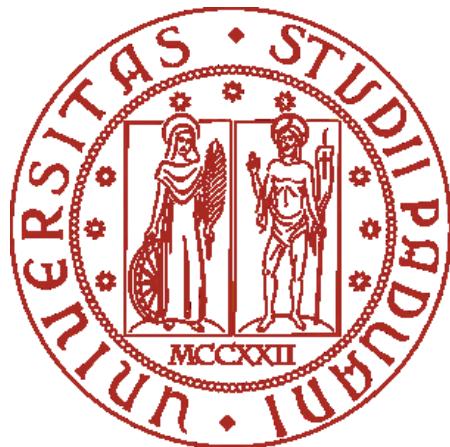


Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze statistiche



Modellazione del PM2.5 negli Stati Uniti

Un'analisi tramite dati spazio-temporali

Riotto Luca, Rudelli Marco

Indice

Introduzione	5
1 Dati e analisi esplorative	7
1.1 I dati	7
1.2 Pre-processing	7
1.3 Lisciamenti ed analisi esplorative	8
2 Modellazione	11
2.1 Modello funzionale simultaneo	11
2.1.1 Regolazione	12
2.1.2 Risultati	12
2.2 Modello state-space	14
2.2.1 Gestione della correlazione tra città	15
2.2.2 Regolazione	16
2.2.3 Risultati	16
2.3 Modello spazio-temporale	17
2.3.1 Risultati	20
2.3.2 Lisciamento dei coefficienti	20
3 Conclusioni	25
A Simulazioni correlazione atmosferica	27
A.1 Fattore latente incorrelato con la variabile osservata	27
A.2 Fattore latente correlato con la variabile osservata	28
Bibliografia	31

Introduzione

L'inquinamento atmosferico rappresenta una seria minaccia per le condizioni ambientali sostenibili del XXI secolo. La sua importanza nel determinare la salute e gli standard di vita nelle aree urbane è destinata ad aumentare nel tempo. Diversi fattori, dalle emissioni artificiali ai fenomeni naturali, sono noti come agenti causali o influenze primarie dell'aumento dei livelli di inquinamento atmosferico.

In questo progetto si è affrontato il problema della modellazione del particolato pm 2.5 in relazione ad alcune variabili Meteorologiche e legate all'attività umana.

Capitolo 1

Dati e analisi esplorative

1.1 I dati

Al fine di modellare il pm 2.5 si utilizza un ampio insieme di dati che presenta misurazioni a livello giornaliero su più città. Complessivamente, l'insieme di dati contiene un totale di 35.596 osservazioni misurati in 54 città e 24 mesi (anni 2019/2020), ogni osservazione rappresenta una combinazione unica (data, città).

Le covariate a disposizione si suddividono in due categorie:

- rilevazioni atmosferiche: *temperatura, pressione, umidità, vento*.
- covariate legate all'attività umana, come ad esempio le *miglia percorse dalle auto* (in milioni) .

Tutte le misurazioni atmosferiche fanno riferimento alla mediana del giorno di rilevazione.

Per approfondimenti sui dati usati si veda Bhattacharyya et al. (2021).

1.2 Pre-processing

Il dataset presenta alcune criticità, tra cui la presenza di dati mancanti e di outlier. Osserviamo in tabella 1.1 le variabili che presentano dati mancanti e le proporzioni di NA associati. Si è deciso di interpolare i dati al fine di gestire tali valori mancanti.

Tabella 1.1: Proporzione di osservazioni mancanti nelle diverse variabili.

	Proporzione di NA
pressure_median	0.0256
pm25_median	0.0026
humidity_median	0.0255
temperature_median	0.0258
wind.speed_median	0.0254

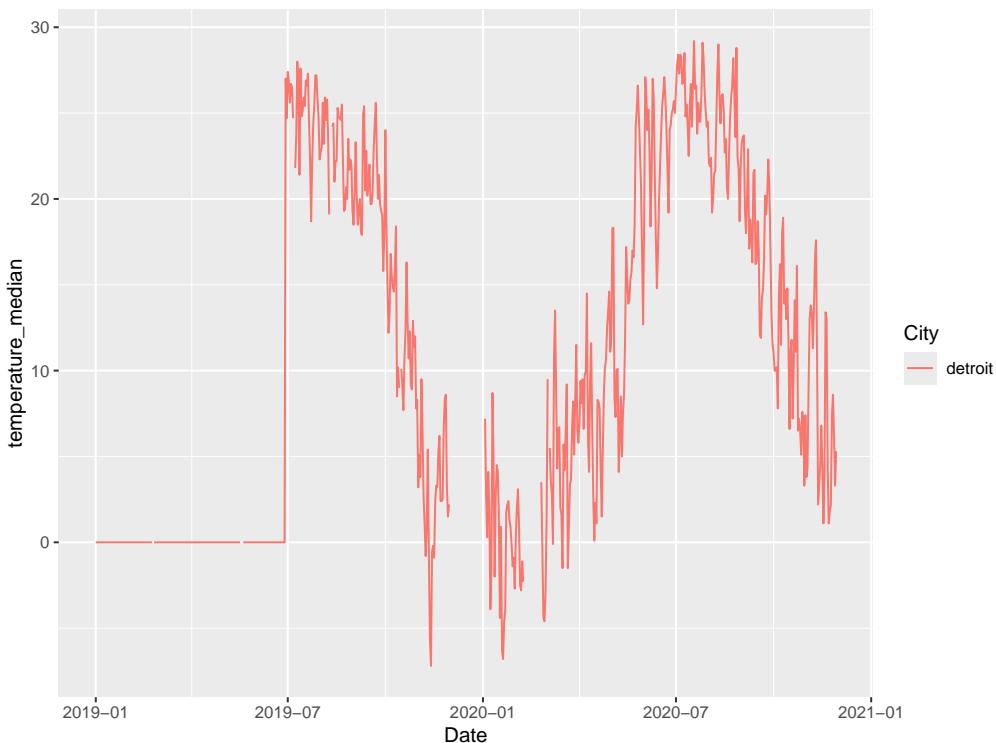


Figura 1.1: Variabile temperatura per Detroit.

Il problema dei tati mancanti non si limita tuttavia alla sola presenza di osservazioni codificate come NA, infatti, come possiamo ad esempio osservare in figura 1.1, vi sono variabili che nell’anno 2019 assumono molti zeri (che si considerano essere dei dati mancanti). Per tale ragione si è deciso, nella suddivisione dei dati in insieme di stima ed insieme di verifica, di utilizzare l’anno 2020 come insieme di stima e l’anno 2019 come insieme di verifica. L’insieme di verifica è stato a sua volta suddiviso nel ”validation set”, che è stato utilizzato per individuare l’ottimo dei parametri di regolarizzazione, e il ”test set” che è stato utilizzato per confrontare la capacità predittiva dei diversi modelli. La scelta di utilizzare l’anno 2020 come insieme di stima ed il 2019 come insieme di verifica è inoltre giustificata dal fatto che nella risposta non si osserva un trend tra un anno e l’altro.

Un’altra criticità che abbiamo dovuto affrontare riguarda la presenza di picchi concentrati in alcuni periodi (agosto-ottobre) nella variabile risposta. Tali picchi sono dovuti alla presenza di incendi, che tuttavia non è un’informazione che abbiamo a disposizione tra le covariate. Al fine di migliorare l’adattamento dei modelli e di renderli più generalizzabili alla previsione in altri anni, si è deciso di porre un limite superiore alla risposta pari a 120.

1.3 Lisciamenti ed analisi esplorative

In seguito alla prima fase di preprocessing dei dati si è proceduto con una breve analisi esplorativa dei dati, al fine di poter sfruttare le variabili al meglio.

Come visto precedentemente, tra le covariate a disposizione abbiamo delle variabili legate all’attività umana, in particolare disponiamo delle variabili:

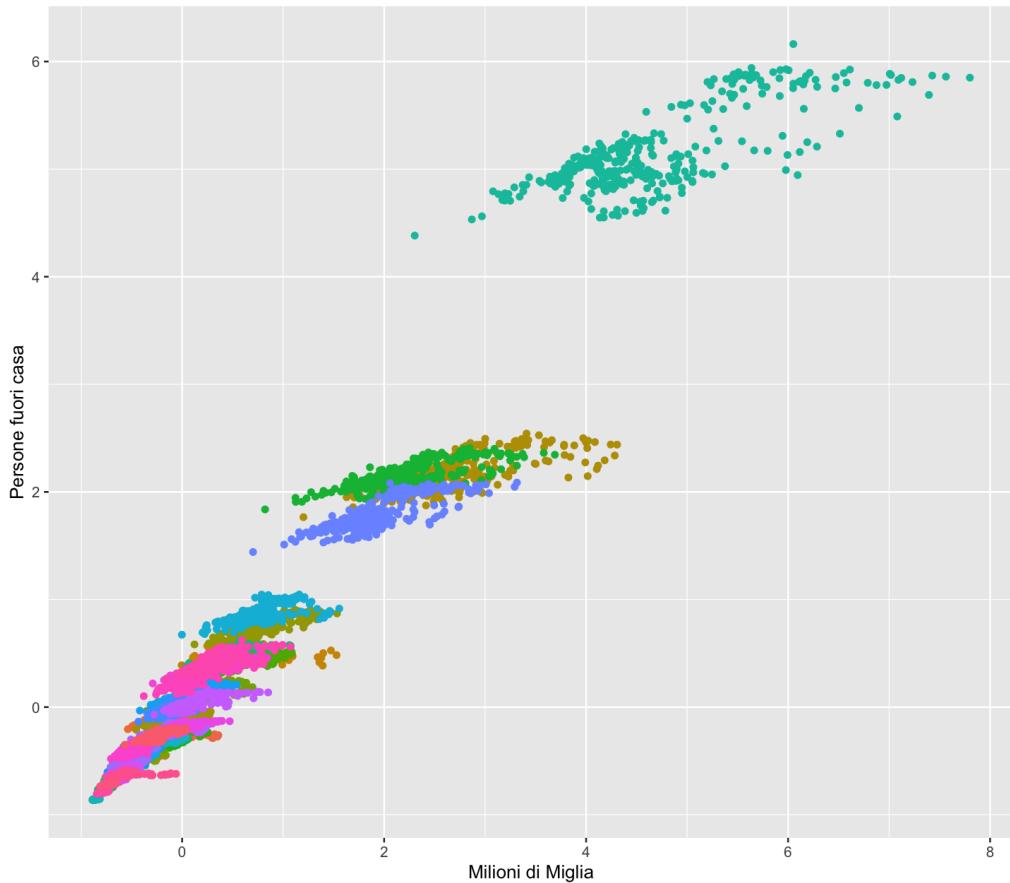


Figura 1.2: Scatterplot tra le milioni di miglia percorse e la popolazione che esce di casa.

- miglia percorse da veicoli,
- numero di abitanti che esce di casa,
- numero di abitanti che rimane a casa.

Possiamo osservare (figura 1.2) che tra le variabili miglia percorse e abitanti usciti dall’abitazione vi è una correlazione globale del 98% circa, ciò avviene in quanto entrambe le variabili sono strettamente legate alla dimensione delle città. I modelli utilizzerebbero entrambe le variabili come proxy della data dimensione e non coglierebbero l’informazione dell’andamento relativo alla data città. Si decide quindi di inserire una variabile (costante nel tempo) che indica la popolazione totale, aggiungendo una seconda che indica la frazione di persone fuori casa nel dato giorno.

Anche stratificando per città le variabili numero di persone fuori casa e numero di miglia percorse risultano abbastanza correlate (75% circa), si è quindi deciso (anche alla luce di valutazioni legate all’adattamento e alla capacità previsiva) di tenere una delle due opzioni per favorire la stabilità dei modelli.

Si è poi proceduto con un lisciamento di tutte le variabili. Al fine di poter cogliere tutta la variabilità nella risposta e nelle covariate, si è deciso di interpolare i dati utilizzando delle B-splines di primo grado, con un nodo in corrispondenza di ogni osservazione. Tale scelta è giustificata dal fatto che si crede che la variazione delle covariate tra un giorno e l’altro possa essere un’importante informazione per cogliere variazioni del pm 2.5. Tuttavia se si

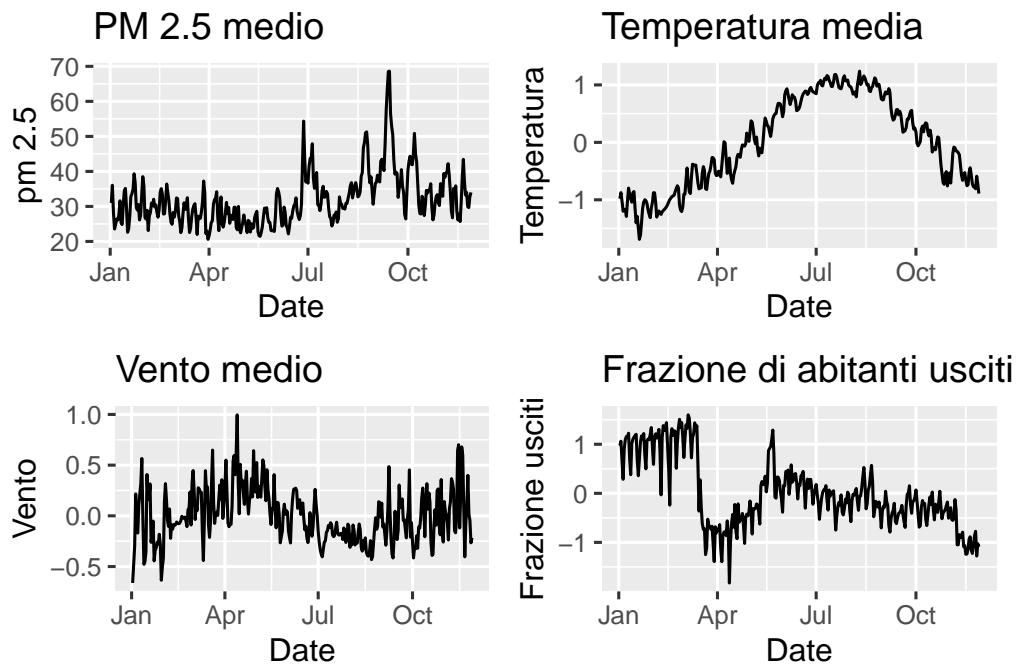


Figura 1.3: Medie funzionali per alcune variabili.

rendono lisci i dati (ad esempio selezionando il livello di rugosità mediate GCV) si perde parte dell'informazione data dalla variabilità giornaliera.

In figura 1.3 possiamo osservare i grafici delle medie funzionali di alcune variabili. Nella variabile risposta, pm 2.5, possiamo osservare i picchi legati agli incendi nel periodo tra agosto ed ottobre.

Contrariamente a quanto ci si poteva aspettare, le medie funzionali ottenute risultano essere poco lisce. Ciò risulta essere dovuta alla presenza di correlazione tra le osservazioni, infatti, le osservazioni presentano una correlazione spaziale oltre a quella temporale. Tale correlazione dovrà essere modellata in modo adeguato al fine di ottenere dei buoni modelli.

Capitolo 2

Modellazione

Data la natura delle variabili in esame, si sono considerati modelli per dati funzionali e modelli state-space per dati in forma di serie storiche. In prima battuta si è applicato un approccio di tipo *concurrent*, per poi esplorare un modello ad effetto non simultaneo.

2.1 Modello funzionale simultaneo

Il primo approccio considerato è un modello a risposta e covariate funzionali simultaneo.

Il modello è così definito:

$$y_i(t) = \beta_0(t) + \sum_{j=1}^p \beta_j(t)x_{ij}(t) + \epsilon_i(t)$$
$$\epsilon_i(t_1) \perp \epsilon_w(t_2); \forall t_1, t_2; i \neq w$$
$$i \in \{1, \dots, 50\} \quad t \in \{1, \dots, 333\}$$

Dove le covariate funzionali inserite sono: temperatura, vento, pressione e umidità per quanto riguarda i fattori metereologici; numero di abitanti e frazione di abitanti usciti dall'abitazione come fattori legati all'attività umana.

Questo tipo di legame tra risposta e covariate è detto simultaneo in quanto valore delle esplicative in un dato tempo t è legato al valore della risposta al medesimo tempo t .

I coefficienti funzionali $\beta_j(t)$ sono modellati tramite B-splines cubiche aventi un numero di nodi equispaziati pari a 10.

$$\beta_j(t) = \sum_{k=1}^{10} \gamma_k(t)\alpha_k$$

Si noti come, in questo modello, il termine di errore $\epsilon_i(t)$ sia considerato incorrelato tra osservazioni (città) i diverse. Questo aspetto, che verrà approfondito in seguito, non si è coerente con i dati osservati, in quanto la risposta, anche al netto delle covariate, si dimostrerà correlata tra città diverse.

Il modello è stato implementato tramite la libreria *fda*; in particolare la stima è implementata dalla funzione *fRegress*.

2.1.1 Regolazione

Per migliorare la stima degli effetti del modello, evitando di rilevare andamenti spuri, si è effettuata una regolazione basata sulla minimizzazione dell'errore quadratico medio nel *validation set*.

La regolazione si basa sulla penalizzazione delle derivate dei coefficienti temporali $\beta_j(t)$; la fase di stima avviene quindi minimizzando la somma degli integrali delle funzioni residuali al quadrato, penalizzata, per ogni coefficiente j , dai seguenti termini:

- $\lambda_{2j} \cdot \int_1^{333} \beta''(t)^2 dt$
- $\lambda_{1j} \cdot \int_1^{333} \beta'(t)^2 dt$
- $\lambda_{0j} \cdot \int_1^{333} \beta(t)^2 dt$

La funzione di perdita diventa:

$$\text{loss} = \sum_{i=1}^{50} \int_1^{333} (y_i(t) - \hat{y}_i(t))^2 dt + \sum_{j=0}^p \int_1^{333} \beta''(t)^2 \lambda_{2j} + \beta'(t)^2 \lambda_{1j} + \beta(t)^2 \lambda_{0j} dt$$

Le penalità legate alla derivata seconda regola la liscezza degli effetti temporali; le penalità legate alla derivata prima regola la presenza di effetti temporali rispetto a quelli fissi; le penalità legate alla funzione stessa regolano invece la presenza/assenza dell'effetto nel suo complesso, che sia questo temporale o fisso.

Si ha così una regolazione simultanea di che variabili vengono inserite nel modello, di quali hanno effetto fisso o temporale, e del grado di liscezza degli effetti temporali.

Dato l'elevato numero di parametri di regolazione (in tutto 21), si è applicato un algoritmo iterativo che, partendo da un set di inizializzazione, ottimizzi numericamente l'errore sull'insieme di validazione, regolando una variabile alla volta (3 parametri) per un fissato numero di passi di ottimizzazione, ripetendo l'operazione fino a convergenza.

Algorithm 1 Algoritmo di regolazione

```

1: abs.tol =← 0.001
2: error_prev ← ∞
3: error_current ← 0
4: while error_prev – error_current > abs.tol do
5:   error_prev ← error_current
6:   for j ← 1 to 7 do
7:     ottimizza  $\lambda_{j0}, \lambda_{j1}, \lambda_{j2}$  e aggiorna modello
8:   end for
9:   error_current ← errore modello aggiornato
10: end while

```

L'ottimizzazione numerica nel ciclo interno è implementata dalla funzione *optim*, tramite metodo *Nelder-Mead*, per un numero massimo di passi di ottimizzazione fissato a 10.

2.1.2 Risultati

Si riportano in figura 2.1 le stime ottenute dopo la fase di regolazione.

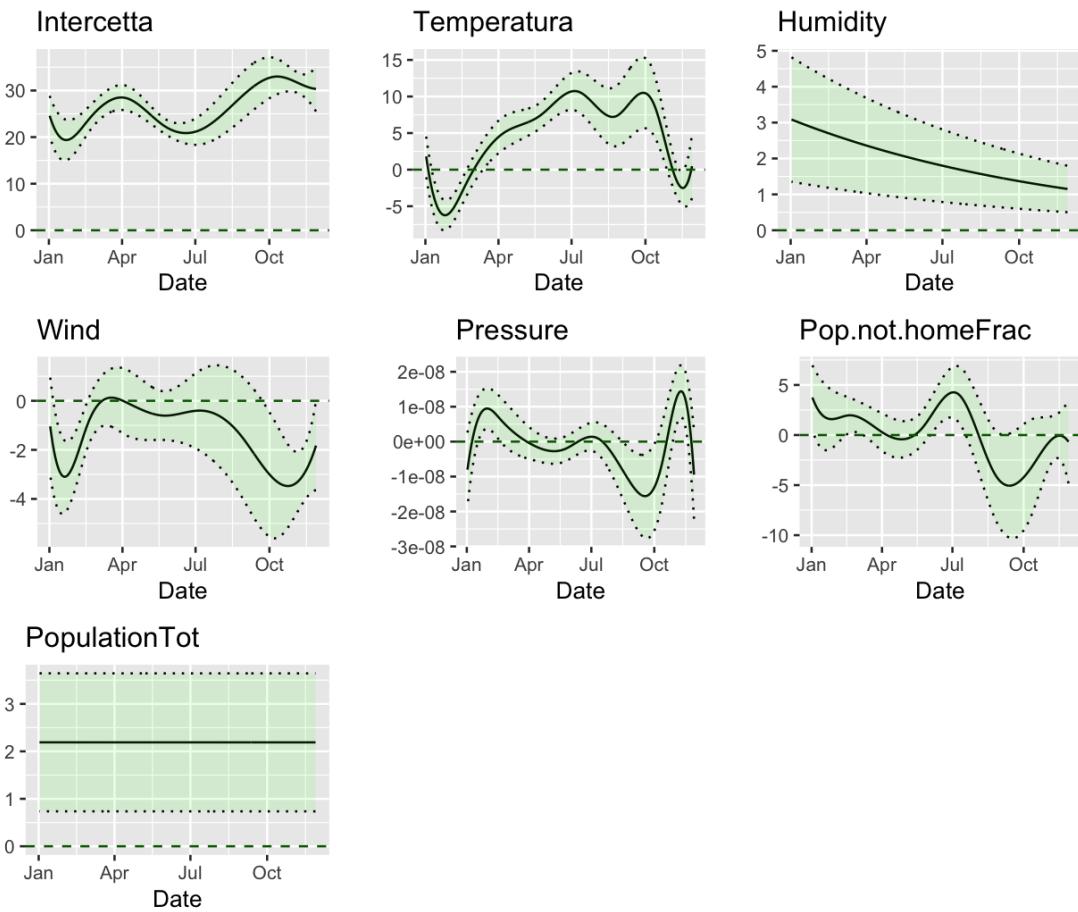


Figura 2.1: Coefficienti modello funzionale simultaneo

Si può notare come il coefficiente relativo a *pressure* è stato sostanzialmente annullato grazie alla penalità legata a λ_0 . Ciò porta a dedurre che la pressione non abbia effetto reale sulla risposta.

La variabile *PopulationTot* è stata regolata fino ad avere un effetto fisso grazie alla penalità legata a λ_1 , mentre le restanti variabili sembrano avere un effetto temporale con diversi gradi di liscezza.

Commenti più approfonditi e considerazioni di carattere interpretativo riguardo gli effetti si riportano a seguito nella sezione conclusiva.

In figura 2.2 si riportano le previsioni per le città nell'insieme di *test*; possiamo notare come il modello colga (anche se non in tutte le città) l'andamento di lungo periodo del pm 2.5, ed in alcuni casi anche variazioni più dettagliate. L'errore quadratico medio nel *test set* vale 164.3094. L'indice R^2 complessivo del modello è di 0.153.

Un aspetto critico sembra essere legato all'incapacità del modello di cogliere il livello medio della risposta per particolari città (ad esempio *Miami*). Questo aspetto verrà risolto in seguito tramite effetti casuali legati appunto alle città. Una seconda criticità è legata alla correlazione geo-spatiale possibilmente presente tra gli errori; anche questo aspetto verrà approfondito in seguito.

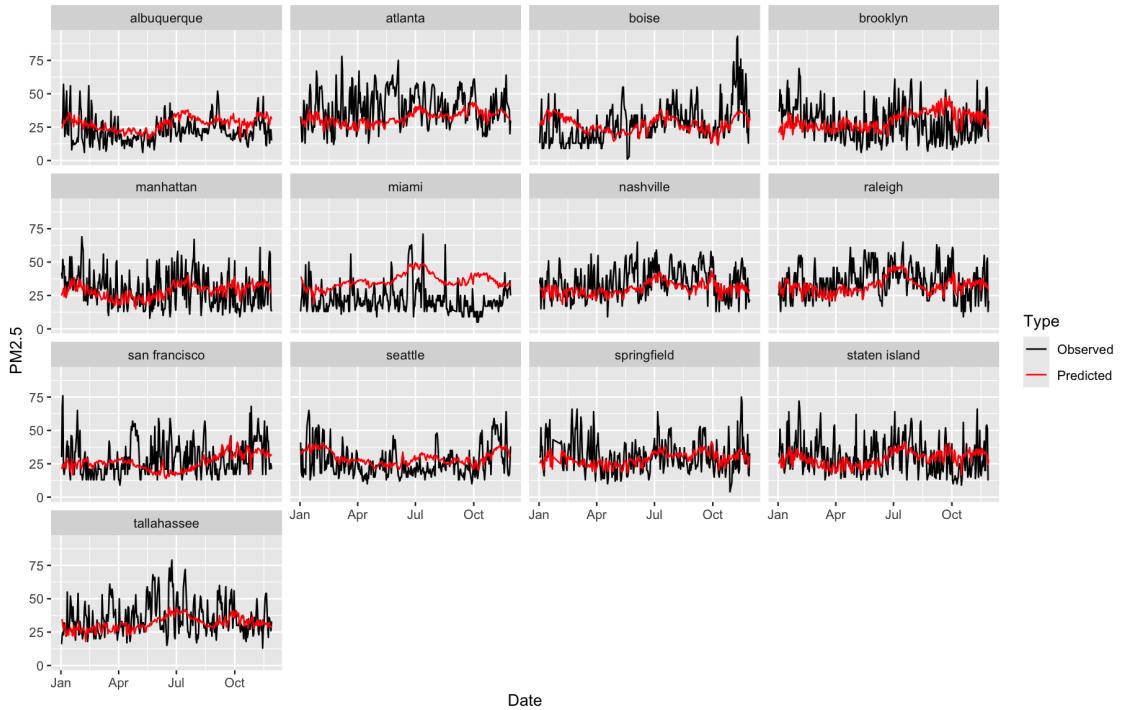


Figura 2.2: Previsioni modello funzionale simultaneo

2.2 Modello state-space

Il secondo approccio alla modellazione implementa un modello di tipo state-space ad effetto simultaneo.

La definizione del modello state space gaussiano è la seguente:

$$\begin{cases} y_t = X_t \beta_t + \varepsilon_t \\ \beta_t = \beta_{t-1} + \eta_t, \quad t = 2, \dots, 333 \\ \beta_1 \sim \mathcal{N}_d(b_0, P_1) \end{cases}$$

e

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim \mathcal{N}_{n+p} \left(0, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\eta \end{pmatrix} \right),$$

$$i \in \{1, \dots, 50\} \quad t \in \{1, \dots, 333\}$$

Anche in questo caso, le covariate inserite sono: temperatura, vento, pressione e umidità per quanto riguarda i fattori metereologici; numero di abitanti e frazione di abitanti usciti dall'abitazione come fattori legati all'attività umana.

I parametri di questo modello sono legati alle matrici di varianza covarianza del termine di errore ε_t e dell'innovazione dello stato latente η_t .

Il punto di partenza b_0 è stato stimato a priori, attraverso un modello lineare semplice utilizzando i primi 5 giorni disponibili, inserendo poi in P_1 la variabilità di tale stima.

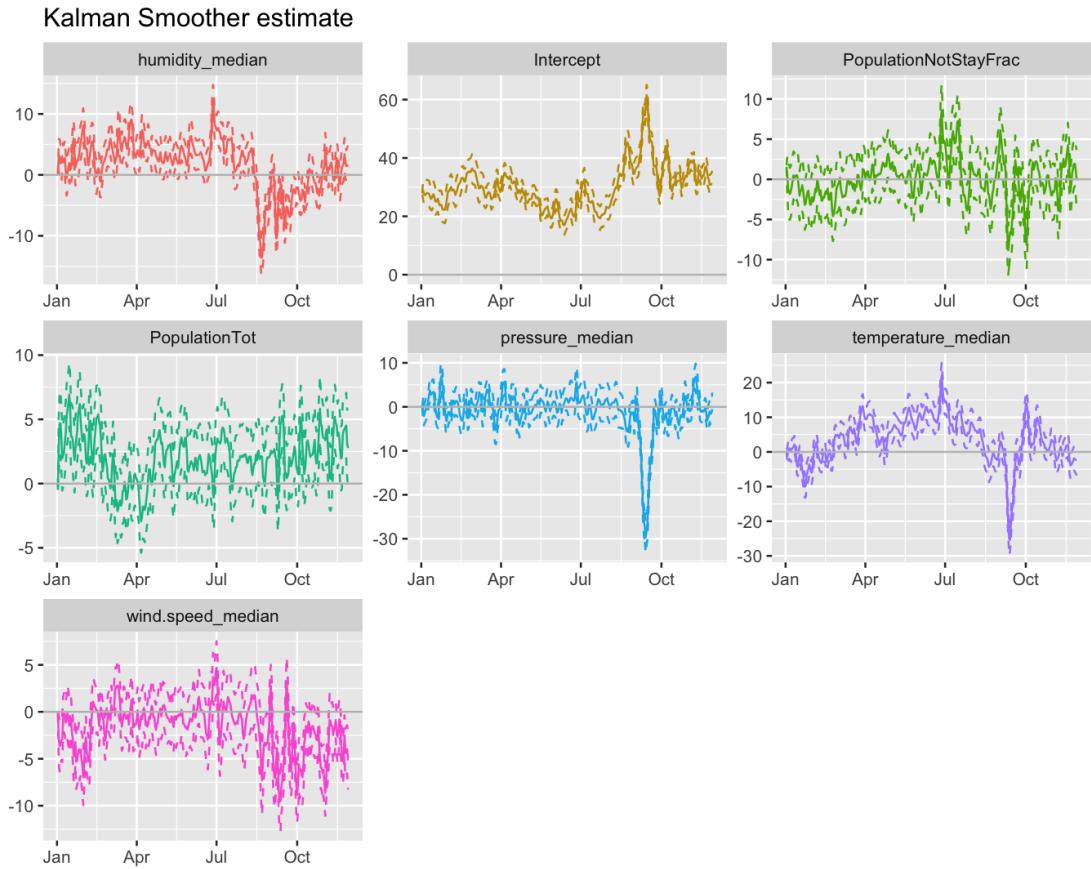


Figura 2.3: Coefficienti state-space (Σ_η diagonale)

La struttura della matrice Σ_η va a definire le possibili correlazioni tra i termini di errore delle varie città, in modo da cogliere potenziali dipendenze di tipo geo-spaziale. Nella seguente sezione vedremo come, abbandonando una struttura di tipo diagonale, le stime degli effetti miglioreranno notevolmente.

Il modello è stato implementato tramite la libreria *KFAS*; in particolare la stima è implementata dalla funzione *KFS*.

2.2.1 Gestione della correlazione tra città

In una prima versione stimata del modello, le matrici di covarianza Σ_η e ε_t sono state assunte diagonali, ed entrambe stimate tramite la massimizzazione della verosimiglianza. Il modello dispone così di ($p + 1$ parametri per la diagonale di Σ_η e 1 parametro per tutta la diagonale di ε_t). Come si può vedere in figura 2.3, ciò porta a delle stime molto frastagliate.

La motivazione è probabilmente la correlazione geo-spaziale tra le città, che presentano a gruppi movimenti simili anche al netto delle covariate. Concettualmente questo può essere rappresentato da fattori latenti (potenzialmente correlati alle variabili inserite nel modello) che influenzano la risposta. I coefficienti delle variabili esplicative osservate (o dell'intercetta) diventano frastagliati in quanto cercano di cogliere l'effetto dovuto ai suddetti fattori latenti.

Per depurare quindi le stime si è deciso di dare una particolare struttura alla matrice varianza covarianza dell'errore Σ_η .

In particolare, la struttura che ha portato i risultati più soddisfacenti è la seguente:

$$\Sigma_\eta = (0.5 \cdot M(\phi, \kappa) + 0.5 \cdot A) \cdot \sigma_\eta$$

dove $M(\phi, \kappa)$ è una matrice di correlazione di Matérn, che considera quindi solo le distanze euclidee tra le città, mentre A è una matrice che considera le correlazioni degli eventi atmosferici, ottenuta come media delle matrici di correlazione tra città stimate utilizzando le 4 covariate metereologiche disponibili $A = 0.25 \cdot (\text{cov}(\text{temperature}) + \text{cov}(\text{wind}) + \text{cov}(\text{humidity}) + \text{cov}(\text{pressure}))$. Questa specificazione permette di sopperire ad eventuali distorsioni legate all'utilizzo delle sole distanze euclidee; si pensi a città spazialmente vicine ma tra le quali c'è una barriera naturale (es. montagne), o a città distanti collegate da macro-correnti atmosferiche. Essendo l'errore del modello probabilmente legato ad effetti latenti di carattere metereologico, supponiamo che questo abbia, almeno in parte, una struttura di correlazione tra città simile a quella stimata in A . Si noti inoltre che, nonostante le covariate possano depurare la risposta da correlazioni dovute agli eventi atmosferici in esse osservati (o ad essi fortemente correlati), i fattori latenti sono potenzialmente, almeno in una loro componente, incorrelati alle variabili inserite nel modello; è proprio la correlazione degli effetti latenti (o la sua parte che non viene catturata dalle esplicative) ad essere inglobata nel termine di errore, per il quale ipotizziamo avere una correlazione simile a quella descritta in A . Si spera inoltre di limitare che l'effetto stimato delle variabili osservate catturi anche l'effetto di ipotetici fattori latenti ad esse correlate, riducendo il *bias* delle stime.

Questa scelta è stata esplorata più accuratamente in appendice, con qualche esempio di simulazione.

La matrice di correlazione stimata tramite massima verosimiglianza è rappresentata in figura 2.4.

I coefficienti del modello ottenuti da questa specificazione sono riportati nella figura 2.5

2.2.2 Regolazione

La stima tramite massima verosimiglianza delle varianze delle innovazioni η_t porta comunque a traiettorie dei coefficienti casuali β_t molto frastagliate e di difficile interpretazione. Si è quindi effettuata una regolazione delle suddette varianze attraverso la minimizzazione del MSE sul *validation-set* tramite ottimizzazione numerica mediante l'algoritmo *BFGS*.

2.2.3 Risultati

Si riportano in figura 2.6 le stime ottenute dopo la fase di regolazione.

Si osserva che mediante la regolazione delle componenti di varianza delle innovazioni η_t le stime dei parametri dinamici risultano essere molto più lisce e interpretabili (l'errore nell'insieme di test è inoltre notevolmente migliorato, da 203.41 a 159.21)

Si ha inoltre che il coefficiente della variabile *temperature* risulta essere stimato in modo coerente con quello ottenuto dal modello funzionale simultaneo.

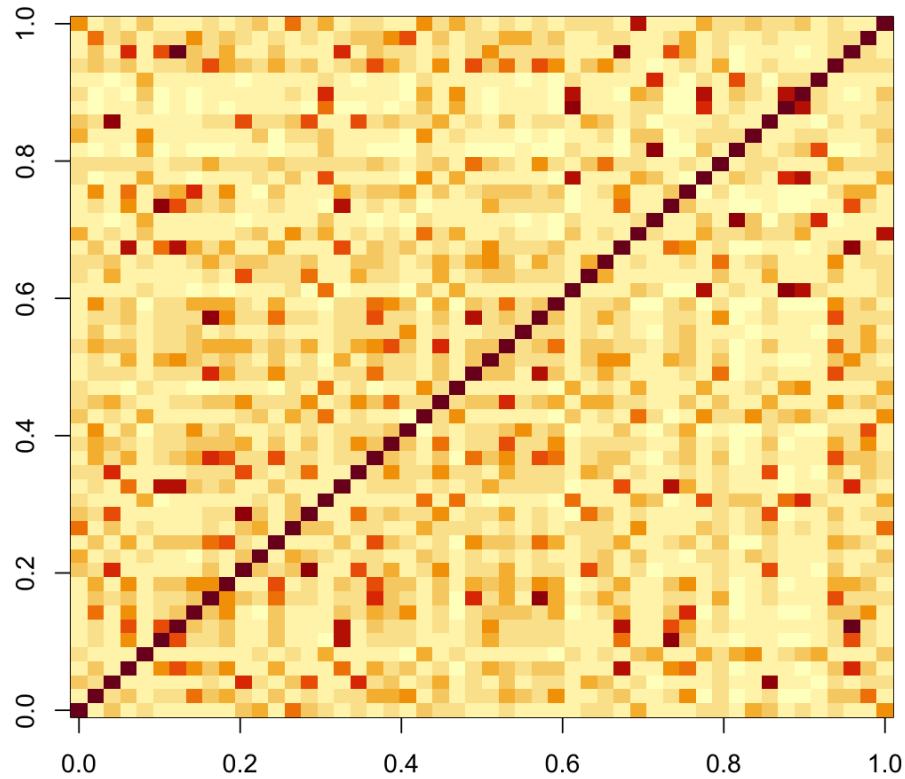


Figura 2.4: Matrice di correlazione Σ_η stimata

L’interpretazione dei coefficienti si lascia alla sezione conclusiva.

In figura 2.7 si riportano le previsioni per le città nell’insieme di *test*; si possono fare considerazioni simili a quelle del modello precedente. L’errore quadratico medio nel *test set* vale 159.2071.

Anche in questo caso, una criticità sembra legata all’incapacità del modello di cogliere il livello medio della risposta per alcune città. Questo aspetto verrà affrontato tramite la definizione esplicita di effetti casuali spazio-temporali nel modello successivo.

2.3 Modello spazio-temporale

Al fine di cogliere i livelli medi di città si è deciso di adattare un modello spazio-temporale.

Sia $y_i(t)$ il $PM_{2.5}$ misurato nella città s_i ($i = 1, \dots, n$) e giorno $t = 1, \dots, T$

Si assume

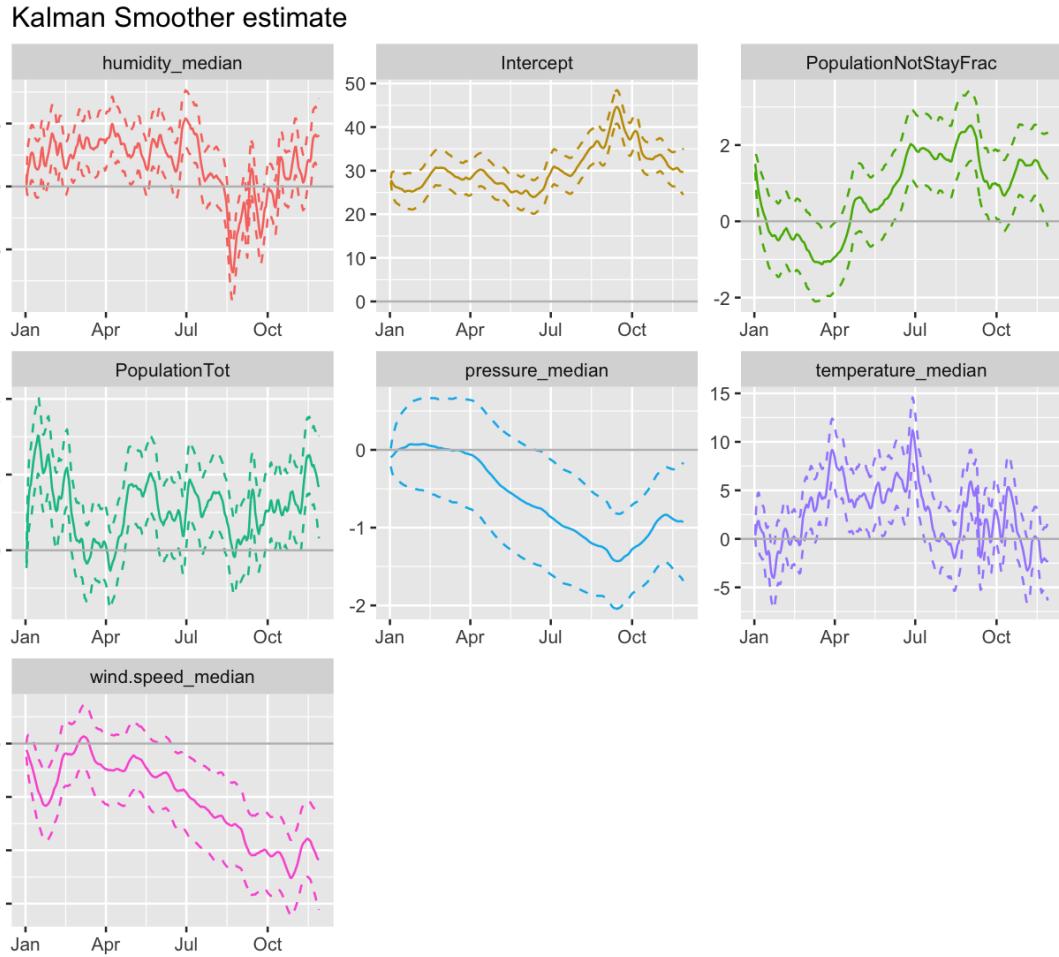


Figura 2.5: Coefficienti state-space (Σ_η non diagonale)

$$Y_i(t) \sim \mathcal{N}(\mu_i(t), \sigma^2),$$

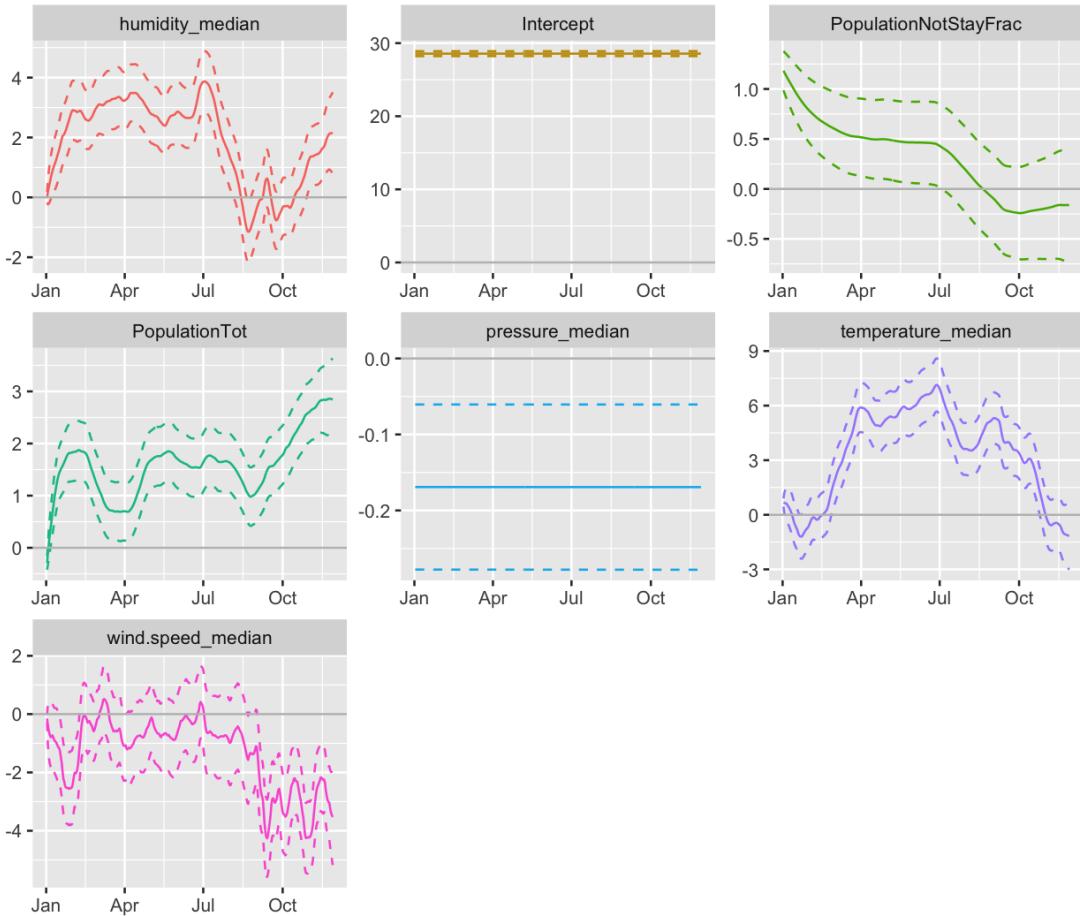
con σ_e^2 la varianza dell'errore, sia serialmente che spazialmente non correlato, e

$$\begin{cases} \mu_i(t) = \beta_0 + \sum_{j=1}^p \beta_j(t)x_{ji}(t) + \omega_i(t) \\ \beta(t) = \beta(t-1) + \eta(t) \end{cases}$$

dove β_0 è l'intercetta fissa, $\eta(t) \sim \mathcal{N}(0, W)$ (con W tipicamente diagonale) e $\beta(t) = (\beta_1(t) \dots \beta_p(t))$ sono gli effetti relativi alle covariate meteorologiche e di popolazione x_1, \dots, x_p . Il termine ω_{it} si riferisce al processo spazio-temporale latente (cioè il vero livello non osservato di inquinamento) che cambia nel tempo con dinamiche autoregressive di primo ordine e innovazioni correlate spazialmente:

$$\omega_{it} = \rho\omega_{i(t-1)} + \xi_{it},$$

Kalman Smoother estimate

Figura 2.6: Coefficienti state-space regolato tramite *validation*

con $t = 2, \dots, T$, $|\rho| < 1$ e $\omega_{i1} \sim \mathcal{N}(0, \sigma^2/(1 - \rho^2))$. Inoltre, ξ_{it} è un processo gaussiano a media zero, assunto temporalmente indipendente e caratterizzato dalla seguente funzione di covarianza spazio-temporale:

$$\text{Cov}(\xi_{it}, \xi_{ju}) = \begin{cases} \text{Cov}(\xi_i, \xi_j) & \text{se } t = u, \\ 0 & \text{se } t \neq u. \end{cases}$$

per $i \neq j$, dove $\text{Cov}(\xi_i, \xi_j)$ è data dalla funzione di covarianza spaziale Matérn. Per approfondimenti si veda Blangiardo and Cameletti (2015).

Il modello è stato stimato con le covariate temperatura e vento con effetti variabili nel tempo e le covariate umidità, pressione, popolazione totale e frazione della popolazione che esce di casa come effetti fissi. Si è scelta tale specificazione del modello in quanto si è visto essere la migliore in termini di previsione sul *test-set*. Per stimare il modello si è utilizzato il pacchetto INLA, e si sono utilizzate le priori (non informative) di default del pacchetto.

In seguito alla stima del modello si sono lisciate le intercette spazio-temporali di ogni città mediante B-splines di grado 3, si è regolata la rugosità delle stime mediante il set di

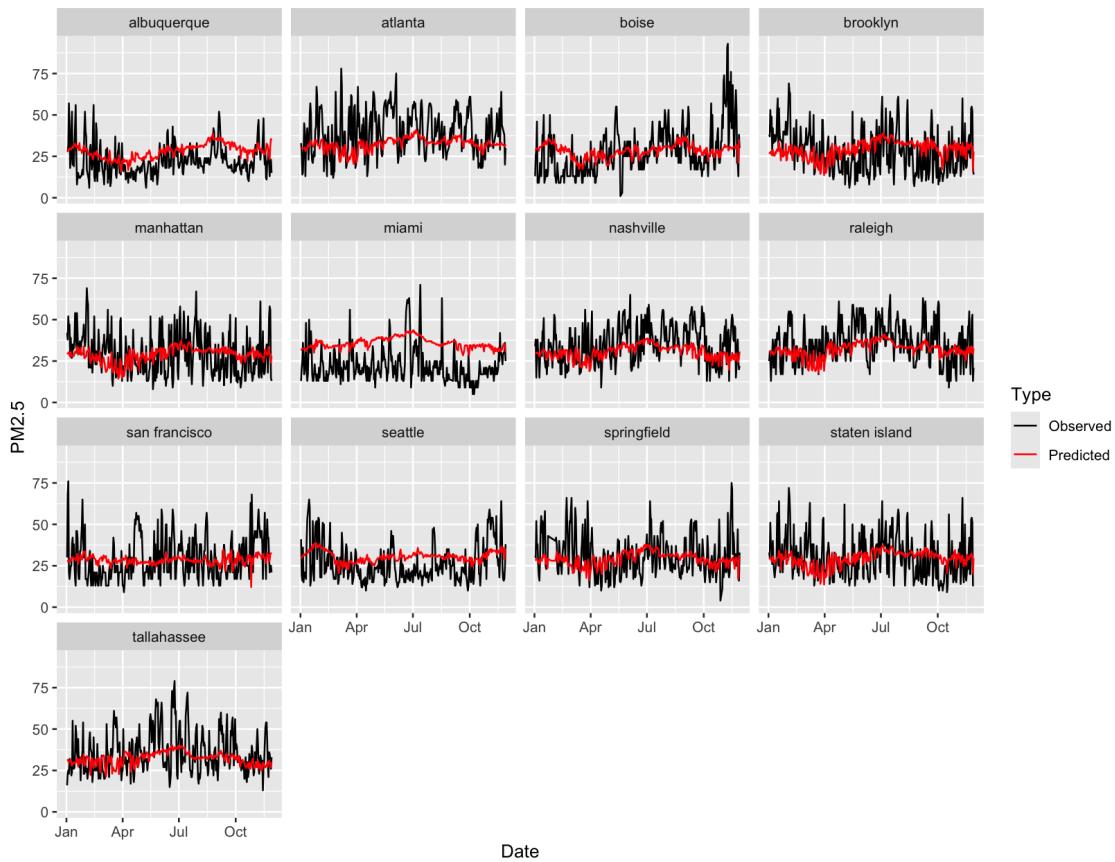


Figura 2.7: Previsioni modello state-space

validation.

2.3.1 Risultati

Il modello adattato produce un errore quadratico medio sul *test-set* pari a 153.33.

Si riportano in figura 2.8 le stime dei parametri dinamici e le intercette lisciate. Si osserva che le stime del coefficiente associato alla variabile temperatura risultano essere coerenti con quelle stimate mediante i modelli visti precedentemente, tuttavia risulta essere molto variabile. Il parametro associato alla variabile vento risulta essere negativo, con un andamento leggermente decrescente. Rispetto ai modelli visti precedentemente, tale coefficiente viene stimato in modo molto più liscio.

In tabella 2.1 si riportano le stime dei parametri fissi nel modello.

Coerentemente con la procedura di selezione delle variabili fatta per il modello funzionale simultaneo, in cui il coefficiente della variabile pressione veniva fortemente compresso a zero, l'intervallo di credibilità al 95% per tale variabile include lo zero.

2.3.2 Lisciamento dei coefficienti

Come si è visto in figura 2.8, le stime del parametro dinamico della covariante *temperature* risultano essere molto frastagliate. Tuttavia ci aspettiamo che una stima liscia del parametro possa portare ad un miglioramento del modello, in quanto ne riduce la variabilità.

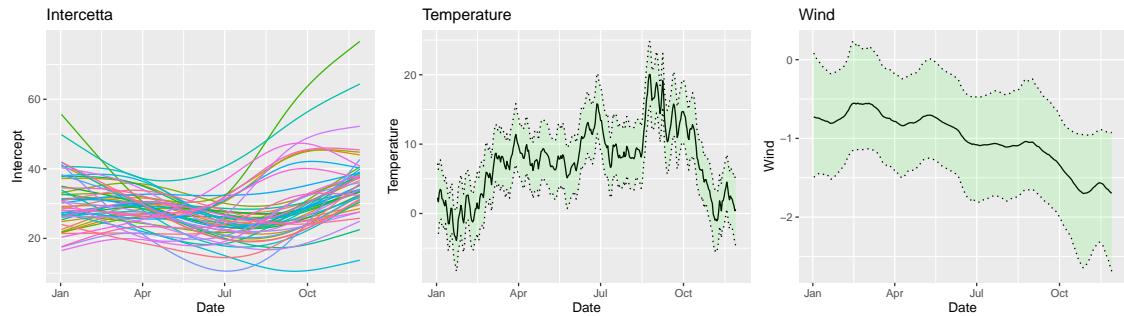


Figura 2.8: Coefficienti RW1 del modello spazio tempo.

Tabella 2.1: Stime degli effetti fissi del modello

	mean	sd	0.025quant	0.975quant
Intercept	28.73	0.93	26.90	30.56
Temperature	7.59	0.34	6.92	8.26
Wind	-1.01	0.10	-1.21	-0.80
Humidity	2.87	0.18	2.52	3.22
PopulationTot	0.36	0.16	0.05	0.67
Pressure	0.14	0.15	-0.16	0.44
PopulationFrac	0.36	0.15	0.07	0.66

Specifichiamo quindi un modello del tipo

$$Y_i(t) \sim \mathcal{N}(\mu_i(t), \sigma^2),$$

con σ_e^2 la varianza dell'errore, sia serialmente che spazialmente non correlato, e

$$\mu_i(t) = \beta_0 + \sum_{j=1}^p \beta_j(t)x_{ij}(t) + \omega_i(t)$$

dove β_0 è l'intercetta fissa, e $\omega_i(t)$ ha la medesima definizione del modello visto precedentemente.

Per i parametri $\beta_j(t)$ specifichiamo un'espansione in basi B-splines di secondo grado con 20 nodi equispaziati, del tipo:

$$\beta_j(t) = \phi_j(t)b$$

Regoliamo la complessità del lisciamento specificando per i parametri b una struttura di covarianza di Matérn. Per approfondimenti si veda Gómez-Rubio (2020).

Come per il modello a coefficienti dinamici stimato (RW1), si specificano due coefficienti funzionali, *temperature* e *wind*, e tutti gli altri vengono inseriti come parametri fissi. Ancora una volta lisciamo le intercette di ogni città mediante B-splines di grado 3, si è regolata la rugosità delle stime mediante il set di validation.

Risultati

Il modello adattato produce un errore quadratico medio sul *test-set* pari a 148.15.

Si riportano in figura 2.9 le stime dei parametri di *temperature* e *wind*.

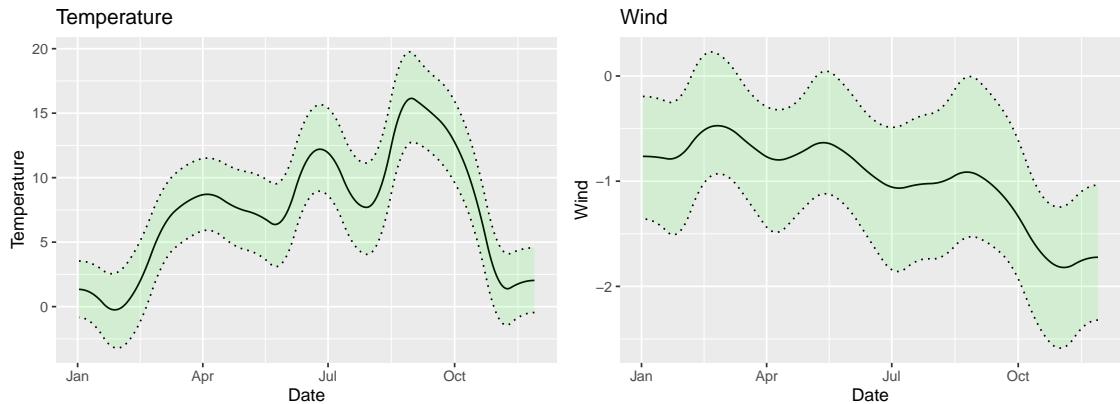


Figura 2.9: Coefficienti stimati mediante B-splines nel modello spazio-tempo.

Osserviamo che ora anche le stime dei parametri associati alla variabile temperatura risultano essere lisce.

In figura 2.9 si possono osservare le previsioni dei due modelli spazio-temporali. Si può notare che mediante tali modelli riusciamo a cogliere meglio il livello delle città rispetto al modello state space visto nella sezione precedente.

Si riporta in figura 2.11 un grafico che mostra la dinamica dell'intercetta nello spazio per quattro tempi differenti.

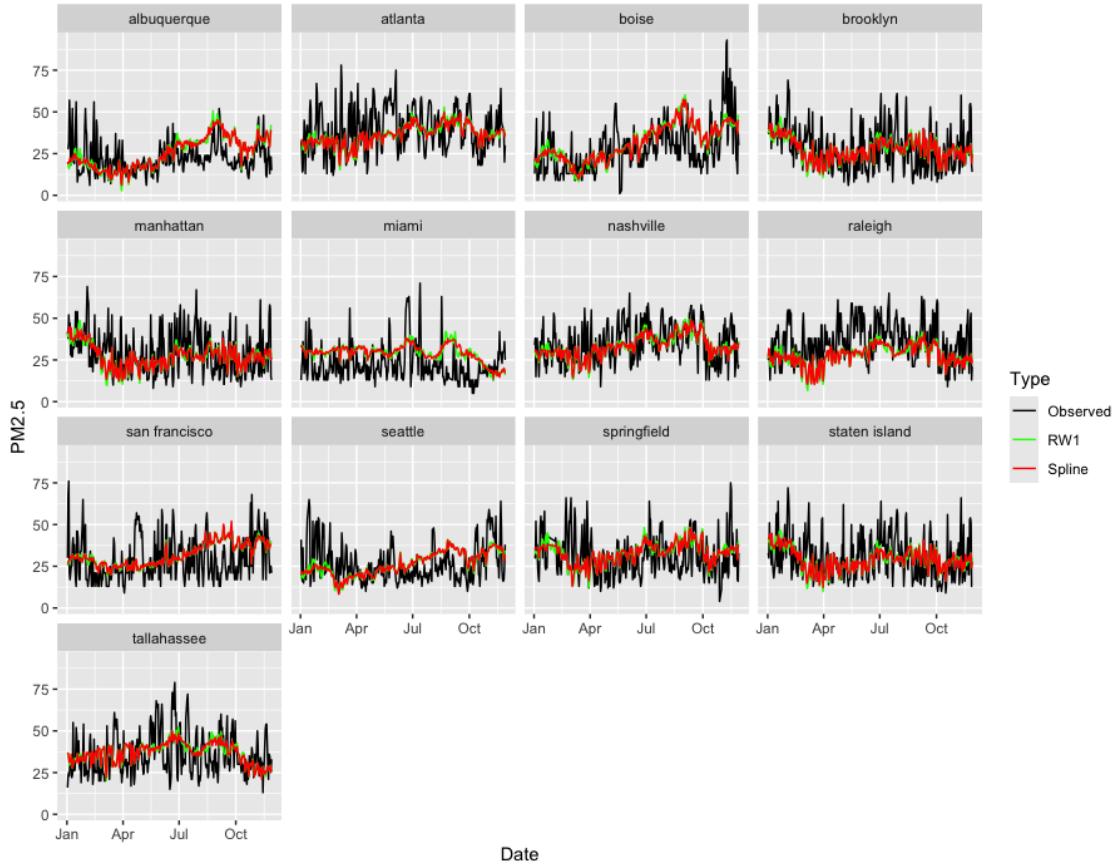


Figura 2.10: Previsioni dei due modelli spazio-tempo.

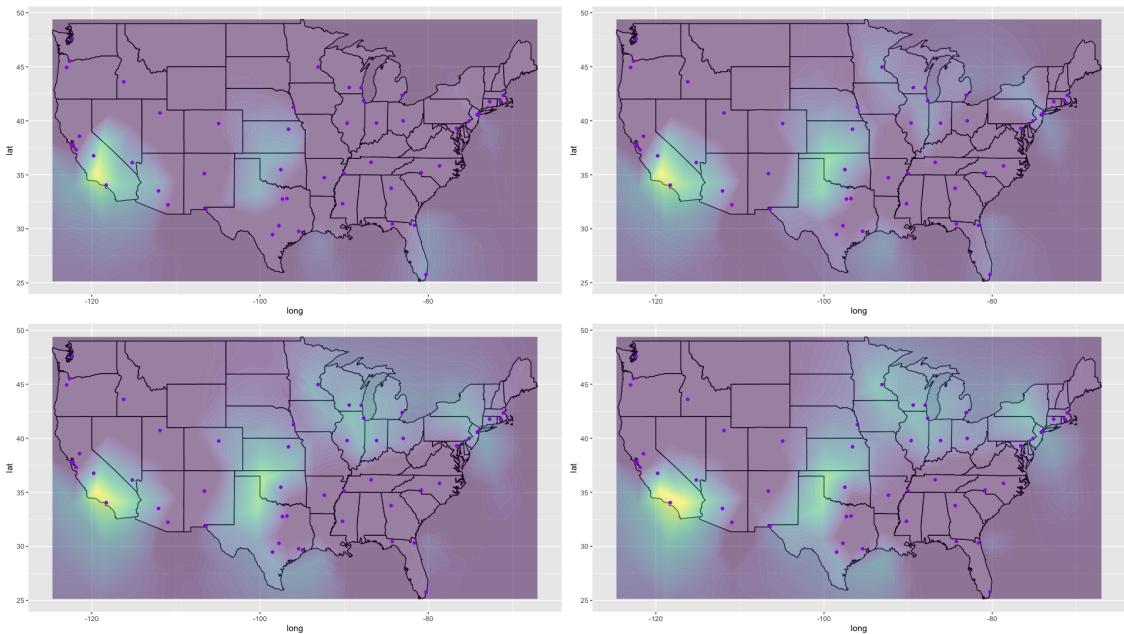


Figura 2.11: Intercette spaziali per 4 diversi tempi consecutivi

Capitolo 3

Conclusioni

L'obiettivo di questo lavoro è di analizzare la relazione di agenti atmosferici e variabili legate all'attività umana con il pm 2.5, e come tali relazioni possano variare nell'arco dell'anno.

Per perseguire tale obiettivo si sono adattati diversi modelli, quali:

- Modello funzionale simultaneo;
- Modello state-space;
- Modello spazio-tempo.

Ognuno dei modelli adattati è stato poi valutato in termini di capacità previsiva nel *test-set*. Si riporta in tabella 3.1 gli MSE per i diversi modelli adattati.

Tabella 3.1: MSE per i modelli adattati.

	MSE
Funzionale simultaneo	164.30
State space (regolato)	159.20
Spazio-tempo RW1	153.33
Spazio-tempo B-splines	148.15

Il modello che minimizza l'MSE sul *test-set* risulta essere il modello spazio-temporale. Per tale modello, i parametri variabili nel tempo sono stimati mediante un'espansione in basi B-splines di secondo grado.

Inoltre, si ritiene tale modello il più affidabile anche in termini di qualità delle stime degli effetti, per quanto riguarda la loro interpretazione, in quanto si ritiene che l'aggiunta di effetti latenti spazio-temporali "pulisca" le stime da eventuali fattori latenti che governano il fenomeno reale. Si è scelto quindi di basare le interpretazioni degli effetti delle covariate sulla risposta, riportate in seguito, sui risultati di tale modello.

Le stime parametri associati alle variabili *temperature* e *wind* (si veda figura 2.9) presentano un andamento temporale.

Il parametro della variabile *temperature*, nei primi mesi dell'anno viene stimato prossimo a zero, successivamente presenta un andamento crescente, fino al mese di aprile, seguito da un periodo oscillatorio, per poi tornare a decrescere verso lo zero negli ultimi mesi dell'anno.

Si può quindi affermare che un aumento di temperatura nei mesi caldi dell'anno sembra essere associato ad una crescita del pm 2.5. Si osserva inoltre che l'evoluzione del parametro nel tempo presenta un pattern molto simile in tutti i modelli stimati.

Il parametro della variabile *wind* viene stimato dal modello con segno negativo, inoltre presenta un'evoluzione temporale decrescente con delle piccole oscillazioni. Ciò è coerente con l'idea che il vento possa disperdere il particolato presente in aree urbane, diminuendone così la concentrazione.

Il coefficiente legato alla variabile *humidity* risulta positivo. Questo potrebbe sembrare controintuitivo rispetto all'effetto della pioggia, che notoriamente abbassa le concentrazioni di pm 2.5. Si noti che però alcuni sistemi di rilevazione del particolato sono inficiati dalla presenza di umidità, che si lega alle particelle di pm 2.5 rendendole più pesanti.

Il coefficiente legato alla variabile *pressure* non sembra significativo; ciò è coerente con la regolazione del modello funzionale simultaneo, che collassava l'effetto della suddetta variabile fino, sostanzialmente, ad annullarlo.

In ultimo, il coefficienti legati alle variabili *popolazione totale* e *frazione della popolazione uscita di casa* risultano positivi.

Appendice A

Simulazioni correlazione atmosferica

Si riportano dei semplici esempi, ottenuti da scenari simulati, per esplorare la scelta di impiegare la correlazione stimata tramite variabili inserite nel modello per stimare la correlazione del termine di errore. Gli scenari esplorati non sono in alcun modo esaustivi, e rappresentano in qualche modo le condizioni ideali perché questo metodo possa funzionare, ma sono utili a rappresentare l’idea di fondo che ha portato alla scelta fatta nella sezione 2.2.1. In tutti gli scenari si mantengono 50 osservazioni rilevate in 100 tempi ciascuna.

A.1 Fattore latente incorrelato con la variabile osservata

Il primo, e più semplice, scenario, implementa una generazione della risposta come:

$$y_i(t) = 5 \cdot o_i(t) + 5 \cdot l_i(t) + \epsilon_i(t)$$

dove sia il fattore osservato $o_i(t)$, che quello latente $l_i(t)$ sono generati da una normale multivariata a media nulla. Tale normale multivariata avrà la stessa struttura di correlazione tra le osservazioni per entrambi i fattori (uguale alla correlazione stimata per la variabile temperatura nei dati originali). Il fattore latente e quello osservato saranno però del tutto indipendenti sia tra loro, che con se stessi a tempi diversi.

Nella figura A.1 si riportano, in ordine: la matrice di correlazione (50×50) utilizzata per la generazione delle covariate; le covariate generate (fattore osservato e latente); la risposta; gli effetti stimati dal modello state space, per il quale si è inserita solo la variabile osservata tra le covariate (in nero utilizzando una matrice diagonale per il termine di covarianza dell’errore, in rosso utilizzando la correlazione stimata dalla variabile osservata).

Si nota come stimare la correlazione dell’errore tramite la correlazione tra le osservazioni della variabile osservata migliori la stima, non in termini di bias, ma in termini di variabilità. Ciò non sorprende in questo caso, in quanto il fattore latente (che presenta la stessa struttura di correlazione tra osservazioni della variabile osservata) può essere inglobato a tutti gli effetti nel termine di errore.

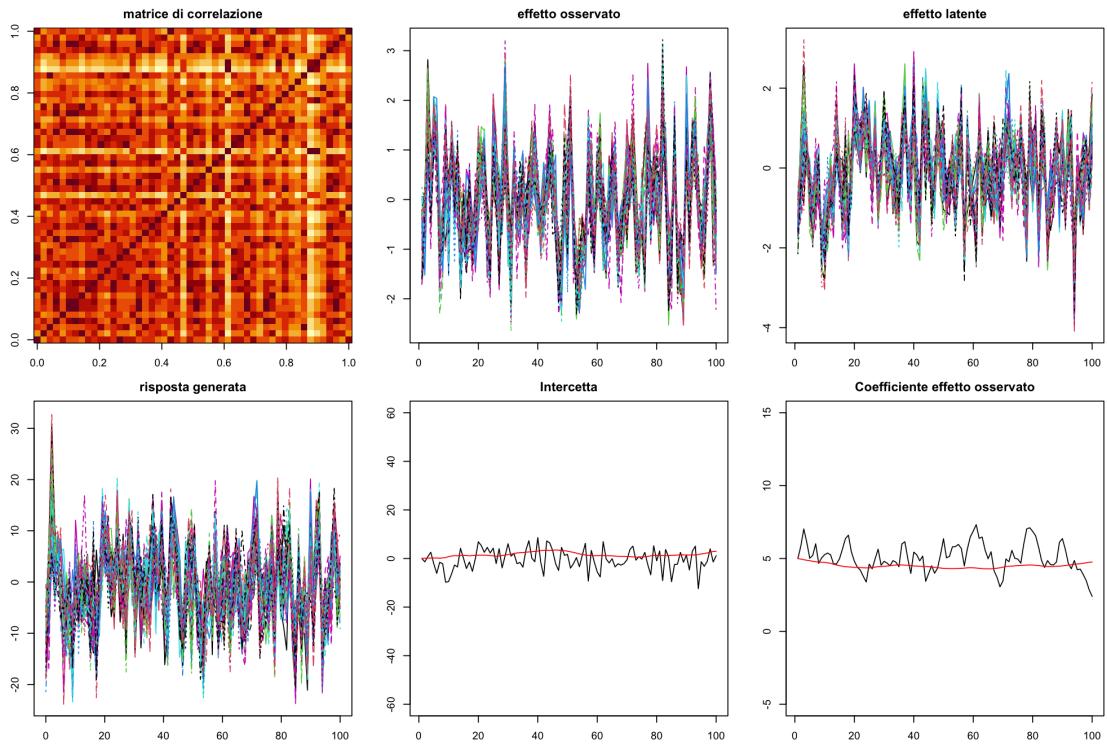


Figura A.1: Risultati primo scenario

A.2 Fattore latente correlato con la variabile osservata

Il secondo scenario implementa la medesima generazione della risposta dello scenario precedente, ma gli effetti latenti ed osservati sono ora correlati. In particolare, provengono ancora entrambi da normali multivariate a media nulla con la stessa struttura di correlazione tra città, sono ancora incorrelati se si considerano tempi diversi, ma sono correlati tra loro a livello simultaneo (con una correlazione pari a circa 0.84 all'interno della medesima città).

Lo scenario è riassunto nella figura A.2

Anche qui, la stima della correlazione dell'errore tramite correlazione della variabile osservata migliora il risultato delle stime del modello state-space. In particolare, per quanto riguarda il coefficiente dell'effetto osservato, il modello con matrice di covarianza diagonale assorbe una parte dell'effetto latente. Imporre la struttura non diagonale alla matrice di covarianza porta ad un miglioramento in termini la riduzione del bias (e di varianza), in quanto il modello riesce a non assorbire del tutto l'effetto latente nel coefficiente legato all'effetto osservato (ciò avviene perchè questi sono correlati). Per gli obiettivi interpretativi della nostra analisi, si preferisce che l'effetto delle variabili osservate non catturi (per quanto possibile) l'effetto di fattori latenti ad esse correlati. A livello previsivo questa scelta non è del tutto ovvia, in quanto se la correlazione fosse generalizzabile ad altri anni

Per altri scenari di simulazione, che esplorano la presenza di fattori latenti e osservati correlati temporalmente e caratterizzati da trend simili (quindi non più a media nulla), si veda il file R allegato.

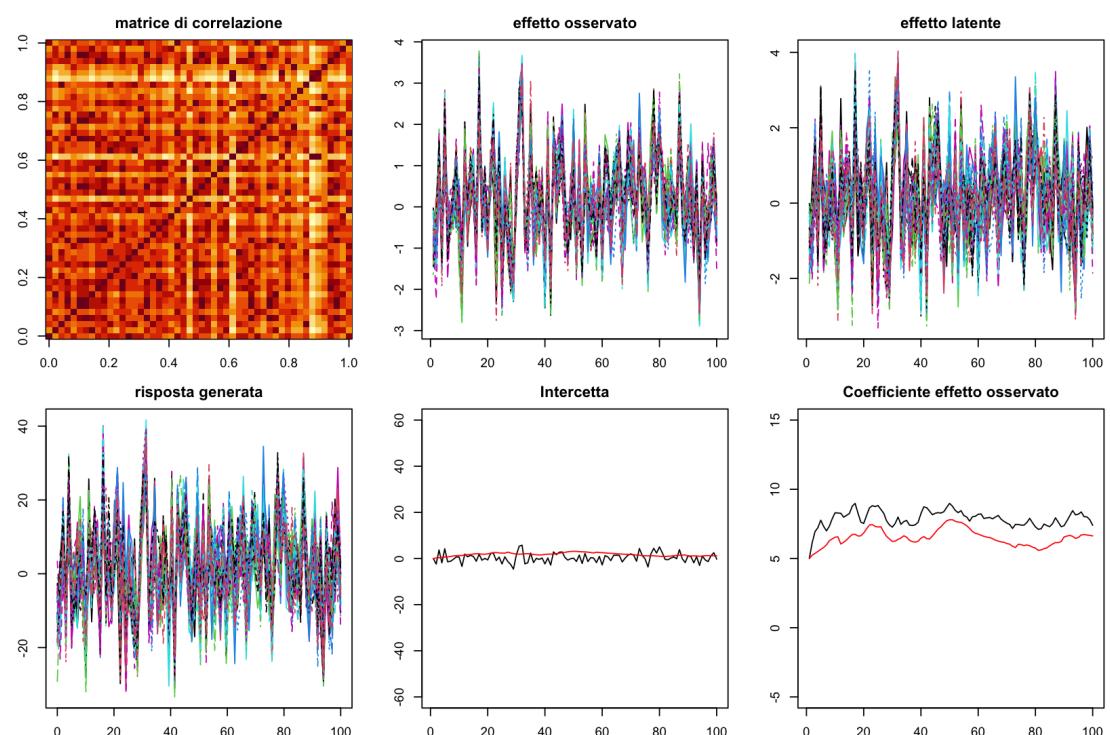


Figura A.2: Risultati secondo scenario

Bibliografia

Mayukh Bhattacharyya, Sayan Nag, and Udita Ghosh. Deciphering environmental air pollution with large scale city data. *arXiv preprint arXiv:2109.04572*, 2021.

Marta Blangiardo and Michela Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

Virgilio Gómez-Rubio. *Bayesian inference with INLA*. Chapman and Hall/CRC, 2020.