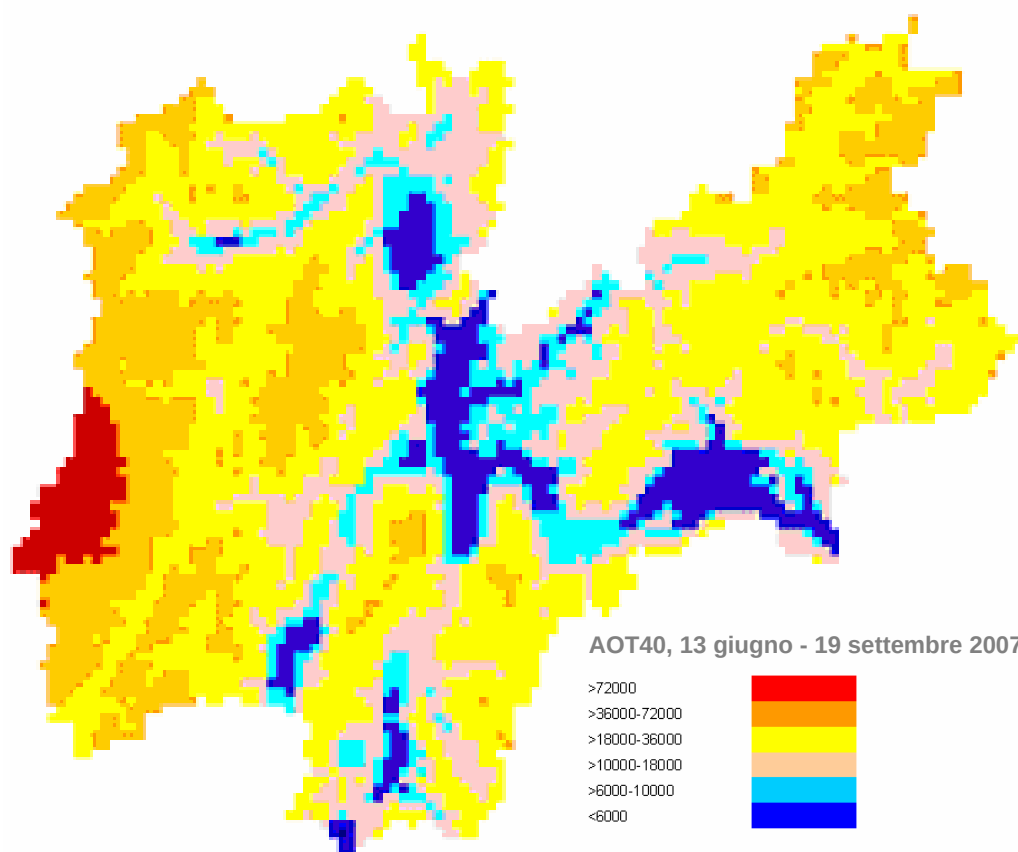


STUDIO DI UN MODELLO PER LA STIMA DELL'OZONO TROPOSFERICO IN PROVINCIA DI TRENTO



STUDIO DI UN MODELLO PER LA STIMA DELL'OZONO TROPOSFERICO IN PROVINCIA DI TRENTO

Gottardini E., Cristofolini F., Cristofori A.

Fondazione Edmund Mach, Centro di ricerca - Via E. Mach 1, 38010 San Michele all'Adige (Trento)

Bacaro G., Calderisi M., Rocchini D., Ferretti M.

TerraData environmetrics - Via P.A. Mattioli 4, 53100 Siena

Integrazione della relazione relativa allo studio realizzato nel 2007 nell'ambito della convenzione stipulata tra la Fondazione Edmund Mach - Istituto Agrario di S. Michele a/A (IASMA) e Servizio Foreste e Fauna della Provincia Autonoma di Trento per lo svolgimento di attività relative al monitoraggio e alla difesa fitosanitarie delle aree boscate del Trentino per l'anno 2007 (Determinazione del sostituto Dirigente del Servizio Foreste e Fauna N. 407 del 29/11/07; Convenzione N. 4297 del 12/3/08).

Forma raccomandata di citazione:

Bacaro G., Calderisi M., Cristofolini F., Cristofori A., Gottardini E., Rocchini D., Ferretti M., 2008. Studio di un modello per la stima dell'ozono troposferico in provincia di Trento. Relazione tecnica dello studio realizzato nell'ambito della convenzione tra IASMA e Servizio Foreste e Fauna della Provincia Autonoma di Trento (Convenzione N. 4297 del 12/3/08). 45 pp.

Nota

Questa relazione descrive l'attività svolta in collaborazione tra FEM-IASMA e TerraData environmetrics. Due relazioni precedenti a firma comune rappresentano la base di questo documento:

Gottardini E., Cristofori A., Cristofolini F., Ferretti M., 2007. Bio-monitoraggio integrato dell'ozono in provincia di Trento. Istituto Agrario di San Michele all'Adige (IASMA). Centro Sperimentale. Relazione dello studio realizzato nell'ambito della convenzione tra APPA Trento e IASMA nel campo della tutela e studio della qualità dell'aria (Det. N. 174/2 del 09.07.2007).

Gottardini E., Cristofori A., Cristofolini F., Ferretti M., 2008. Rilevamento di Ozono presso i siti forestali UN/ECE Level I nella Provincia di Trento nel 2007. Istituto Agrario di San Michele all'Adige (IASMA). Centro Sperimentale. Relazione dello studio realizzato nell'ambito della convenzione tra Servizio Foreste e IASMA: 26 pp.

Ringraziamenti

Gli autori ringraziano le organizzazioni e le persone che a vario titolo hanno contribuito alla realizzazione di questo lavoro.

Tra le istituzioni:

il Servizio Foreste e Fauna della Provincia Autonoma di Trento per il contributo economico;

l'Agenzia Protezione Ambiente della Provincia di Trento per il contributo economico.

Tra le persone:

Mauro Confalonieri (Servizio Foreste e Fauna, PAT) per il sostegno al lavoro e la collaborazione organizzativa;

Gabriele Tonidandel (APPA TN) per i dati delle centraline automatiche;

Giambattista Toller (FEM - IASMA, Centro Agro-Meteorologico) per i dati meteorologici;

Fabio Zottele (FEM - IASMA) per l'aiuto nelle applicazioni GIS;

Ultimo, ma non in ordine di importanza, un particolare ringraziamento al personale delle Stazioni Forestali della Provincia per il prezioso contributo nell'installazione delle stazioni di monitoraggio dell'ozono presso i siti di Livello I e la puntuale sostituzione settimanale dei dosimetri.

RIASSUNTO	6
1 INTRODUZIONE.....	7
2 MATERIALI E METODI	7
2.1 Concetto.....	7
2.2 Dataset e completezza dei dati	8
2.2.1 Concentrazioni di ozono.....	8
2.2.2 Dati meteo	10
2.2.3 Modello digitale del terreno	14
2.2.4 Uso del suolo.....	15
2.3 Metodologie di sviluppo del modello	16
2.3.1 Approccio statistico	16
2.3.2 Approccio geo-statistico	18
2.3.3 Spazializzazione e rappresentazione.....	24
2.3.4 Validazione.....	24
3 RISULTATI	25
3.1 Selezione di variabili e modelli.....	25
3.1.1 Approccio statistico	25
3.1.2 Approccio geostatistico	27
3.3 Risultati in blind validation.....	33
3.4 Distribuzione spaziale dei livelli di ozono in Trentino nell'estate 2007.....	35
5. CONCLUSIONI E PROSPETTIVE.....	41
6. BIBLIOGRAFIA ED ULTERIORI LETTURE.....	43

Riassunto

Nel 2007 sono state condotte due campagne di misura dell'ozono troposferico in provincia di Trento, una a scala locale (Maggio-Giugno) ed una provinciale (Giugno-Settembre). I dati così raccolti, unitamente a quelli della rete di centraline automatiche APPA, ai dati meteoroclimatici (Temperatura e Precipitazioni) ed ai dati territoriali (quote, uso del suolo) sono stati considerati per valutare la possibilità di modellare le concentrazioni di ozono sul territorio provinciale.

Sono stati seguiti due approcci, uno di tipo statistico, mediante l'uso della PLSR (Partial Least Square Regression), ed uno di tipo geostatistico, che ha considerato la dipendenza spaziale dei dati. Questi due approcci sono stati dapprima studiati relativamente ad un periodo di due settimane, tra il 13 ed il 27 Giugno 2007, considerando separatamente ciascuna settimana e poi la media delle due. Questo periodo è stato scelto perché era quello comune alle due campagne di misurazione e permetteva di massimizzare la numerosità delle osservazioni. Entrambi gli approcci hanno identificato la quota, la temperatura e le precipitazioni come i principali predittori dei livelli di ozono in Trentino, seppur con pesi diversi a seconda del periodo considerato. La validazione su siti indipendenti ha evidenziato errori medi assoluti pari al 28.6% (approccio statistico) e 16.1% (approccio geostatistico). L'RMSEP (Root- Mean-Square-Error of Prediction, una misura delle differenze tra il valore vero ed il valore predetto, espressa nella stessa unità di misura della variabile di interesse) è risultato leggermente migliore nel modello derivante dall'approccio geostatistico, per il quale anche la varianza spiegata è risultata superiore.

In virtù dei minori errori, l'approccio geostatistico è stato quindi applicato per le stime a scala provinciale e per la previsione delle concentrazioni medie sul periodo Giugno-Settembre. Quota e precipitazioni sono risultati i predittori significativi che permettono di massimizzare la prestazione del modello con il minor numero di predittori. Le stime così ottenute sono state validate mediante Cross-Validation attraverso la procedura di Leave-One-Out, che ha evidenziato scostamenti medi dai dati misurati pari a $13.98 \mu\text{g}\cdot\text{m}^{-3}$. Le stime indicano valori di ozono superiori (fino anche ad oltre $180 \mu\text{g}\cdot\text{m}^{-3}$) nella parte occidentale della provincia, a ridosso dell'Adamello, e nel settore nord-orientale, dalla Marmolada al Primiero ($120\text{-}180 \mu\text{g}\cdot\text{m}^{-3}$). In queste zone l'AOT40 (stimato sul periodo metà Giugno-metà Settembre) arriva a valori 3-4 volte superiori al Target stabilito dalla Direttiva EU per la protezione della vegetazione. Merita ricordare tuttavia che esistono aree con misurazioni assenti (es. area centro-occidentale della provincia) che possono avere influenza nella formulazione del modello e sulla precisione delle stime.

I risultati ottenuti sono promettenti e dimostrano la praticabilità dell'approccio seguito, che potrebbe portare a sviluppi applicativi interessanti: con l'aggiunta di una componente dinamica (tempo), si potrebbero ottenere stime e previsioni più robuste; sovrapponendo le stime dei livelli di ozono con altri layer informativi sarebbero possibili stime di rischio, oltreché per la popolazione, anche per ecosistemi e vegetazione; diventerebbero fattibili una mappatura routinaria delle aree a rischio e studi correlativi sul possibile ruolo dell'ozono sulle condizioni dei boschi trentini.

1 Introduzione

L'ozono è l'inquinante gassoso di più ampia diffusione spaziale ed i cui livelli sono tendenzialmente crescenti, sia in termini di proiezioni (e.g. Vingarzan, 2004) che di dati misurati (e.g. Mangoni e Buffoni, 2008). Salute umana, materiali e vegetazione sono potenziali bersagli dell'azione dell'ozono (e.g. UN/ECE, 2004). Ai fini di proteggere salute, materiali e vegetazione vengono stabilite soglie, valori obiettivo e livelli critici (Direttiva 2008/50/CE; UN/ECE 2004 e revisioni). Nell'ottica della Direttiva comunitaria per la valutazione del rischio risulta quindi essenziale poter conoscere le concentrazioni e l'esposizione (in termini di AOT40¹) a cui si trovano soggette popolazioni, manufatti e vegetazione. Dato che non è possibile misurare ovunque le concentrazioni di ozono, la Direttiva prevede il ricorso a stime oggettive e modellistica.

Nell'estate 2007 sono state svolte due campagne di misurazione dell'ozono a due distinti livelli spaziali: una campagna di dettaglio in una definita area localizzata (15*15 km) ed una campagna a livello provinciale. Struttura, metodi e risultati sono stati esposti in dettaglio da Gottardini et al. (2007, 2008). I dati rilevati in queste due campagne, unitamente ad altre serie di dati già disponibili, sono stati utilizzati per lo studio presentato in questa relazione e che mira alla definizione di un modello deterministico delle concentrazioni di ozono in provincia di Trento.

2 Materiali e metodi

2.1 Concetto

Il concetto seguito per questo studio è stato quello di utilizzare al massimo i dati a disposizione. Per questo motivo il modello è stato studiato riferendosi al periodo di maggior disponibilità di dati di concentrazione di ozono, compreso tra il 13 ed il 27 giugno (vedi Tab. 1). Su questo periodo sono stati studiati due approcci, uno statistico ed uno geostatistico (descritti in seguito, par. 2.3), i cui risultati sono stati comparati. Sulla base dei risultati della comparazione, è stato scelto uno dei due approcci per proseguire lo studio per la stima delle concentrazioni medie nel periodo 13 giugno-19 settembre ed a scala provinciale.

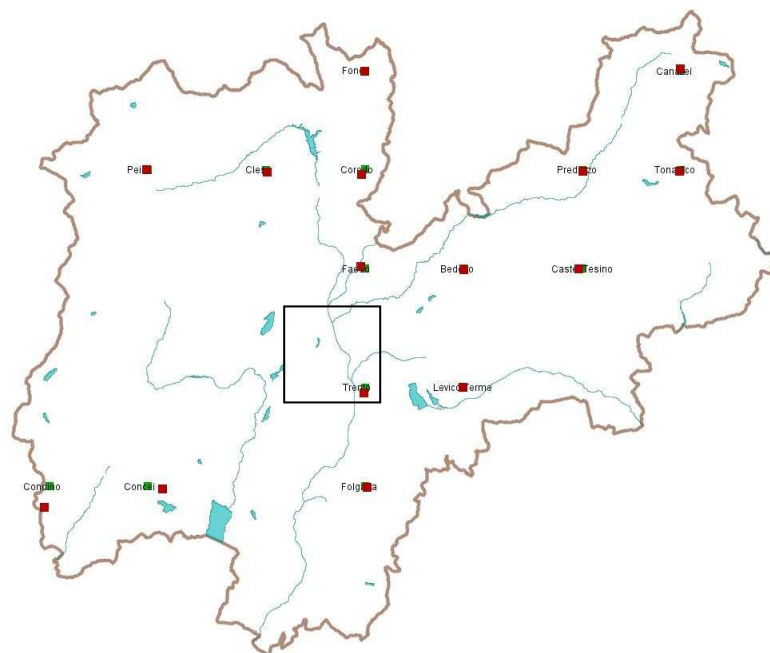
¹Accumulated Over Theshold 40, somma delle concentrazioni di ozono eccedenti le 40 ppb ($\sim 80 \mu\text{g}\cdot\text{m}^{-3}$), misurate tra le 8 e le 20 e cumulate nel periodo Maggio-Luglio.

2.2 Dataset e completezza dei dati

2.2.1 Concentrazioni di ozono

Le concentrazioni di ozono utilizzate nel presente studio sono state rilevate nel 2007 con i dettagli riportati di seguito:

- concentrazioni misurate mediante dosimetri passivi nel periodo 30 maggio-27 giugno 2007 in 18 punti (3 siti * 6 località) situati all'interno di un'area di 15*15 km nella parte centrale della provincia (Gottardini et al., 2007) e presso 3 stazioni automatiche APPA (Fig. 1; Tab. 1);
- concentrazioni misurate mediante dosimetri passivi nel periodo 13 giugno-19 settembre 2007 in 15 punti situati in corrispondenza della rete EU-UN/ECE per il rilevamento delle condizioni dei boschi sul territorio provinciale (densità nominale: 1 punto ogni 225 km², corrispondente ad una rete di 15*15 km) (Gottardini et al., 2008) (Fig. 1; Tab. 1);
- concentrazioni misurate mediante analizzatori automatici nelle stazioni APPA nel periodo 30 maggio-12 settembre 2007 (Tab. 1).



*Fig. 1. Aree di studio. Il quadrato 15*15 km dello studio di dettaglio, i siti di monitoraggio di Livello I (in verde) e dei siti di misura dell'ozono (in rosso). In caso di completa sovrapposizione appare solo il colore rosso.*

Tabella 1. Sintesi dei rilevamenti di ozono misurati nel 2007.

Nome stazione	x	y	Quota	Ambito spaziale	Tipo di misurazione	Periodo	[O3], media (ug*m-3)	
Condino	1617253	5081830	1856	Provincia	Dosimetri passivi	13 giugno-12 settembre	131.5	± 23.00
Concei	1635204	5084579	1309	Provincia	Dosimetri passivi	14 giugno-12 settembre	122.0	± 20.32
Tonadico	1713788	5132795	1854	Provincia	Dosimetri passivi	15 giugno-12 settembre	122.4	± 20.11
Cles	1651186	5132648	1698	Provincia	Dosimetri passivi	16 giugno-12 settembre	92.2	± 23.14
Predazzo	1699060	5132798	1464	Provincia	Dosimetri passivi	17 giugno-12 settembre	91.1	± 16.85
Canazei	1713980	5148362	1588	Provincia	Dosimetri passivi	18 giugno-12 settembre	88.9	± 18.63
Fondo	1665949	5148025	1375	Provincia	Dosimetri passivi	19 giugno-12 settembre	90.5	± 17.40
Folgaria	1666290	5084879	1210	Provincia	Dosimetri passivi	20 giugno-12 settembre	88.0	± 12.62
Coredo	1665450	5132359	1389	Provincia	Dosimetri passivi	21 giugno-12 settembre	84.7	± 29.78
Peio	1632826	5133037	1675	Provincia	Dosimetri passivi	22 giugno-12 settembre	76.7	± 18.63
Bedollo	1680976	5117859	1400	Provincia	Dosimetri passivi	23 giugno-12 settembre	70.0	± 19.16
Levico Terme	1680918	5100063	1105	Provincia	Dosimetri passivi	24 giugno-12 settembre	70.0	± 20.25
Castello Tesino	1698517	5117954	1549	Provincia	Dosimetri passivi	25 giugno-12 settembre	66.0	± 18.55
Faedo	1665324	5118424	306	Provincia	Dosimetri passivi	26 giugno-12 settembre	61.0	± 20.26
Trento	1665808	5099179	359	Provincia	Dosimetri passivi	27 giugno-12 settembre	59.4	± 17.63
Cadine	1659728	5105269	537	Dettaglio 15*15	Dosimetri passivi	30 maggio-27 giugno	90.1	± 18.62
Trento Gardolo	1663127	5107995	195	Calibrazione 15*15	Dosimetri passivi	30 maggio-27 giugno	45.4	± 13.57
M.te Gazza	1651408	5105102	1608	Calibrazione 15*15	Dosimetri passivi	30 maggio-27 giugno	136.9	± 26.45
Lasino N	1653445	5098393	459	Dettaglio 15*15	Dosimetri passivi	30 maggio-27 giugno	81.3	± 14.28
Lasino S	1653624	5097638	469	Dettaglio 15*15	Dosimetri passivi	30 maggio-27 giugno	86.0	± 20.15
Margone	1652157	5103618	918	Dettaglio 15*15	Dosimetri passivi	30 maggio-27 giugno	112.8	± 18.73
Maso Ariol	1656473	5107571	638	Dettaglio 15*15	Dosimetri passivi	30 maggio-27 giugno	94.5	± 27.00
Monte Terlago	1657519	5107861	694	Dettaglio 15*15	Dosimetri passivi	30 maggio-27 giugno	96.3	± 38.77
Grumo S. Michele	1664123	5118120	209	Calibrazione	Dosimetri passivi	30 maggio-27 giugno	54.1	± 10.09
Trento PSC	1664477	5103260	204	Centraline APPA	Analizzatori automatici	13 giugno-19 settembre	66.5	± 20.83
Trento Gardolo	1663127	5107995	195	Centraline APPA	Analizzatori automatici	14 giugno-19 settembre	57.2	± 14.01
Grumo S. Michele	1664123	5118120	209	Centraline APPA	Analizzatori automatici	15 giugno-19 settembre	63.1	± 13.78
Rovereto LGP	1658386	5084048	195	Centraline APPA	Analizzatori automatici	16 giugno-19 settembre	82.6	± 23.17
Borgo Vals.	1689884	5102745	390	Centraline APPA	Analizzatori automatici	17 giugno-19 settembre	61.4	± 21.44
Riva del Garda	1643123	5083661	85	Centraline APPA	Analizzatori automatici	18 giugno-19 settembre	62.4	± 21.23
M.te Gazza	1651408	5105102	1608	Centraline APPA	Analizzatori automatici	19 giugno-19 settembre	111.8	± 17.99

2.2.2 Dati meteo

I dati meteo utilizzati sono stati forniti da FEM-IASMA e rilevati in 133 stazioni di misura distribuite sul territorio provinciale (Fig. 2). Durante la fase di analisi esplorativa, sono stati rilevati i seguenti errori all'interno dei dataset da utilizzare per le successive analisi: record contenenti valori non utilizzabili codificati come "NA" (Fig. 3, in alto); record contenenti celle vuote (Fig. 3, al centro). La somma di entrambi è riportata nella Fig. 3 (in basso).

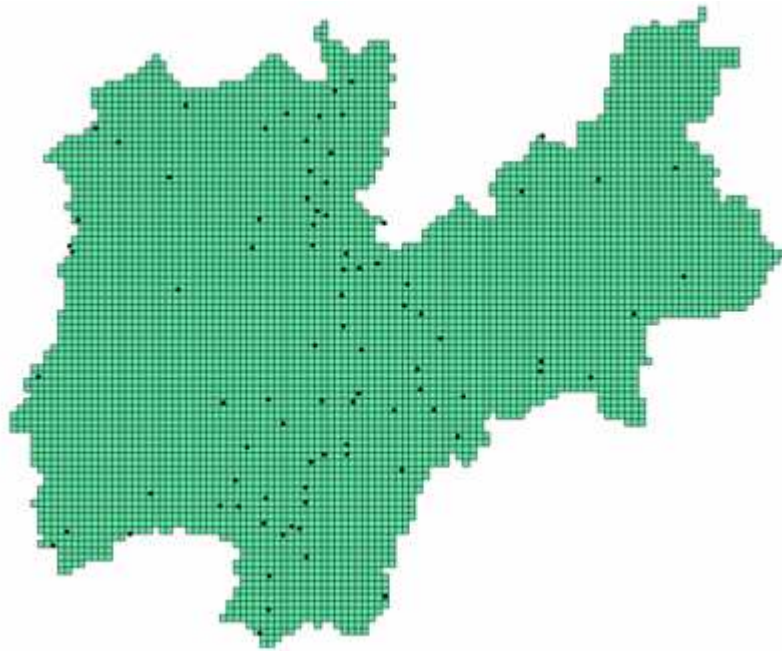


Fig. 2. Localizzazione delle stazioni meteorologiche utilizzate per l'estrazione dei dati meteo.

L'eliminazione di interi record in funzione dell'assenza di dati avrebbe comportato un totale di 1221 eliminazioni su 1978 record. Per questo è stato deciso di eliminare i record valutando l'assenza di dati per ciascuna variabile separatamente. In particolare, sono stati eliminati: per temperatura media (Tmed) 262 record; per temperatura minima (Tmin) 267 record; per temperatura massima (Tmax) 267 record eliminati; per Umidità relativa media (Urmed) 356 record; per pioggia (Pio) 256 record; per Radiazione globale (Radg) 1189 record.

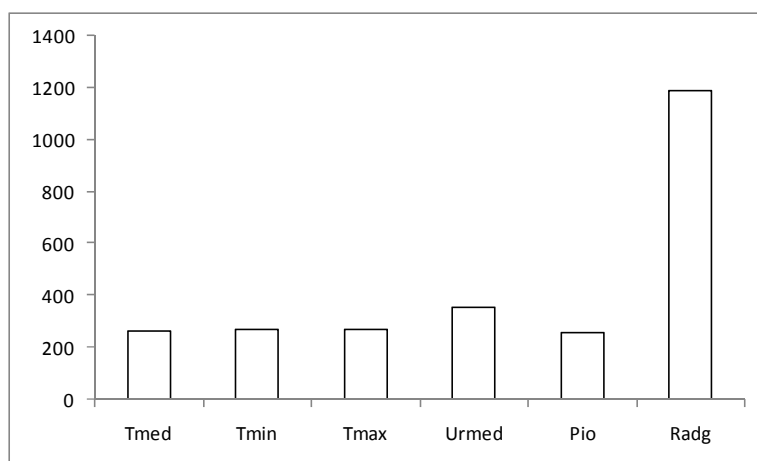
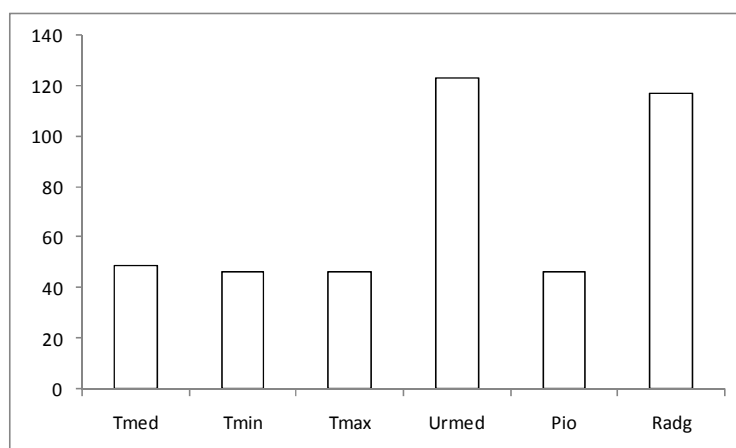
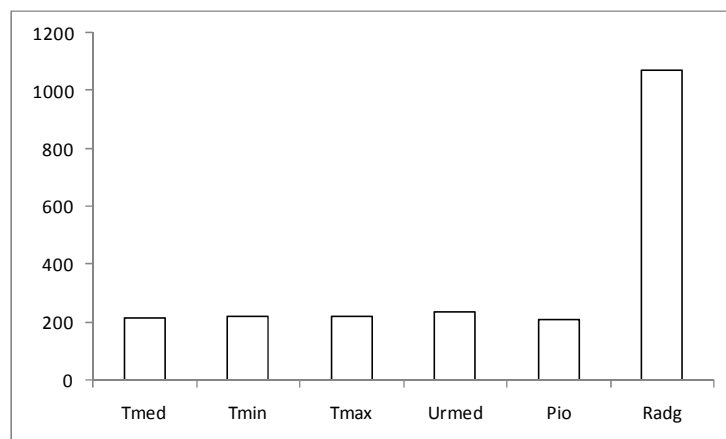


Fig. 3. Numero di record contenenti valori nulli codificati come “NA” (in alto), contenenti celle vuote (al centro) e loro somma (in basso) in funzione di ciascuna variabile climatica rilevata. Abbreviazioni nel testo.

Inoltre, tutti i record con valori non congruenti sono stati eliminati (es. Tmin: valori -534.6, -152.3: 2 record eliminati).

I dati così ripuliti sono stati utilizzati per la derivazione di Temperatura media e Piovosità per ogni cella chilometrica della provincia. Le mappe sono state create tramite Inverse Distance Weighting (IDW) quadratico, una funzione inversa rispetto alla distanza spaziale tra i punti (equazione 1, Fig. 4, Longley et al., 2003).

$$z_i = \frac{\sum z_j \times w}{\sum w} \quad (1)$$

dove z_i = valore stimato, z_j = valore misurato, w = peso della distanza d_{ij} , con esponente $-x$. All'aumentare dell'esponente x aumenta il decadimento della distanza assunto dal modello. In altri termini i valori dei punti da stimare si avvicineranno ai valori dei punti di input più vicini. Se $x=1$, ovvero si assume una variazione lineare dei valori di input nello spazio. L'applicazione di un esponente $x=2$ assume una variazione quadratica nello spazio, ovvero il valore stimato mostra una ancor maggiore autocorrelazione con i punti spazialmente più vicini (Burrough & McDonnell 1998).

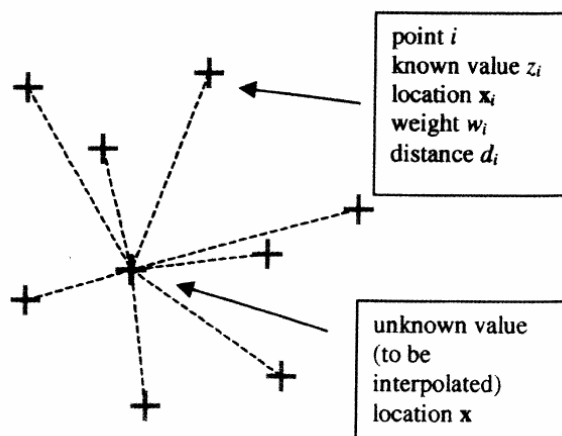


Fig. 4. Rappresentazione spaziale del metodo IDW per l'interpolazione spaziale.

Il metodo funziona come segue: dati due punti a valore noto (20 ed 80) ed uno a valore incognito, situato a distanza 10 m dal punto a modalità 20 e 20 m dal punto a modalità 80, applicando un IDW lineare:

$$Z_i = \frac{\sum Z_j^x w}{\sum w}$$

$$w = d_{ij}^{-1}$$

Sostituendo i valori nella formula, otterremo:

$$Z_i = \frac{(20 \times (1/10)) + (80 \times (1/20))}{1/10 + 1/20} = 40$$

È possibile variare l'esponente da dare alla distanza nel calcolo del peso w , in modo da pesare in modo maggiore l'inverso della distanza (Lam, 1983). Ad esempio, un trend quadratico presuppone:

$$Z_i = \frac{\sum Z_j^x w}{\sum w}$$

$$w = d_{ij}^{-2}$$

$$Z_i = \frac{(20 \times (1/10^2)) + (80 \times (1/20^2))}{1/10^2 + 1/20^2} = 32$$

All'aumentare dell'esponente x il valore del punto incognito si avvicinerà maggiormente ai valori spazialmente più vicini. Le Figg. 5 e 6 riportano le stime di temperature e piovosità medie per il periodo 13-20 giugno 2007.

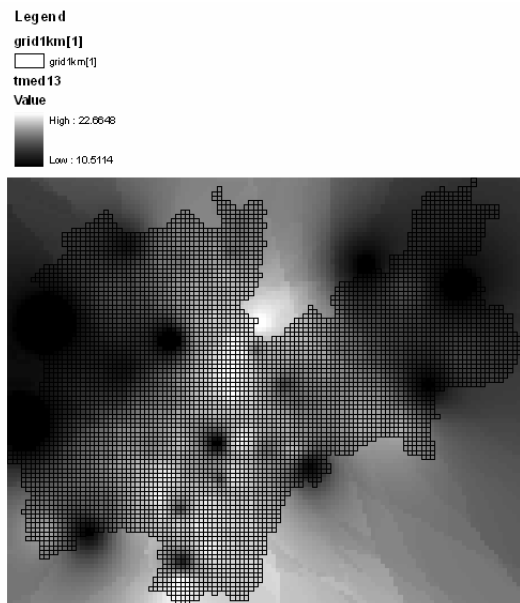


Fig. 5. Esempio della mappa di temperatura media relativa alla settimana dal 13 al 20 Giugno. La scala dei valori va da 10.5 °C (nero) a 22.7 °C (bianco).

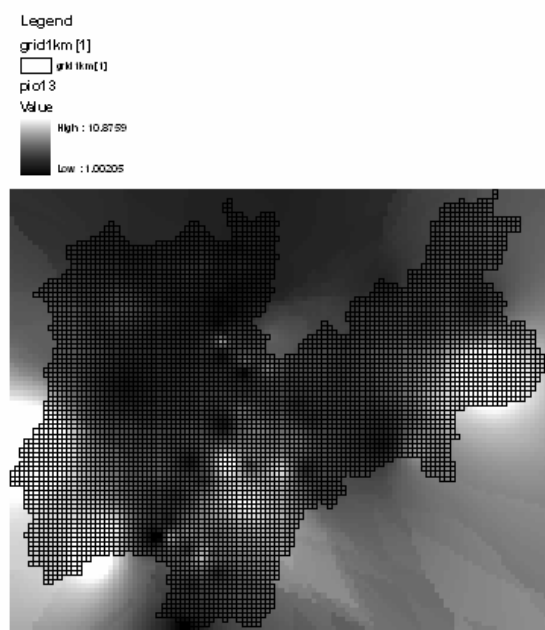


Fig. 6. Esempio della mappa di piovosità relative alla settimana dal 13 al 20 Giugno. La scala va da 1.0 (nero) a 10.9 mm (bianco).

2.2.3 Modello digitale del terreno

La quota per ogni cella chilometrica è stata estratta da un modello digitale del terreno a grana 1x1km (Fig. 7).

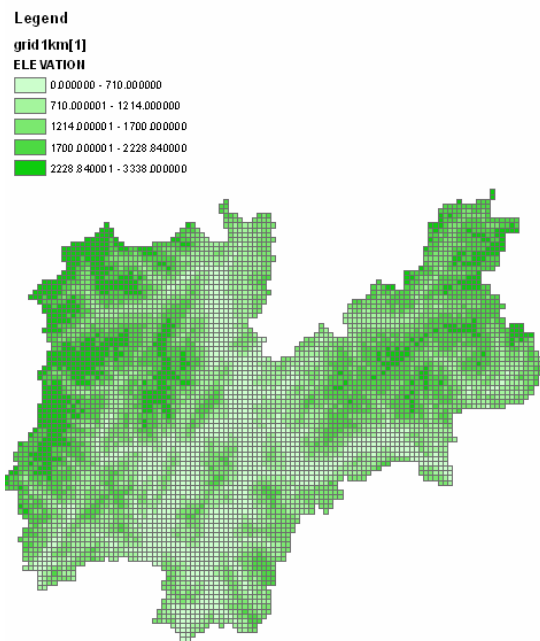


Fig. 7. Modello digitale del terreno a scala 1x1km.

2.2.4 Uso del suolo

Mappe di uso del suolo sono state create utilizzando sia una classificazione booleana (Fig. 8), che una classificazione di tipo *fuzzy* attraverso l'uso di un'immagine satellitare Landsat ETM+ (Fig. 9), in cui l'appartenenza di ogni singolo pixel ad una classe viene espressa da una funzione continua piuttosto che da una funzione booleana 1/0 (si veda Rocchini & Ricotta, 2007). Al momento, tali mappe non sono state utilizzate nel modello, ma potranno essere inserite come variabili ancillari in futuro.

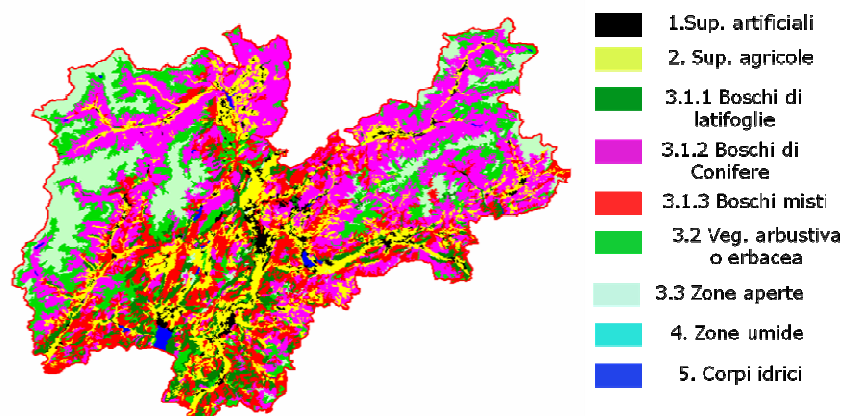


Fig. 8. Uso del suolo, Provincia di Trento (CORINE Land Cover)

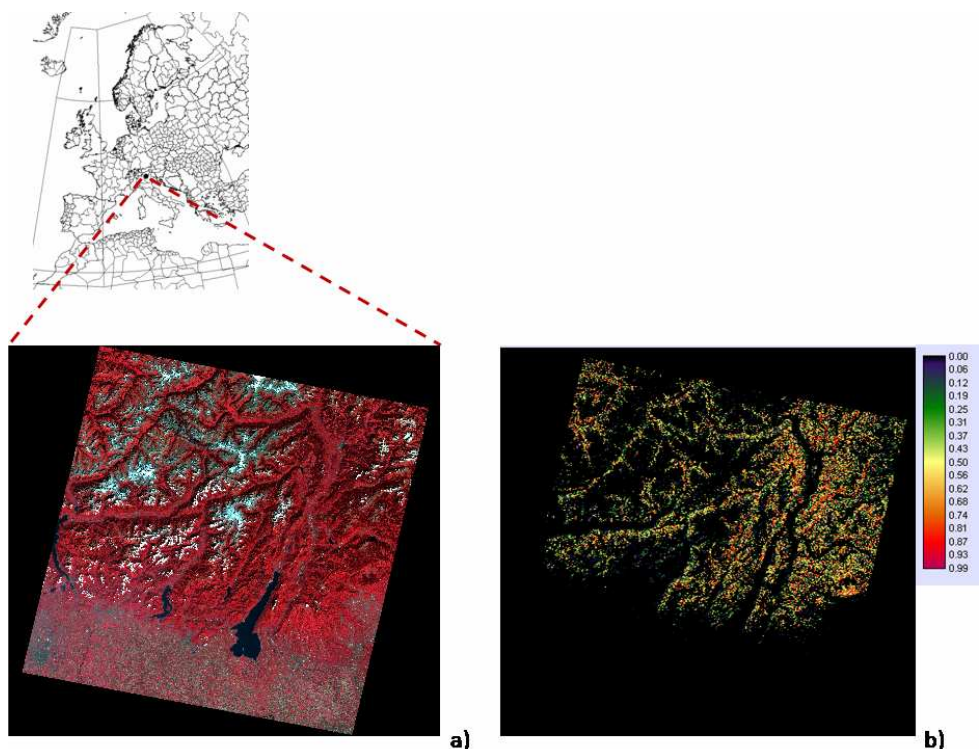


Fig. 9. Esempio di classificazione fuzzy. Immagine Landsat ETM+ (a) e grado di appartenenza (fuzzy) di ogni pixel alla classe “Foreste”.

2.3 Metodologie di sviluppo del modello

Sono stati seguiti due diversi approcci per la selezione delle variabili da utilizzare nel modello: un approccio statistico ed uno geostatistico.

2.3.1 Approccio statistico

La PLSR (Partial Least Square Regression) (Wold et al., 2001) è una tecnica statistica multivariata di regressione. La PLSR, che può esser considerata come uno sviluppo recente del classico approccio MLR (Multiple Linear Regression), è un metodo per relazionare due matrici di dati, X ed Y, usando un modello lineare multivariato. Va al di là delle normali tecniche di regressione tradizionali in quanto:

- 1) modella entrambe le matrici;
- 2) è in grado di funzionare anche in presenza di molte variabili collineari e/o particolarmente affetti da *noise*;

- 3) può creare un modello anche in presenza di un numero di variabili predittrici molto grande e, soprattutto, anche in casi in cui il numero di variabili sia superiore rispetto al numero di osservazioni disponibili;
- 4) è in grado anche di gestire più variabili dipendenti (risposta) insieme.

La PLSR lavora su variabili latenti create sfruttando le componenti di massima correlazione tra le variabili predittrici e le variabili risposta. Lavorando su variabili latenti si ha, da una parte una rimozione della ridondanza dell'informazione presente sulle variabili predittrici correlate, dall'altra una sintesi del dataset ed una riduzione del numero delle variabili predittrici stesse. In ultima analisi, la PLSR è in grado di lavorare considerando anche più variabili risposta allo stesso tempo. Lavorando infatti su strutture latenti, opera una trasformazione adeguata anche sulle variabili risposta.

Le origini della tecnica PLSR risalgono al 1975, grazie al lavoro di Herman Wold (Wold, 1982). Per spiegare il modello matematico su cui si basa, si consideri una matrice X di N osservazioni e K variabili (predittori) ed una matrice Y con M variabili dipendenti. Il criterio seguito dal modello PLSR è quello di massimizzare la covarianza spiegata, considerando al contempo sia la varianza della matrice X che quella della matrice delle risposte, Y .

$$\max[\text{cov}(t,y) | Xw=t \text{ e } ||w||=1]$$

Il modello lineare PLSR ricerca un numero ridotto, a , di “nuove” variabili “latenti”, chiamate X -scores (indicate con t) che modellano la matrice dei predittori X e spiegano la matrice delle variabili dipendenti Y . Gli X -scores sono combinazioni lineari, ricavate mediante l'uso di appropriati pesi (w) delle variabili predittrici originali (matrice X) e sono ortogonali tra loro. In forma matriciale ciò si può esprimere con la seguente formula:

$$T=(XW)$$

Gli *scores* hanno la seguente importante caratteristica, ovvero che possono essere moltiplicati per i rispettivi *loadings* (p) per dare la matrice X originale, a meno di un errore (e).

$$X=(TP')+E$$

Nel caso in cui si abbia una matrice Y con più variabili dipendenti, che vogliamo trattare simultaneamente, si hanno i corrispondenti Y-scores (u) ed i rispettivi *weights* (c):

$$Y=(UC')+G$$

dove G è il termine di errore.

Gli X-scores, come precedentemente detto, sono predittori delle variabili dipendenti, pertanto si può dire che

$$Y=(TC')+F$$

dove F è il termine di errore.

Da cui è possibile tornare indietro al modo classico di esprimere una funzione, ovvero alla forma chiusa

$$Y=XW*C'+F=XB+F$$

L'output dell'analisi e la conseguente interpretazione dei risultati della regressione sono quindi pressoché analoghi a quelli di un tradizionale modello di regressione lineare multipla.

Usualmente, prima dell'analisi, i dati possono essere trasformati, centrati e scalati in modo opportuno. La tipologia di pre-trattamento dipende dal tipo di dati e dall'obiettivo preposto. In ogni caso, il pretrattamento più tipico è il cosiddetto autoscaling, che prevede la trasformazione delle variabili originali in elementi con media zero e varianza unitaria.

2.3.2 Approccio geo-statistico

statistici e metodi messi a punto per l'analisi di dati spazialmente referenziati (per una revisione di gran parte di questi metodi si veda Cressie 1993). All'interno delle statistiche spaziali, il termine "geostatistica" si riferisce a quei modelli e metodi sviluppati per dati che presentano le seguenti caratteristiche:

i) i valori di una variabile $Y_i : i = 1, \dots, n$ sono osservati per un insieme (set) discreto di punti di osservazioni x_i all'interno di una determinata area A (comune, provincia, regione o nazione, ad esempio);

ii) ogni valore osservato Y_i rappresenta una misura diretta, o comunque risulta statisticamente relazionato, al valore di un certo fenomeno continuo nello spazio $S(x)$ ("segnale", in termini geostatistici; ad esempio si può annoverare come fenomeno continuo nello spazio la concentrazione di ozono, oppure la concentrazione di un certo inquinante nel suolo) , per quella specifica località x_i (Diggle & Ribeiro 2007).

Gli obiettivi principali di un'analisi geostatistica sono fondamentalmente rappresentati dal processo di stima e di predizione:

1) il processo di stima si riferisce alla necessità di fare inferenza relativamente ai parametri di un processo stocastico partendo dai dati raccolti. Vengono quindi considerati e stimati quei parametri di interesse diretto, come ad esempio quelli che definiscono una relazione regressiva tra una variabile di risposta (la concentrazione di ozono) e una variabile di tipo esplicativo (la temperatura), oppure i parametri di interesse indiretto, per esempio quelli che definiscono la struttura di covarianza per il modello del "segnale" in analisi $S(x)$.

2) il processo previsionale si riferisce invece all'inferenza che viene effettuata relativamente alla realizzazione di un "segnale" $S(x)$ non osservato. Da un punto di vista applicativo, l'obiettivo specifico di una previsione potrebbe includere la predizione di un certo fenomeno $S(x)$ per una arbitraria località x all'interno di una certa area A . Il risultato di questo processo viene classicamente rappresentato da una mappa dell'area A in cui vengono riportati i valori predetti di $S(x)$ per tutte le possibili località x .

Date queste premesse concettuali, si è proceduto, attraverso le metodologie geostatistiche, a creare delle previsioni spazialmente continue della concentrazione di ozono, relativamente all'intera superficie della Provincia di Trento. In particolare, partendo dai dati puntiformi e discreti disponibili per le concentrazioni di ozono (si vedano i paragrafi precedenti) è stato, in prima istanza, necessario formulare un modello formale di cui, successivamente, sono stati stimati tutti i parametri necessari alla sua definizione. Sulla base dei parametri utilizzati e stimati, sono stati poi successivamente spazializzati (predetti) i valori di ozono su tutta l'area di interesse. Si procede alla descrizione dettagliata del metodo di formulazione del modello, di stima dei parametri e, infine, di spazializzazione utilizzata. Ovviamente,

come per ogni analisi statistica che si effettua, è necessario, in fase preliminare, realizzare un'analisi esplorativa dei dati grezzi, così da essere in grado di selezionare le adeguate trasformazioni e i diversi accorgimenti per rendere i modelli previsionali il più robusti possibili. Queste analisi esplorative vengono quindi descritte come primo punto del procedimento complessivo.

2.3.2.1 Analisi esplorative preliminari

i) Non-Spaziali: per ciascun set di dati sono stati sviluppati degli istogrammi relativi alla distribuzione delle frequenze delle concentrazioni di ozono. L'obiettivo di tale operazione è di fornire uno strumento semplice ed immediato per comprendere l'esistenza di asimmetrie evidenti nella distribuzione dei valori (deviazione dalla Normalità), così come di evidenziare la presenza di valori eccessivamente lontani dal trend distributivo (outliers). Sulla base degli istogrammi ottenuti (e anche attraverso un procedimento di tipo quantitativo basato su stime di massima verosimiglianza, si veda in seguito), si è provveduto, dove necessario, alla normalizzazione dei dati attraverso la famiglia di trasformazioni Box-Cox (Box & Cox 1964).

Successivamente sono state esaminate, in termini univariati, le singole relazioni esistenti tra le variabili esplicative (quota, temperatura e piovosità) rispetto alle concentrazioni di ozono registrate. Queste analisi sono state ripetute per ciascun set di dati, ovvero per le concentrazioni di ozono relative a due singole settimane (dal 13 al 20 giugno e dal 21 al 27 giugno), per le concentrazioni medie delle due settimane (dati mediati relativi alle concentrazioni di ozono dal 13 al 27 giugno) e, infine, per l'arco temporale relativo alle 9 settimane di dati disponibili (si vedano i paragrafi precedenti).

ii) Spaziali: il metodo più semplice, per analizzare quali-quantitativamente, la dipendenza spaziale nella distribuzione della variabile di interesse, è di effettuare una semplice regressione tra variabile di risposta e le singole componenti spaziali (latitudine e longitudine). Inoltre, per verificare se i dati di concentrazione di ozono misurati presentano una forte dipendenza spaziale (rispondendo, in altri termini, alla ben nota legge dell'autocorrelazione spaziale), è stato calcolato un semivariogramma empirico per ogni settimana (o media di settimane) considerata. Da un punto di vista generale, per ogni set di dati spazialmente definiti come (x_i, y_i) : $i = 1, \dots, n$, le ordinate del semivariogramma empirico rappresentano le quantità

$$v_{ij} = \frac{1}{2}(y_i - y_j)^2.$$

Se i valori y_i presentano una media e una varianza spazialmente costante, allora i valori v_{ij} sono proporzionali a

$$\sigma^2 \{1 - \rho(x_i, x_j)\}$$

dove σ^2 rappresenta la varianza, mentre $\rho(x_i, x_j)$ quantifica la correlazione esistente tra i valori di y_i e y_j . Se gli y_i sono generati da un processo spaziale stazionario, allora si avrà che il valore $\rho(\cdot)$ dipende esclusivamente dalla distanza tra x_i e x_j e tende allo 0 all'aumentare della distanza considerata. Ne consegue che i valori attesi di v_{ij} tendono ad un valore costante rappresentato da σ^2 , allorchè la distanza u_{ij} tra x_i e x_j aumenta. Al contrario, se i valori y_i non sono tra loro spazialmente correlati, allora tutti i valori v_{ij} assumeranno il valore σ^2 . Il semivariogramma viene quindi rappresentato, per comodità di lettura, come i valori di v_{ij} (solitamente il valore medio di semivarianza mediato per la classe di distanza considerata) verso la corrispondente distanza u_{ij} . Da un punto di vista pratico, ciò significa che se la semivarianza v_{ij} in un semivariogramma tende ad aumentare in modo costante, questo risultato potrà essere interpretato come se i valori osservati della variabile di risposta sono stati generati da un processo stocastico di tipo stazionario. In altre parole, l'intervallo della correlazione spaziale esistente tra i valori della variabile si estenderà anche al di fuori dell'area di studio esaminata. Al contrario, se il semivariogramma presenta un classico andamento in cui inizialmente i valori di v_{ij} crescono per poi raggiungere un andamento asintotico al crescere della distanza tra punti di osservazione, l'autocorrelazione spaziale esistente tra le osservazioni potrà essere correttamente stimata e parametrizzata (attraverso la stima dei parametri del semivariogramma, si veda in seguito) e inclusa nel modello predittivo finale, per ottenere stime maggiormente accurate. Nel caso in questione, i 4 semivariogrammi empirici calcolati, sono stati poi successivamente impiegati per la stima, tramite massimizzazione della verosomiglianza, della struttura di covarianza delle concentrazioni di ozono in funzione della loro distribuzione spaziale.

2.3.2.2 Formulazione del modello

Classicamente, un modello geostatistico viene specificato nel seguente modo:

in assenza di qualsiasi variabile esplicativa (e.g., quota), e considerando $(x_i, y_i): i = 1, \dots, n$ le località x_i in cui sono stati misurati i valori della variabile di interesse y_i , si ha che il modello stazionario Gaussiano è specificato dalle due componenti:

1) $\{S(x) : x \in \mathcal{R}^2\}$ rappresenta un processo Gaussiano con media μ , varianza $\sigma^2 = \text{Var}\{S(x)\}$ e funzione di correlazione $\rho(u) = \text{Corr}\{S(x), S(x')\}$, dove, come in precedenza, $u = \|x - x'\|$ e $\|\cdot\|$ esprimono la distanza tra i punti di misurazione.

2) condizionate rispetto a $\{S(x) : x \in \mathcal{R}^2\}$, le misurazioni y_i rappresentano misurazioni mutualmente indipendenti della variabile casuale Y_i , normalmente distribuite con media condizionale $E[Y_i|S(\cdot)] = S(x_i)$.

In forma più semplice, il modello può essere specificato utilizzando la notazione

$$Y_i = S(x_i) + Z_i : i = 1, \dots, n$$

dove $\{S(x) : x \in \mathcal{R}^2\}$ è definito dall'assunto al punto uno precedentemente descritto, mentre Z_i sono variabili random con distribuzione $N(0, \sigma^2)$ mutualmente indipendenti.

2.2.2.3 Stima dei parametri del modello

Il modello geostatistico appena descritto, necessita dunque di diverse specificazioni e della stima di molti parametri. In particolare:

Componente spaziale: al fine di specificare un modello corretto, la funzione di correlazione $\rho(u)$ (quindi la componente spaziale in senso stretto) deve essere positivamente definita. Questa componente, che tende a 0 al crescere della distanza tra le osservazioni, viene graficamente descritta come la componente del semivariogramma che esprime il tasso di crescita della semivarianza al crescere della distanza (quando il semivariogramma assume un andamento asintotico al crescere della distanza u_{ij} , significa che la funzione di correlazione tra le osservazioni $\rho(u)$ tende al valore 0). La funzione $\rho(u)$ viene stimata fittando al semivariogramma empirico una funzione definita Matern, esplicitata come:

$$\rho(u) = \{2^{\kappa-1} \Gamma(\kappa)\}^{-1} (u/\phi)^{\kappa} K_{\kappa}(u/\phi)$$

e di cui deve essere stimato il parametro di scala ϕ (per la spiegazione degli altri parametri, si rimanda direttamente a Diggle & Ribeiro 2007). Poiché la componente spaziale è rappresentata anche dagli altri parametri che descrivono il semivariogramma, sempre utilizzando il semivariogramma empirico costruito sui dati misurati, dovranno essere stimati l'intercetta τ^2 (definita anche *nugget*, rappresenta il valore di semivarianza alla distanza pari a 0, e se maggiore di 0, rappresenta la variabilità nei valori osservati dovuta alla varianza presente a scale di osservazione inferiori rispetto a quella considerata e che non è possibile descrivere) e il valore di semivarianza raggiunto nella parte asintotica del semivariogramma, pari a $\tau^2 + \sigma^2$ (questo valore viene definito anche *sill*, che corrisponde alla somma della varianza *nugget* e alla varianza del processo osservato σ^2 , come precedentemente definita). Riassumendo, dunque, per descrivere la struttura di covarianza spazialmente esplicita delle concentrazioni di ozono considerate, sono stati stimati 3 diversi parametri a partire dal semivariogramma empirico costruito sui dati: ϕ , τ^2 e σ^2 .

Predittori ambientali: la variazione della variabile di interesse rispetto alle diverse località in cui questa viene misurata, può direttamente essere dipendente dalla contestuale variazione di altri parametri fisico-chimici, così come anche topografici. Al fine di considerare questi effetti all'interno di un modello geostatistico, si deve considerare che il valore medio μ del “segnale” in analisi $S(x)$ è dato da:

$$\mu = \beta_0 + \sum \beta_i d_i(x)$$

dove, come in un classico modello regressivo di tipo lineare, β_0 rappresenta l'intercetta, mentre di $i = 1, \dots, n$ rappresentano gli *i-esimi* predittori e β_i ; $i = 1, \dots, n$ i rispettivi coefficienti. L'introduzione di predittori all'interno di un modello geostatistico dà vita a un modello definito “a media non costante”. Per questo tipo di modelli, dovranno dunque essere stimati tutti i parametri (coefficienti) relativi a ciascun predittore più, ovviamente, il parametro che esprime l'intercetta.

Per quanto riguarda le concentrazioni di ozono (relativamente a ciascun periodo considerato), sono stati utilizzati, come già ricordato, i seguenti predittori continui: 1) *quota*, 2) *temperatura*, 3) *piovosità*, 4) *latitudine*, 5) *longitudine*. L'analisi ha avuto come obiettivo quello di sviluppare un modello “parsimonioso”, ovvero in grado di spiegare il massimo della varianza con il minor numero di predittori. Per questo scopo, oltre il fit individuale dei predittori rispetto alle concentrazioni di ozono (per

valutare la varianza spiegata da ogni singolo predittore), è stato costruito un modello nullo (solo con i parametri spaziali, come definiti al punto precedente) e un modello totale (considerando invece congiuntamente tutti i predittori all'interno del modello). Dal confronto di questi due modelli, e attraverso una procedura di eliminazione progressiva di predittori dal modello totale (*stepwise* di tipo *backward*) si è giunti alla formulazione del miglior modello parsimonioso.

Tutti i parametri stimati sono stati ottenuti attraverso la massimizzazione della verosimiglianza (*maximum likelihood*) per il cui calcolo si rimanda a testi specialistici (Pawitan 2001). Tutte le analisi sono state effettuate utilizzando il software R (R Development Core Team 2008)

2.3.3 Spazializzazione e rappresentazione

Una volta controllata l'accuratezza del modello utilizzato, i parametri stimati attraverso l'utilizzo del metodo geo-statistico sono stati successivamente utilizzati per la predizione spaziale dei valori delle concentrazioni di ozono per tutta la provincia di Trento. In particolare, per ottenere i valori di ozono per ciascuna delle celle di 1x1 km ottenute coerenti con quelle utilizzate dal modello CALGRID (TECHNE 2006), è stato applicato il metodo definito di "kriging" semplice. Le concentrazioni di ozono, per ogni località di interesse, sono state dunque calcolate come:

$$S(x) = \mu + \sum_{i=1}^n w_i(x)(y_i - \mu)$$

dove $w_i(x)$ sono funzioni esplicite dei parametri di covarianza ϕ , r^2 e σ^2 .

2.3.4 Validazione

Per la validazione di un modello di regressione è consuetudine basarsi su tecniche di *cross validation*, ovvero su metodiche di validazione che prevedono la costruzione del modello su di un dataset privato di una o più osservazioni, che poi vengono successivamente predette dal modello stesso. In base alla capacità del modello di predire le osservazioni escluse è possibile valutare l'efficienza dello stesso. Il metodo di validazione più comune è il cosiddetto *leave one out* (LOO), che consiste nell'escludere un punto per volta. Sono altresì disponibili metodi che prevedono di escludere più osservazioni o elementi della matrice X

contemporaneamente (*leave more out*, *venetian blind*, ecc...), condizione per il loro impiego è di avere un numero sufficientemente alto di osservazioni. Inoltre, per avere una effettiva validazione del modello può altresì essere opportuno testarlo su osservazioni che non sono state usate per costruire il modello stesso. Si crea quindi un *training set* sul quale ci si basa per costruire il modello (impiegando le tecniche di *cross validation* esposte precedentemente) e lo si usa per predire la risposta di osservazioni raccolte nel *validation set*.

Per i modelli sviluppati in questo lavoro, entrambe le tecniche di validazione (*leave-one-out* e *blind validation*) sono state applicate. In particolare, in fase di analisi preliminare dei dati, si è provveduto ad eliminare dal data-set totale, tre stazioni APPA di rilevamento automatico delle concentrazioni di ozono (Monte Gazza, Riva del Garda e Borgo Valsugana). Successivamente, i valori osservati per queste tre centraline sono stati comparati con i valori predetti da entrambe le metodologie adottate (statistica e geostatistica) e per ciascuno dei tre modelli sviluppati (settimana del 13 e del 20 giugno e media delle due settimane). Ai fini della comparazione sono stati utilizzati sia l'RMSEP (Root-Mean-Square Error of Prediction: radice quadrata della discrepanza quadratica media fra i valori osservati ed i valori stimati), la pendenza della regressione lineare dei valori predetti sui valori misurati ed il valore di varianza spiegata dei valori misurati da parte dei valori stimati. .

3 Risultati

3.1 Selezione di variabili e modelli

3.1.1 Approccio statistico

Il modello PLSR è stato eseguito su tre dataset: la settimana del 13-20 di giugno; la settimana 21- 27 di giugno, la media delle due. Le variabili predittrici studiate sono state:

1. coordinate (X ed Y)
2. quota
3. temperatura media sul periodo
4. piovosità media sul periodo

L'analisi è stata eseguita seguendo l'approccio logico di seguito riportato:

1. regressione su tutto il dataset;

2. individuazione delle osservazioni “anomale” (osservazioni con elevato *leverage*, es. osservazioni caratterizzata da una coppia x, y molto diversa dalle altre e che “forza” il modello; *outliers*);
3. esclusione delle suddette osservazioni anomale e nuova regressione sul dataset ridotto, per trovare le variabili più importanti;
4. modello di regressione finale calcolato sul dataset ridotto, contenente solo le variabili selezionate;

Per la valutazione dell'affidabilità del modello sono stati impiegati diversi criteri quali:

- il valore dell' R^2 , ovvero la percentuale di variazione dei valori misurati spiegata dai valori modellati;
- il valore dell' R^2 , calcolato in cross validation mediante metodo *Leave-One-Out* (LOO);
- la distribuzione dei residui, ovvero delle differenze tra valori misurati e valori predetti, che deve approssimare la distribuzione normale.
- la corrispondenza dei valori reali con quelli calcolati in *cross validation*

Nella Tab. 2 sono riportati i siti risultati anomali, che sono quindi stati esclusi dall'elaborazione del modello finale.

Tab. 2. Siti esclusi dai modelli finali. Tra parentesi le variabili risultate anomale.

Periodo	Stazione
13-20 giugno	Peio (bassa T), Coredò (bassa P)
20-27 giugno	Peio (bassa T), Castello Tesino (bassa T), Gazza (bassa T, bassa P), Monte Terlago (bassa P)
13-27 giugno	Peio (bassa T), Castello Tesino (bassa T), Gazza (bassa P), Monte Terlago (bassa P), Margone (bassa P), Maso Ariol (bassa P)

In Tab. 3 si riportano le variabili selezionate per ciascun modello ed i rispettivi i valori dei coefficienti. Longitudine (X, solo per la settimana 13-20 giugno), quota e temperatura media (Tmed) sono risultati i predittori significativi selezionati.

Tab. 3 Variabili selezionate e coefficienti. 13, 20 e 13_20 indicano i periodi di riferimento per le variabili.

Periodo	constant	X	Y	Quota	Tmed13	Tmed20	Tmed13_20	Pio13	Pio20	Pio13_20
13-20 giugno	1675.42	-0.00087		0.0047	-7.34					
20-27 giugno	201.025			0.017		-5.03				
13-27 giugno	160.784			0.013			-4.07			

Nella Tab. 4 sono riportati i dati di R^2 e R^2_{cv} , che danno una idea, sintetica ma efficace, dell'affidabilità dei modelli calcolati.

Tab. 4. Performance dei modelli espressa come varianza spiegata (%). CV=cross-validation.

Periodo	R^2	R^2_{cv}
13-20 giugno	53.42	29.2
20-27 giugno	59.1	45.9
13-27 giugno	55.2	39.2

La varianza spiegata da ciascun modello si attesta generalmente tra il 50% ed il 60%. Il valore di varianza spiegata in cross validation, ovviamente inferiore, è decisamente basso solo per il primo modello. Negli altri due casi si attesta su valori confrontabili con quelli dei corrispondenti R^2 .

3.1.2 Approccio geostatistico

3.1.2.1 Modelli sviluppati per il periodo 13-20 giugno, 21-27 giugno e per i valori mediati delle due settimane

Dalle analisi esplorative effettuate, risulta che la distribuzione dei valori della concentrazione di ozono per la prima settimana considerata è pressoché normale. Tale andamento è stato verificato utilizzando la classica trasformazione Box-Cox e osservando che la verosimiglianza è massimizzata utilizzando i dati originari (parametro $\lambda = 1$, equivale a non effettuare trasformazioni). Per la seconda settimana analizzata e per il periodo di due settimane, si è osservata una leggera asimmetria nella distribuzione di frequenze dei valori di ozono misurati. Il calcolo della massima verosimiglianza rispetto alle diverse tipologie di normalizzazioni effettuate, suggerisce l'utilizzo di una trasformazione di tipo radice quadrata ($\lambda=0.5$).

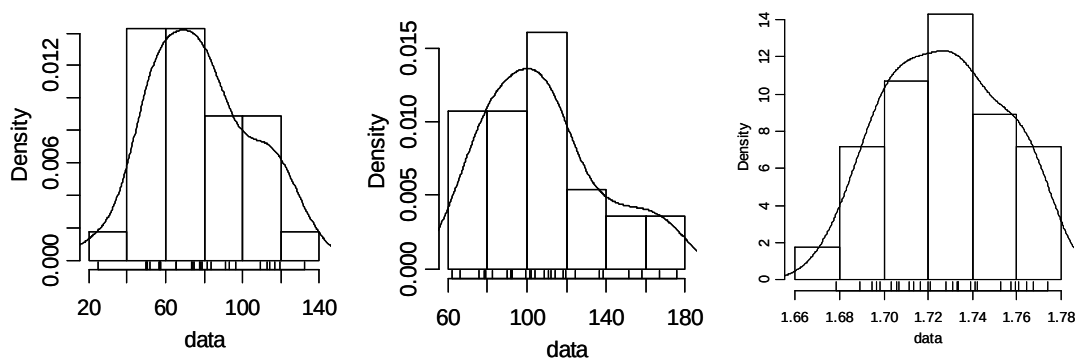


Fig. 10. Analisi della distribuzione delle concentrazioni di ozono (a sx.: settimana 13-20 giugno; al centro: 20-27 giugno; a dx: 13-27 giugno).

Il fit individuale dei singoli predittori ha messo in luce come (Tab. 5):

- per la prima settimana, le temperature, ma soprattutto la quota, concorrono maggiormente a spiegare i valori misurati di ozono. Questo andamento viene confermato anche dal modello ridotto ottenuto in ultima analisi, in cui sia le temperature che la quota risultano incluse (Tab. 6);
- per la seconda settimana, il fit individuale dei predittori rispetto alle concentrazioni di ozono ha mostrato come la quota rappresenti, anche in questo caso, la migliore variabile esplicativa tra quelle utilizzate. Questa osservazione viene confermata dai risultati del modello parsimonioso ottenuto in seguito a selezione stepwise dei predittori (Tab. 6). E' risultato infatti che la sola quota produce una massimizzazione significativa dei valori della verosimiglianza (Tab. 5);
- il modello sviluppato sul dato medio di ozono delle due settimane precedentemente analizzate, mostra caratteristiche molto simili a quello sviluppato per la sola prima settimana (13-20 giugno). Sia il fit individuale che, in definitiva, il modello ridotto, indicano nuovamente che le temperature e soprattutto la quota hanno un elevato peso nel determinare i valori di ozono rilevati (Tabb. 5, 6).

Tab. 5. Stima della massima verosimiglianza per i singoli predittori nei tre periodi considerati (i valori stimati delle componenti spaziali non vengono mostrati).

Fit Individuali	Likelihood		
	13-20 giugno	20-27 giugno	13-27 giugno
Coordinate	-126.7	-134.2	-134.2
Quota	-116.8	-126.1	-120.3
Temperature	-118.2	-135.7	-123.8
Piovosità	-119	-138.3	-133.9

Tab. 6. Descrizione sintetica dei modelli spaziali ridotti ottenuti dopo le procedure di selezione dei predittori. Nel caso della prima settimana, sono state incluse nel modello finale la quota e le temperature medie; per la seconda settimana solo la quota; per il periodo di due settimane la quota e le temperature medie.

Periodo	Modello	Intercetta	X	Y	Quota	T(°)	P(mm)	σ^2	Parametri Spaziali			logL
13-20 giugno	Totale	3895.51	-0.0002	-0.0007	0.0135	-4.778	-2.361	432.04	7607.89	0		-114.9
13-20 giugno	Ridotto	80.88			0.018	-1.464		649.52	13893.15	0		-116.7
20-27 giugno	Totale	2670.98	0.0001	-0.0005	0.02	-2.55	-1.456	461.4609	8978.131	105.6671		-124.6
20-27 giugno	Ridotto	60.69			3.84E-02			9.84E+02	3.67E+04	1.23E+02		-126.1
13-27 giugno	Totale	3608.99	0.027	-3.46	-7.36	0	-0.001	456.57	14945.21	28.31		-117.8
13-27 giugno	Ridotto	86.08			0.028	-1.45		541.09	15193.41	27		-119.5

La cross-validation mediante LOO ha permesso di evidenziare come ci sia una buona corrispondenza tra valori predetti ed osservati, con uno scarto medio (calcolato come media dei valori assoluti degli scarti) pari a 14.14 (prima settimana), 15.45 (seconda settimana) e 12.91 $\mu\text{g}\cdot\text{m}^{-3}$ (periodo complessivo 13-27 giugno?). La Fig. 11 riporta i confronti tra valori stimati e misurati e l'andamento spaziale delle differenze.

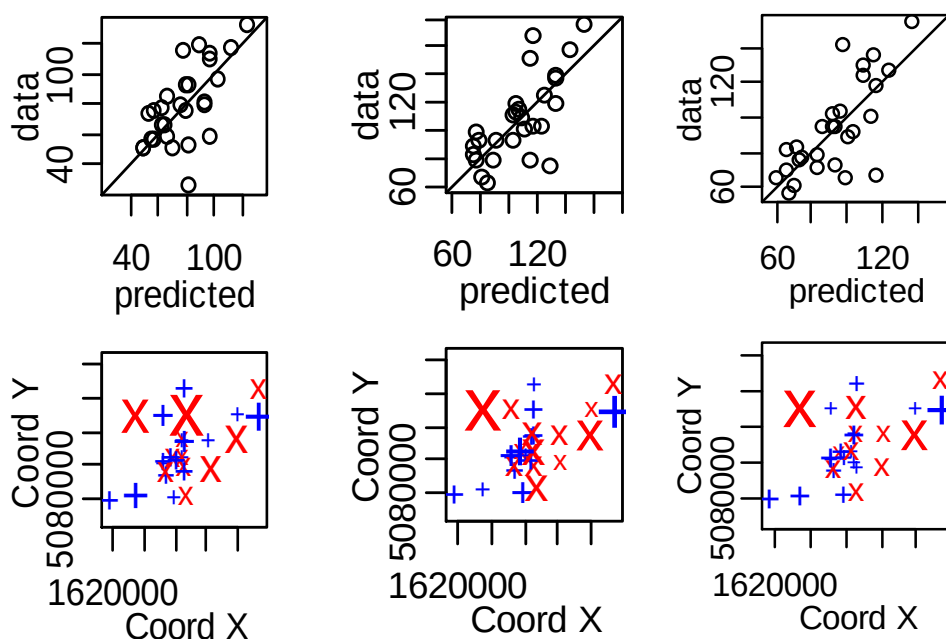


Figura 11. In alto: confronto tra valori predetti (predicted) dal modello spaziale in seguito a LOO e valori realmente osservati (data) per le settimane 13-20 giugno, 20-27 giugno e 13-27 giugno (da sx a dx, rispettivamente). In basso: differenza tra valori osservati e predetti nello spazio (la grandezza delle croci esprime il valore assoluto dello scostamento, il blu esprime un errore positivo, il rosso negativo)

Osservando l'andamento degli scarti, si osserva come le predizioni siano in genere maggiormente precise nel centro dell'area di studio, lì dove sono disponibili un elevato numero di misurazioni (accuratezza maggiore).

3.1.2.2 Modello sviluppato per il periodo 13 giugno – 12 settembre

L'ultimo modello sviluppato è relativo ad un intervallo temporale molto più ampio (9 settimane in totale) e considera un sotto-insieme di tutte le unità di rilevamento (ovvero i 15 siti provinciali e le 7 centraline APPA). La distribuzione delle frequenze per questi dati evidenzia una forte asimmetria e l'analisi dei parametri da adottare per normalizzarli suggerisce l'applicazione di una trasformazione di tipo radice quadrata ($\lambda=0.5$) o l'utilizzo dell'inverso del valore originale ($\lambda=-1$). Si è comunque deciso di optare per l'applicazione della trasformazione in radice quadrata, per ottenere dati facilmente comparabili con i precedenti modelli.

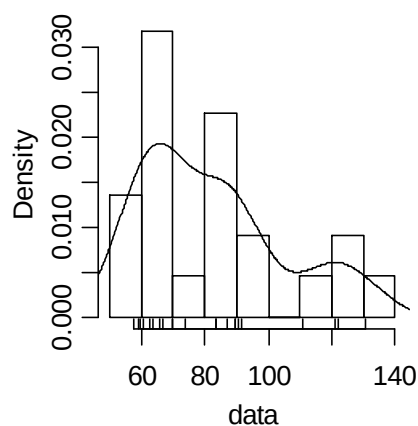


Figura 12. Analisi della distribuzione delle concentrazioni di ozono.

Il dato trasformato è stato successivamente utilizzato, come di consueto, per il fit individuale rispetto ai predittori e successivamente per la costruzione del modello finale parsimonioso. In questo ultimo modello, sebbene la quota rappresenti ancora un predittore altamente correlato con la concentrazione di ozono (come per tutti gli altri modelli sviluppati), si è osservato l'alto potere esplicativo della variabile relativa alle precipitazioni atmosferiche (Tab. 7). Tale variabile è infatti stata inclusa anche all'interno del modello ridotto (Tab. 8).

Tab. 7. Stima della massima verosimiglianza per i singoli predittori (i valori stimati delle componenti spaziali non vengono mostrati).

Fit Individuali	Likelihood
Coordinate	-98.38
Quota	-91.4
Temperature	-96.25
Piovosità	-92.91

Tab. 8. Descrizione sintetica del modello spaziale ridotto ottenuto dopo le procedure di selezione dei predittori. In questo caso sono state incluse nel modello finale la quota e le precipitazioni.

Modello	Intercetta	X	Y	Quota	T(°)	P(mm)	σ^2	Parametri Spaziali $\square$	τ^2	logL
Totale										-88.7
Ridotto	8.919			0.022		1.598	1.668	4.561e+04	1.200	-89.15

In seguito alla procedura di LOO, lo scarto medio tra valori osservati e predetti è risultato pari a $13.98 \mu\text{g}\cdot\text{m}^{-3}$, anche se è possibile notare una minore accuratezza nelle singole previsioni (la dimensione dei simboli nel grafico del LOO risulta notevolmente superiore rispetto ai precedenti modelli).

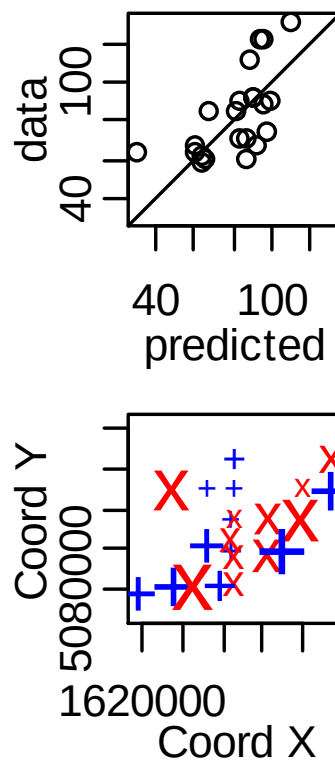


Figura 13. Nel primo grafico, confronto tra valori predetti dal modello spaziale in seguito a LOO e valori realmente osservati. Nel secondo grafico, differenza tra valori osservati e predetti (la grandezza delle croci esprime il valore assoluto dello scostamento, il blu esprime un errore positivo, il rosso negativo).

3.3 Risultati in blind validation

Nella costruzione dei modelli del periodo 13-27 giugno non sono state deliberatamente incluse le stazioni APPA di Monte Gazza, Riva del Garda e Borgo Valsugana allo scopo di utilizzare queste stazioni come validazione “cieca”. In altre parole, il modello è stato costruito facendo finta che queste stazioni non esistessero; in un secondo momento le stesse sono state usate come siti indipendenti dove confrontare previsioni e dati misurati. La scelta di queste stazioni è stata dettata da considerazioni riguardanti il range di ozono (differenze tra stazioni superiori al 100% dei valori minimi) e di quota (da stazioni di bassa quota sino ai circa 1700 m s.l.m. del Monte Gazza) e la dislocazione spaziale (parte meridionale, centro-occidentale e orientale della provincia).

Per i tre modelli in cui è stato possibile effettuare la “blind validation”, la Fig. 14 e la Tab. 9 riportano i confronti tra i valori predetti dai due approcci adottati.

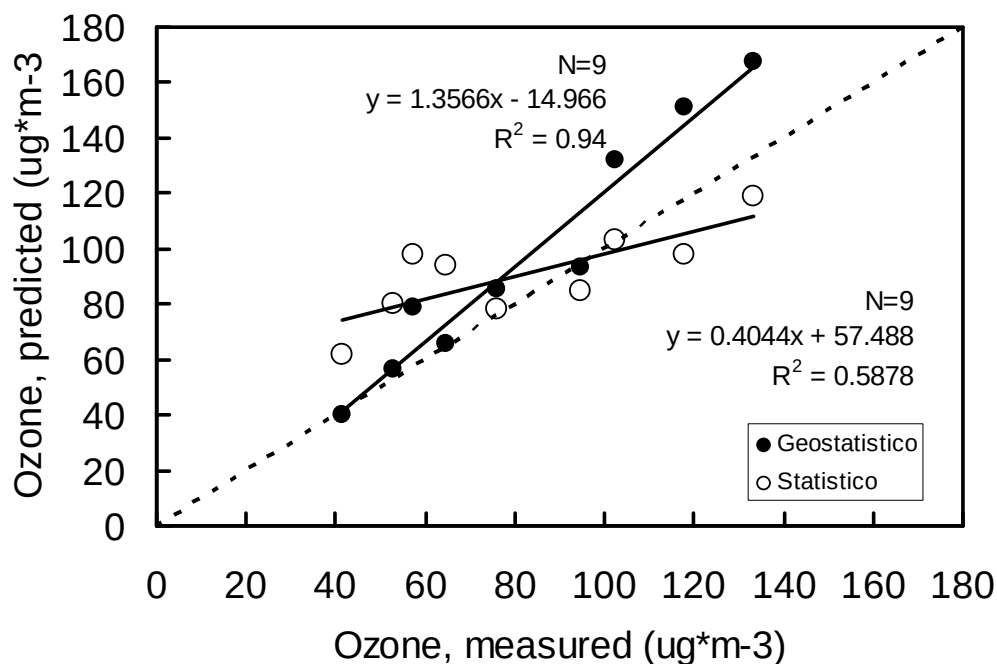


Fig. 14. Ozono misurato ed ozono predetto dai due approcci adottati per le stazioni di Monte Gazza, Borgo Valsugana e Riva del Garda.

Tab. 9. Confronto tra valori misurati dalle centraline di Monte Gazza, Borgo Valsugana e Riva del Garda e valori predetti dal modello statistico (S) e geostatistico (GS). Sono riportati anche gli scarti in % del valore misurato.

Periodo	Stazione	Misurato APPA	Predetto S	Predetto GS	Scarto S, %	Scarto GS, %
13-20 giugno	MONTE GAZZA	102.3	103.0	132.0	0.7	29.0
13-20 giugno	RIVA GAR	57.4	98.0	78.9	70.6	37.4
13-20 giugno	BORGO VAL	41.4	62.0	40.4	49.6	-2.5
21-27 giugno	MONTE GAZZA	133.3	119.0	167.4	-10.7	25.6
21-27 giugno	RIVA GAR	94.9	85.0	93.1	-10.4	-1.9
21-27 giugno	BORGO VAL	64.6	94.0	65.8	45.6	1.9
13-27 giugno	MONTE GAZZA	117.8	98.0	151.1	-16.8	28.3
13-27 giugno	RIVA GAR	76.2	78.0	85.3	2.4	12.0
13-27 giugno	BORGO VAL	53.0	80.0	56.4	50.9	6.4

Osservando i valori riportati in Tab. 9 e Fig. 14, si nota che il modello statistico per la settimana del 13-20 giugno è quello che ha dato prestazioni peggiori. I due modelli sono stati infine confrontati utilizzando il RMSEP (Root mean square error in prediction). Come si può vedere in tabella 10, il modello geostatistico ha offerto performance predittive migliori per la prima settimana, mentre le prestazioni sono equivalenti per la seconda settimana e per la media delle due.

Tab 10. RMSEP per confrontare i due modelli.

Modello	13-20 giugno	20-27 giugno	13-27 giugno
Modello statistico	45.5	33.6	33.4
Modello geostatistico	35.4	33.8	34.2

3.4 Distribuzione spaziale dei livelli di ozono in Trentino nell'estate 2007

Le Figg. 15-18 riportano i valori di ozono “in continuo” (senza ripartizione in classi) stimati sulla base del modello geostatistico nei differenti periodi presi in considerazione e la relativa varianza. In sintesi, si può dire che le concentrazioni medie più elevate vengono stimate per le zone occidentali e nord-orientali della provincia, mentre i fondovalle e la parte centrale mostrano concentrazioni più basse. Come si nota dalle figure, questo è anche l'andamento della varianza associata alle stime che corrisponde in parte alla densità delle misurazioni di ozono, praticamente assenti nel settore centro-occidentale della provincia.

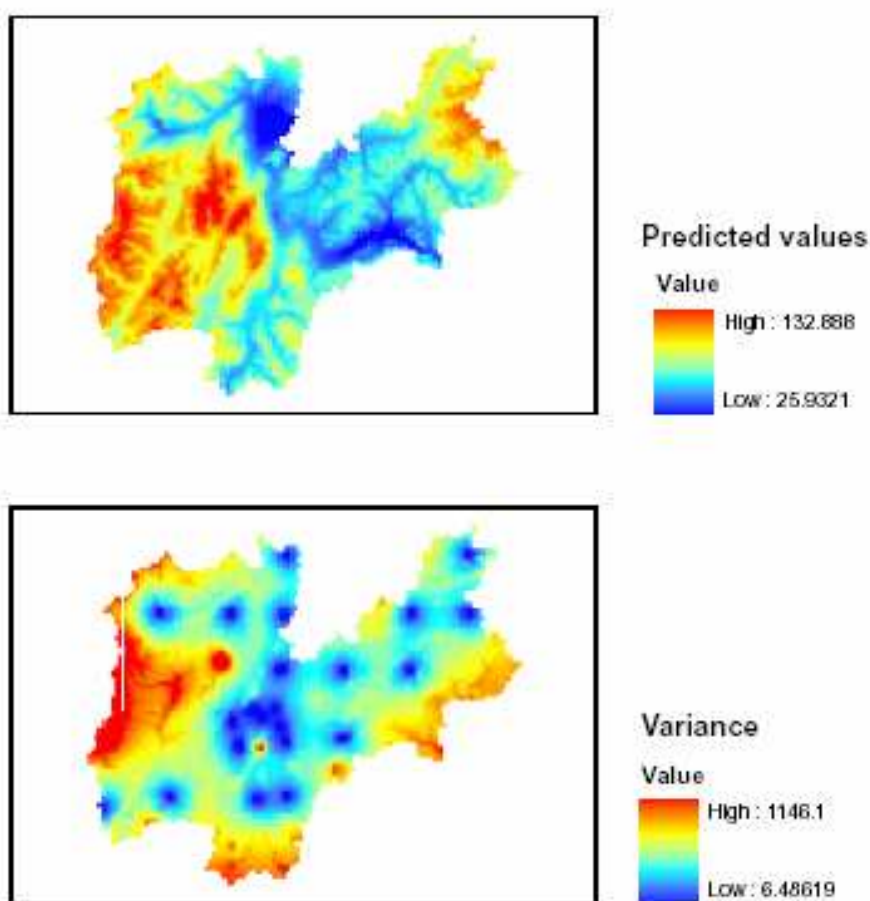


Fig. 15. Modello previsionale per la settimana dal 13 al 20 Giugno.

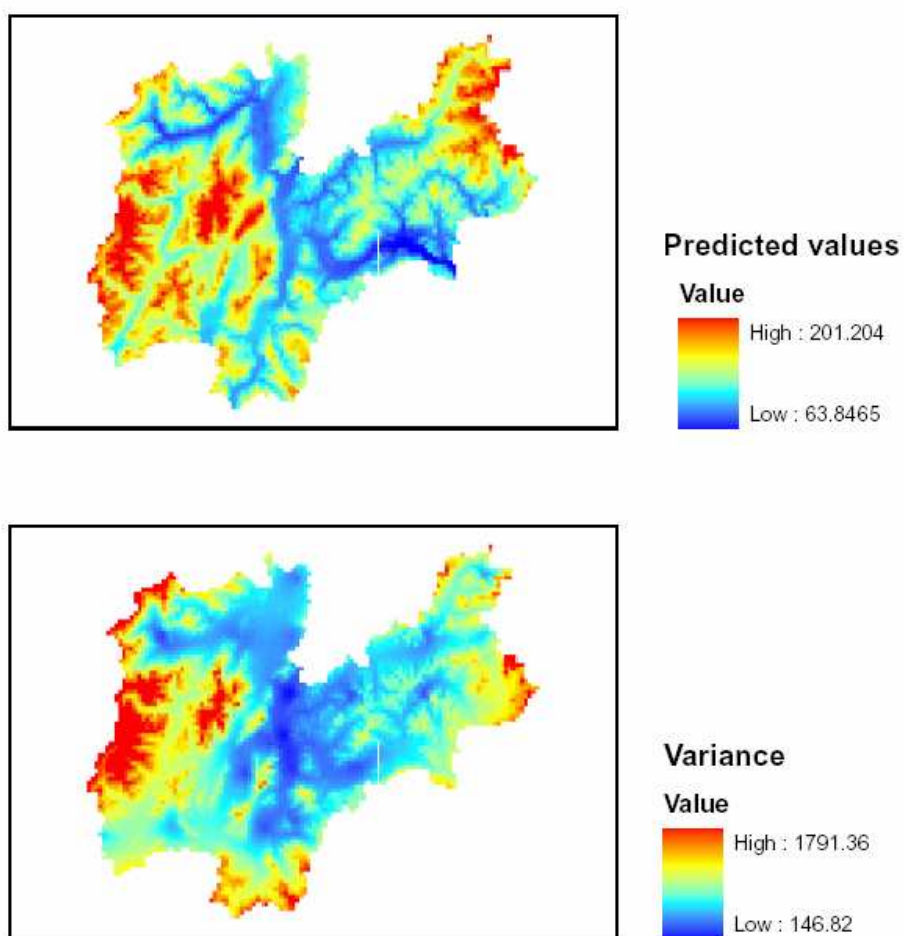


Fig. 16. Modello previsionale per la settimana dal 21 al 27 Giugno.

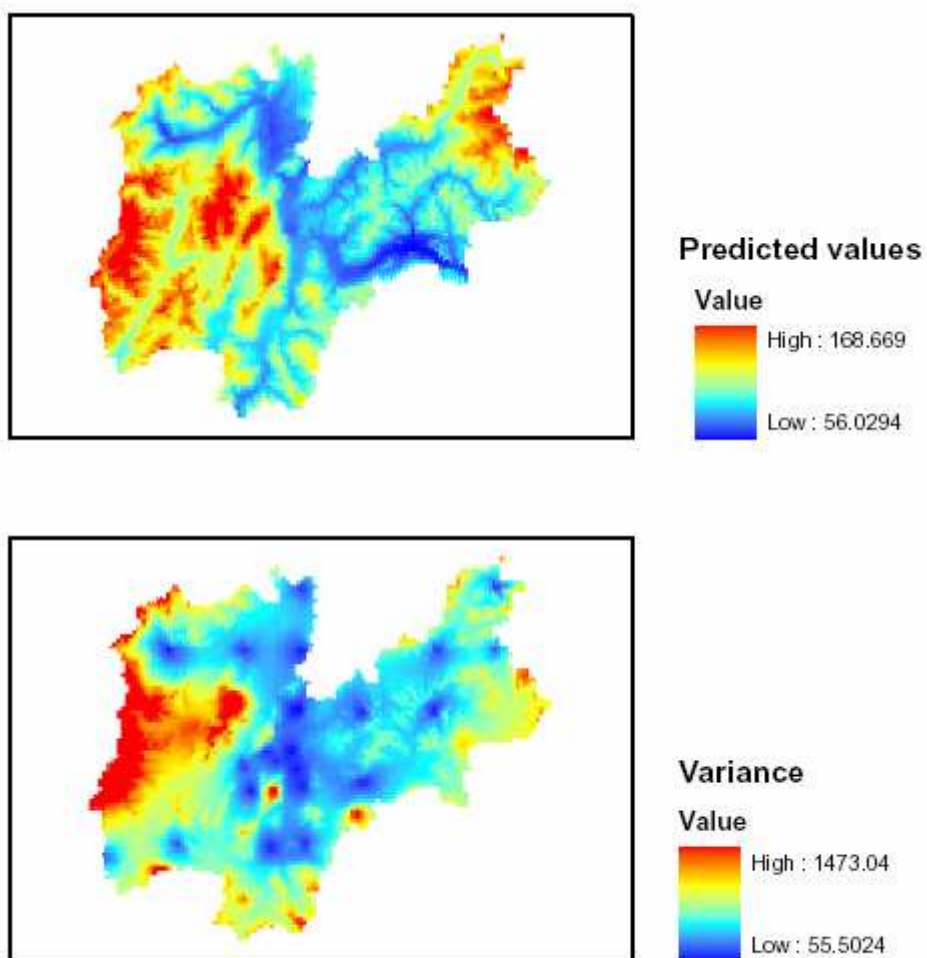


Fig. 17. Modello previsionale basato sulla media delle due settimane tra il 13 ed il 27 giugno.

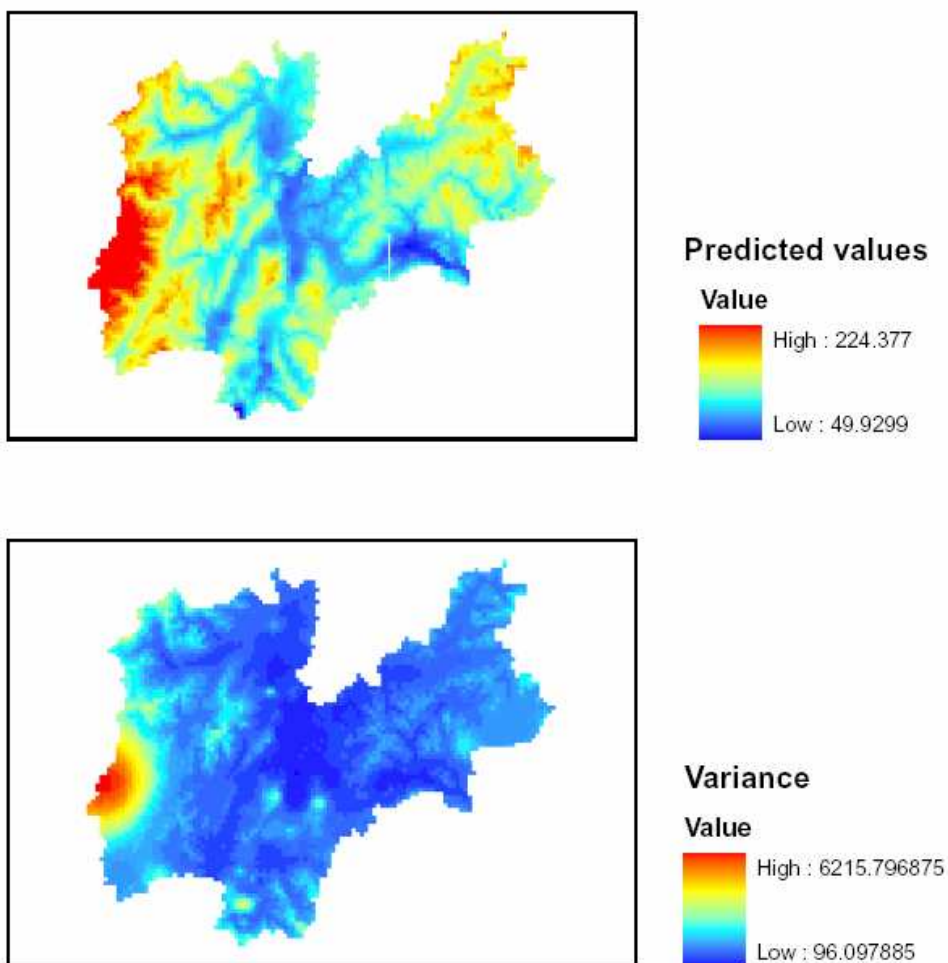


Fig. 18. Modello previsionale per le concentrazioni medie tra il 13 giugno ed il 19 settembre.

Al fine di rendere comparabili le mappe e di tener conto dell'incertezza delle stime ($15.0 \mu\text{g}\cdot\text{m}^{-3}$, come media degli scarti assoluti in blind validation), la distribuzione spaziale dei livelli di ozono è stata poi riportata in classi di concentrazione (Fig. 19). E' evidente la notevole differenza tra le settimane del 13-20 e 20-27 giugno 2007, rilevata del resto anche dalle centraline APPA (Fig. 20), e che evidenzia, oltre ad un piccolo episodio di ozono, anche il potenziale informativo che porterebbe con se l'adozione di una componente temporale nel modello: sarebbe possibile realizzare una sequenza continua di mappe settimanali dell'ozono in provincia di Trento.

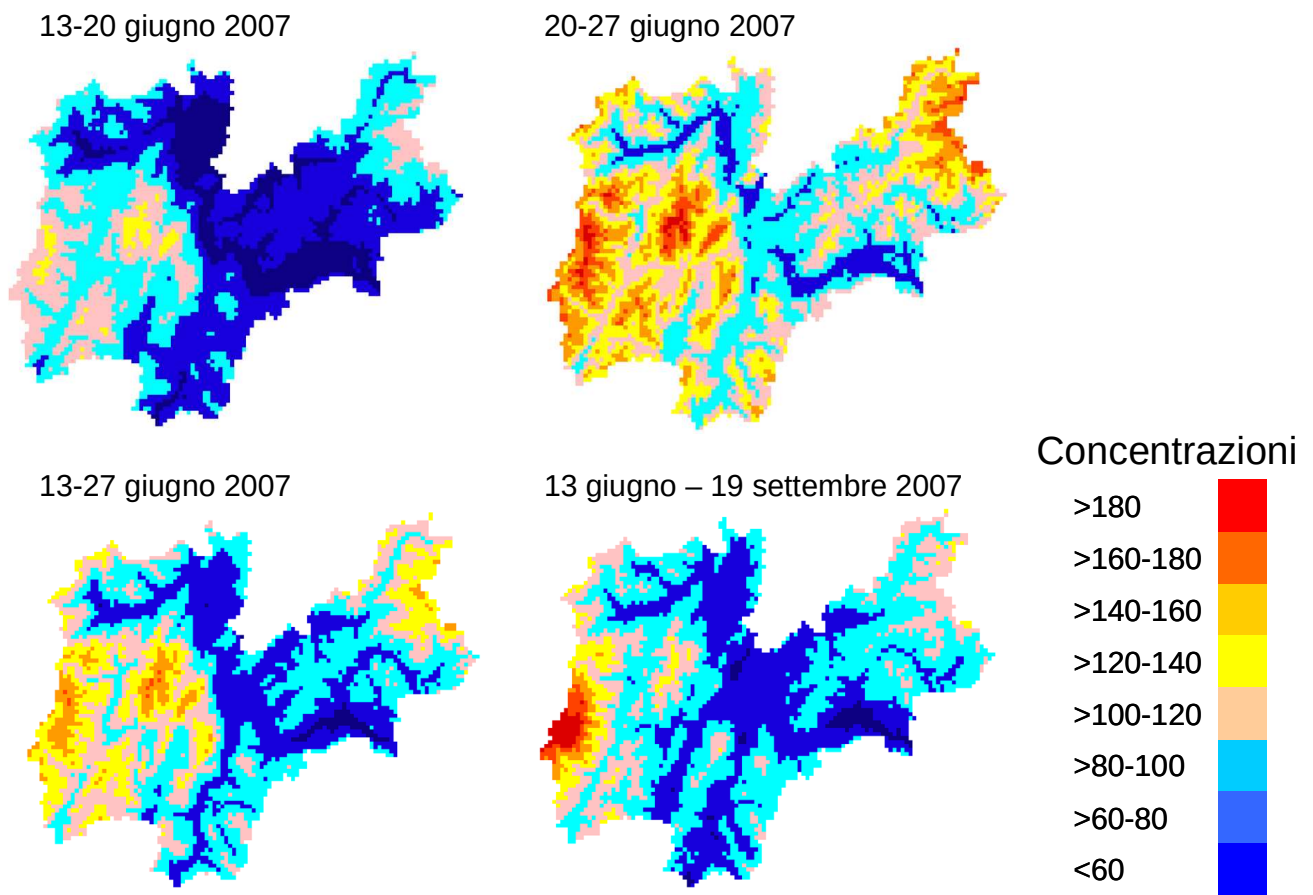


Fig. 19. Visualizzazione delle mappe della distribuzione spaziale delle concentrazioni di ozono ($\mu\text{g}\cdot\text{m}^{-3}$) in Trentino nell'estate 2007.

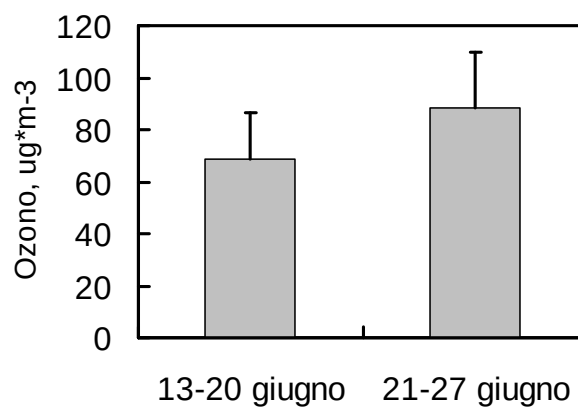


Fig. 20. Concentrazione di ozono media registrata dalle centraline APPA nelle settimane 13-20 e 21-27 giugno 2007.

Le stime effettuate indicano valori di ozono più elevati (fino anche ad oltre 180 $\mu\text{g}\cdot\text{m}^{-3}$) nella parte occidentale della provincia, a ridosso dell'Adamello, e nel settore nord-orientale, dalla Marmolada al Primiero (120-180 $\mu\text{g}\cdot\text{m}^{-3}$). Le zone a quota inferiore, i fondovalle, sembrano essere soggette a concentrazioni più basse, probabilmente in virtù della rimozione chimica dell'ozono operata dagli altri inquinanti. Questo *pattern* sembra piuttosto costante nel tempo.

Un modello lineare semplice lega le concentrazioni medie stagionali ed AOT40 (Fig. 21). Nella Fig. 22 è riportato un primo tentativo di stima dell'AOT40 per la provincia di Trento. Assumendo il modello lineare utilizzato, nelle zone a concentrazione di ozono più elevata, l'AOT40 (stimato sul periodo metà Giugno - metà Settembre) arriva a valori 3-4 volte superiori al valore obiettivo stabilito dalla Direttiva EU per la protezione della vegetazione (18000 $\mu\text{g}\cdot\text{m}^{-3}\cdot\text{h}$) ed eccede di 6-7 volte il Livello Critico UN/ECE (10000 $\mu\text{g}\cdot\text{m}^{-3}\cdot\text{h}$) per la vegetazione sensibile ed in condizioni sensibili. Occorre tuttavia ricordare che esistono aree con misurazioni assenti (es. area centro-occidentale della provincia) che possono avere influenza nella formulazione del modello e nella precisione delle stime.

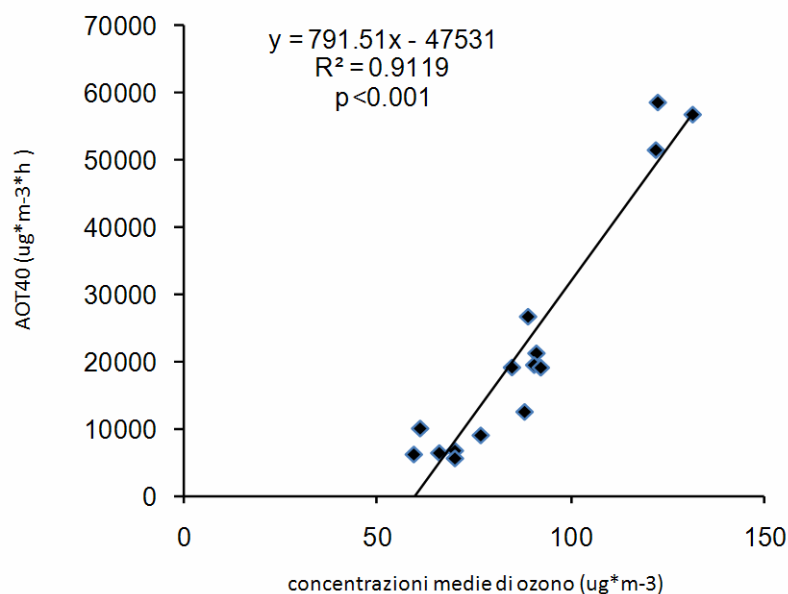


Fig. 21. Relazione tra AOT40 e le concentrazioni medie di ozono. Attraverso la relazione empirica $y=791.51x-47531$ sono stati spazializzati i valori di AOT40 a partire dai valori delle concentrazioni medie stagionali di ozono.

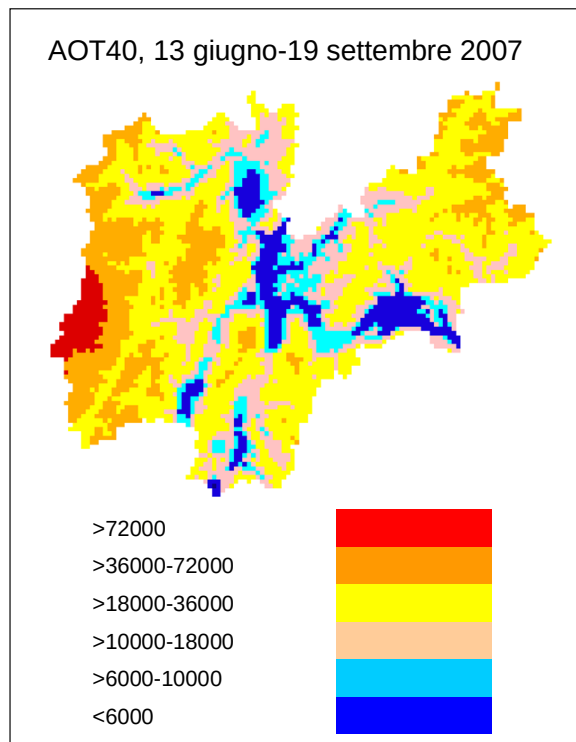


Fig. 22. Spazializzazione dell'AOT40 (espresso in $\mu\text{g}\cdot\text{m}^{-3}\cdot\text{h}$) derivato secondo il modello di regressione lineare. Dal colore salmone in poi si evidenzia un superamento del Livello Critico UN/ECE per le foreste (specie sensibili in condizioni sensibili), anche se il tempo di accumulo considerato (metà giugno-metà settembre) è la metà di quello stabilito dall'UN/ECE (1 aprile-30 settembre); dal giallo in poi si evidenzia un superamento dei valori obiettivo della Direttiva UE (vegetazione).

5. Conclusioni e prospettive

E' stato realizzato un modello geostatistico per stimare le concentrazioni di ozono in provincia di Trento. Il modello adottato è stato scelto dopo aver confrontato i risultati di con quelli forniti da un approccio statistico. Entrambi gli approcci avevano individuato gli stessi predittori (quota, temperatura e precipitazioni), ma il modello geostatistico ha complessivamente offerto *performance* predittive migliori, con errori più contenuti ed un andamento più coerente con i valori misurati. Tra i predittori, la quota è risultata sempre significativa, mentre il ruolo di temperatura e precipitazioni varia a seconda del dominio temporale considerato. A livello stagionale (giugno-

settembre), quota e precipitazioni risultano i predittori che ottimizzano il modello (modello parsimonioso).

Le stime effettuate indicano concentrazioni più elevate ($180 \mu\text{g}\cdot\text{m}^{-3}$) nella parte occidentale della provincia, a ridosso dell'Adamello. Valori elevati ($120\text{-}180 \mu\text{g}\cdot\text{m}^{-3}$) si riscontrano anche nel settore nord-orientale, dalla Marmolada al Primiero. Assumendo un modello lineare semplice che lega concentrazioni medie stagionali ed AOT40 (verificato sui dati rilevati), in queste zone l'AOT40 (stimato sul periodo metà giugno-metà settembre) arriva a valori 3-4 volte superiori al valore obiettivo stabilito dalla Direttiva EU per la protezione della vegetazione ($18000 \mu\text{g}\cdot\text{m}^{-3}\cdot\text{h}$) e 6-7 volte superiore al Livello Critico UN/ECE. Merita ricordare tuttavia che esistono aree con misurazioni assenti (es. area centro-occidentale della provincia) che possono avere influenza nella formulazione del modello e nella precisione delle stime.

I risultati dimostrano che è possibile modellizzare per via statistica/geostatistica le concentrazioni medie settimanali di ozono a scala provinciale, con livelli di incertezza noti ed accettabili. Alla prova della validazione cieca, la differenza tra dati misurati (medie settimanali) e dati stimati è risultata mediamente del 16.1%, ca. $15 \mu\text{g}\cdot\text{m}^{-3}$. A causa della definizione diversa del parametro di incertezza questo dato non è direttamente comparabile con i requisiti di incertezza per la modellistica e per le stime stabilito dalla Direttiva (incertezza del 50% sui valori orari o sulla media delle otto ore; 75% sui valori obiettivo²). Tuttavia, è prevedibile il rispetto di questi requisiti, come evidenziato dalle modellizzazioni dei dati orari, basata sui profili orari giornalieri e portata avanti nel 2007 (Gottardini et al., 2007; vedi anche Gerosa et al., 2007). E' quindi un obiettivo realistico la stima dei valori orari con incertezze accettabili e quindi anche l'irrobustimento della stima dell' AOT40.

I risultati del lavoro aprono quindi la strada a prospettive di sviluppo interessanti, tra cui:

- completamento della rete anche nelle zone non ancora coperte, in modo da diminuire gli errori e le incertezze;

² "L'incertezza per la modellizzazione è definita come lo scarto massimo dei livelli di concentrazione misurati e calcolati del 90 % dei singoli punti di monitoraggio, nel periodo considerato, dal valore limite (o dal valore obiettivo in caso di ozono), a prescindere dalla tempistica degli eventi. L'incertezza per la modellizzazione va interpretata come applicabile nell'intorno dell'opportuno valore limite (o valore obiettivo in caso di ozono). Le misurazioni in siti fissi selezionate ai fini di un raffronto con i risultati della modellizzazione sono rappresentative della scala coperta dal modello. L'incertezza per la stima obiettiva è definita come lo scarto massimo dei livelli di concentrazione misurati e calcolati, nel periodo considerato, dal valore limite (o dal valore obiettivo in caso di ozono), a prescindere dalla tempistica degli eventi."

- inserimento della componente dinamica (tempo) nella modellistica e stabilizzazione del modello, fino al mappaggio su base settimanale;
- stima di rischio potenziale per i vari recettori: foreste, coltivazioni e - non ultimo - salute umana;
- mappe annuali del rischio potenziale per i vari recettori e relative incertezze;
- un sistema routinario in grado di misurare i progressi verso il raggiungimento di riduzioni di territorio provinciale soggetto a eccedenze di ozono;
- basi per lo studio dei possibili effetti dell'ozono sugli ecosistemi forestali, specialmente in considerazione della dotazione di dati già disponibile per la provincia.

Come si vede, le possibilità applicative dello studio svolte sono molte e tali da interessare diversi enti ed agenzie preposte alla tutela della qualità dell'ambiente.

6. Bibliografia ed ulteriori letture

- Box G. E. P., Cox D. R., 1964. An analysis of transformations (with discussion).
Journal of the Royal Statistical Society, Series B, 26: 211-252.
- Burrough P. A., McDonnell, R. A., 1998. Principles of Geographic Information Systems. Oxford Press University, New York.
- Cressie N., 1993. Statistics for Spatial Data. Wiley, New York.
- Diggle P. J., Ribeiro J. P. Jr., 2007. Model-based Geostatistics. Springer.
- DIRETTIVA 2008/50/CE del 21 maggio 2008. Gazzetta ufficiale dell'Unione europea, 11.6.2008
- Gerosa G., Ferretti M., Bussotti F., Rocchini D., 2007. Estimates of ozone AOT40 from passive sampling in forest sites in South-Western Europe. Environmental Pollution, 145: 629-635.
- Goodchild M.F., 2004. GIScience, Geography, Form, and Process. Annals of the Association of American Geographers 94, 709-714
- Gottardini E., Cristofori A., Cristofolini F., Ferretti M, 2008. Rilevamento di Ozono presso i siti forestali UN/ECE Level I nella Provincia di Trento nel 2007. Istituto Agrario di San Michele all'Adige (IASMA). Centro Sperimentale. Relazione dello studio realizzato nell'ambito della convenzione tra Servizio Foreste e IASMA: 26 pp.

- Gottardini E., Cristofori A., Cristofolini F., Ferretti M., 2007. Bio-monitoraggio integrato dell'ozono in provincia di Trento. Istituto Agrario di San Michele all'Adige (IASMA). Centro Sperimentale. Relazione dello studio realizzato nell'ambito della convenzione tra APPA Trento e IASMA nel campo della tutela e studio della qualità dell'aria (Det. N. 174/2 del 09.07.2007).
- Lam, N. S. N. 1983. Spatial interpolation methods: a review. *The American Cartographer* 10: 129-149.
- Longley, P. A., Goodchild, M. F., Maguire, D. J. & Rhind, D. W. 2003. *Geographic Information Systems and Science*. John Wiley & Sons, Chichester, New York, Weinheim, Brisbane, Singapore, Toronto.
- Mangoni M., Buffoni A., 2008. Status and trend of ground-level ozone at the CONECOFOR plots, 1996-2005. In: Ferretti M., Bussotti F., Fabbio G., Petriccione B., (Eds.), 2008. *Ecological condition of selected forest ecosystems in Italy. Status and changes 1995-2005. Fourth report of the Task Force on Integrated and Combined (I&C) evaluation of the CONECOFOR programme*. *Annali Istituto Sperimentale per la Selvicoltura*, Vol. 34, 2005-2006: 85-100.
- Pawitan Y., 2001. In *All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, Oxford.
- R Development Core Team, 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rocchini D., Ricotta C. 2007. Are landscapes as crisp as we may think? *Ecological Modelling* 204, 535-539.
- UN-ECE, 2004. Mapping critical levels for vegetation (Chapter 3). In: *Revised Manual on Methodologies and Criteria for Mapping Critical Levels/Loads and Geographical Areas Where They are Exceeded*. Umweltbundesamt, Berlin, Germany, 53 p.
http://www.oekodata.com/icpmapping/pub/manual_2004/mapman_3.pdf.
- TECHNE Consulting s.r.l, 2006. Studio modellistico sulla generazione e trasporto dell'ozono. APPA TN, N.Raccolta 13/05, pp.106.
- Vingarzan R., 2004. A review of surface ozone background levels and trends. *Atmospheric Environment*, 38: 3431–3442
- Wold H, 1982. Soft modelling, The basic design and some extensions, in: K.-G. Joreskog, H. Wold Eds. , *Systems Under Indirect Observation*, vols. I and II, North-Holland, Amsterdam.

Wold S., et al., 2001. PLS-regression: a basic tool of chemometrics Chemometrics and Intelligent Laboratory Systems 58: 109 – 130 115

