Alma Mater Studiorum
Università di Bologna

Scuola di Economia, Management e Statistica
Corso di Laurea in Scienze Statistiche
Curriculum Statistico-Matematico

Tesi di laurea

Interpolazione spaziale del PM_{10} mediante l'impiego di mappe di uso del suolo

Relatore:

Prof. Fedele Greco

Presentata da:

Carlo Cavalieri

Sessione I
Anno Accademico 2016/2017

Sommario

1. Introduzione

In questa tesi si affronterà il problema di effettuare previsioni spaziali accurate della concentrazione media giornaliera di PM_{10} nel territorio della regione Emilia-Romagna utilizzando in modo efficiente i valori acquisiti da 46 centraline facenti parte della rete di monitoraggio della qualità dell'aria gestita dall'Agenzia regionale per la prevenzione, l'ambiente e l'energia dell'Emilia-Romagna (Arpae).
Il PM_{10} (Particulate Matter) è un inquinante presente in atmosfera sotto forma di particelle microscopiche aventi diametro aerodinamico inferiore o uguale a 10 micrometri, la cui unità di misura è microgrammi/m3. Si tratta di una delle frazioni di maggiore interesse con cui viene classificato il particolato.
Secondo quanto riportato nell'inventario delle emissioni della Regione Emilia-Romagna nell'anno 2010 (INEMAR - Arpa Emilia-Romagna 2013), circa il 40% del totale delle emissioni deriva da processi di combustione non industriale (attività di riscaldamento, produzione di acqua calda e cottura cibi), mentre il 34% dai trasporti stradali.
Il PM_{10} presenta una forte componente stagionale, con valori medi più alti osservati in corrispondenza del periodo invernale attribuibili all'attività degli impianti di riscaldamento.
La distribuzione delle emissioni e il fatto che la maggior parte delle stazioni di rilevamento è situata nei centri abitati lasciano intendere che per ottenere una buona previsione spaziale non si può prescindere dal tenere in considerazione le caratteristiche locali di uso e copertura del suolo onde evitare che le concentrazioni di inquinanti rilevate nei centri urbani vengano estrapolate nelle zone rurali e alpine. Al riguardo è utile ricordare che qualsiasi metodo di previsioni spaziale richiede come prerequisito l'omogeneità spaziale dei valori di input. Per fronteggiare questa necessità è stato utilizzato e adattato alla realtà regionale il modello statistico di interpolazione spaziale RIO sviluppato da Janssen et al. (2008) in risposta alle esigenze dell'agenzia interregionale belga per l'ambiente IRCEL.

Distribuzione percentuale delle emissioni in atmosfera dei principali inquinanti per macrosettore (anno 2010)
Fonte: Regione Emilia-Romagna, Arpa Emilia-Romagna

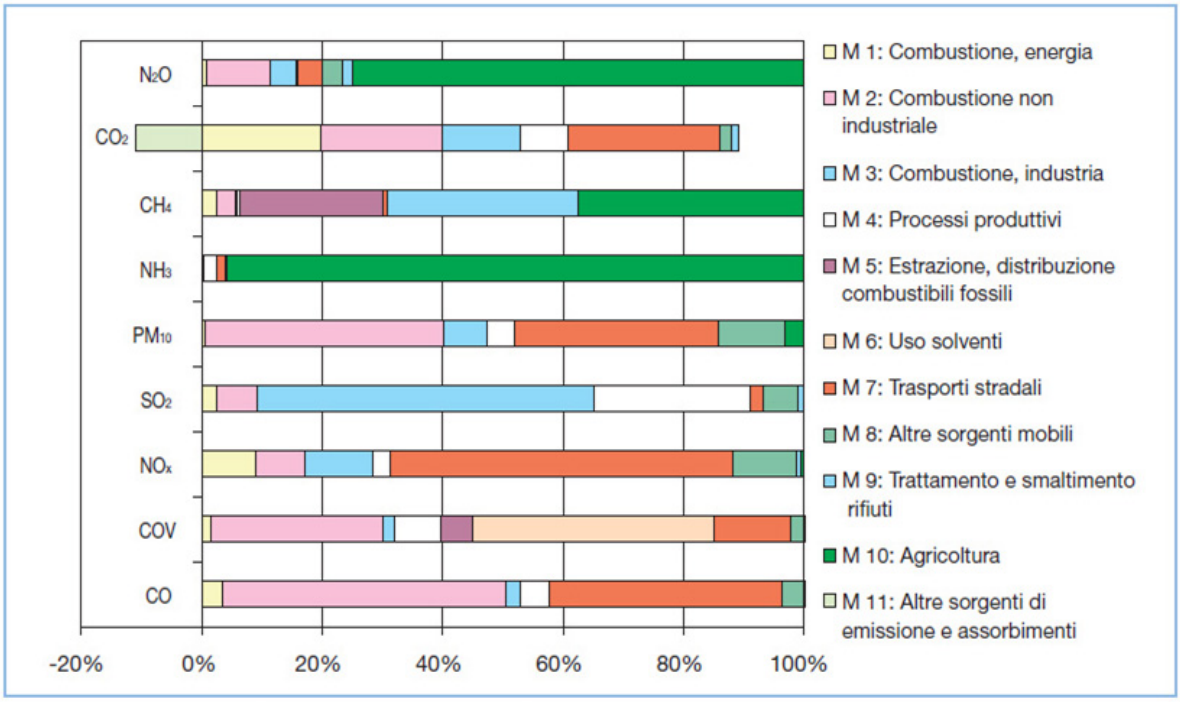

Figura 1

2. Acquisizione ed analisi preliminari dei dati

I dati delle rilevazioni sulla qualità dell'aria raccolti ed elaborati dalla rete di monitoraggio Arpae sono disponibili sul portale https://*dati.arpae.it*.
Una volta scaricati, i dataset contenenti i dati storici 2010-2016 relativi al PM_{10} (concentrazione media giornaliera) sono stati uniti in un unico file csv contenente 97719 osservazioni.
Visto che nel corso del tempo sono state installate nuove stazioni di monitoraggio mentre altre sono state dismesse, si è ritenuto opportuno circoscrivere l'analisi ai soli dati relativi agli anni 2015-2016 e alle 46 centraline attive nell'intero periodo dei due anni (sulle 48 totali definite in "Anagrafe stazioni 2015"), che ammontano a 32525 osservazioni. In questo lasso di tempo i dati mancanti sono mediamente 12 giorni annuali per centralina, una cifra contenuta che non complica l'analisi.

Giorno	Media PM_{10}	STD PM_{10}
Lunedì	24.63	13.93
Martedì	26.93	15.58
Mercoledì	27.15	16.26
Giovedì	26.45	16.73
Venerdì	27.70	16.82
Sabato	26.59	16.27
Domenica	23.28	12.62

Tabella 1 – Media e deviazione standard dei valori della concentrazione di PM_{10} per giorno della settimana

Di seguito vengono presentati una tabella e alcuni grafici descrittivi del dataset: la Tabella 1 e la Figura 2 mostrano la media e la deviazione standard dei valori della concentrazione di PM_{10} suddivisi rispettivamente per giorno della settimana e per stazione di monitoraggio. La Figura 3 mostra le serie storiche complete dei dati 2010-2016 mentre la figura 4 quelle impiegate per l'analisi. Nel grafico a barre si osserva che le localizzazioni campionate sono caratterizzate da valori medi e variabilità differenti, laddove dai grafici delle serie storiche emerge la presenza di una componente stagionale.

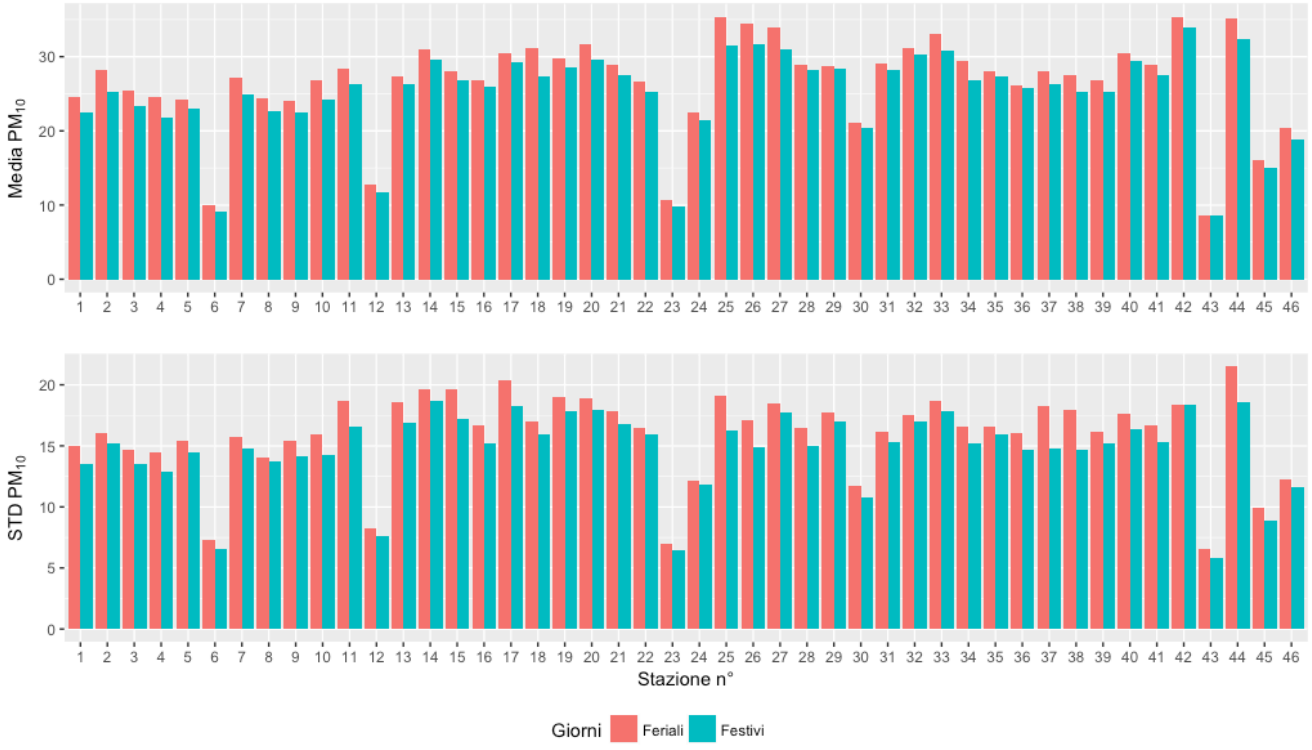

Figura 2 – Media e deviazione standard dei valori di PM_{10} nelle 46 stazioni di monitoraggio impiegate per l'analisi. Sono mostrati fianco a fianco i valori relativi ai giorni feriali e festivi (sabato e domenica).

Figura 3 – Serie storiche 2010-2016 della concentrazione media giornaliera del PM_{10} rilevata nelle stazioni di monitoraggio Arpae (97719 osservazioni totali). Nell'intestazione è riportato il relativo codice Arpae.

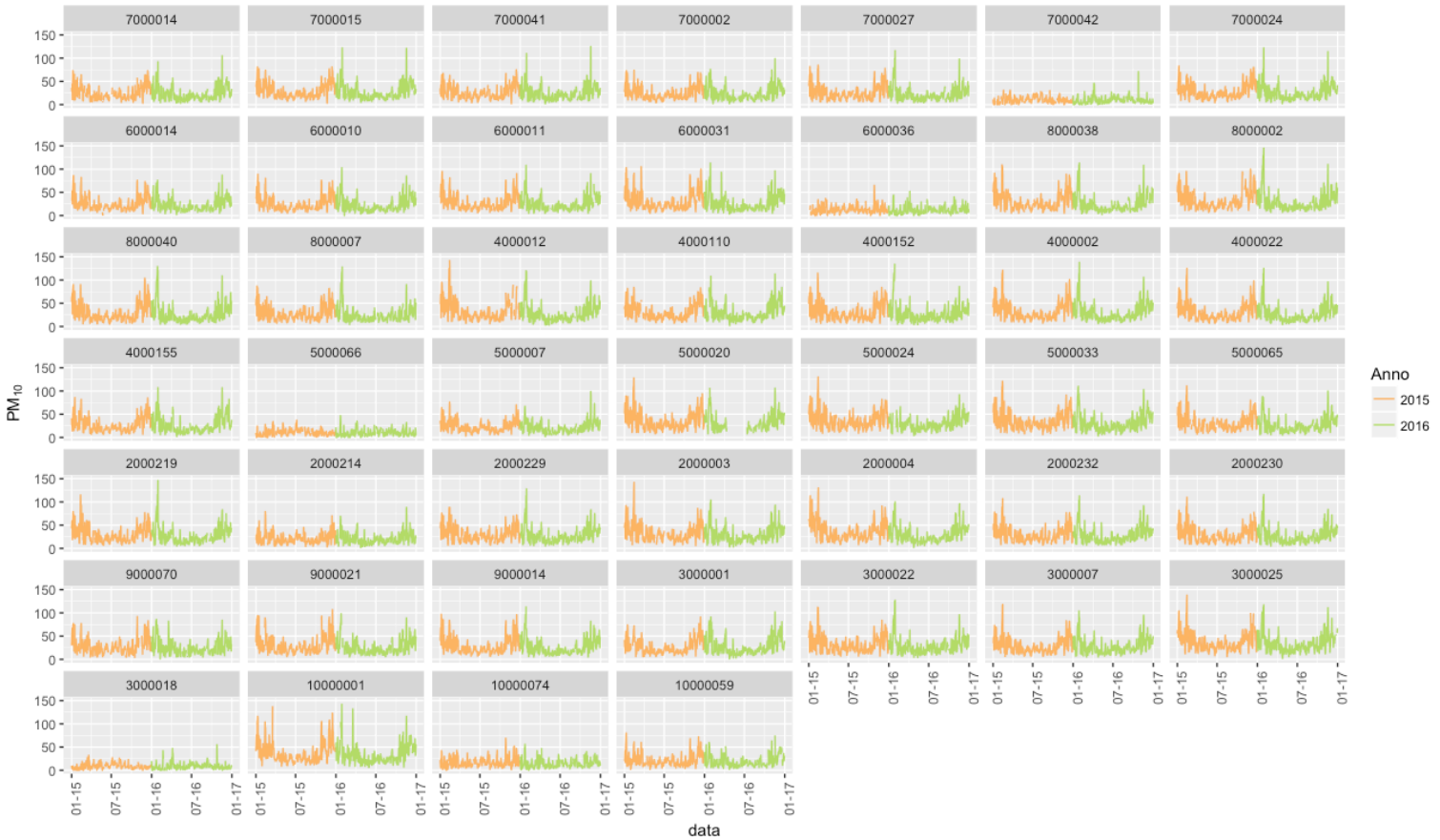

Figura 4 – Serie storiche impiegate per l'analisi (32525 osservazioni totali)

3. Mappe di uso e copertura del suolo

Per la creazione di una mappa adeguata si è deciso di utilizzare come punto di partenza il database di uso del suolo 2008 – edizione 2011 – realizzato dalla regione Emilia-Romagna. Si tratta di una base di dati georeferenziata di tipo vettoriale contenente 84.358 poligoni definiti mediante un codice numerico di quattro cifre che indica la categoria di uso del suolo associata. Il sistema di proiezione adottato è denominato ETRS89/UTM zone 32N (EPSG:25832), un sistema di riferimento in un unico fuso, l'UTM 32 Nord, che prevede l'estensione del suddetto fuso anche sul territorio che ricade nel 33 Nord (Virgilio Cima et al.).
Essa presenta un dettaglio di molto superiore a quello delle mappe del progetto europeo Corine Land Cover (CLC) impiegate da Janssen et al. (2008), ed è utilizzabile senza particolari problemi perché i primi tre livelli del sistema di classificazione derivano dalle specifiche CLC, le quali prevedono un sistema di nomenclatura a 44 classi su 3 livelli tematici.
Per poter attuare la metodologia RIO è necessario eseguire la conversione della mappa da formato vettoriale a raster durante la quale a ciascuna cella viene assegnata la tipologia di copertura del suolo in essa dominante. Per preservare il più possibile il maggiore dettaglio della mappa regionale si è scelto di utilizzare pixel di 50m rispetto ai 100m di CLC.

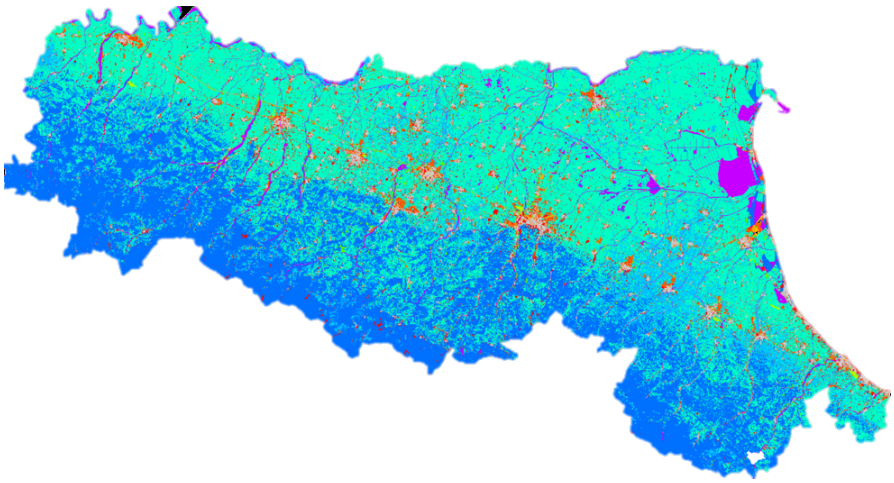

Figura 5 – Mappa RCL Emilia-Romagna

Il modello statistico di interpolazione RIO prevede come prima cosa che le 44 classi di copertura del suolo siano raggruppate in 11 tipologie più generali, denominate classi RIO (RCL). Per ottenere un unico indicatore β che metta in relazione i livelli medi di inquinamento da PM_{10} di un sito con le caratteristiche di uso del suolo della zona limitrofa viene considerata una zona circolare di raggio costante (generalmente 2 o 5 km) centrata sul sito, all'interno della quale ciascun pixel del raster viene individuato e classificato secondo le classi RCL. È ora possibile identificare l'indicatore di uso del suolo β come il logaritmo della media ponderata e normalizzata dei coefficienti a_i corrispondenti alle classi RIO con relativi pesi il numero dei pixel presenti dentro al cerchio ascrivibili alla *i-esima* classe RIO (che chiamiamo n_{RCLi}):

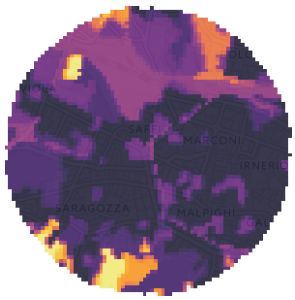

Figura 6 – Buffer circolare di raggio 2 km intorno alla centralina nr 7000015 (Bologna Porta San Felice)

$$\beta = \log\left[1 + \frac{\sum_{i=1}^{11} a_i * n_{RCLi}}{\sum_{i=1}^{11} n_{RCLi}}\right] \quad (1)$$

Tale indicatore β può essere ottimizzato specificatamente per il PM_{10} scegliendo il miglior insieme dei coefficienti a_i. L'ottimizzazione viene effettuata nei punti su cui giacciono le stazioni di monitoraggio, ovvero gli unici in cui si conoscono le i reali livelli di concentrazione

dell'inquinante, e consiste nel minimizzare la somma dei quadrati dei residui del seguente modello di regressione polinomiale di secondo grado:

$$\bar{x}_i = a_0 + a_1\beta_i + a_2{\beta_i}^2 + \epsilon_i \quad (i = 1, 2, \dots, 46) \tag{2}$$

dove $\bar{x}_i$ è la media di lungo periodo dei valori di PM_{10} rilevati dalla i-esima stazione di rilevamento, β_i è il valore dell'indicatore di uso del suolo.

Il coefficiente a_{10} relativo alla classe RCL10 viene posto uguale a 0 poiché la copertura del suolo corrispondente (territori boscati e ambienti seminaturali) non è associata ad alcun tipo di emissione non trascurabile, invece quello relativo alla classe RCL2 (zone urbanizzate a tessuto discontinuo) viene posto uguale ad 1 affinché l'indicatore β sia indotto a variare tra 0 e un valore superiore accettabile (circa 1.43).
I valori da stimare si sono così ridotti a 9, sui quali per prevenire l'eccessivo adattamento (overfitting) sono stati imposti alcuni vincoli lineari sulla base di considerazioni di buon senso.
L'ottimizzazione vincolata è stata eseguita dal software R utilizzando il metodo di Nelder-Mead. Una scelta adeguata dei valori iniziali per i coefficienti aiuta a limitare il rischio di confondere minimi locali con minimi globali; a questo scopo è stata considerata la distribuzione delle emissioni di PM_{10} per macrosettore (vedi figura 1) ed è stata stabilita una relazione di massima tra gli 11 macrosettori EMEP e le 11 classi RIO.
Una volta ricavati i coefficienti a_i, è possibile calcolare il valore assunto da β non solo in corrispondenza delle stazioni di monitoraggio, ma anche su una griglia regolare che copra l'intero territorio regionale.

4. Stima e rimozione del trend spaziale

Ciascuna delle 46 centraline considerate possiede ora un valore dell'indicatore di uso del suolo β che può essere impiegato per la stima e la rimozione del trend spaziale relativo alla media e alla deviazione standard della concentrazione media giornaliera di PM_{10}, entrambe positivamente correlate con β. Il trend rappresenta la componente strutturale del campo aleatorio $\{Z(\boldsymbol{u}): \boldsymbol{u} \in D \subset \mathcal{R}^2\}$, con $Z(\boldsymbol{u})$ variabile casuale nel sito $\boldsymbol{u}$, e descrive la variabilità di larga scala del fenomeno; la sua identificazione e rimozione è funzionale ad ottenere valori omogenei nello spazio da passare in input all'algoritmo di interpolazione spaziale e viene eseguita separatamente per i giorni feriali e festivi, ove si osservano differenti livelli di inquinamento da PM_{10}.
Le due funzioni trend vengono ricavate per mezzo della stima dei modelli di regressione polinomiale di secondo grado che descrivono la relazione della media e della deviazione standard delle serie storiche dei dati di monitoraggio con il valore assunto da β.
In accordo con quanto presentato in Janssen et al. (2008), prima viene rimosso il trend del valore medio attraverso una semplice traslazione lineare dei dati di ciascuna delle serie storiche:

$$x_{detr} = x - \mu(\beta) + \mu_{rif} \tag{3}$$

dove $\mu(\beta)$ è la media teorica corrispondente al valore assunto dall'indicatore di uso del suolo β nei pressi della relativa stazione di rilevamento (cioè il valore della funzione trend valutata in β), mentre μ_{rif} è un valore di riferimento arbitrario e costante.

Successivamente viene eliminata la discrepanza tra le varianze delle serie storiche dei dati di monitoraggio mediante un cambiamento di scala:

$$x_{detr} = (x - \bar{x})\frac{\sigma_{rif}}{\sigma(\beta)} + \bar{x} \tag{4}$$

dove $\sigma(\beta)$ è il valore della funzione trend per la deviazione standard valutata in β, σ_{rif} è un valore di riferimento arbitrario costante e $\bar{x}$ la nuova media della serie storica a seguito della traslazione (la quale coincide con la somma tra il corrispondente residuo del modello polinomiale per la media e il valore di riferimento che è stato scelto).

Come anzidetto, i valori risultanti dall'attuazione delle sopraccitate trasformazioni vengono utilizzati come dati di input dell'algoritmo di interpolazione spaziale posteriormente illustrato, sui risultati del quale verrà eseguita una procedura di reintroduzione del trend su tutti i punti della griglia (quindi anche nei luoghi dove non sono disponibili dati di monitoraggio) in modo tale da introdurre le caratteristiche locali nel livello di concentrazione del PM_{10} sulla base dei valori di β specifici di ogni cella.
Per ripristinare il cambiamento di scala associato alla deviazione standard:

$$x_{fin} = \mu_{rif} + \left(x_{detr} - \mu_{rif}\right) * \frac{\sigma(\beta)}{\sigma_{rif}} \tag{5}$$

Infine per ripristinare la traslazione lineare relativa al valore medio:

$$x_{fin} = x_{detr} + \mu(\beta) - \mu_{rif} \tag{6}$$

NB: La procedura di *detrending* viene eseguita sui dati delle sole localizzazioni campionate (gli unici disponibili prima dell'interpolazione), mentre la reintroduzione del trend su ciascuna delle celle della griglia di previsione. Questo spiega per quale ragione nell'equazione (5) compaia la media teorica μ_{rif} (desunta dal modello di figura 7) anziché il valore reale $\bar{x}$, noto per i soli punti su cui giacciono le stazioni di monitoraggio.

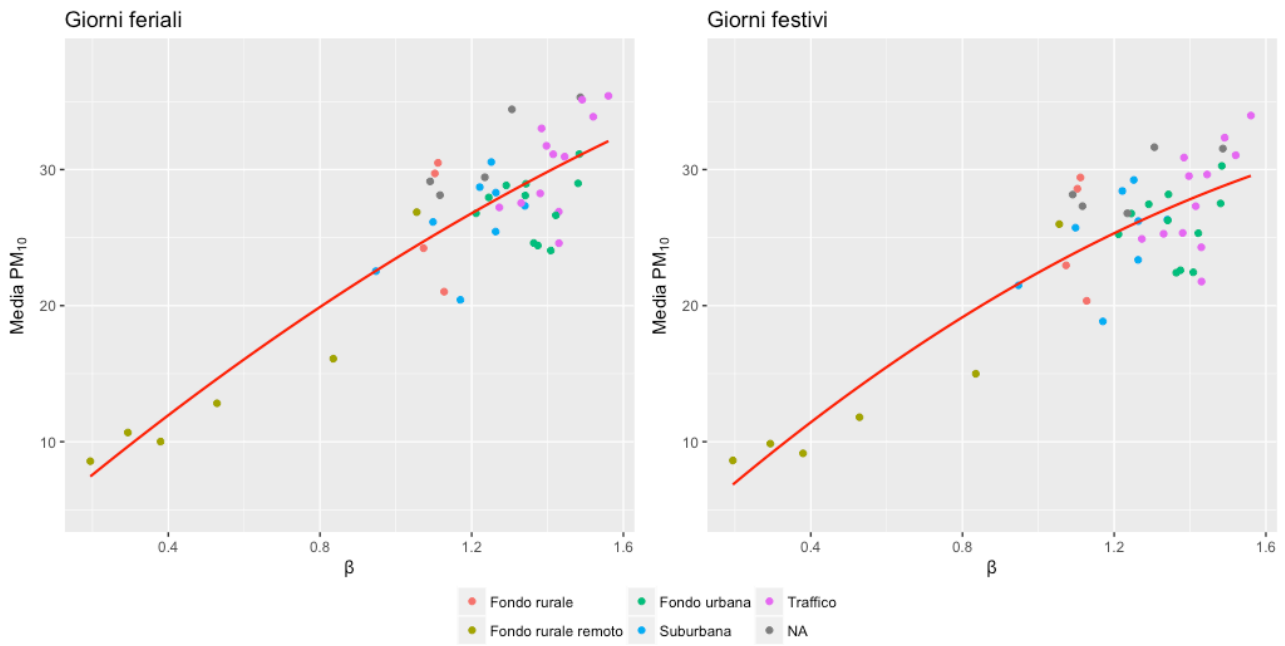

Figura 7 – Funzione trend per la media

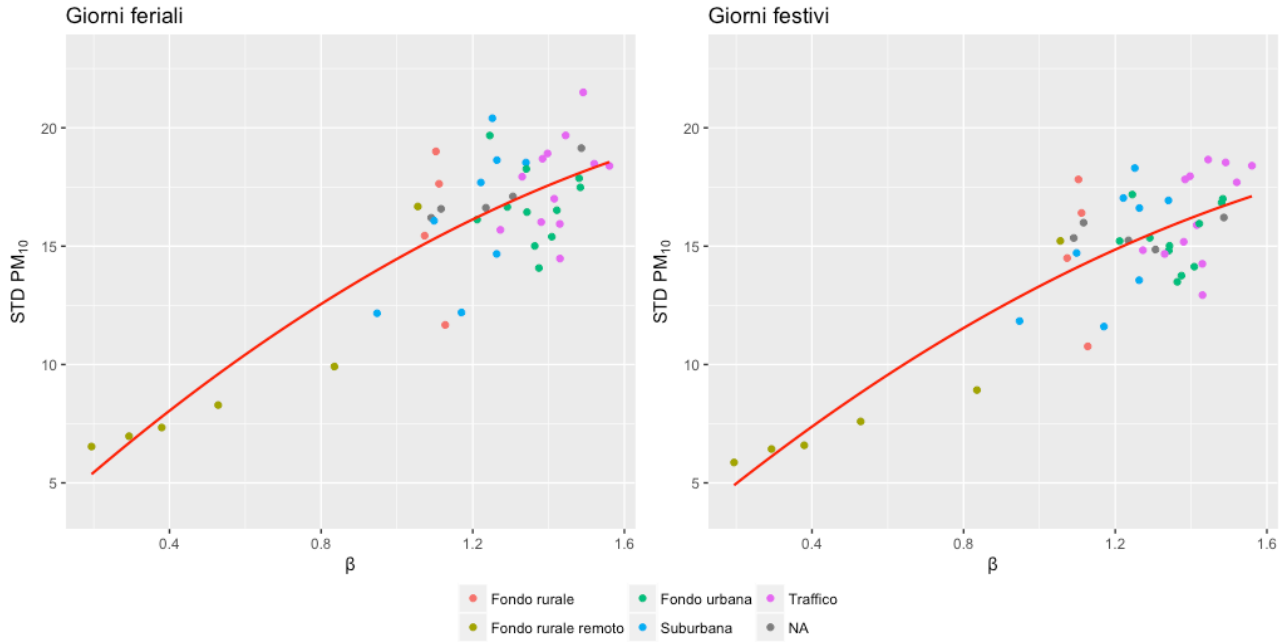

Figura 8 – Funzione trend per la deviazione standard. Si deduce che il campo casuale è eteroschedastico.

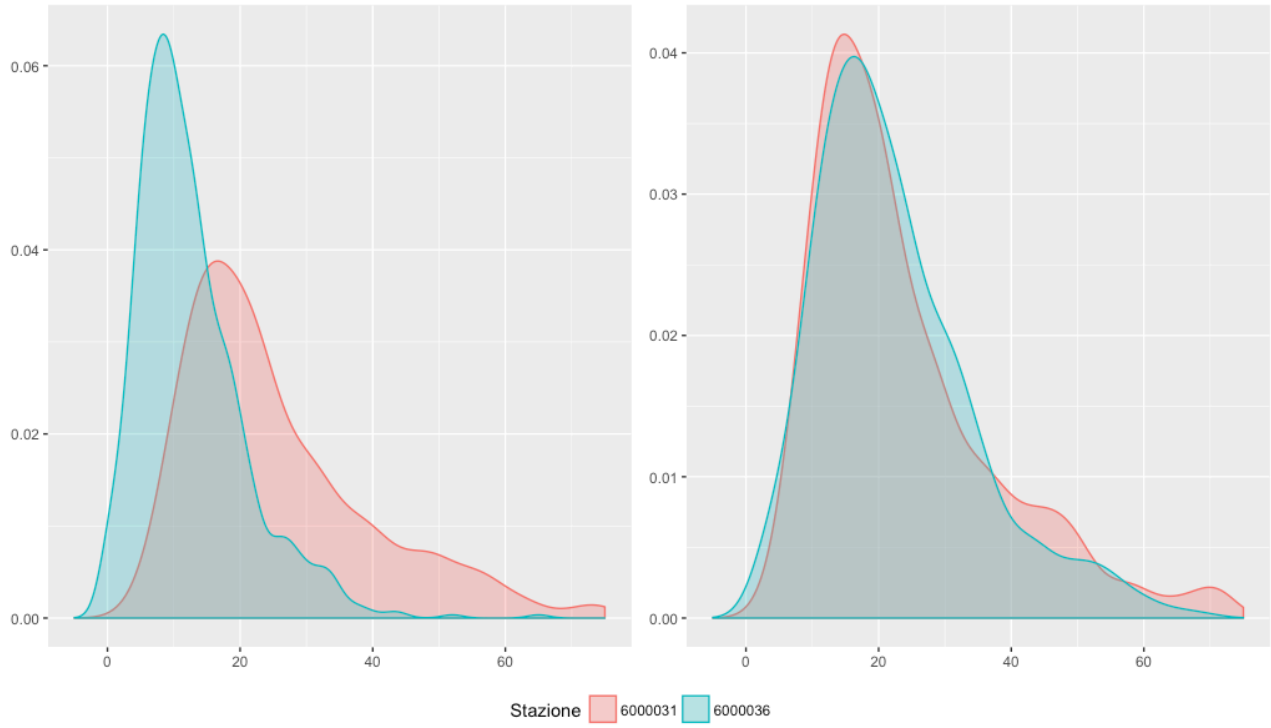

Figura 9 – Stima kernel della densità della concentrazione di PM_{10} per la stazione suburbana 6000031 e la stazione fondo rurale remoto 6000036 prima e dopo il detrending per la media e la deviazione standard

5. Stima del variogramma

L'obiettivo pratico del *detrending* è quello di rimuove dai dati di monitoraggio una parte importante delle caratteristiche locali del fenomeno in modo da poter assumere che il campo aleatorio soddisfi le ipotesi di stazionarietà del secondo ordine. Esse prevedono fondamentalmente che la media e la varianza del processo stocastico spaziale non dipendano dalla localizzazione $\boldsymbol{u} \in D$ e che la funzione di covarianza esista e dipenda solo dal vettore di separazione **h**.
Il principio fondante della geostatistica è l'autocorrelazione spaziale, un concetto che si basa sulla prima legge della geografia formulata da Tobler: "ogni cosa è correlata con le altre, ma le cose più vicine sono più correlate tra loro di quelle distanti". (Tobler 1970)
Dinnanzi a un campo aleatorio stazionario del secondo ordine, l'autocorrelazione spaziale può essere caratterizzata equivalentemente da due funzioni alternative e in relazione tra di loro: il variogramma ed il covariogramma. Sotto assunzione di isotropia (indipendenza dalla direzione), la prima mette in relazione la distanza tra due punti del dominio con la varianza della differenza tra i valori del campo casuale nei 2 siti, la seconda invece non è altro che un appellativo per la classica funzione di covarianza quando essa dipende solo dalla distanza.

Considerato l'esiguo numero di stazioni di monitoraggio e l'elevato livello di dispersione della variabile regionalizzata, risulta preferibile stimare la funzione di autocorrelazione spaziale sulla base delle serie storiche complete piuttosto che dei soli dati relativi al giorno di osservazione esaminato.
La procedura originale proposta in Janssen et al. (2008) prevede la stima di un modello log-lineare $\boldsymbol{\log(C(h)) = \beta_0 + \beta_1 h}$ dove il logaritmo della covarianza tra le serie storiche è una funzione lineare della distanza tra le localizzazioni campionate.
Stimati i parametri, è possibile effettuare previsioni spaziali con il metodo del Kriging impiegando il seguente covariogramma esponenziale:

$$C(h) = Cov(u, u+h) = \sigma^2 e^{-\frac{h}{R}} \tag{7}$$

dove
$\sigma^2 = \exp(\beta_0)$; $R = -\frac{1}{\beta_1}$

Oppure in alternativa è possibile utilizzare il variogramma esponenziale ottenuto mediante la relazione $\Upsilon(h) = C(0) - C(h)$, con il quale si ottengono le medesime previsioni spaziali.

Il modello di variogramma esponenziale è definito come segue:

$$\gamma(h) = C[1 - \exp(-h/a)] \tag{8}$$

dove il parametro C è il *valore di sella* mentre a viene indicato come *range*, ma nel caso del modello esponenziale NON è interpretabile come la distanza oltre la quale non si osserva più autocorrelazione spaziale dato che quest'ultima non decade mai a 0 e il valore di sella viene raggiunto solo asintoticamente: in corrispondenza di un *range* effettivo $a' = 3a$ il modello raggiunge il 95% del valore di sella.

Questo modo di procedere presenta tuttavia un problema: è inconsistente perché ignora che la covarianza tra le serie storiche include anche una componente temporale.
Stante che per fare interpolazione spaziale è opportuno considerare la sola componente spaziale, diversamente dalla metodologia proposta in Janssen et al. (2008) è stato calcolato con il pacchetto *gstat* di *R* il variogramma empirico spazio-temporale, dal quale è stata estratta la componente di ritardo zero così da ottenere un variogramma *pooled* che metta insieme tutti i dati considerando le osservazioni di ogni giorno come se fossero una realizzazione dello stesso processo stocastico spaziale. In questo modo il calcolo del variogramma empirico si basa su 2 anni di dati, quindi gli errori casuali sono minimi e risulta appropriato stimare i parametri del modello teorico definito con il metodo dei minimi quadrati ordinari, giacché eventuali pesi attribuiti ai quadrati dei residui risulterebbero fuorvianti.

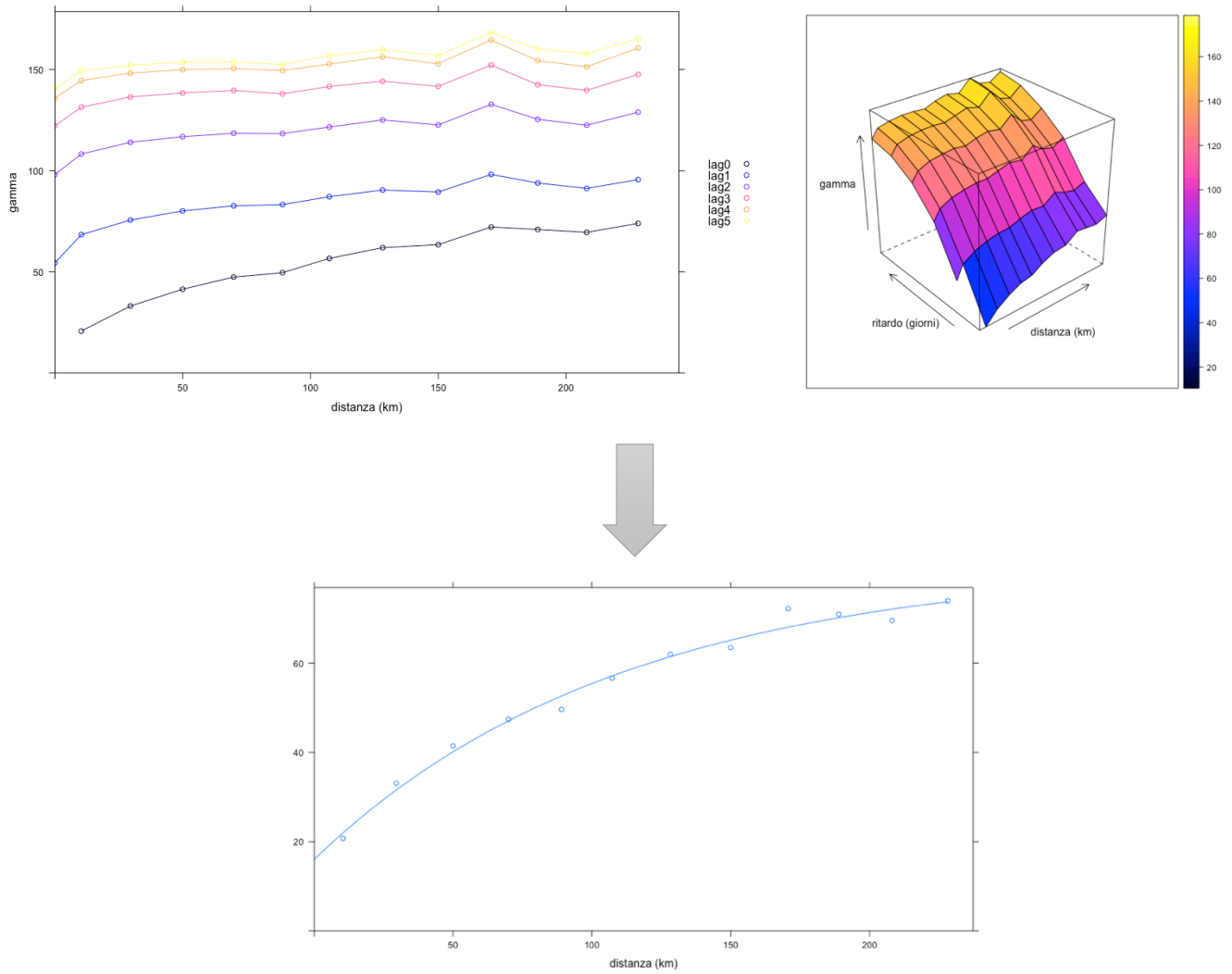

Figura 10 – Dal variogramma empirico spazio-temporale (in alto 2 rappresentazioni equivalenti) al variogramma pooled (in basso) con relativo modello

Il modello di variogramma che meglio si adatta ai dati e che restituisce previsioni spaziali più accurate è risultato comunque quello esponenziale, ma come mostrato in tabella 2 i parametri stimati differiscono ampiamente da quelli ottenuti in precedenza.

Metodo	Modello	Effetto nugget	Sella (parziale)	Range
Janssen et al. (2008)	Esponenziale	Non previsto	194.18	473.81 km
Variogramma *pooled*	Esponenziale	16.04	66.01	110.19 km

Tabella 2

L'estrazione della sola componente spaziale dell'autocorrelazione come atteso occasiona una stima del Range inferiore e più coerente con il reale comportamento del fenomeno oggetto di studio. Inoltre il secondo metodo tiene in conto che il variogramma sperimentale (vedi figura 10) presenta un evidente effetto nugget dovuto alle variazioni del fenomeno di piccolissima scala ovvero il cui range è inferiore alla più piccola distanza disponibile per il calcolo del variogramma.

Come segnalato nell'introduzione, la concentrazione media giornaliera di PM_{10} mostra una forte componente stagionale. Nonostante ciò la stima e l'impiego di modelli di variogramma differenti secondo il periodo dell'anno non produce miglioramenti apprezzabili nell'accuratezza della previsione (verificata con la convalida incrociata, vedi paragrafo 6.1).
Ad ogni modo si osserva che la correlazione spaziale è mediamente più debole nei periodi invernali ed autunnali laddove il range è minore.

Al contrario, una riduzione significativa dell'errore medio di stima si ottiene rimpiazzando l'ipotesi di isotropia con quella di anisotropia geometrica, ovvero definendo due variogrammi direzionali uguali in tutto eccetto che per il range che varia disegnando un'ellissi. Nello specifico la direzione NordOvest-SudEst (120°) è stata considerata come direzione di massima continuità spaziale, quindi in corrispondenza del range massimo.
La ragione di questa scelta risiede nel fatto che, a parità di distanza, l'autocorrelazione spaziale è maggiore fra i valori del PM_{10} osservati in luoghi posti in zone altimetriche omogenee piuttosto che situati in zone altimetriche diverse, e nel territorio della regione Emilia-Romagna la linea pedecollinare si trova proprio in direzione NordOvest-SudEst, così come la via Emilia che attraversa i maggiori centri urbani.
Le previsioni spaziali più accurate si ottengono con un rapporto di Anisotropia (rapporto tra Range minimo e Range massimo) pari 0.4.

6. Analisi dei risultati

La validazione del modello viene effettuata mediante la cross validazione Leave-one-out (LOOCV) e l'analisi delle mappe delle previsioni spaziali.

6.1 Convalida incrociata

La convalida incrociata (chiamata anche cross validazione) è una tecnica statistica in grado di misurare la bontà di un modello tramite l'informazione campionaria confrontando valori osservati con valori della previsione.

In particolare, la cross validazione Leave-one-out (LOOCV) considera una localizzazione campionata alla volta, ne predice il valore utilizzando le restanti osservazioni relative a quell'intervallo di tempo, dopodiché prosegue con la successiva localizzazione. La procedura viene ripetuta per ogni valore delle serie storiche dei dati di monitoraggio: si tratta di 32525 valori osservati nei 731 giorni del biennio 2015-2016.
Una volta ottenuto il vettore contenente i valori della previsione, le capacità predittive del modello in esame vengono valutate mediante l'ausilio dei seguenti indicatori statistici: RMSE (radice quadrata dell'errore quadratico medio), MAE (errore medio assoluto), Bias (distorsione) e coefficiente di correlazione lineare tra valori osservati e valori della previsione.

La convalida incrociata viene inizialmente eseguita sul modello da validare, ovvero il modello RIO a cui sono state apportate le modifiche descritte in questa tesi (imposizione di vincoli lineari nell'ottimizzazione di β, impiego di un variogramma *pooled*, assunzione di anisotropia geometrica), di seguito menzionato come RIO2 per semplicità di trattazione.
Successivamente, per finalità di confronto, la cross validazione viene replicata su 3 diverse varianti del modello RIO2 che prevedono l'uso di funzioni di correlazione differenti e su 2 tecniche standard di interpolazione: il Kriging Universale e il Kriging Ordinario.
La prima variante esclude l'anisotropia geometrica ed adotta un unico variogramma omnidirezionale, la seconda prevede l'impiego di una funzione di correlazione ricavata così come descritto in Janssen et al. (2008) e commentato nel capitolo 5, mentre la terza l'utilizzo di un modello di variogramma differente per ogni intervallo di tempo e definito sulla base delle sole osservazioni relative a quel giorno: tra diversi modelli teorici standard (sferico, esponenziale, gaussiano e i modelli della famiglia Matern nella parametrizzazione di *M. Stein*) viene selezionato quello che, di volta in volta, si adatta meglio alla distribuzione spaziale dei dati.
Nel Kriging Universale il calcolo della funzione trend viene effettuato su base giornaliera tramite un modello lineare generalizzato (GLM) assumendo una dipendenza lineare tra le concentrazioni medie giornaliere di PM_{10} e il valore dell'indicatore di uso del suolo β nelle localizzazioni campionate.
Infine nel Kriging Ordinario si assume che il valore atteso del campo aleatorio sia costante e l'interpolazione viene quindi effettuata direttamente sui dati di monitoraggio.

Indicatori di performance dei modelli

METODO	RMSE	MAE	BIAS	COR
Modello RIO2	4.83	3.44	-0.03	0.96
Con variogramma omnidirezionale (ipotesi di isotropia)	5.19	3.66	-0.02	0.95
Con covariogramma (Janssen et al. 2008)	5.24	3.77	-0.15	0.95
Con stima variogramma giornaliera	5.43	3.77	-0.06	0.95
Kriging Universale	5.72	3.86	0.02	0.94
Kriging Ordinario	6.87	4.45	0.64	0.92

Tabella 3 – Indicatori di performance dei modelli di interpolazione/previsione spaziale

Il modello statistico di interpolazione spaziale RIO2 risulta essere tra quelli considerati quello che genera previsioni più accurate. Un valore del RMSE inferiore a 5 è un buon risultato per una variabile di qualità dell'aria quale la concentrazione di PM_{10}.
La variante che prevede la stima giornaliera del variogramma registra come intuibile capacità predittive peggiori: in effetti, causa l'esiguo numero di dati di monitoraggio disponibili, in corrispondenza di alcuni giorni il modello di variogramma teorico stimato con i dati disponibili descrive in modo fortemente inaccurato il reale grado di dipendenza spaziale del campo casuale; questo spiega perché le peggiori performance rispetto al modello RIO2 si manifestino soprattutto nell'indicatore RMSE che attribuisce un peso relativamente alto a grandi errori di previsione.

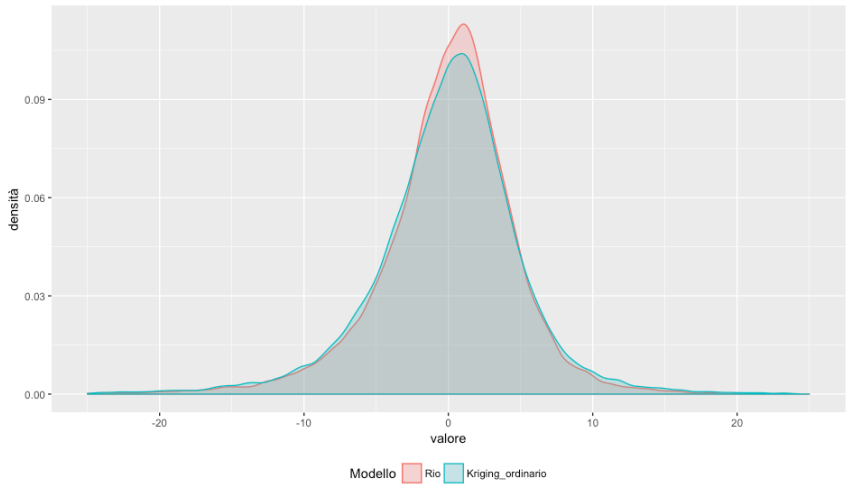

Figura 11 – Stima kernel della densità dell'errore assoluto di previsione valutato con la convalida incrociata. Può essere utile per calcolare gli intervalli di confidenza delle stime effettuate.

Per il medesimo motivo è inevitabile che la stima del modello per il trend su base giornaliera propria del Kriging Universale sovente restituisca previsioni spaziali di molto distanti dal valore reale e che quindi incidono pesantemente soprattutto nel calcolo del RMSE.

6.2 Mappe delle previsioni spaziali

In questa sezione il modello di interpolazione RIO2 viene esaminato e confrontato con il Kriging Ordinario sulla base dell'osservazione diretta delle mappe di interpolazione spaziale ottenute effettuando previsioni spaziali con *gstat* su una griglia regolare fine con celle 150x150mq.

Di seguito a titolo di esempio viene mostrata la superficie statistica di predizione della concentrazione media di PM_{10} per il giorno 19/11/2016 ottenuta mediante l'impiego del metodo RIO2 con variogramma omnidirezionale, preceduta dalla mappa della funzione trend per la media e dalla mappa del Kriging eseguito sui valori estratti dalle serie storiche frutto della detrendizzazione.
Nel giorno considerato sono stati rilevati livelli di inquinamento da PM_{10} molto alti nella zona nord-ovest della regione.

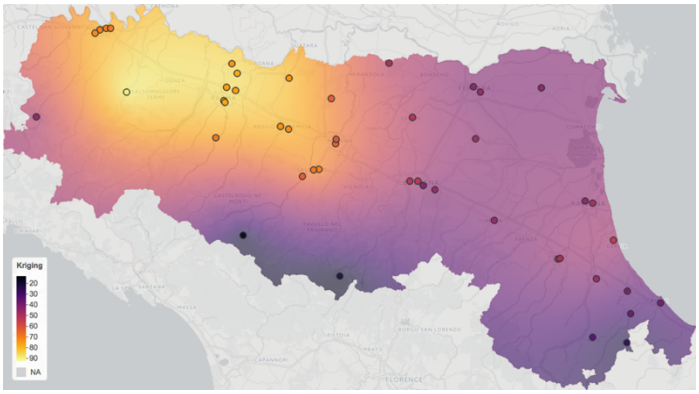

Figura 12 – Mappa del Kriging sui valori risultanti dal detrending relativi al giorno 19/11/2016

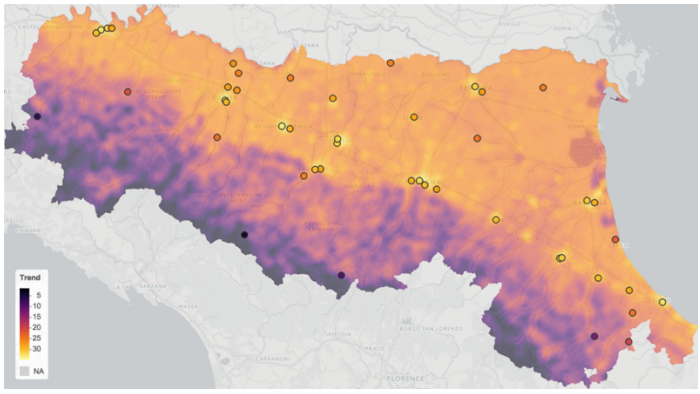

Figura 13 – Valore della funzione trend a cui viene sommato il valore di riferimento arbitrario (=25)

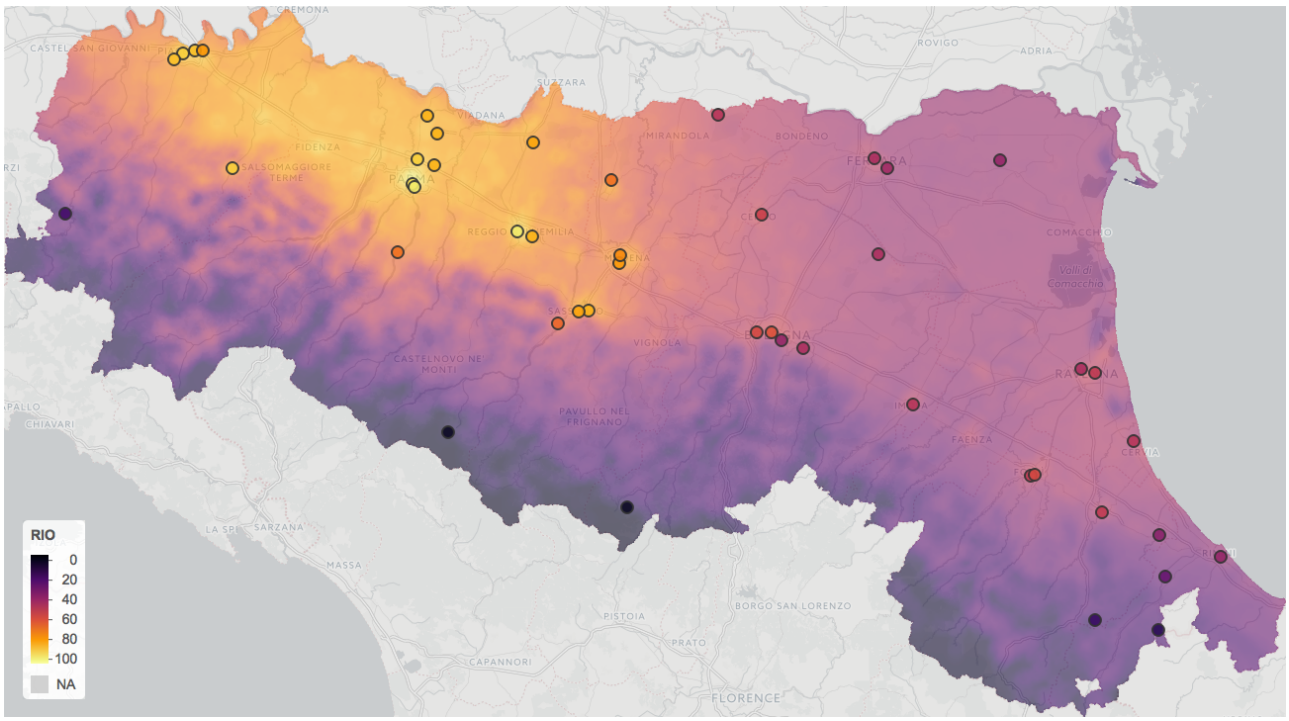

Figura 14 – Mappa finale RIO per la concentrazione media di PM_{10} del giorno 19/11/2016 (valori in microgrammi/m3)

Alla mappa di previsione spaziale è possibile associare una carta degli errori standard che fornisca un'indicazione sul grado di incertezza dei valori stimati.
In ciascun punto della griglia di previsione, nella risoluzione delle corrispondenti equazioni del Kriging, è possibile ottenere il valore assunto dalla varianza dell'errore di previsione in corrispondenza della soluzione di minimo sotto l'ipotesi implicita che la struttura di autocorrelazione spaziale della variabile regionalizzata corrisponda esattamente a quella descritta dal variogramma teorico impiegato nel Kriging. Oltre a ignorare quindi l'incertezza dei parametri del variogramma, tale valore, che chiamiamo s_K^2, nel caso del modello RIO presenta due problemi rilevanti:

1. La scala complessiva della varianza dipende unicamente e linearmente dal valore di sella del variogramma teorico impiegato. In realtà la varianza della concentrazione del PM_{10} è strettamente legata al valore dell'indicatore di uso del suolo β, come illustrato in figura 8.

2. Il valore non tiene conto dell'errore che scaturisce dalla procedura di reintroduzione del trend.

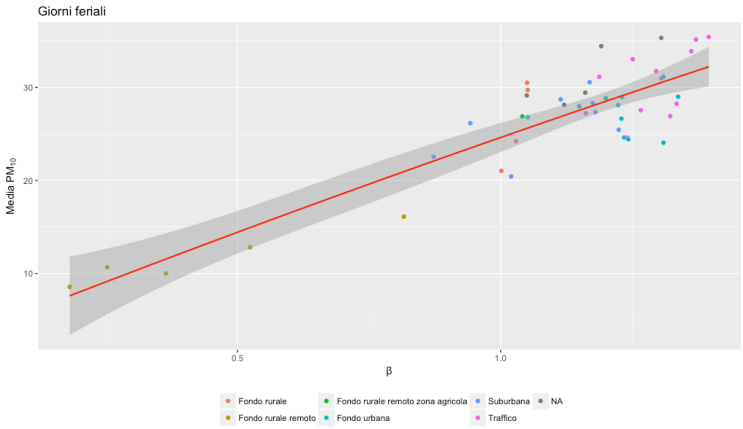

Figura 15 – Funzione trend per la media con relativo intervallo di confidenza al 95% (+/- 1.96 SEM)

Per risolvere il primo problema gli autori del modello RIO propongono di utilizzare come fattore di scala per s_K^2 un valore che dipenda da β sulla base del trend presentato nella *figura 8*, cosicché l'incertezza determinata dalla distanza dei punti dalle localizzazioni campionate si coniuga con le variabilità del fenomeno associata alle caratteristiche locali di copertura del suolo.
Infine il secondo problema viene risolto sommando a $s_K^2(\beta)$ il quadrato dell'errore standard di previsione della media (SEM) relativo alla funzione trend per la media (vedi figura 15), $s_{trend}^2(\beta)$, così da tenere in considerazione anche l'incertezza dei parametri stimati del trend.

In sintesi la varianza totale dell'errore di previsione viene stimata nel seguente modo:

$$s_{tot}^2(\beta) = s_K^2(\beta) + s_{trend}^2(\beta) \tag{9}$$

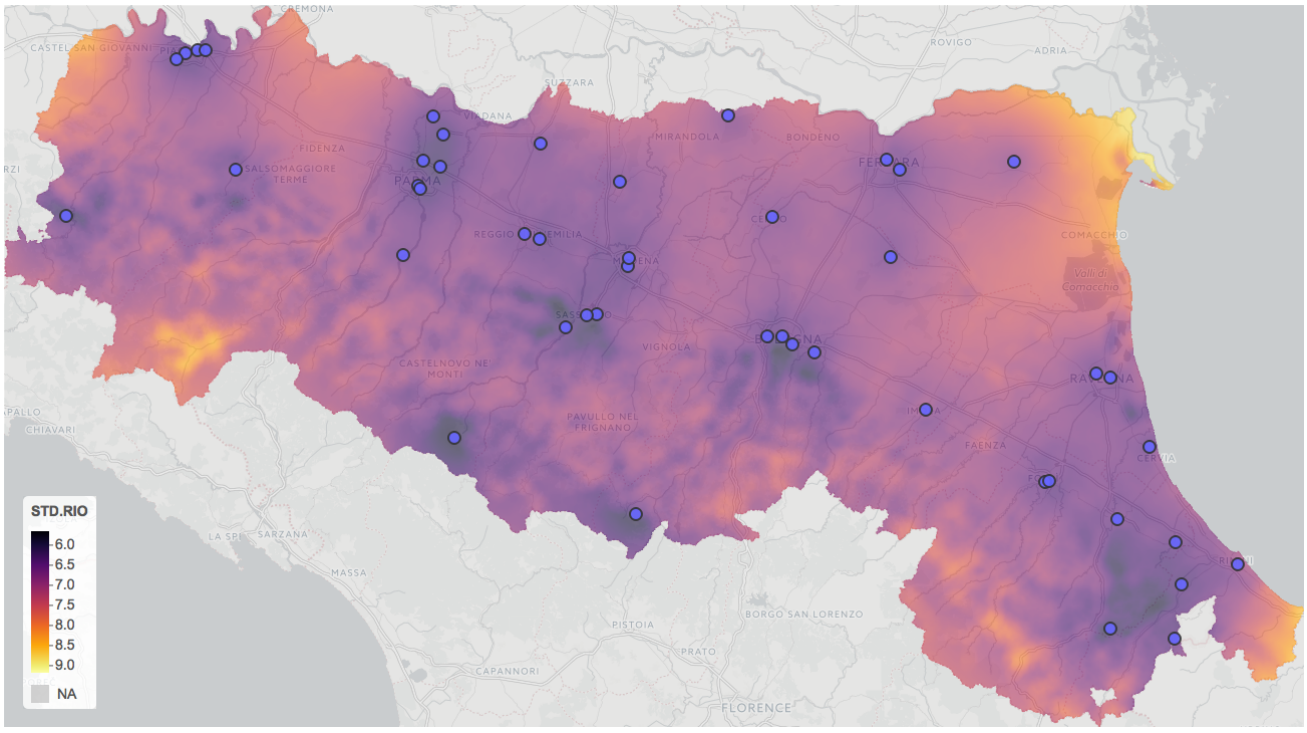

Figura 16 – Mappa dell'errore standard $s_{tot}(\beta)$ relativo alle previsioni di figura 12 (del giorno 19/11/2016)

Di seguito i modelli RIO2 e Kriging Ordinario vengono messi a confronto mediante le mappe della concentrazione media del PM_{10} nell'anno 2016 ottenute eseguendo l'interpolazione spaziale per tutti i giorni dell'anno con i dati di monitoraggio di volta in volta disponibili e alla fine calcolando per ciascuna cella la media di tutte le previsioni giornaliere.
Si osserva che il Kriging Ordinario tende ad estendere alle zone rurali i livelli di inquinamento da PM_{10} misurati dalle stazioni di rilevamento situate in zone urbane/suburbane, viceversa i modelli RIO/RIO2 sono in grado di cogliere le caratteristiche locali del territorio esibendo variazioni del fenomeno anche in luoghi dove non sono disponibili dati di monitoraggio.

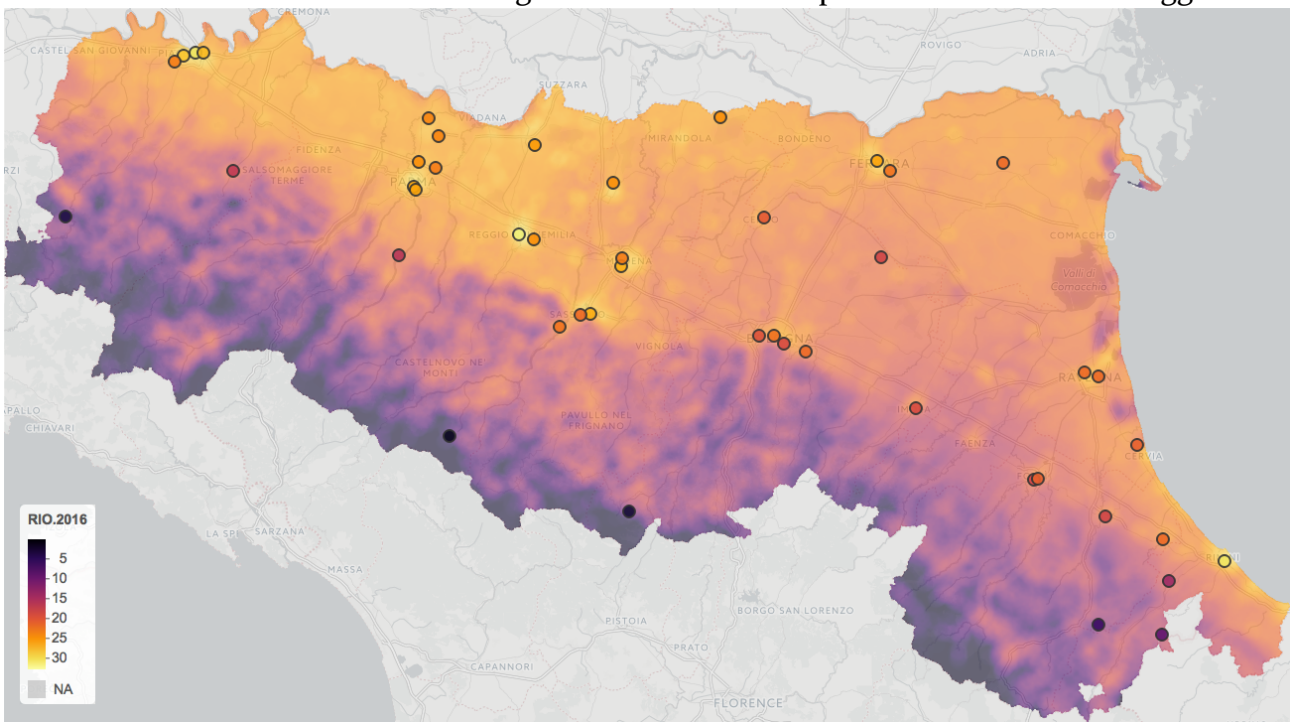

Figura 17 - Mappa RIO della concentrazione media di PM_{10} nell'anno 2016 (valori in microgrammi/m3)
Mappa interattiva disponibile su http://carlocavalieri.it/RIO_2016.html

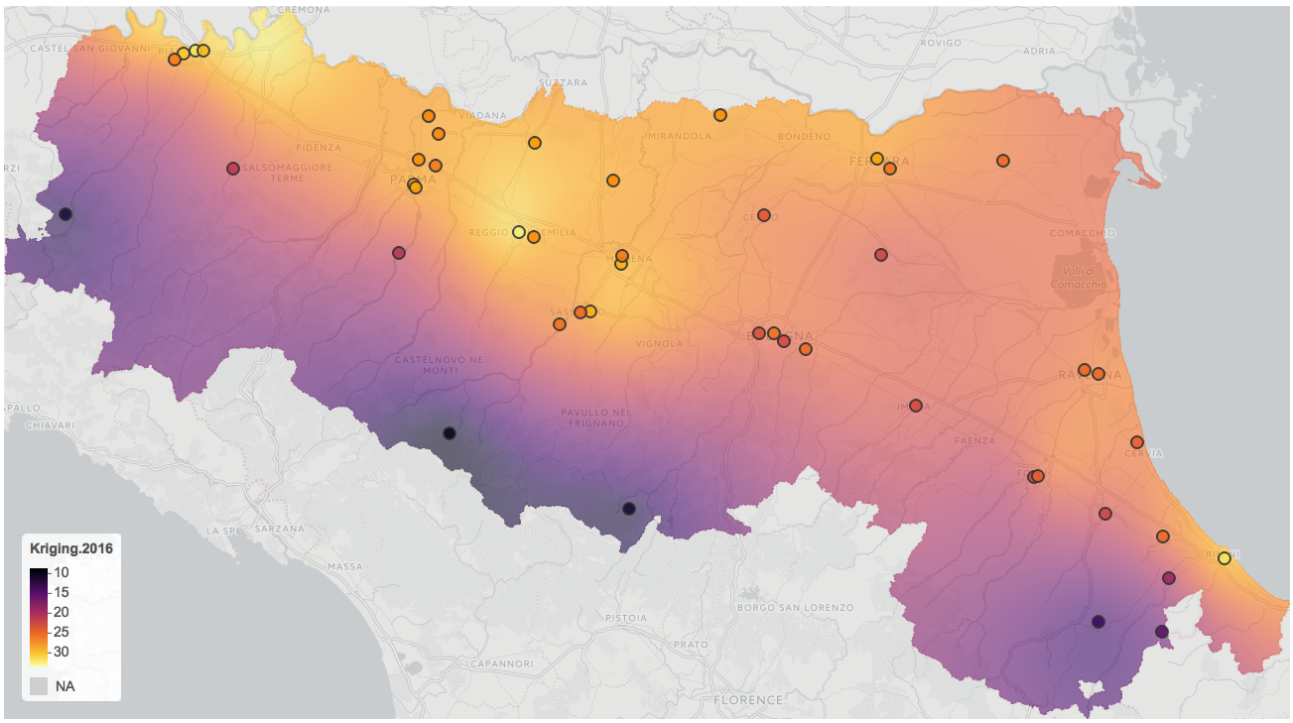

Figura 18 – Mappa Kriging ordinario della concentrazione media di PM_{10} nell'anno 2016 (valori in microgrammi/m3)
Mappa interattiva disponibile su http://carlocavalieri.it/Kriging_Ordinario_2016.html

7. Conclusioni

Il metodo di interpolazione spaziale illustrato consente di effettuare previsioni spaziali accurate e con bassi costi computazionali utilizzando in modo intelligente l'informazione campionaria. Il principale vantaggio rispetto alle tecniche tradizionali consiste nel fatto che la stima delle funzioni trend e di un modello di variogramma teorico idoneo a descrivere la correlazione spaziale viene effettuata sulla base delle serie storiche dei dati di monitoraggio invece che con l'impiego dei soli dati relativi a un singolo intervallo di tempo, i quali, come detto in precedenza, risultano insufficienti in considerazione dell'esiguo numero di localizzazioni campionate e dell'elevato livello di dispersione della variabile regionalizzata PM_{10}.

In caso di impiego per dati relativi ad altre sostanze contaminanti presenti nell'aria, è necessario effettuare l'ottimizzazione di β separatamente per ciascun inquinante poiché essi sono caratterizzati da fonti di emissioni differenti (vedi figura 1) e quindi da una diversa associazione con le caratteristiche locali di copertura del suolo.

Il modello non è in grado di rilevare la maggiore concentrazione del PM_{10} in corrispondenza delle zone maggiormente trafficate. Infatti, nonostante sia presente una classe RIO dedicata al traffico (RCL4), nella conversione da formato vettoriale a raster ciascun pixel di 50m x 50m viene assegnato a quella classe soltanto se è prevalente in quell'area. Inoltre essa non distingue tra reti stradali e reti ferroviarie, ma si tratta di un problema facilmente superabile evitando di accorpare le due classi che nella mappa vettoriale di uso del suolo della regione Emilia-Romagna si presentano separate. Resta però da capire come valutare l'effetto sull'inquinamento della vicinanza di un punto da strade con alto volume di traffico, forse studiando le differenze tra i valori rilevati dalle stazioni di misurazione di tipo traffico e quelle di tipo fondo.

È possibile che il processo di ottimizzazione vincolata dell'indicatore di uso del suolo β restituisca in output un insieme migliore di parametri se condotto su tutta la rete di monitoraggio gestita da Arpae, comprese quelle stazioni di rilevamento non incluse in questa analisi poiché i relativi dati non sono disponibili sul portale https://dati.arpae.it.

La struttura di correlazione spaziale viene descritta da un'unica funzione comune a tutti i periodi dell'anno. La stima e l'impiego di 4 variogrammi diversi in relazione alla stagione meteorologica non determina miglioramenti nelle capacità predittive. Un approccio alternativo potrebbe essere quello di utilizzare funzioni di correlazione differenti in base ai livelli medi di inquinamento da PM_{10} osservati giornalmente piuttosto che in conseguenza della semplice stagione meteorologica.

In presenza di dataset caratterizzati da una maggiore frequenza temporale dei dati (ad esempio 1h anziché 24h) è possibile ottenere migliori performance predittive, seppur a costi computazionali decisamente superiori, con il Kriging Spazio-Temporale e un modello per descrivere l'autocorrelazione adeguato al variogramma empirico mostrato in figura 10. Con una risoluzione temporale di 1 giorno è invece probabile che la correlazione temporale sia troppo debole rispetto a quella spaziale per ottenere miglioranti percepibili.

Bibliografia

Arpae. s.d. *Qualità dell'Aria - Dati di monitoraggio.* https://dati.arpae.it.

Arpae. 2013. «Rapporto finale dell'inventario emissioni 2010.»

Benedikt Gräler, Edzer Pebesma and Gerard Heuvelink. 2016. «Spatio-Temporal Interpolation using gstat.» *The R Journal* 8/1 : 204-218.

Donato Posa, Sandra De Iaco. 2009. *Geostatistica teoria e applicazioni.* Torino: G. Giappichelli.

Janssen S, Dumont G, Fierens F and Mensink C. 2008. «Spatial interpolation of air pollution measurements using CORINE land cover data.» *Atmospheric Environment* 42: 4884 – 4903.

Pebesma, E.J. 2004. «Multivariable geostatistics in S: the gstat package.» *Computers & Geosciences* 30: 683-691.

Tobler, W. R. 1970. «A computer movie simulating urban growth in the Detroit region.» *Economic Geography* 46: 234–240.

Virgilio Cima, Stefano Olivucci, Luca Zennaro. s.d. «Sistemi di Coordinate Geografiche e Cartografiche in Regione Emilia Romagna e loro trasformazioni.» Servizio Statistica ed Informazione Geografica, Regione Emilia-Romagna. http://geoportale.regione.emilia-romagna.it/it/contenuti-allegati/Trasformazione_Sistemi_Coordinate_v1.0.2.pdf.

Pacchetti R utilizzati

maptools
sp
raster
rgdal
chron
ggplot2
reshape2
gstat
spacetime
MonoPoly
mapview
gridExtra

Appendice

Classi RIO	Descrizione
RCL1	Zone urbanizzate – Tessuto continuo
RCL2	Zone urbanizzate – Tessuto discontinuo + Aree verdi artificiali non agricole
RCL3	Insediamenti industriali o commerciali
RCL4	Reti ed aree infrastrutturali stradali e ferroviarie e spazi accessori
RCL5	Aree portuali
RCL6	Aree aeroportuali ed eliporti
RCL7	Aree estrattive, discariche, cantieri e terreni artefatti e abbandonati
RCL8	Seminativi
RCL9	Colture permanenti, Prati stabili, Zone agricole eterogenee
RCL10	Territori boscati e ambienti seminaturali
RCL11	Ambiente umido + Ambiente delle acque

Tabella 4 – Descrizione delle 11 classi RIO (RCL) ottenute mediante accorpamenti delle 44 classi CLC

CODICE ARPAE	ALTEZZA	**INDICATORE β**	MEDIA PM_{10}		DEVIAZIONE STANDARD PM_{10}	
			GIORNI FERIALI	GIORNI FESTIVI	GIORNI FERIALI	GIORNI FESTIVI
7000014	43	0.90	24.60	22.42	15.01	13.49
7000015	54	0.97	28.25	25.34	16.02	15.18
7000041	56	0.92	25.44	23.37	14.67	13.56
7000002	42	0.94	24.59	21.77	14.48	12.93
7000027	11	0.90	24.22	22.96	15.44	14.50
7000042	811	0.23	10.01	9.15	7.34	6.59
7000024	64	0.93	27.22	24.90	15.69	14.83
6000014	41	0.88	24.42	22.60	14.08	13.75
6000010	29	0.93	24.05	22.45	15.40	14.13
6000011	25	0.95	26.90	24.29	15.94	14.25
6000031	32	0.92	28.31	26.22	18.63	16.61
6000036	615	0.44	12.82	11.80	8.29	7.60
8000038	15	0.95	27.33	26.32	18.53	16.93
8000002	8	0.97	30.95	29.64	19.68	18.66
8000040	8	0.91	27.95	26.77	19.67	17.18
8000007	-2	0.94	26.87	25.99	16.68	15.22
4000012	25	0.87	30.55	29.23	20.40	18.30
4000110	131	0.94	31.13	27.31	17.00	15.89
4000152	4	0.95	29.71	28.59	19.00	17.82
4000002	39	0.96	31.74	29.52	18.92	17.95
4000022	30	1.00	28.99	27.51	17.87	16.84
4000155	118	0.91	26.64	25.32	16.52	15.95
5000066	765	0.18	10.68	9.86	6.97	6.43
5000007	210	0.70	22.54	21.49	12.16	11.84
5000020	61	1.06	35.31	31.54	19.14	16.21
5000024	61	1.00	34.42	31.64	17.10	14.86
5000033	61	1.03	33.88	31.05	18.48	17.70
5000065	61	0.99	28.95	28.18	16.44	15.01
2000219	30	0.96	28.71	28.43	17.69	17.03
2000214	202	0.89	21.02	20.35	11.67	10.77
2000229	25	0.95	29.13	28.16	16.20	15.34
2000003	60	0.99	31.14	30.27	17.48	17.00
2000004	55	0.96	33.02	30.88	18.69	17.82
2000232	40	0.97	29.44	26.78	16.62	15.24
2000230	35	0.95	28.12	27.31	16.57	15.99
9000070	0	0.80	26.15	25.73	16.07	14.71
9000021	4	0.96	28.09	26.27	18.26	14.81
9000014	4	0.98	27.55	25.28	17.93	14.67
3000001	150	0.86	26.79	25.24	16.12	15.22
3000022	22	0.95	30.49	29.42	17.63	16.40
3000007	55	0.95	28.84	27.45	16.66	15.35
3000025	59	1.03	35.41	33.97	18.39	18.40
3000018	1121	0.08	8.57	8.62	6.53	5.86
10000001	5	1.01	35.13	32.34	21.50	18.54
10000074	540	0.71	16.10	15.00	9.92	8.92
10000059	78	0.81	20.43	18.84	12.20	11.60

Tabella 5 – Stazioni di monitoraggio considerate nell'analisi con relative statistiche