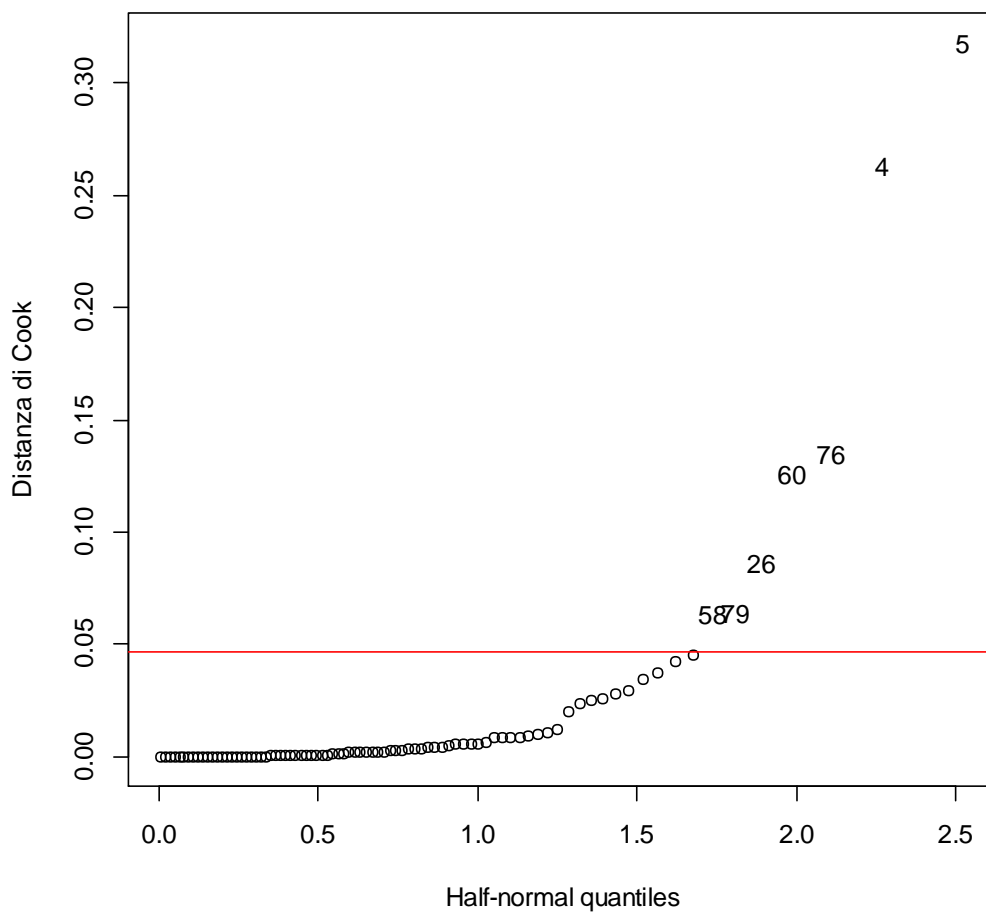
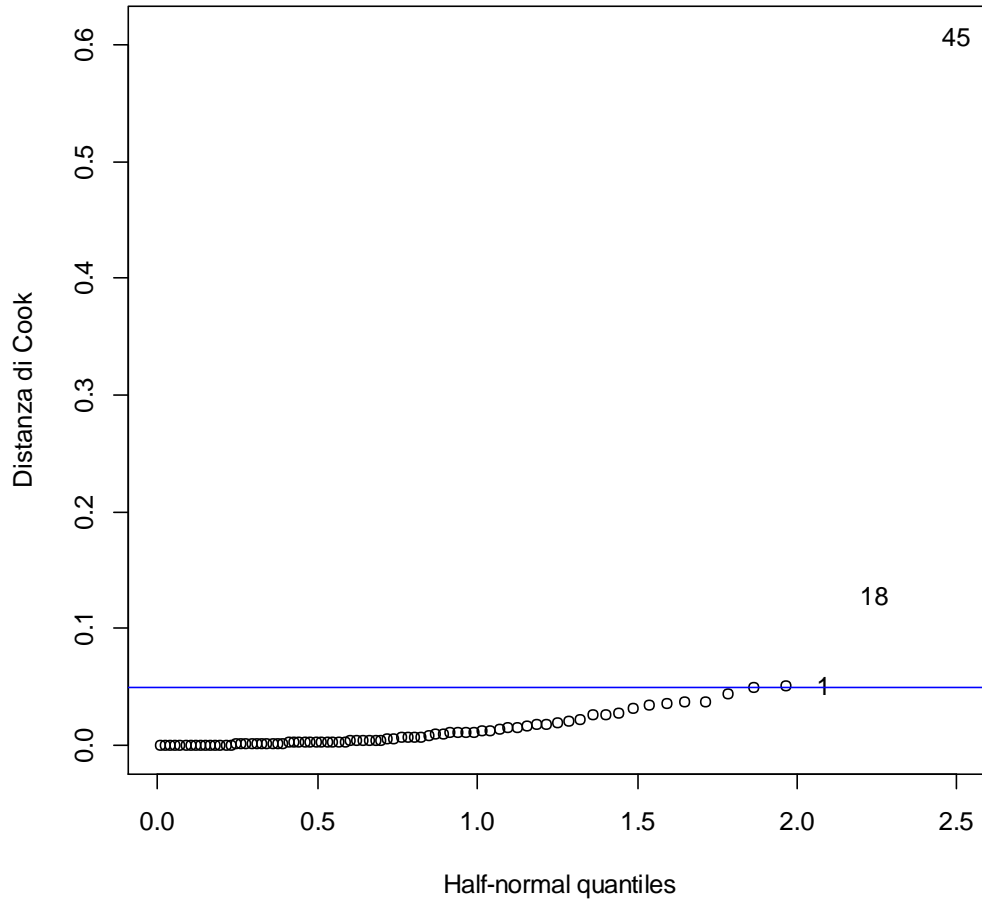


Figura 38: Distanza di Cook, Via Malcontenta



In Figura 39 e Figura 40 viene riportato il comportamento dei residui dei due nuovi modelli additivi, generati omettendo i valori anomali.

Figura 39: Grafici dei residui del modello GAM per bias, Via Lissa

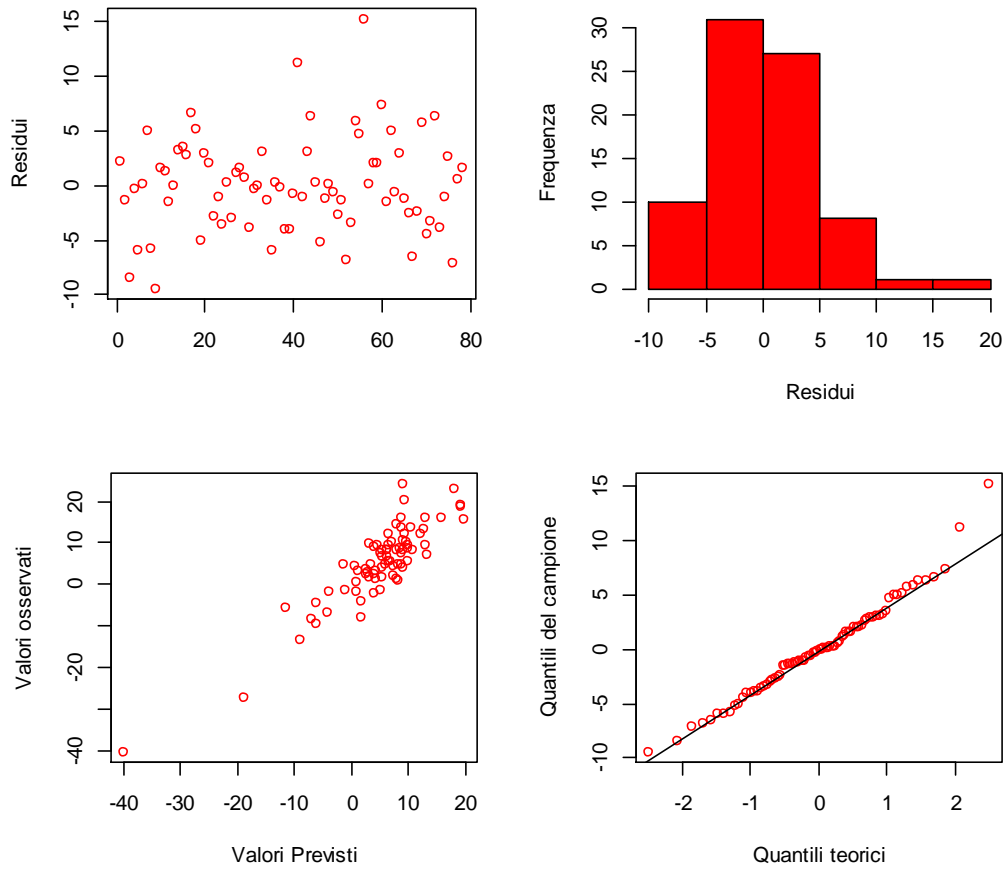
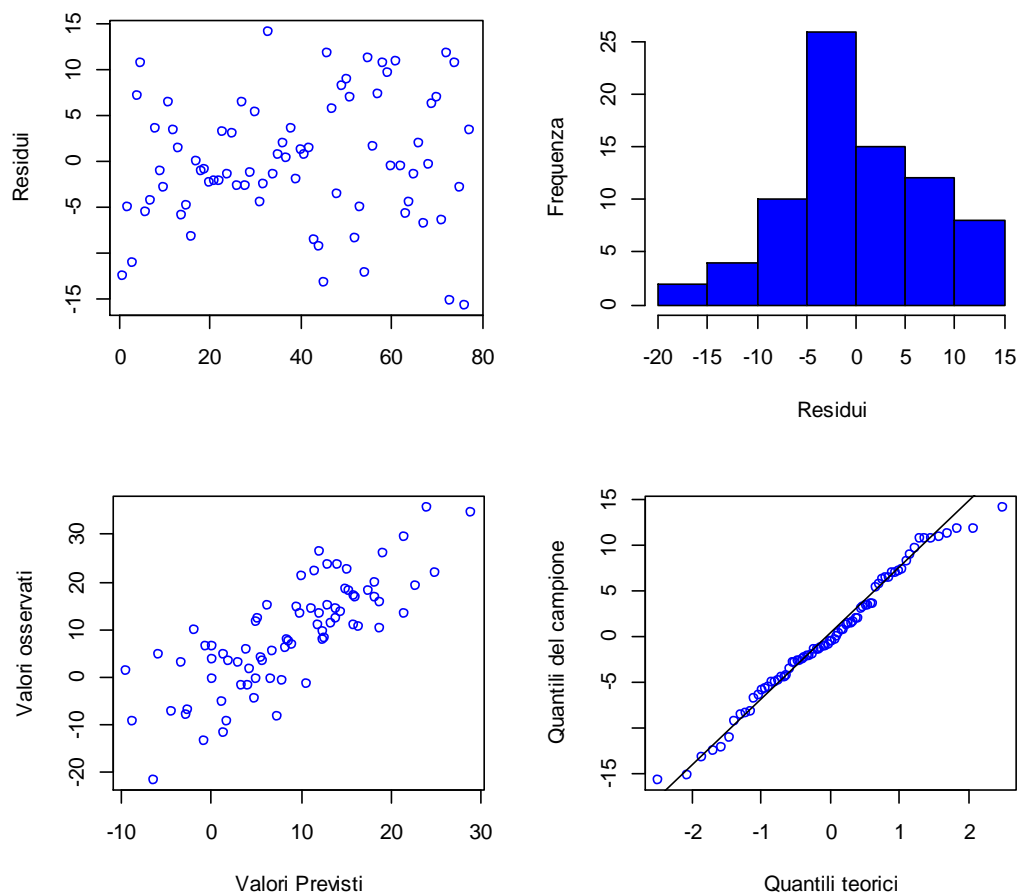


Figura 40: Grafici dei residui del modello GAM per *bias*, Via Malcontenta



Nonostante si trovino ancora dei residui che si discostano leggermente dalla normalità, l'andamento generale migliora sensibilmente in entrambi i siti.

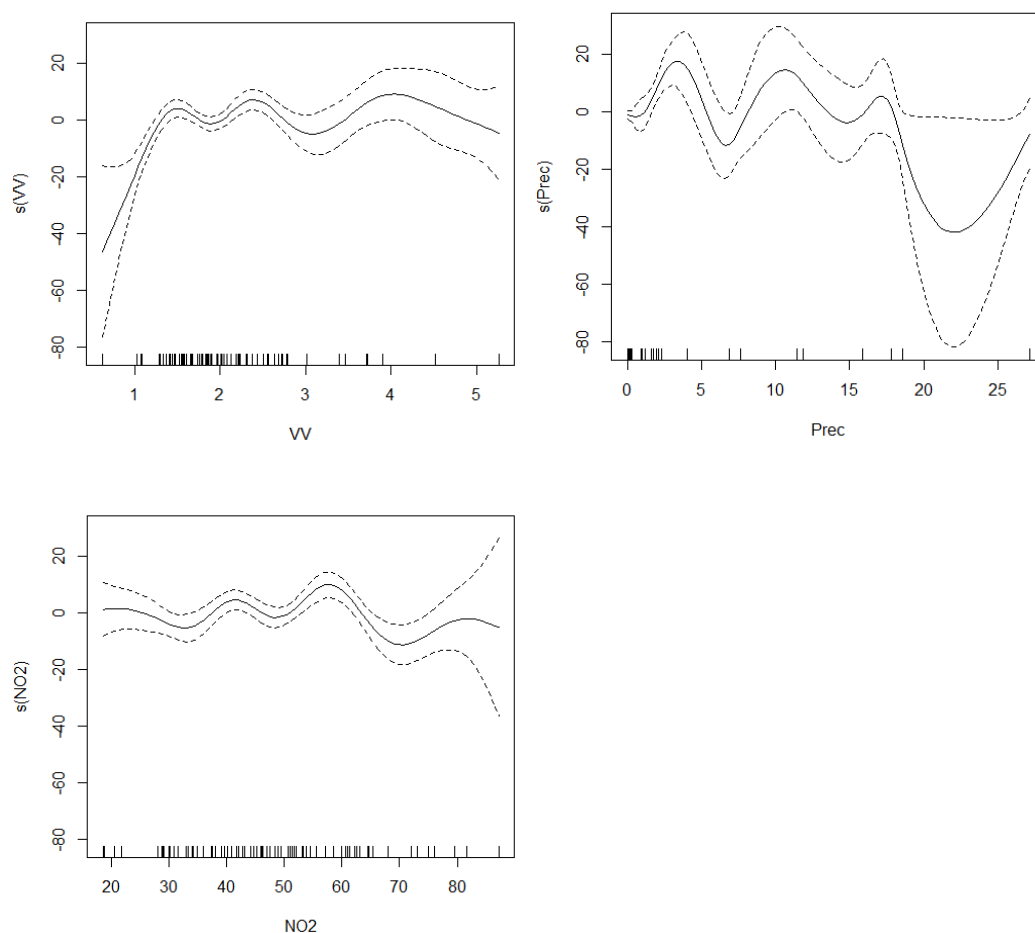
In Tabella 17 vengono infine restituite le stime delle funzioni di lisciamiento. Laddove il fattore entra in forma lineare, viene riportata la stima del suo coefficiente.

Tabella 17: Stime modello GAM per *bias*, Via Lissa e Via Malcontenta

BIAS							
Via Lissa				Via Malcontenta			
Parametri	stima	Deviazione standard	Signif 5%	Parametri	stima	Deviazione standard	Signif 5%
Intercetta	-1.714	8.160	No	Intercetta	36.122	7.384	Si
Temp	0.357	0.206	No	VV	-2.519	1.353	No
Rsm	0.001	0.007	No	NO2	-0.616	0.161	Si
Umidità	-0.094	0.100	No	Smoothers	stima gdl	Test F	Signif 5%
SO2	0.668	0.216	Si	s(Umidità)	4.549	1.733	No
Smoothers	stima gdl	Test F	Signif 5%	s(Temp)	2.09	1.389	No
s(VV)	7.905	4.535	Si	s(Rsm)	4.133	0.873	No
s(Precip)	8.612	3.200	Si	s(Precip)	1.861	1.375	No
s(NO2)	8.353	4.347	Si	s(SO2)	3.692	12.046	Si

Ciò che genera la distorsione appare, nei due siti, di natura discordante. Per il sito urbano, il *bias* è correlato linearmente alla concentrazione del diossido di zolfo (che detiene coefficiente positivo), e alle funzioni relative alla velocità del vento, alle precipitazioni e alla concentrazione del diossido di azoto. Queste ultime non entrano nel modello in forma lineare e la stima delle loro funzioni di lisciamento sono visibili in Figura 41.

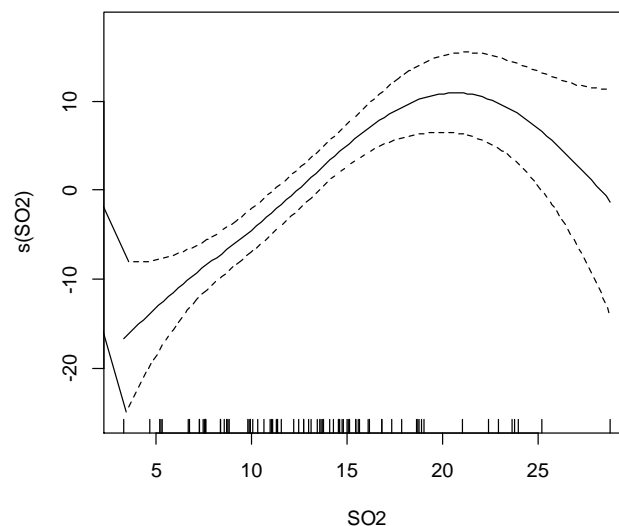
Figura 41: Stima delle funzioni di lisciamento, *bias* Via Lissa



Nonostante la poca frequenza di alcuni valori, si osserva come per una velocità del vento che aumenta da 1 a 3 ms^{-1} , la tendenza a sovrastimare aumenta. Lo stesso all'aumentare dei millimetri di pioggia per le piccole precipitazioni; passati i 4mm la stima della funzione diventa poco affidabile per la rarità dei casi. La funzione stimata per il diossido di azoto presenta un effetto oscillante simile a quella delle precipitazioni. Tendenzialmente per valori attorno a 40 o 60 $\mu g m^{-3}$, si tende ad una sovrastima, per i valori estremi il modello distorce in negativo ed infine, per concentrazioni comprese tra i 40 e i 60 $\mu g m^{-3}$, l'apporto al *bias* è pressoché nullo.

Anche nella distorsione per il sito industriale è rilevante l'NO₂, che però rientra nella formulazione in forma lineare e con coefficiente negativo. In questo caso l'intercetta assume significatività e segna una tendenza a sovrastimare la concentrazione del particolato. Non sono significative altre funzioni di liscio se non quella riferita al diossido di zolfo, dove in Figura 42 viene riportato il suo andamento. Riscontriamo, nei valori più frequenti di questa variabile, una stima della funzione di liscio praticamente lineare e crescente. Per valori piccoli di concentrazione il modello fotochimico sottopredice, mentre, superato il valore di circa $14 \mu\text{g m}^{-3}$, l'andamento diventa positivo.

Figura 42: Stima della funzione di liscio per l'SO₂, bias Via Malcontenta



L'incapacità della regressione lineare di riuscire a cogliere questi aspetti è dovuta alla natura non lineare delle relazioni tra la variabile risposta e le esplicative.

La possibilità di utilizzare lo strumento dei modelli additivi nello studio della distorsione del modello fotochimico ha permesso invece di identificare, per ogni variabile, il suo rapporto rispetto al *bias*, caratterizzandone la significatività.

Conclusioni

Questo lavoro è stato svolto con l'obiettivo di utilizzare modelli di regressione per meglio comprendere la capacità di simulazione di un modello fotochimico utilizzato per lo studio del particolato atmosferico nella zona orientale della pianura padano-veneta. Sono stati considerati due siti in cui è stato effettuato il rilevamento ed il confronto della concentrazione del $PM_{2.5}$, Via Lissa e Via Malcontenta, che rappresentavano, rispettivamente, un contesto urbano ed uno industriale.

Si è ritenuto opportuno, per comprendere quanto questi temi siano di cruciale rilevanza in questi anni, introdurre un capitolo atto a descrivere, sinteticamente, il problema dell'inquinamento atmosferico e, nello specifico, del particolato. A questo è seguita una descrizione dei principali modelli matematici impiegati in letteratura per lo studio del fenomeno.

Una principale criticità risiede nel fatto che spesso l'analista si trova ad utilizzare degli indicatori sintetici di *performance*, che tuttavia non gli permettono dare una comprensione generale del fenomeno simulato.

Dall'analisi dei modelli lineari si è osservato come la concentrazione del particolato è in relazione con alcune specifiche variabili meteorologiche osservate, e con la concentrazione del diossido di azoto. Si sono riscontrate tuttavia alcune differenze tra questi rapporti, caratteristiche del sito di rilevamento.

Infine è stato osservato un andamento autoregressivo negli errori di entrambi i modelli, più persistente nel sito urbano dove c'è minor capacità di ricambio dell'aria essendo più confinato.

Successivamente, la dispersione del particolato atmosferico è stata stimata attraverso i modelli additivi che, soppesando l'apporto delle singole variabili, hanno stimato delle funzioni di liscio capaci di cogliere relazioni non lineari. Lo sviluppo di questi modelli, ha permesso di osservare come ci sia una forte componente stagionale alla base della concentrazione del particolato, e come questo continui a dipendere dal suo passato, sottolineando il mancato riciclo dell'aria che caratterizza il territorio di Venezia-Mestre.

Tra i fattori che più incidono sull'aumento del $PM_{2.5}$ in atmosfera, si trovano: bassa velocità del vento, alte concentrazioni di NO_2 , assenza di precipitazioni e temperature estreme, alte o basse.

Il successivo confronto tra le previsioni del modello fotochimico ed i modelli statistici sviluppati e poi testati attraverso il *training test e test set*, hanno permesso di caratterizzare, in maniera più adeguata, le differenze nelle simulazioni dei due siti. Il sito urbano si mostra più facile da simulare, con un indice di correlazione di Spearman pari a 0.83 per il CTM e 0.87 per il modello GAMM. Il modello matematico predice, di fatto, relativamente bene la dispersione del particolato, presentando una leggera sovrastima che il modello additivo riesce a ridurre. Nel sito industriale la situazione è differente, la correlazione diminuisce a 0.71 per il fotochimico, e 0.82 per il GAMM.

La differenza di correlazione tra i due siti è dovuta a diversi fattori, quali la distribuzione delle sorgenti di emissione e la particolarità meteorologica del sito che entrambi i modelli non riescono a rappresentare bene. Nel caso del modello additivo, per una tendenza ad omogeneizzare. Diversamente, per il modello matematico, è probabilmente dovuta all'impossibilità di descrivere perfettamente le sorgenti emmissive.

Per meglio comprendere quali potessero essere i fattori legati alla distorsione del modello fotochimico, si è infine implementata un'analisi sul *bias*, la differenza tra i valori predetti e quelli

osservati. Per questo esame, sviluppato attraverso i modelli additivi, sono stati utilizzati non più i valori osservati, ma l'*input* del modello additivo, calcolato attraverso un pre-processore meteorologico. Con questa scelta, si è cercato di porre l'attenzione sulla sola simulazione operata dal CTM.

Seppur con peso diverso, sono risultati significativi, sia nel sito urbano che industriale, le concentrazioni del diossido di zolfo e di azoto, entrambe derivanti principalmente dalle combustioni fossili. Inoltre, se nel sito urbano si presentavano significative anche la velocità del vento e le precipitazioni, per il sito industriale si osserva invece una forte tendenza alla sovrastima, con un'intercetta dal valore di 36.12.

In quest'ultima analisi si sono osservati alcuni valori anomali, che sono stati tolti per non compromettere le stime. Tali valori sono riconducibili, verosimilmente, a delle sorgenti emissive straordinarie, poiché si riconducono, principalmente, ad effetti di forte sottostima in condizioni climatiche stabili.

Rispetto agli obiettivi preposti, l'analisi ha permesso di individuare e misurare i fattori che risultano essere più correlati alla concentrazione del particolato nei due siti di campionamento, non esaminando esclusivamente delle relazioni lineari. Inoltre la comparazione della capacità prognostica, grazie anche all'immediatezza grafica dei diagrammi di Taylor, ha evidenziato la buona capacità del modello fotochimico nella simulazione anche in zone più critiche e dinamiche, come il sito urbano.

Lo studio sulla distorsione ha infine permesso di apprendere come siano di notevole importanza il diossido di zolfo e di azoto all'interno di quei processi di formazione e sviluppo del particolato atmosferico, rispondendo all'interrogativo sorto in partenza su cosa causasse, all'interno del CTM, una più grande difformità nelle previsioni rispetto alla concentrazione realmente osservata.

I risultati hanno chiaramente confermato gli esiti riscontrati anche con analisi più semplici. L'aspetto innovativo risiede nella velocità di individuare contemporaneamente le variabili coinvolte nel fenomeno, consentendo al modellista di ricorrere meno ad interpretazioni personali e soggettive.

Un aspetto inoltre rilevante è la possibilità di individuare il legame tra i precursori, le variabili meteo e le concentrazioni predette dal modello matematico. Questo infatti, oltre a spiegare la distorsione del modello come qui evidenziato, diventa ancora più importante nell'analisi della predizione delle componenti del particolato. Quest'aspetto, qui non considerato perché richiederebbe ulteriori approfondimenti, verrà sviluppato in futuro. Migliorie dell'analisi svolta possono riguardare inoltre uno studio sul superamento di soglia, con l'obiettivo di prevedere i giorni in cui possa essere più probabile oltrepassare i limiti delle normative vigenti, ed agire in prevenzione.

Allegati

Allegato 1: indici per il confronto dei modelli calcolati sull'intero dataset

Tabella 18: Coefficienti di correlazione di Spearman per Via Lissa e Via Malcontenta

Via Lissa					
ρ	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.705	0.898	0.914	0.907	0.933
Estate	0.728	0.832	0.858	0.841	0.842
Autunno	0.764	0.893	0.860	0.909	0.846
Inverno	0.406	0.188	0.188	0.285	0.285
Totale	0.830	0.863	0.840	0.894	0.890

Via Malcontenta					
ρ	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.708	0.646	0.719	0.589	0.707
Estate	0.709	0.726	0.708	0.762	0.749
Autunno	0.699	0.757	0.732	0.776	0.726
Inverno	0.657	0.964	0.976	0.930	0.964
Totale	0.708	0.879	0.847	0.904	0.889

Tabella 19: Mean fractional bias per Via Lissa e Via Malcontenta

Via Lissa					
MFB	Fotochimico	LM	GLM	GAM	GAMM
Primavera	-0.539	0.443	-0.051	0.895	-0.494
Estate	15.675	0.474	3.387	-0.461	-0.867
Autunno	10.355	7.255	7.818	2.479	3.682
Inverno	0.167	-11.779	-16.996	-9.302	-10.639
Totale	8.912	0.701	1.315	-0.471	-0.809

Via Malcontenta					
MFB	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.643	-3.956	-5.072	-1.948	-2.567
Estate	16.304	1.924	3.557	0.277	0.629
Autunno	14.233	1.320	0.966	-1.251	-2.206
Inverno	0.206	0.442	-3.289	0.369	-1.005
Totale	11.295	0.668	0.597	-0.523	-0.940

Tabella 20: Root mean square error per Via Lissa e Via Malcontenta

Via Lissa					
RMSE	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.426	0.254	0.278	0.257	0.270
Estate	0.494	0.288	0.295	0.232	0.228
Autunno	0.496	0.408	0.442	0.305	0.380
Inverno	0.331	0.694	0.777	0.611	0.636
Totale	0.463	0.382	0.415	0.322	0.348

Via Malcontenta					
RMSE	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.250	0.253	0.263	0.229	0.210
Estate	0.608	0.331	0.346	0.242	0.244
Autunno	0.622	0.412	0.447	0.373	0.424
Inverno	0.328	0.253	0.271	0.233	0.252
Totale	0.544	0.340	0.362	0.287	0.309

Tabella 21: Normalized mean square error per Via Lissa e Via Malcontenta

Via Lissa					
NMSE	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.0134	0.0047	0.0056	0.0048	0.0053
Estate	0.0298	0.0117	0.0119	0.0077	0.0075
Autunno	0.0221	0.0152	0.0178	0.0089	0.0137
Inverno	0.0078	0.0382	0.0505	0.0290	0.0318
Totale	0.0201	0.0146	0.0172	0.0105	0.0123

Via Malcontenta					
NMSE	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.0046	0.0049	0.0054	0.0040	0.0034
Estate	0.0423	0.0143	0.0154	0.0078	0.0079
Autunno	0.0340	0.0167	0.0197	0.0141	0.0183
Inverno	0.0077	0.0045	0.0054	0.0038	0.0046
Totale	0.0273	0.0117	0.0133	0.0084	0.0098

Allegato 2: indici per il confronto dei modelli calcolati sul test set

Tabella 22: Coefficienti di correlazione di Spearman per Via Lissa e Via Malcontenta

Via Lissa					
ρ	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.705	0.857	0.879	0.771	0.853
Estate	0.728	0.502	0.465	0.654	0.697
Autunno	0.764	0.702	0.675	0.738	0.704
Inverno	0.406	0.596	0.818	0.644	0.665
Totale	0.830	0.847	0.847	0.839	0.873

Via Malcontenta					
ρ	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.708	0.742	0.740	0.667	0.754
Estate	0.709	0.652	0.610	0.708	0.733
Autunno	0.699	0.784	0.787	0.562	0.661
Inverno	0.657	0.930	0.906	0.906	0.903
Totale	0.708	0.792	0.792	0.760	0.819

Tabella 23: Mean fractional bias per Via Lissa e Via Malcontenta

Via Lissa					
MFB	Fotochimico	LM	GLM	GAM	GAMM
Primavera	-0.539	-3.186	-2.750	-5.582	-4.930
Estate	15.675	2.022	5.053	3.164	0.769
Autunno	10.355	4.479	4.598	8.329	4.818
Inverno	0.167	-1.757	-0.848	-6.443	-0.049
Totale	8.912	1.020	2.502	1.355	0.399

Via Malcontenta					
MFB	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.643	-1.545	-2.912	-4.149	-2.500
Estate	16.304	5.383	8.046	0.976	2.121
Autunno	14.233	8.622	8.477	8.686	7.223
Inverno	0.206	-1.491	-1.318	-4.260	2.172
Totale	11.295	4.497	5.366	1.962	3.028

Tabella 24: Root mean square error per Via Lissa e Via Malcontenta

Via Lissa					
RMSE	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.426	0.378	0.446	0.452	0.469
Estate	0.494	0.275	0.301	0.260	0.228
Autunno	0.496	0.466	0.487	0.499	0.454
Inverno	0.331	0.269	0.251	0.353	0.276
Totale	0.463	0.354	0.384	0.387	0.361

Via Malcontenta					
RMSE	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.250	0.211	0.236	0.247	0.218
Estate	0.608	0.422	0.455	0.359	0.361
Autunno	0.622	0.487	0.503	0.555	0.482
Inverno	0.328	0.198	0.237	0.261	0.232
Totale	0.544	0.399	0.424	0.409	0.374

Tabella 25: Normalized mean square error per Via Lissa e Via Malcontenta

Via Lissa					
NMSE	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.0134	0.0109	0.0151	0.0159	0.0171
Estate	0.0298	0.0106	0.0124	0.0094	0.0074
Autunno	0.0221	0.0208	0.0226	0.0228	0.0196
Inverno	0.0078	0.0053	0.0045	0.0095	0.0055
Totale	0.0201	0.0127	0.0148	0.0151	0.0133

Via Malcontenta					
NMSE	Fotochimico	LM	GLM	GAM	GAMM
Primavera	0.0046	0.0034	0.0043	0.0047	0.0036
Estate	0.0423	0.0227	0.0258	0.0171	0.0172
Autunno	0.0340	0.0222	0.0237	0.0286	0.0220
Inverno	0.0077	0.0028	0.0041	0.0051	0.0038
Totale	0.0273	0.0157	0.0176	0.0168	0.0140

Bibliografia/sitografia

Agabiti N., Davoli M., Fusco D., Stafoggia M., Perucci C. A., 2011. *Comparative evaluation of health services outcomes*, Epidemiologia & Prevenzione, Anno 35, Supplemento 1

Agostini C., 2012. *Confronto di dati geochimici del particolato atmosferico in due aree urbane della pianura padana*. Tesi di laurea a.a. 2011/2012, Università Ca' Foscari, Venezia

Azzalini A., 2000. *Inferenza Statistica. Una presentazione basata sul concetto di verosimiglianza*. 2 edizione, Springer

Azzalini A., Scarpa B., 2004. *Analisi dei dati e data mining*. Springer

Banerjee, T., Barmanand S. C., Srivastava R. K., 2011. *Application of air pollution dispersion modeling for source contribution assessment and model performance evaluation at integrated industrial estate - Pantnagar*. Environmental Pollution, 159, 4: 865-875

Barnaba F., Gobbi G. P., de Leeuw G., 2007. *Aerosol stratification, optical properties and radiative forcing in Venice (Italy) during (ADRIEX)*. Quarterly Journal of the Royal Meteorological Society, 133: 47-60.

Bellman, R. E., 1961. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ

Benassi A., Dalan F., Gnocchi A., Maffei G., Malvasi G., Liguori F., et al., 2011. *A one-year application of the Veneto air quality modelling system: regional concentrations and deposition on Venice lagoon*. International Journal of Environment and Pollution, 44: 32-42

Bollen K. A. and Jackman R. W., 1985. *Regression diagnostics: An expository treatment of outliers and influential cases*. Sociological Methods & Research, 13: 510-542

Boylan J. and Russel A., 2006. *PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models*. Atmospheric environment, 40: 4946-4959.

Camuffo D., Zampieri F., Zambon G., 1979. *Local mesoscale circulation over Venice as a result of the mountain-sea interaction*. Boundary-layer Meteorology, 16: 83-92.

Carnevale C., Finzi G., Pisoni E., Volta M., 2008. *Modelling assessment of PM₁₀ exposure control policies in Northern Italy*. Ecological Modelling, 217: 219-29

Carnevale C., Pisoni E., Volta M., 2010. *A non-linear analysis to detect the origin of PM₁₀ concentrations in Northern Italy*. Science of the Total Environment, 409: 182-91

- Clark M., 2012. *Generalized Additive Models, getting started with additive models in R*. Center for Social Research, University of Notre Dame
- Cryer J. D., Chan K. S., 2008. *Time Series Analysis, With Applications in R*. Second Edition. Springer
- Di Fonzo T., Lisi F., 2005. *Serie storiche economiche. Analisi statistiche e applicazioni*. Roma, Carocci
- DLgs n.155, 13 agosto 2010. “Attuazione della direttiva 2008/50/CE relativa alla qualità dell’aria ambiente e per un’aria più pulita in Europa”, Gazzetta Ufficiale n.216 del 15 settembre 2010 - Suppl. Ordinario n. 217
- DM n.60, 2 aprile 2002. “Recepimento della direttiva 1999/30/CE del Consiglio del 22 aprile 1999 concernente i valori limite di qualità dell’aria ambiente per biossido di zolfo, il biossido di azoto, gli ossidi di azoto, le particelle e il piombo e della direttiva 2000/69/CE relativa ai valori limite di qualità dell’ambiente per il benzene ed il monossido di carbonio”. Gazzetta Ufficiale n.87 del 13 aprile 2002 - Suppl. Ordinario n. 77
- EEA, 2012. *AirBase - The European Air Quality Database*. European Environment Agency, disponibile nel sito: <http://www.eea.europa.eu/themes/air/airbase>. Ultimo accesso: settembre 2013
- Englert, N., 2004. *Fine particles and human health. A review of epidemiological studies*. Toxicology Letters, 149: 235-242
- EPA, 2007. *Guidance on the use of models and other analyses for demonstrating attainment of air quality goals for ozone, PM_{2.5}, and regional haze*. EPA-454/B-07-002
- European Commission, 2008. *Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe*. Technical Report 2008/50/EC, L152. Official Journal of the European Communities
- Hamilton J. D., 1994. *Time Series Analysis*, Princeton University Press
- Hanna, S. R., Chang J. C. and Strimaitis D. G., 1993. *Hazardous gas model evaluation with field observations*. Atmospheric Environment, 27A: 2265-2285
- Hastie, T. and Tibshirani, R., 1990. *Generalized Additive Models*. New York, Chapman and Hall
- Hastie T., Tibshirani R., Friedman J., 2001. *The Elements of Statistical Learning*. Springer Verlag, Berlin
- Juda, K., 1986. *Modeling of the air pollution in the Cracow area*. Atmospheric Environment, 20: 2449-2458
- Kauhaniemi M., Karppinen A., Härkönen J., Kousa A., Alaviippola B. and Koskentalo T., 2008. *Evaluation of a modeling system for predicting the concentrations of PM_{2.5} in an urban area*. Atmospheric Environment, 42: 4517-4529
- Kousa, J., Kukkonen A., Karppinen P., Aarnio T. and Koskentalo T., 2001. *Statistical and diagnostic evaluation of a new generation urban dispersion modelling system against an extensive dataset in the Helsinki area*. Atmospheric Environment, 35: 4617-4628

- Kunzli, N., Kaiser, R., Medina S., Studnicka M., Chanel O., Filliger P., Herry M., Horak, F., Puybonnieux-Textier, V., Quénel P., Schneider J., Seethaler R., Vergnaud, J-C., Sommer, H, 2000. *Public-health impact of outdoor and traffic-related air pollution*. European assessment, 356: 795-801
- Lonati G., Pirovano G., Sghirlanzoni G. A., Zanoni A., 2010. *Speciated fine particulate matter in Northern Italy: a whole year chemical and transport modelling reconstruction*. Atmos Res, 95: 496-514
- Marconi, A., 2003. *Materiale particellare aerodisperso: definizioni, effetti sanitari, misura e sintesi delle indagini ambientali effettuate a Roma*. Istituto Superiore di Sanità, 39: 329-342
- Masiol M., Squizzato S., Ceccato D., Rampazzo G., Pavoni B., 2012. *A chemometric approach to determine local and regional sources of PM₁₀ and its geochemical composition in a coastal area*. Atmos Environ, 54: 127-33
- Mitchell T. M., 1997. *Machine Learning*. McGraw-Hill, New York
- Molinaroli, E., Masiol, M., 2006. *Particolato atmosferico e ambiente mediterraneo. Il caso delle polveri sahariane*. Aracne editrice, Roma
- Pasquill, F., 1961. *The estimation of the dispersion of windborne material*. Meteorological Magazine, 90: 33-49
- Pecorari E., Squizzato S., Masiol M., Visin F., Rampazzo G. and Pavoni B., 2011. *Testing a dispersion model representing PM_{2.5} spatial and temporal distribution: a statistical approach*. Journal of Aerosol Science
- Poffe S., 2003. *Inquinamento e salute: un approfondimento dell'analisi a breve termine per la città di Verona*. Tesi di laurea a.a. 2002/2003, Università degli Studi di Padova, Padova
- Rampazzo G., Masiol M., Visin F., Rampado E., Pavoni B., 2008. *Geochemical characterization of PM₁₀ emitted by glass factories in Murano, Venice (Italy)*. Chemosphere, 71: 2068-75
- Rampazzo G., Masiol M., Visin F., Pavoni B., 2008. *Gaseous and PM₁₀-bound pollutants monitored in three environmental conditions in the Venice area (Italy)*. Water Air Soil Pollut, 195: 161-76
- Scarpa B., 1998. *Modelli additivi generalizzati con autocorrelazione tra le osservazioni*. Atti Riunione SIS, 855-864
- Schwartz J, Laden F, Zanobetti A., 2002. *The concentration-response relation between PM_{2.5} and daily deaths*. Environ Health Perspect, 110: 1025-9
- Seigneur C., Moran. M., 2004. *Chemical transport models, in Particulate Matter Science for Policy Makers: A NARSTO Assessment*. Cambridge University Press, United Kingdom, 8: 283-323
- Smith R. L., Davis J. M., Sacks J., Speckman P., Styer P., 2000. *Regression models for air pollution and daily mortality: analysis of data from Birmingham, Alabama*. Environmetrics, 11:719-743
- Speckman P., 1988. *Kernel Smoothing in Partial Linear Models*. Journal of the Royal Statistical Society. Vol. 50, 3: 413-436

- Taylor K. E., 2001. *Summarizing multiple aspects of model performance in a single diagram*. Journal of Geophysical Research, 106: 7183-7192,
- Trimarchi F., 1990. *L'imputazione dei dati mancanti nelle indagini campionarie: un'applicazione delle tecniche di regressione*. Banca d'Italia, Roma
- Weil, J. C., Sykes R. I., Venkatram A., 1992. *Evaluating airquality models: review and outlook*. Journal of Applied Meteorology, 31: 1121-1145
- Willmot C. J., 1981. *Validation of models*. Physical Geography, 2: 184-194
- Wood S. N., 2006. *Generalized Additive Models: an introduction with R*. Chapman & Hall/CRC, 391

Siti consultati

<http://www.arpa.veneto.it/temi-ambientali/aria>, ultimo accesso settembre 2013

<http://www.ecoedility.it/index.php>, ultimo accesso ottobre 2013

http://www.minambiente.it/home_it/menu.html?mp=/menu/menu_attivita/&m=argomenti.html|Inquinamento_atmosferico.html|Qualita_dellaria.html|Gli_inquinanti.html, ultimo accesso settembre 2013

<http://www.r-project.org/>, ultimo accesso settembre 2013

Ringraziamenti

Vorrei ringraziare tutti coloro che mi hanno permesso di giungere a conclusione di questo importante cammino:

i professori Guido Masarotto e Carlo Gaetan, per la cordialità e la disponibilità sempre dimostrata ad ogni mia richiesta o dubbio;

la dottoressa Eliana, per avermi dato questa opportunità e aver avuto la grande pazienza di seguirmi in tutto il lavoro;

il professor Giancarlo Rampazzo, le dottoresse Elena e Stefania, e Flavia, per avermi accolto fin da subito nella loro grande famiglia;

Chiara, per essermi stata sempre accanto in ogni istante, incoraggiandomi nei momenti difficili ed esultando in quelli felici;

i miei genitori, per i tanti sacrifici fatti in questi anni pur di farmi arrivare fin qui.