



**Università degli Studi di Bergamo**

---

DIP. DI INGEGNERIA GESTIONALE, DELL'INFORMAZIONE E DELLA PRODUZIONE  
Corso di Laurea Magistrale in Ingegneria Informatica  
L-32 — Classe delle Lauree in Ingegneria dell'Informazione (D.M. 270/04)

## **Un modello statistico funzionale per la stima di potenziali spazio-temporali in presenza di interazione tra le misure**

Relatore

**Prof. Francesco FINAZZI**

Candidati

**Lorenzo LEONI**

Matricola 1053379

**Nicola ZAMBELLI**

Matricola 1053015



# Indice

<b>Introduzione</b>	<b>1</b>
<b>1 Il modello f-HDGM e l'algoritmo EM</b>	<b>3</b>
1.1 Cenni sull'analisi funzionale dei dati . . . . .	3
1.1.1 Differenza tra modelli parametrici e modelli funzionali . . . . .	3
1.1.2 Basi di Fourier . . . . .	4
1.2 Il modello f-HDGM . . . . .	5
1.2.1 Equazioni del modello . . . . .	5
1.2.2 Errore di misura . . . . .	6
1.2.3 Componente deterministica . . . . .	6
1.2.4 Componente latente . . . . .	7
1.2.5 Modellazione della correlazione spaziale tramite un processo gaussiano . . . . .	7
1.2.6 Parametri da stimare . . . . .	8
1.3 Cenni sull'algoritmo EM . . . . .	8
1.3.1 Derivazione del passo E . . . . .	10
1.3.2 Derivazione del passo M . . . . .	12
<b>2 Il concetto di geo-potenziale condizionato e il modello fp-HDGM</b>	<b>13</b>
2.1 Il geo-potenziale condizionato . . . . .	13
2.1.1 Differenza tra geo-potenziale e geo-potenziale condizionato . .	15
2.2 Il modello fp-HDGM . . . . .	15
2.2.1 Equazioni del modello . . . . .	16
2.2.2 Parametri da stimare . . . . .	16
2.2.3 Simulazione di una mappa di geo-potenziale . . . . .	17
<b>3 Stima EM del modello fp-HDGM</b>	<b>21</b>
3.1 La funzione di verosimiglianza . . . . .	21
3.1.1 Rappresentazione matriciale del modello fp-HDGM . . . . .	21
3.1.2 Distribuzioni delle variabili casuali . . . . .	22
3.1.3 Derivazione della funzione di verosimiglianza . . . . .	22

## *Indice*

3.2 Il valore atteso condizionato . . . . .	24
3.2.1 Derivazione di $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$ . . . . .	25
3.2.2 Espressione della matrice $\Omega_t$ . . . . .	25
<b>4 Caso di studio</b>	<b>27</b>
4.1 Stato dell'arte . . . . .	27
4.2 Ricerca e acquisizione dati . . . . .	29
4.3 Descrizione del dataset . . . . .	30
4.3.1 Numero di ritiri . . . . .	30
4.3.2 Variabili meteorologiche . . . . .	31
4.3.3 Variabili spaziali . . . . .	33
4.3.4 Variabili dummy . . . . .	34
4.4 Metodologia . . . . .	35
4.5 Analisi dei risultati . . . . .	37
4.5.1 Scelta del parametro $\rho$ tramite la cross-validation . . . . .	37
4.5.2 Scelta delle covariate $\beta$ . . . . .	37
4.5.3 Validazione del modello definitivo mediante la LOOCV . . . . .	38
4.5.4 Previsione spaziale utilizzando il kriging . . . . .	39
4.6 Discussione sui risultati . . . . .	40
<b>5 Conclusioni</b>	<b>43</b>
<b>6 Sitografia</b>	<b>45</b>
6.1 Caso di studio . . . . .	45
<b>7 Appendice</b>	<b>47</b>
<b>Bibliografia</b>	<b>49</b>

# Elenco delle figure

1.1	Confronto tra basi di Fourier con $n = 2$ e tra diverse combinazioni lineari . . . . .	4
1.2	Esempio di combinazione lineare di basi di Fourier con $n = 2$ , i cui coefficienti sono diversi da 1 . . . . .	5
1.3	Simulazione di un processo spaziale gaussiano su un piano cartesiano al variare di $\lambda$ . . . . .	9
1.4	Confronto tra la vera funzione di verosimiglianza $L(\theta)$ e la sua approssimazione in $\theta_n$ . . . . .	11
2.1	Collocazione spaziale dei punti di misura della simulazione. . . . .	17
2.2	Previsione spaziale tramite Kriging di un processo fp-HDGM al variare di $\rho$ . . . . .	19
4.1	Punto di ritiro delle biciclette e veduta del quartiere Jersey . . . . .	27
4.2	Confronto tra l'andamento orario delle spline di Fourier per i coefficienti $\hat{\beta}_{Work}$ (a) e $\hat{\beta}_{Home}$ (b) . . . . .	29
4.3	Distribuzione spaziale del numero medio di ritiri giornaliero presso le stazioni al variare della stagione, anno 2020 . . . . .	31
4.4	Andamento giornaliero e orario del numero medio di noleggi al variare della stagione . . . . .	32
4.5	Confronto tra il numero medio di noleggi giornaliero, la temperatura percepita e la piovosità . . . . .	33
4.6	Mappe delle distanze dei punti di interscambio dalla stazione ferroviaria più vicina e della densità abitativa nei pressi dei punti di ritiro . . . . .	34
4.7	Confronto tra il numero medio di prelievi giornaliero, il periodo di lockdown, i weekend e le festività federali . . . . .	35
4.8	Andamento del $MSE$ in cross-validation con e senza arrotondamento al variare del parametro $\rho$ . . . . .	37
4.9	Spline di Fourier per i coefficienti $\beta$ e rispettivi intervalli di confidenza	38
4.10	Mappa della distribuzione del $RMSE_s$ . . . . .	39

*Elenco delle figure*

4.11 Mappa del volume del numero di ritiri di biciclette previsti nei mesi di giugno e luglio, raggruppati per fasce orarie . . . . .	41
7.1 Istogrammi e box-plot delle variabili meteorologiche, parte 1 . . . . .	47
7.2 Istogrammi e box-plot delle variabili meteorologiche, parte 2 . . . . .	48
7.3 Confronto tra il numero medio di noleggi giornaliero, la visibilità orizzontale e la velocità del vento . . . . .	48

# Elenco delle tabelle

4.1	Statistiche principali riguardanti le variabili meteorologiche . . . . .	32
4.2	Statistiche principali riguardanti le variabili spaziali . . . . . . . . .	34
4.3	Statistiche principali riguardanti il $RMSE_s$ con e senza arrotondamento	39



# Introduzione

I modelli spazio-temporali sono un fondamentale strumento concettuale utilizzato in una vasta gamma di discipline scientifiche per comprendere e analizzare fenomeni che si sviluppano nello spazio e nel tempo. Questi sistemi statistici forniscono un framework indispensabile per la descrizione e la predizione di eventi fisici, sociali, economici e naturali, permettendo agli studiosi di esplorare le relazioni tra le varie variabili in contesti complessi.

L'importanza di tali modelli nella vita quotidiana è evidente in molteplici contesti. Ad esempio, nel campo delle scienze ambientali, consentono di monitorare e modellare i cambiamenti climatici, le variazioni negli ecosistemi e la diffusione di inquinanti. Nell'ambito della pianificazione urbana, invece, sono utilizzati per ottimizzare la distribuzione delle risorse e dei servizi, migliorando la qualità della vita nelle città. Inoltre, nel settore dei trasporti, contribuiscono a ottimizzare le reti di trasporto e a gestire il traffico in tempo reale, riducendo congestioni e tempi di percorrenza.

Uno dei modelli spazio-temporali proposti in letteratura è il *Functional Hidden Dynamic Geostatistical Model (f-HDGM)* (Wang et al., 2021). Esso, oltre a essere un modello funzionale, è anche *state-space*; una variabile latente consente di rappresentare e modellare fenomeni non direttamente osservabili, ma che influenzano il comportamento del sistema. Si ipotizzi, per esempio, di voler descrivere la concentrazione di anidride carbonica rilevata nel tempo da una stazione di misura in funzione dell'umidità relativa. Probabilmente il loro legame cambia a seconda dell'ora del giorno, un effetto che, inoltre, potrebbe mutare da un giorno all'altro; infatti, non è detto che l'umidità relativa descriva la concentrazione dell'inquinante in oggetto sempre nello stesso modo, sia in estate che in inverno, in tutti i punti spaziali. Per modellare questa variabilità spazio-temporiale, non direttamente osservabile dai dati, subentra la componente latente.

Uno dei limiti del modello f-HDGM è l'impossibilità di tener conto dell'interazione che potrebbe esistere tra punti di misura. Si prenda come esemplificazione il potenziale di mercato spaziale; esso rappresenta il volume di vendite previsto quando un determinato prodotto viene commercializzato in un'area di scambio. Le vendite

## *Introduzione*

sono previste essere elevate se un negozio viene aperto in una posizione caratterizzata da un alto potenziale di mercato spaziale, mentre si prevede una loro riduzione ove il potenziale mercato risulta essere basso. Ad esempio, se si ipotizzasse di aprire un nuovo punto vendita nelle vicinanze di altri negozi concorrenti ugualmente attrattivi<sup>1</sup>, allora essi si spartirebbero la domanda in quell'area; ciò si riflette in un basso potenziale. Al fine di ottenere una stima riguardo al volume degli scambi commerciali, la valutazione del potenziale di mercato spaziale emerge come un aspetto cruciale.

La reciproca influenza tra i negozi, dove il volume delle vendite di ciascun punto è influenzato dalla presenza degli altri, rende imprescindibile considerare tale interazione al fine di stimare correttamente il potenziale di mercato spaziale. Da tale necessità deriva l'obiettivo di questo lavoro: estendere il modello f-HDGM affinché tenga conto anche di quest'aspetto.

Dopo aver introdotto nel primo capitolo il concetto di analisi funzionale e di stima Expectation-Maximization (EM), nel secondo e nel terzo si entra in medias res illustrando prima l'estensione del modello e poi le formule di stima dei suoi parametri. Nel quarto capitolo, invece, le nozioni teoriche vengono applicate a un caso di studio riguardante il fenomeno del bike sharing in una metropoli statunitense. Infine, il quinto capitolo chiude il lavoro traendo le considerazioni finali e illustrando i possibili sviluppi futuri.

---

<sup>1</sup>la decisione da parte del cliente di scegliere un negozio piuttosto che un altro è oggettiva, ossia non è influenzata da aspetti personali. Di conseguenza, si assume che l'unico fattore decisionale sia la distanza.

# 1 Introduzione all'analisi funzionale di dati spazio-tempo tramite il modello f-HDGM e cenni sull'algoritmo

## Expectation-Maximization

In questo primo capitolo, dopo aver presentato il concetto di *Functional Data Analysis (FDA)*, viene illustrato il modello spazio-temporale oggetto dell'estensione presentata in questo studio, ossia il *Functional Hidden Dynamic Geostatistical Model (f-HDGM)*. Infine, vengono fornite alcune nozioni per comprendere il funzionamento e, in primis, l'idea su cui si basa l'*algoritmo Expectation-Maximization (EM)*.

### 1.1 Cenni sull'analisi funzionale dei dati

#### 1.1.1 Differenza tra modelli parametrici e modelli funzionali

Fornito un campione di dati  $D = (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$  e un modello  $y = f(x, \boldsymbol{\theta}_0)$ , l'obiettivo della statistica parametrica è quello di stimare i parametri  $\boldsymbol{\theta}_0$  andando a minimizzare una funzione di costo  $J(\boldsymbol{\theta})$ , ossia individuare la combinazione di valori stimati dei parametri  $\hat{\boldsymbol{\theta}}_N$  tale che  $\hat{\boldsymbol{\theta}}_N = \arg \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, D)$ . La funzione di costo cambia a seconda del modello da identificare e della tipologia di stima che viene utilizzata, per esempio *Ordinary Least Squares (OLS)* o *Maximum Likelihood Estimation (MLE)*. Inoltre, non sempre è possibile risolvere il problema di minimo in forma chiusa<sup>1</sup>; spesse volte è necessario ricorrere a tecniche di ottimizzazione numerica, in particolare quando la funzione  $J(\boldsymbol{\theta})$  non è convessa.

---

<sup>1</sup>impiego di un'espressione matematica, esente da variabili libere, per calcolare il valore del parametro incognito. La ricerca iterativa dell'ottimo tramite un algoritmo non è necessaria.

## 1 Il modello $f$ -HDGM e l'algoritmo EM

Nei modelli funzionali, invece, l'oggetto della stima non è  $\boldsymbol{\theta}$ , ma la funzione  $f$  continua. Le spline sono una delle classi di funzioni più utilizzate in FDA; la possibilità di regolare la *smoothness* rende loro facilmente generalizzabili.

### 1.1.2 Basi di Fourier

Le basi di Fourier sono una classe di funzioni che viene impiegata per descrivere segnali periodici. Il modello che il toolbox FDA per MATLAB (Ramsay and Dalzell, 1991) implementa è:

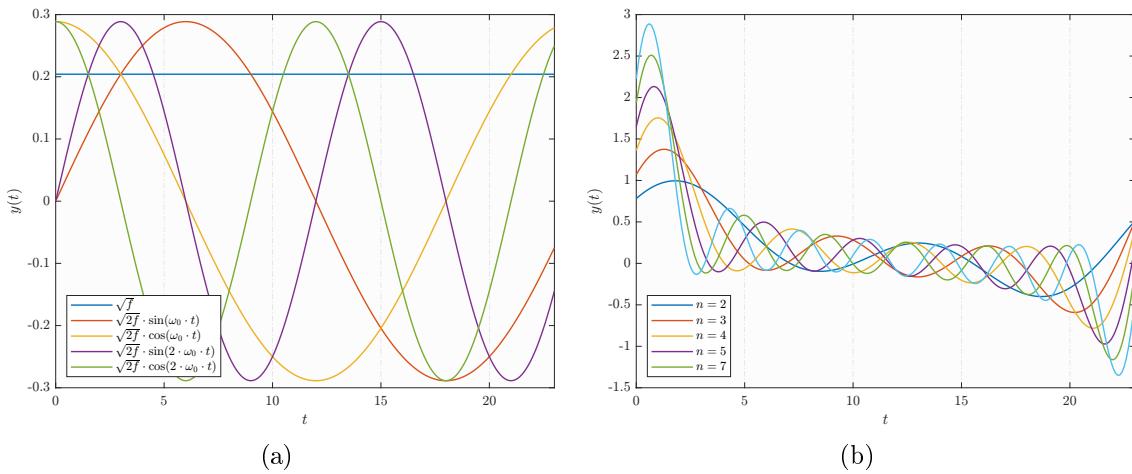
$$y(t) = \sqrt{2\nu} \cdot w(t);$$

$$w(t) = \frac{a_0}{\sqrt{2}} + \mathbf{a}_1 \cdot \mathbf{k}_1(t)^\top + \cdots + \mathbf{a}_i \cdot \mathbf{k}_i(t)^\top + \cdots + \mathbf{a}_n \cdot \mathbf{k}_n(t)^\top;$$

$$\mathbf{a}_i \in \mathbb{R}^2 = \begin{bmatrix} a_{i,\sin} & a_{i,\cos} \end{bmatrix};$$

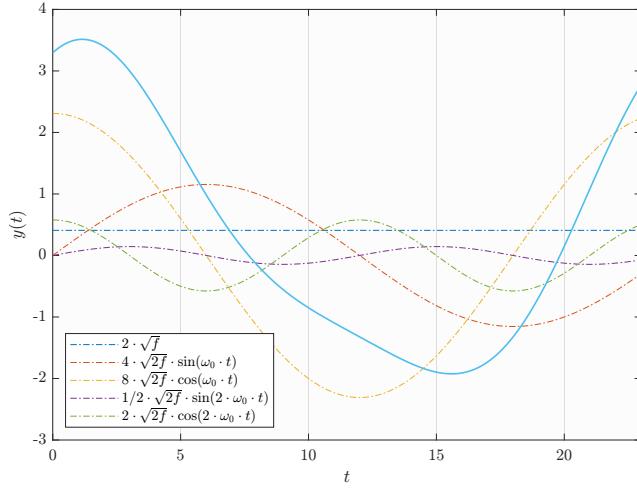
$$\mathbf{k}_i(t) \in \mathbb{R}^2 = \begin{bmatrix} \sin(i \cdot \omega_0 \cdot t) & \cos(i \cdot \omega_0 \cdot t) \end{bmatrix}.$$

$\nu$  è la frequenza,  $n$  indica il numero di armoniche<sup>2</sup>,  $\omega_0 = 2\pi \cdot \nu$  rappresenta la pulsazione dell'armonica fondamentale, mentre  $a_0, a_1, \dots, a_n$  sono i coefficienti combinatori. Oltre alla parte costante, ciascuna funzione seno e coseno costituisce una base. Una volta scelto il valore di  $n$  a seconda del livello di complessità desiderato, l'obiettivo dell'analisi funzionale è quello di calcolare i coefficienti. Nelle Figure 1.1 e 1.2 sono riportati degli esempi sia di basi di Fourier sia di alcune loro combinazioni lineari.



**Figura 1.1:** confronto tra basi di Fourier con  $n = 2$  (a) e tra diverse combinazioni lineari (b) i cui coefficienti sono unitari. Da notare come all'aumentare di  $n$  la funzione risultante tenda al livello basso di un'onda quadra.

<sup>2</sup>il numero di basi è  $n \cdot 2 + 1$ .



**Figura 1.2:** esempio di combinazione lineare di basi di Fourier con  $n = 2$  (in azzurro), i cui coefficienti sono diversi da 1. Nello specifico:  $a_0 = 2$ ,  $a_{1,\sin} = 4$ ,  $a_{1,\cos} = 8$ ,  $a_{2,\sin} = 1/2$  e  $a_{2,\cos} = 2$ .

## 1.2 Il modello f-HDGM

Il Functional Hidden Dynamic Geostatistical Model è stato illustrato in Wang et al. (2021), un paper finalizzato a presentare, oltre al modello, anche D-STEM v2, un software il cui scopo è stimare tramite l'algoritmo EM una serie di modelli spazio-temporali. Esiste anche una versione multi-variata non funzionale del modello in questione (Calculli et al., 2015).

### 1.2.1 Equazioni del modello

Sia  $\mathbf{s} = (s_{lon}, s_{lat})^\top$  un generico punto spaziale sulla sfera  $\mathbb{S}^2$  e  $t \in \mathbb{N}$  un indice temporale discreto. Si assume che la funzione di interesse  $u(\mathbf{s}, l, t)$ , con dominio  $\mathcal{L} = [l_1, l_q] \subset \mathbb{R}$ , possa essere osservata in ogni  $(\mathbf{s}, t)$  ed  $l \in \mathcal{L}$  attraverso delle misurazioni con incertezza  $y(\mathbf{s}, l, t)$  che seguono il seguente modello statistico:

$$y(\mathbf{s}, l, t) = u(\mathbf{s}, l, t) + \epsilon(\mathbf{s}, l, t); \quad (1.1)$$

$$u(\mathbf{s}, l, t) = \mathbf{x}(\mathbf{s}, l, t)^\top \cdot \boldsymbol{\beta}(l) + \Phi_z(l)^\top \cdot \mathbf{z}(\mathbf{s}, t); \quad (1.2)$$

$$\mathbf{z}(\mathbf{s}, t) = G \cdot \mathbf{z}(\mathbf{s}, t - 1) + \boldsymbol{\eta}(\mathbf{s}, t). \quad (1.3)$$

Prima di entrare in medias res, è bene introdurre la nomenclatura che successivamente verrà utilizzata per definire nel modello le dimensioni degli oggetti matriciali e vettoriali:

- $n$  indica il numero di posizioni spaziali (o punti di misura)  $\mathbf{s} \in \mathbb{S}^2$  prese in

esame;

- $q$  rappresenta il numero massimo di osservazioni disponibili per il dominio (funzionale)  $\mathcal{L}$ ;
- $T$  indica il numero di intervalli di tempo, ad esempio i giorni. In assenza di dati mancanti, il singolo elemento del campione è costituito da  $n \cdot q = N$  osservazioni (per ognuna delle  $n$  posizioni spaziali sono note tutte le  $q$  rilevazioni della variabile  $y$ );
- $b$  rappresenta il numero di variabili esplicative, ovvero  $\text{card}(\boldsymbol{\beta}(l)) = b$ ;
- $n_\epsilon$ ,  $n_\beta$  ed  $n_z$  indicano il numero di funzioni base utilizzate per modellare rispettivamente  $\sigma_\epsilon(l)$ , ogni  $\beta(l)$  e  $\Phi(l)$  per  $\mathbf{z}(\mathbf{s}, t)$ .

### 1.2.2 Errore di misura

Nell'equazione 1.1,  $\epsilon(\mathbf{s}, l, t) \in \mathbb{R}$  è una variabile causale con distribuzione gaussiana avente media nulla ( $\mu_\epsilon = 0$ ), indipendente sia nello spazio sia nel tempo. La sua varianza  $\sigma_\epsilon(l)$  è *eteroschedastica*<sup>3</sup> nel dominio  $\mathcal{L}$ ; essa è modellata nel modo seguente:

$$\log \sigma_\epsilon(l) = \Phi(l)^\top \cdot \mathbf{c}_\epsilon.$$

$\Phi_z(l)$  è un vettore contenente le  $n_\epsilon$  funzioni base valutate in  $l$ , mentre  $\mathbf{c}_\epsilon$  contiene i rispettivi  $n_\epsilon$  coefficienti combinatori da stimare. La finalità della modellazione funzionale della varianza consiste nel far variare il livello di incertezza nel dominio funzionale.

### 1.2.3 Componente deterministica

Nell'equazione 1.2,  $\mathbf{x}(\mathbf{s}, l, t)$  è un vettore di dimensione  $b$  di variabili esplicative, i cui coefficienti  $\boldsymbol{\beta}(l) = (\beta_1(l), \dots, \beta_b(l))^\top$  sono così espressi:

$$\beta_j(l) = \Phi_\beta(l)^\top \cdot \mathbf{c}_{\beta,j}.$$

$\Phi_\beta(l)$  è un vettore contenente le  $n_\beta$  basi valutate in  $l$ , mentre  $\mathbf{c}_{\beta,j}$  contiene i rispettivi  $n_\beta$  coefficienti da stimare della  $j$ -esima variabile esplicativa. Piuttosto che

---

<sup>3</sup>dato un campione di variabili casuali, al suo interno esistono delle sotto-popolazioni che hanno diverse varianze.

considerare un effetto globale tempo-invariante, l'utilizzo della modellazione funzionale consente di far variare il peso attribuito a ciascun regressore nel dominio  $\mathcal{L}$ .

### 1.2.4 Componente latente

Nell'equazione 1.3,  $z(\mathbf{s}, t)$  è un vettore contenente i valori assunti dalle  $n_z$  variabili latenti spazio-temporali al tempo  $t$ ; esse evolvono seguendo una dinamica markoviana descritta da  $G$ , una matrice di transizione diagonale<sup>4</sup> di ordine  $n_z$ . Invece,  $\boldsymbol{\eta}(\mathbf{s}, t)$  è il vettore delle  $n_z$  innovazioni che si ottengono da un processo gaussiano multi-variato, indipendente nel tempo ma correlato nello spazio mediante la seguente matrice di covarianza funzionale:

$$\Gamma(\mathbf{s}, \mathbf{s}'; \boldsymbol{\lambda}) = \text{diag}(v_1 \cdot \rho(\mathbf{s}, \mathbf{s}'; \lambda_1), \dots, v_{n_z} \cdot \rho(\mathbf{s}, \mathbf{s}'; \lambda_{n_z})).$$

$\mathbf{v}$  è un vettore contenente le  $n_z$  varianze, mentre  $\rho(\mathbf{s}, \mathbf{s}'; \lambda_j)$  è una funzione di correlazione spaziale idonea per i punti spaziali  $\mathbf{s}, \mathbf{s}' \in \mathbb{S}^2$ , modellata tramite gli  $n_z$  parametri del vettore  $\boldsymbol{\lambda}$ . Per concludere:

- il vettore  $\mathbf{v}$  è la diagonale della matrice diagonale  $V$  di ordine  $n_z$ . Infatti, si assume che le  $n_z$  variabili latenti siano tra di loro indipendenti (nullità dei termini extra-diagonale), ossia  $\eta_1(\mathbf{s}, t) \perp \eta_2(\mathbf{s}, t) \perp \dots \perp \eta_{n_z}(\mathbf{s}, t)$ ;
- un esempio di funzione di correlazione spaziale  $\rho$  è quella esponenziale, ovvero  $\rho(\mathbf{s}, \mathbf{s}'; \lambda_j) = \exp(-\frac{\|\mathbf{s} - \mathbf{s}'\|}{\lambda_j})$ .

### 1.2.5 Modellazione della correlazione spaziale tramite un processo gaussiano

In un contesto spaziale, è possibile interpretare un processo gaussiano come un insieme (o campo) di variabili casuali che mostrano correlazioni spaziali tra di loro. Si ipotizzi di trovarsi in un piano cartesiano regolare di dimensioni  $100 \times 100$  ( $n = 10^4$ ) e che  $n_z = 2$ , il quale corrisponde al numero di processi (tra di loro indipendenti se  $V$  è diagonale). Il processo può essere modellato tramite la normale multi-variata:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \sim \mathcal{N}_N \left\{ \boldsymbol{\mu} = [\mathbf{0}]_N, \Gamma = \begin{bmatrix} v_1 \cdot I_n \cdot [e^{-\frac{D}{\lambda_1}}]_{n \times n} & [\mathbf{0}]_{n \times n} \\ [\mathbf{0}]_{n \times n} & v_2 \cdot I_n \cdot [e^{-\frac{D}{\lambda_2}}]_{n \times n} \end{bmatrix}_{N \times N} \right\}.$$

---

<sup>4</sup>si assume che la cross-covarianza sia nulla.

## 1 Il modello f-HDGM e l'algoritmo EM

$D \in \mathbb{R}^{n \times n}$  è la matrice delle distanze,  $N = n_z \cdot n = 2 \cdot 10^4$  rappresenta il numero totale di variabili aleatorie.

I coefficienti  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^\top$  descrivono la velocità con la quale, fissato  $\mathbf{s}$ , le v.c. collocate nelle altre posizioni spaziali  $\mathbf{s}'$ , tendono ad assumere un valore diverso da quello assunto dal processo gaussiano in  $\mathbf{s}$  all'aumentare della distanza  $\|\mathbf{s} - \mathbf{s}'\|$ . Nella Figura 1.3 è possibile notare come la variabilità del processo  $\boldsymbol{\eta}_1$  tenda a diminuire all'aumentare di  $\lambda_1$  a causa della maggiore correlazione esistente tra punti vicini, un comportamento osservabile anche nei valori assunti dalla densità di probabilità congiunta della normale bi-variata ottenuta scegliendo due variabili causali.

### 1.2.6 Parametri da stimare

In conclusione, viene riportato il vettore  $\boldsymbol{\theta}$  contenente gli  $n_\epsilon + n_\beta \cdot b + 3 \cdot n_z$  parametri da stimare:

$$\boldsymbol{\theta} = (\mathbf{c}_\epsilon^\top, \mathbf{c}_\beta^\top, \mathbf{g}^\top, \mathbf{v}^\top, \boldsymbol{\lambda}^\top)^\top. \quad (1.4)$$

Alcuni di essi possono essere calcolati in forma chiusa, altri richiedono l'ottimizzazione numerica.

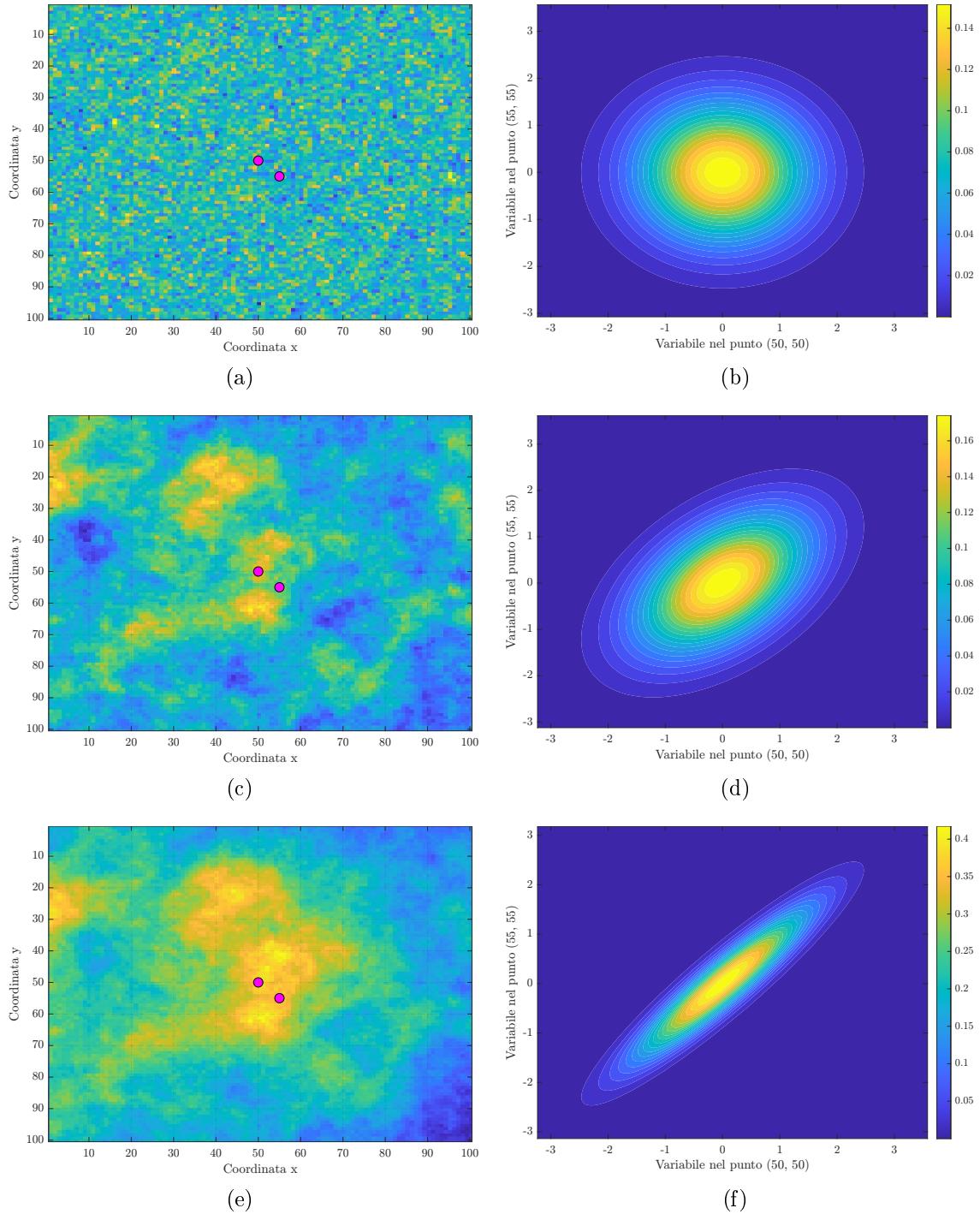
Inoltre, è necessario ricostruire la componente latente  $\mathbf{z}(\mathbf{s}, t)$ , operazione complessa che può essere eseguita integrando il filtro di Kalman (e smoother) nel passo E dell'algoritmo EM (Shumway et al., 2000).

## 1.3 Cenni sull'algoritmo EM

L'algoritmo EM è una tecnica di ottimizzazione iterativa per determinare la stima a massima verosimiglianza **in presenza di dati mancanti o nascosti (variabili latenti)**. Nella stima Maximum Likelihood (ML) si desidera determinare i parametri del modello per i quali i dati osservati risultano essere i più verosimili (Borman, 2012).

Ogni iterazione dell'algoritmo EM consiste in due passi:

- **passo E:** a partire dalla stima corrente dei parametri  $\boldsymbol{\theta}_n$  e dai dati disponibili  $\mathbf{y}$ , quelli mancanti vengono prima stimati e poi impiegati per determinare il *valore atteso condizionato*  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$ , una semplificazione della stima corrente della funzione di verosimiglianza  $l(\boldsymbol{\theta} | \boldsymbol{\theta}_n)$ ;
- **passo M:**  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$  viene ottimizzata per determinare  $\boldsymbol{\theta}_{n+1}$ , assumendo che i dati mancanti siano noti. Le stime di questi, ottenute precedentemente nel passo E, sono utilizzate al posto dei dati mancanti effettivi.



**Figura 1.3:** simulazione del processo spaziale gaussiano  $\eta_1$  su un piano cartesiano regolare e rispettiva densità di probabilità congiunta della normale bi-variata riferita ai punti  $(50, 50)$  e  $(55, 55)$  con  $\lambda_1 = 0.5$  (a, b),  $10$  (c, d) e  $100$  (e, f).

## 1 Il modello f-HDGM e l'algoritmo EM

La convergenza dell'algoritmo è garantita dall'aumento della verosimiglianza a ogni iterazione (Borman, 2012).

### 1.3.1 Derivazione del passo E

Sia  $\mathbf{y}$  un vettore di dati aleatori appartenenti a una famiglia di modelli parametrizzata. Si desidera trovare il valore delle incognite  $\boldsymbol{\theta}$  tale che la verosimiglianza  $L(\boldsymbol{\theta}) = P(\mathbf{y}|\boldsymbol{\theta})$  sia massima. L'algoritmo EM garantisce che  $L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_n)$ ; applicando il logaritmo naturale<sup>5</sup> si ottiene:

$$\ln(P(\mathbf{y}|\boldsymbol{\theta}) - P(\mathbf{y}|\boldsymbol{\theta}_n)) > 0. \quad (1.5)$$

Tale risultato deriva dal fatto che  $L(\boldsymbol{\theta})$  si può dimostrare essere strettamente positiva. Tuttavia, i dati osservati  $\mathbf{y}$  non dipendono soltanto dai parametri da stimare, ma anche dai valori assunti dalle componenti latenti  $\mathbf{z}$ . Infatti, la probabilità totale  $P(\mathbf{y}|\boldsymbol{\theta})$  può essere riscritta in termini di  $\mathbf{z}$  come:

$$P(\mathbf{y}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} P(\mathbf{y}|\boldsymbol{\theta}, \mathbf{z}) \cdot P(\mathbf{z}|\boldsymbol{\theta}).$$

Di conseguenza, l'equazione 1.5 diventa:

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_n) = \ln \sum_{\mathbf{z}} P(\mathbf{y}|\boldsymbol{\theta}, \mathbf{z}) \cdot P(\mathbf{z}|\boldsymbol{\theta}) - \ln P(\mathbf{y}|\boldsymbol{\theta}_n). \quad (1.6)$$

A partire dall'introduzione del termine  $P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n)$  nell'equazione 1.6, è possibile eseguire la seguente derivazione:

$$\begin{aligned} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_n) &= -\ln P(\mathbf{y}|\boldsymbol{\theta}_n) + \ln \sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) \cdot P(\mathbf{z}|\boldsymbol{\theta}) \cdot \frac{P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n)}{P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n)} \\ &= -\ln P(\mathbf{y}|\boldsymbol{\theta}_n) + \ln \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n) \cdot \frac{P(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) \cdot P(\mathbf{z}|\boldsymbol{\theta})}{P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n)} \\ &= \ln \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n) \cdot \frac{P(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) \cdot P(\mathbf{z}|\boldsymbol{\theta})}{P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n) \cdot P(\mathbf{y}|\boldsymbol{\theta}_n)} \\ &\geq \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n) \cdot \ln \frac{P(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) \cdot P(\mathbf{z}|\boldsymbol{\theta})}{P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n) \cdot P(\mathbf{y}|\boldsymbol{\theta}_n)} = \Delta(\boldsymbol{\theta}|\boldsymbol{\theta}_n); \end{aligned} \quad (1.7)$$

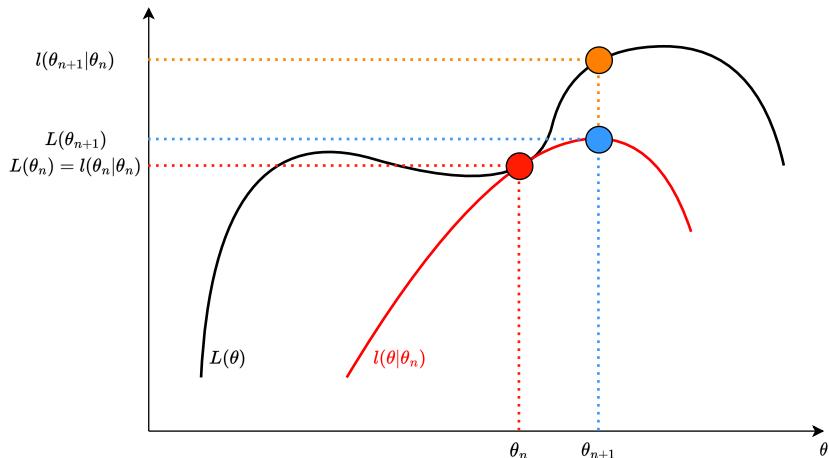
---

<sup>5</sup>semplificazione che spesse volte viene adottata per trasformare il prodotto di densità di probabilità che caratterizza  $L$  (assumendo l'indipendenza tra i campioni) in una somma.

per il corollario della disegualanza di Jensen<sup>6</sup>. Pertanto:

$$L(\boldsymbol{\theta}) \geq L(\boldsymbol{\theta}_n) + \Delta(\boldsymbol{\theta}|\boldsymbol{\theta}_n) = l(\boldsymbol{\theta}|\boldsymbol{\theta}_n). \quad (1.8)$$

La Figura 1.4 materializza il risultato appena ottenuto: l'approssimazione di  $L(\boldsymbol{\theta})$  in un intorno di  $\boldsymbol{\theta}_n$  non sovrastima la vera funzione di verosimiglianza, quindi ottimizzare l'approssimazione comporta la massimizzazione implicita anche di  $L(\boldsymbol{\theta})$ . Si ricorda, inoltre, che  $\boldsymbol{\theta}_n = \arg \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\boldsymbol{\theta}_{n-1})$ .



**Figura 1.4:** confronto tra la vera funzione di verosimiglianza  $L(\theta)$  e la sua approssimazione in un intorno di  $\theta_n$ . Da notare che lo stimatore  $l(\theta|\theta_n)$  non sovrastima la funzione reale.

---

<sup>6</sup>sia  $f$  una funzione concava (ln lo è) e  $\lambda_1, \dots, \lambda_n \in [0, 1]$  tali che  $\sum_{i=1}^n \lambda_i = 1$ , allora  $f(\sum_{i=1}^n \lambda_i \cdot y_i) \geq \sum_{i=1}^n \lambda_i \cdot f(y_i)$ . Il teorema è applicabile perché  $\sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n) = 1$ .

### 1.3.2 Derivazione del passo M

Dopo aver stimato nel passo E l'approssimazione  $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$  della vera funzione di verosimiglianza  $L(\boldsymbol{\theta})$ , è necessario ottimizzarla per determinare  $\boldsymbol{\theta}_{n+1}$ . Nello specifico:

$$\begin{aligned}
 \boldsymbol{\theta}_{n+1} &= \arg \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\boldsymbol{\theta}_n) \\
 &= \arg \max_{\boldsymbol{\theta}} \left\{ L(\boldsymbol{\theta}_n) + \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n) \cdot \ln \frac{P(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) \cdot P(\mathbf{z}|\boldsymbol{\theta})}{P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n) \cdot P(\mathbf{y}|\boldsymbol{\theta}_n)} \right\} \\
 &= \arg \max_{\boldsymbol{\theta}} \left\{ \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n) \cdot \ln P(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) \cdot P(\mathbf{z}|\boldsymbol{\theta}) \right\} \\
 &= \arg \max_{\boldsymbol{\theta}} \left\{ \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n) \cdot \ln P(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) \right\} \\
 &= \arg \max_{\boldsymbol{\theta}} \{ \mathbb{E}_{\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n} \ln P(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) \} \\
 &= \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n).
 \end{aligned} \tag{1.9}$$

I termini  $L(\boldsymbol{\theta}_n)$ ,  $P(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_n)$  e  $P(\mathbf{y}|\boldsymbol{\theta}_n)$  vengono rimossi poiché costanti rispetto a  $\boldsymbol{\theta}$ , quindi non influenzano la ricerca dell'ottimo, mentre al prodotto  $P(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) \cdot P(\mathbf{z}|\boldsymbol{\theta})$  viene applicato il teorema di Bayes.

Attraverso la riscrittura di  $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$  in  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$  mediante la derivazione 1.9, si può comprendere l'idea che sta alla base dell'algoritmo EM: la log-verosimiglianza viene iterativamente corretta tramite la stima corrente della densità di probabilità delle variabili latenti  $\mathbf{z}$ , un'approssimazione ottenuta nel passo E a partire dai dati osservati  $\mathbf{y}$  e dalla stima corrente dei coefficienti  $\boldsymbol{\theta}_n$ .

# 2 Il concetto di geo-potenziale condizionato e definizione del modello fp-HDGM

In questo capitolo, viene introdotto il concetto di geo-potenziale condizionato, delineando le sue distinzioni rispetto al geo-potenziale. Successivamente, si procede con la presentazione del *Functional Potential Hidden Dynamic Geostatistical Model (fp-HDGM)*, il quale rappresenta un'estensione del modello f-HDGM per l'analisi dei dati spazio-temporali con interazioni tra i punti di osservazione. Esso rappresenta il contributo innovativo di questo lavoro. In questo contesto, vengono esplicite le equazioni del modello e i parametri da stimare. Infine, viene condotta la simulazione di un modello fp-HDGM al fine di concretizzare i concetti trattati nel capitolo.

## 2.1 Il geo-potenziale condizionato

Un modello geo-statistico per l'analisi di dati nello spazio per i quali non esiste solo una correlazione spaziale, ma anche un'interazione tra punti di misura, è il *Geostatistical Potential Model (GPM)* (Finazzi, 2013).

Sia  $\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_N\}$  l'insieme dei punti appartenenti alla regione  $D \subset \mathbb{S}^2$  nei quali è stata misurata la variabile d'interesse. La funzione di interazione è così descritta:

$$p(\mathbf{s}|\mathcal{S}) = u(\mathbf{s}) \cdot h_\rho(\mathbf{s}|\mathcal{S}); \quad (2.1)$$

$$h_\rho(\mathbf{s}|\mathcal{S}) = \left( 1 + \sum_{\mathbf{s}' \in \mathcal{S}} f_\rho(\mathbf{s}, \mathbf{s}') \right)^{-1}; \quad (2.2)$$

$$f_\rho(\mathbf{s}, \mathbf{s}') = f_\rho(\|\mathbf{s} - \mathbf{s}'\|) = \exp\left(-\frac{\|\mathbf{s} - \mathbf{s}'\|}{\rho}\right). \quad (2.3)$$

Nello specifico:

- $u(\mathbf{s})$  è il potenziale di un campo casuale spaziale definito in una regione  $D \subset \mathbb{S}^2$ ;

## 2 Il concetto di geo-potenziale condizionato e il modello fp-HDGM

- $f_\rho(\|\mathbf{s} - \mathbf{s}'\|)$  è una funzione non negativa definita da  $\mathbb{S}^2 \times \mathbb{S}^2$  a  $\mathbb{R}^+$ ;
- $\|\mathbf{s} - \mathbf{s}'\|$  è la distanza geodetica tra due punti nello spazio;
- $\rho \in \mathbb{R}^+$  è il parametro che descrive la forza dell'interazione tra i punti di misura nella regione spaziale  $D$ .

Per comprendere appieno il comportamento della funzione in esame, esploriamo i suoi valori limite. Innanzitutto, si consideri il parametro  $\rho$  fissata la distanza geodetica  $\|\mathbf{s} - \mathbf{s}'\|$ ; si osserva che:

$$\lim_{\rho \rightarrow 0} f_\rho(\|\mathbf{s} - \mathbf{s}'\|) = 0 \Rightarrow \lim_{\rho \rightarrow 0} p(\mathbf{s}|\mathcal{S}) = u(\mathbf{s}); \quad (2.4)$$

$$\lim_{\rho \rightarrow \infty} f_\rho(\|\mathbf{s} - \mathbf{s}'\|) = 1 \Rightarrow \lim_{\rho \rightarrow \infty} p(\mathbf{s}|\mathcal{S}) = u(\mathbf{s}) \cdot \frac{1}{N+1}. \quad (2.5)$$

Quando  $\rho$  tende ad assumere il valore minimo del dominio di definizione, espressione 2.4, la funzione di interazione spaziale  $h_\rho(\mathbf{s}|\mathcal{S})$  raggiunge il suo massimo assoluto, ovvero 1, lasciando così il termine  $u(\mathbf{s})$  inalterato. Viceversa, quando  $\rho$  tende a infinito, espressione 2.5,  $h_\rho(\mathbf{s}|\mathcal{S})$  tende a decrescere fino al suo limite inferiore pari a  $\frac{1}{N+1}$ , penalizzando significativamente il potenziale puro  $u(\mathbf{s})$ .

Si consideri ora  $\|\mathbf{s} - \mathbf{s}'\|$  fissato  $\rho$ ; si nota che:

$$\lim_{\|\mathbf{s} - \mathbf{s}'\| \rightarrow 0} f_\rho(\|\mathbf{s} - \mathbf{s}'\|) = 1 \Rightarrow \lim_{\|\mathbf{s} - \mathbf{s}'\| \rightarrow 0} p(\mathbf{s}|\mathcal{S}) = u(\mathbf{s}) \cdot \frac{1}{N+1}; \quad (2.6)$$

$$\lim_{\|\mathbf{s} - \mathbf{s}'\| \rightarrow \infty} f_\rho(\|\mathbf{s} - \mathbf{s}'\|) = 0 \Rightarrow \lim_{\|\mathbf{s} - \mathbf{s}'\| \rightarrow \infty} p(\mathbf{s}|\mathcal{S}) = u(\mathbf{s}). \quad (2.7)$$

Quando  $\|\mathbf{s} - \mathbf{s}'\|$  tende a 0, espressione 2.6, allora si hanno  $N$  stazioni di misura situate nella medesima posizione spaziale; il risultato è l'omogenea ripartizione del potenziale  $u(\mathbf{s})$  disponibile in  $\mathbf{s}$ . Invece, quando  $\|\mathbf{s} - \mathbf{s}'\|$  tende a infinito, espressione 2.7, allora i punti di misura sono tra loro indipendenti (assenza di interazione spaziale), di conseguenza il potenziale  $u(\mathbf{s})$  rimane invariato.

Queste due condizioni estreme forniscono un'importante comprensione del comportamento della funzione di interazione in relazione al parametro  $\rho$  e alla distanza geodetica  $\|\mathbf{s} - \mathbf{s}'\|$ ; esse possono avere profonde implicazioni per l'analisi e l'interpretazione del fenomeno studiato.

### 2.1.1 Differenza tra geo-potenziale e geo-potenziale condizionato

Dopo aver presentato la funzione d’interazione, si è in grado di delineare la seguente divergenza:

- il **geo-potenziale**  $u(\mathbf{s})$  è definito come il valore atteso osservato quando  $y(\mathbf{s})$  viene osservato nella posizione spaziale  $\mathbf{s} \in D$ . Esso non è condizionato dalle rilevazioni svolte presso gli altri punti di misura;
- il **geo-potenziale condizionato**  $p(\mathbf{s}|\mathcal{S})$ , invece, è il valore atteso osservato quando  $y(\mathbf{s}|\mathcal{S})$  viene misurato nella posizione spaziale  $\mathbf{s} \in D$ , dato che viene contemporaneamente rilevato nella collezione di posizioni  $\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_N\}$ , con  $\mathbf{s}_i \in D, N \geq 1$ . Esso è il potenziale spaziale tenente conto anche dell’ influenza reciproca esistente tra i punti di osservazione.

Si prenda come esempio il potenziale di mercato spaziale in una rete di attività commerciali anticipato nel capitolo introduttivo. Il geo-potenziale potrebbe dare delle stime promettenti nell’intorno di un generico punto di misurazione  $\mathbf{s}_i \in \mathcal{S}$ , poiché non tiene in considerazione dell’interazione delle altre stazioni di misurazione  $\mathbf{s}_1, \dots, \mathbf{s}_N$ , in particolare di  $\mathbf{s}_i$ . Viceversa, il geo-potenziale condizionato ottiene valori inferiori nell’intorno di  $\mathbf{s}_i$ , in quanto rappresenta il potenziale che sarebbe osservato con l’aggiunta di una nuova stazione di misura se quest’ultima fosse posizionata nelle prossimità di  $\mathbf{s}_i$  (Finazzi, 2013).

In parole povere, l’apertura di un nuovo punto vendita nelle vicinanze di un altro preesistente comporterà probabilmente una riduzione delle vendite per entrambi, a causa della concorrenza tra attività commerciali. Tale osservazione è fondamentale per comprendere il motivo per il quale il modello f-HDGM dev’essere esteso affinché tenga in considerazione il geo-potenziale condizionato e, di conseguenza, possa essere utilizzato per modellare contesti in cui c’è concorrenza commerciale (e non solo).

## 2.2 Il modello fp-HDGM

Il Functional Potential Hidden Dynamic Geostatistical Model vuole espandere l’applicazione del modello f-HDGM all’analisi di dati spazio-temporali in cui è presente interazione tra i punti di misurazione. Questo modello parte dal presupposto che il fenomeno oggetto di studio abbia, anche se minima, una componente di interazione tra i punti di osservazione.

## 2 Il concetto di geo-potenziale condizionato e il modello fp-HDGM

L'output del processo, indicato come  $y(\mathbf{s}, l, t | \mathcal{S})$  rappresenta la rilevazione al tempo  $t$ , all'istante istante  $l \in \mathcal{L}$  e nella posizione spaziale  $\mathbf{s} \in D$ , simultaneamente misurato nell'insieme di punti  $\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_N\}$ , con  $\mathbf{s}_i \in D, N \geq 1$ .

### 2.2.1 Equazioni del modello

Il nuovo modello proposto viene così definito dalla seguente gerarchia di equazioni:

$$y(\mathbf{s}, l, t | \mathcal{S}) = w(\mathbf{s}, l, t) \cdot h_\rho(\mathbf{s} | \mathcal{S}); \quad (2.8)$$

$$w(\mathbf{s}, l, t) = u(\mathbf{s}, l, t) + \epsilon(\mathbf{s}, l, t); \quad (2.9)$$

$$u(\mathbf{s}, l, t) = \mathbf{x}(\mathbf{s}, l, t)^\top \cdot \boldsymbol{\beta}(l) + \Phi_z(l)^\top \cdot \mathbf{z}(\mathbf{s}, t); \quad (2.10)$$

$$\mathbf{z}(\mathbf{s}, t) = G \cdot \mathbf{z}(\mathbf{s}, t - 1) + \boldsymbol{\eta}(\mathbf{s}, t). \quad (2.11)$$

Un'osservazione interessante emerge esaminando il modello al variare del parametro  $\rho$ , secondo le considerazioni fatte a riguardo nelle equazioni 2.4 e 2.5. Nel dettaglio, si nota che:

$$\lim_{\rho \rightarrow 0} y(\mathbf{s}, l, t | \mathcal{S}) = w(\mathbf{s}, l, t). \quad (2.12)$$

Questa osservazione porta a vedere il modello f-HDGM come un caso particolare del modello fp-HDGM, ovvero con  $\rho = 0$ .

Per fare chiarezza,  $u(\mathbf{s}, l, t)$  è il potenziale di un campo casuale spaziale misurato al tempo  $t$ , all'istante  $l$  e nel punto  $\mathbf{s}$ . Se si volesse calcolare il potenziale condizionato per fare previsione spaziale, allora  $p(\mathbf{s}, l, t | \mathcal{S}) = u(\mathbf{s}, l, t) \cdot h_\rho(\mathbf{s} | \mathcal{S})$ . Per eseguire questa operazione si utilizza il **Kriging**, ovvero un metodo di regressione impiegato in geostatistica che, minimizzando l'errore quadratico medio, permette di interpolare una grandezza nello spazio. Questa tecnica consente di prevedere il valore del potenziale laddove la variabile in uscita non è stata misurata (Wang et al., 2021). A tale scopo, il Kriging ricorre al filtro di Kalman per ricostruire la componente latente nei punti  $\mathbf{s} \in D \setminus \mathcal{S}$ , determina quella deterministica combinando linearmente le spline  $\boldsymbol{\beta}$  precedentemente stimate con i valori delle covariate  $\mathbf{x}$  note in questi punti e applica infine la funzione di interazione  $h_\rho$  per prevedere il potenziale condizionato.

### 2.2.2 Parametri da stimare

In definitiva, si presenta il vettore dei parametri  $\boldsymbol{\theta}$  da stimare di dimensione  $n_\epsilon + n_\beta \cdot b + 3 \cdot n_z + 1$ :

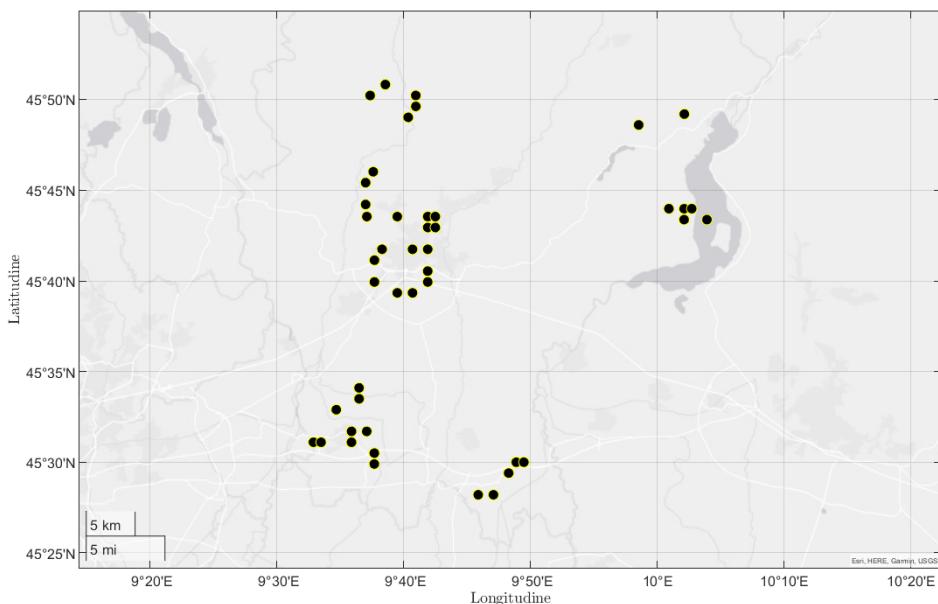
$$\boldsymbol{\theta} = (\mathbf{c}_\epsilon^\top, \mathbf{c}_\beta^\top, \mathbf{g}^\top, \mathbf{v}^\top, \boldsymbol{\lambda}^\top, \rho)^\top. \quad (2.13)$$

Analogamente a come precedentemente discusso per il modello f-HDGM, non tutti i parametri possono essere stimati attraverso formule chiuse, ma necessitano di essere risolti mediante ottimizzazione numerica.

**Differenza tra correlazione spaziale e interazione spaziale** Nonostante  $\lambda$  e  $\rho$  siano parametri dalla stessa funzione<sup>1</sup>, essi differiscono nel loro scopo fondamentale.  $\lambda$  è progettato per modellare l'interazione naturale tra tutti i punti nello spazio, ovvero la correlazione spaziale, mentre  $\rho$  insiste sull'interazione tra le stazioni di misura presenti nello spazio. In termini chiari, il parametro  $\lambda$  condiziona entrambi i geo-potenziali, viceversa  $\rho$  influenza solo il geo-potenziale condizionato.

### 2.2.3 Simulazione di una mappa di geo-potenziale

Al fine di concretizzare le nozioni riportate nel capitolo, con l'utilizzo del software MATLAB, è stata compiuta una simulazione della realizzazione di un processo descritto secondo un fp-HDGM. La simulazione è stata condotta nell'area della provincia di Bergamo. Per garantire la pertinenza del caso di studio con un contesto realistico, sono stati distribuiti i punti di misurazione in cluster localizzati in alcune aree urbane della provincia (Figura 2.1). Di seguito è riportata la numerosità dei



**Figura 2.1:** collocazione spaziale dei punti di misura della simulazione. Sono stati definiti 7 cluster di stazioni di misurazione nei comuni di Bergamo, Treviglio, San Pellegrino Terme, Sovere, Sorisole, Antegnate e Parzanica di numerosità pari a 15, 10, 5, 2, 3, 5 e 5, rispettivamente.

---

<sup>1</sup>  $f_i(\|\mathbf{s} - \mathbf{s}'\|) = \exp(-\frac{\|\mathbf{s} - \mathbf{s}'\|}{i})$ .

## 2 Il concetto di geo-potenziale condizionato e il modello fp-HDGM

parametri scelti per la simulazione:

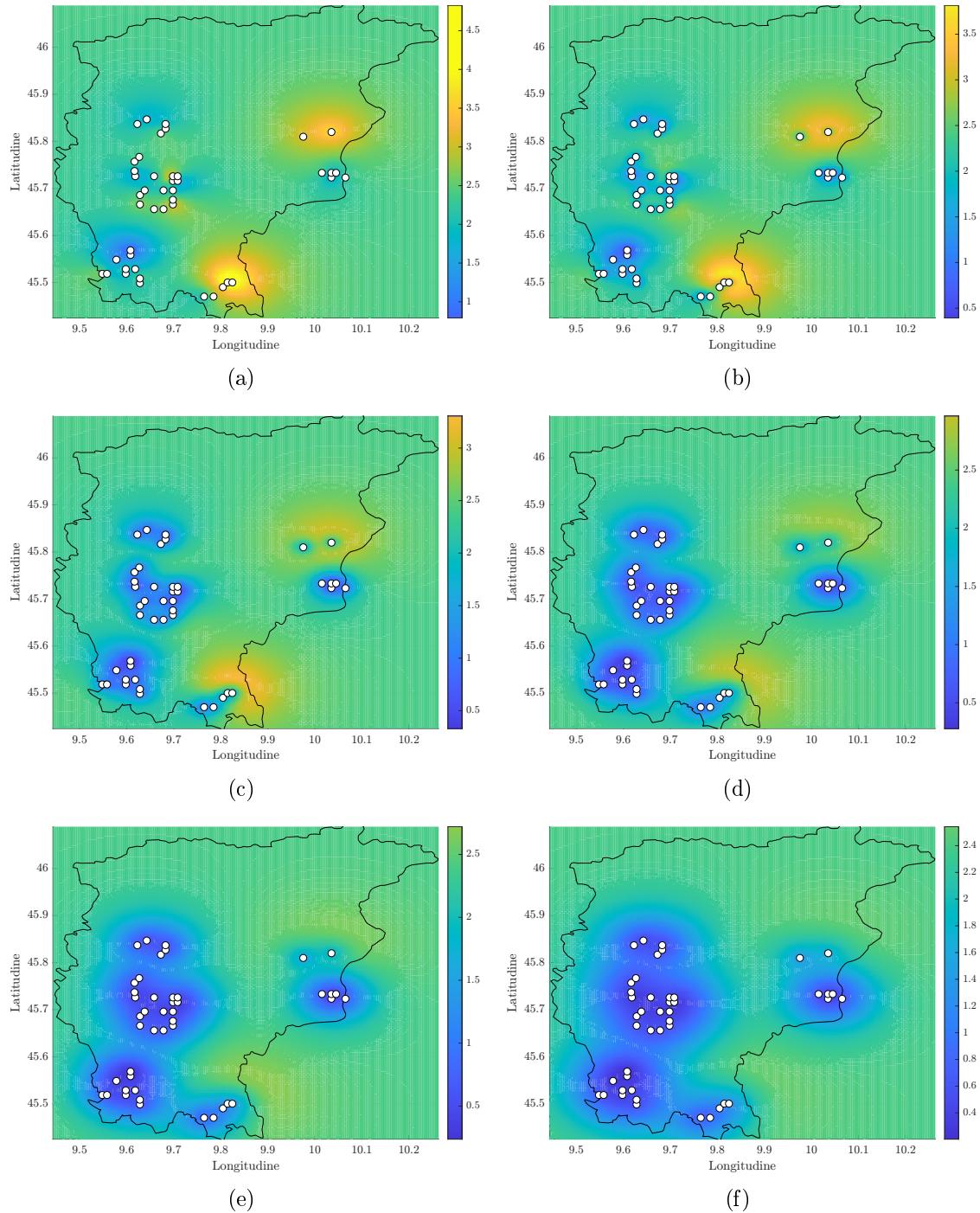
- numero di stazioni di misura,  $n = 45$ ;
- numero di covariante globali,  $b = 1$ ;
- dimensione del dominio funzionale ad alta frequenza,  $q = 24$ ,  $\mathcal{L} = [0, 23]$ ;
- numero di campioni in bassa frequenza,  $T = 365$ ;
- numero di funzioni base utilizzate per modellare  $\sigma_\epsilon(l)$ ,  $n_\epsilon = 5$ ;
- numero di funzioni base utilizzate per modellare  $\beta(l)$ ,  $n_\beta = 5$ ;
- numero di funzioni base utilizzate per modellare  $\Phi_z(l)$ ,  $n_z = 5$ .

Infine, si espongono i valori fissati del set di parametri:

- $\rho = 1100 \text{ m}$ ;
- $\mathbf{c}_\epsilon = \begin{bmatrix} 2 & -2 & -10 & 0.5 & 2 \end{bmatrix}$ ;
- $\mathbf{c}_\beta = \begin{bmatrix} 2 & 4 & 8 & 0.5 & 2 \end{bmatrix}$ ;
- $\boldsymbol{\lambda} = \begin{bmatrix} 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \end{bmatrix}$ ;
- $G = \text{diag}(0.26, 0.41, 0.6, 0.26, 0.6)$ ;
- $V = \text{diag}(8, 3, 2, 3, 4)$ .

Una volta eseguita la simulazione del processo, è stata applicata la tecnica di Kriging allo scopo di rappresentare la mappa del potenziale condizionato al variare del parametro  $\rho$ . Da sottolineare la circostanza con  $\rho = 0$ , ossia il caso in cui il geo-potenziale condizionato è pari al geo-potenziale.

Nella Figura 2.2 si osserva come il geo-potenziale condizionato diminuisca all'aumentare di  $\rho$ , specialmente nelle zone con una forte densità di stazioni di misurazione. Si prenda ora come ipotesi che le punti di misura possano rappresentare dei punti di vendita di un prodotto; valgono le considerazioni fatte riguardo al potenziale di mercato spaziale. Infatti, se si decidesse di aggiungere un punto di vendita nei pressi delle coordinate  $(45.75^\circ\text{N}, 9.70^\circ\text{W})$ , il geo-potenziale previsto in quest'area risulta essere promettente ma forviante in quanto vicino ad altri punti di vendita (Figura 2.2(a)). Con  $\rho \geq 500 \text{ m}$  (Figure 2.2(b), 2.2(c), 2.2(d) e 2.2(e)) è facile osservare che il geo-potenziale condizionato è previsto basso nell'area presa in esame, ovvero il punto di vendita verrebbe collocato in un'area caratterizzata da un basso potenziale di mercato spaziale.



**Figura 2.2:** previsione spaziale tramite Kriging di un processo fp-HDGM, fissati gli indici  $l$  e  $t$ , con  $\rho = 0 \text{ m}$  (a),  $500 \text{ m}$  (b),  $1000 \text{ m}$  (c),  $1500 \text{ m}$  (d),  $2000 \text{ m}$  (e) e  $2500 \text{ m}$  (f).



# 3 Stima EM del modello fp-HDGM

Dopo aver presentato nel capitolo precedente il modello fp-HDGM, in questo vengono illustrate le formule di stima. Innanzitutto viene derivata la funzione di verosimiglianza  $L(\boldsymbol{\theta})$ , poi la componente  $\Omega(t)$  del valore atteso condizionato  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$ ; essa è l'oggetto dell'ottimizzazione svolta nel passo M per stimare il nuovo parametro  $\rho$ . Per la nomenclatura si rimanda alla sezione 1.2.1.

## 3.1 La funzione di verosimiglianza

La stima dei parametri  $\boldsymbol{\theta}$  e dalla variabile latente spazio-temporale  $\mathbf{z}(\mathbf{s}, t)$  è basata sull'approccio a massima verosimiglianza (Wang et al., 2021).

### 3.1.1 Rappresentazione matriciale del modello fp-HDGM

Si misuri la variabile  $y(\mathbf{s}_i, l_j, t)$  per ogni valore  $l_j \in \mathcal{L}^1$  in uno specifico punto nello spazio  $\mathbf{s}_i \in \mathcal{S} = \{(s_{lon,1}, s_{lat,1}), \dots, (s_{lon,n}, s_{lat,n})\}$  e istante temporale  $t$ . Il vettore risultante è:

$$\mathbf{y}(\mathbf{s}_i, t) = [y(\mathbf{s}_i, l_1, t) \ \dots \ y(\mathbf{s}_i, l_j, t) \ \dots \ y(\mathbf{s}_i, l_q, t)]_{q \times 1}^\top.$$

Esso prende il nome di *profilo* osservato. Se si percorrono tutti i punti di misura  $\mathcal{S}$ , allora si costruisce la seguente matrice:

$$\mathbf{y}_t = [y(\mathbf{s}_1, t) \ \dots \ y(\mathbf{s}_k, t) \ \dots \ y(\mathbf{s}_n, t)]_{N \times 1}^\top;$$

dove  $N = n \cdot q$ . Applicando le equazioni alla struttura dati appena definita, si ottiene la seguente rappresentazione matriciale del modello fp-HDGM:

$$\mathbf{y}_t = H \cdot \boldsymbol{\omega}_t; \tag{3.1}$$

---

<sup>1</sup>per semplicità di notazione, si assume l'assenza di dati mancanti, ovvero  $q$  osservazioni per ogni  $\mathbf{s}_i$  e  $t$ .

### 3 Stima EM del modello fp-HDGM

$$\boldsymbol{\omega}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t \quad (3.2)$$

$$\boldsymbol{\mu}_t = \mathbf{X}_t \cdot \Phi_{\beta,t} \cdot \mathbf{c}_{\beta} + \Phi_{z,t} \cdot \mathbf{z}_t; \quad (3.3)$$

$$\mathbf{z}_t = \tilde{G} \cdot \mathbf{z}_{t-1} + \boldsymbol{\eta}_t. \quad (3.4)$$

$H \in \mathbb{R}^{N \times 1}$  è una matrice diagonale contenente i coefficienti  $h_i = \left(1 + \sum_{s \in S/s_i} e^{\frac{\|\mathbf{s}-\mathbf{s}_i\|}{\rho}}\right)^{-1}$  ripetuti  $q$  volte per ogni punto di misura,  $X_t \in \mathbb{R}^{N \times b}$  è la matrice delle covariate,  $\Phi_{\beta,t} \in \mathbb{R}^{b \times (b \cdot n_{\beta})}$  e  $\Phi_{z,t} \in \mathbb{R}^{N \times (n \cdot n_z)}$  contengono i valori delle basi rispettivamente per  $\beta$  e per  $\mathbf{z}$ , mentre  $\tilde{G} \in \mathbb{R}^{(n \cdot n_z) \times (n \cdot n_z)}$  è una matrice diagonale a blocchi costruita con  $G$ , ossia la matrice di transizione. Infine,  $\boldsymbol{\epsilon}_t \in \mathbb{R}^N$  ed  $\boldsymbol{\eta}_t \in \mathbb{R}^{n \cdot n_z}$  sono i vettori delle variabili casuali: il primo descrive il rumore sull'uscita, il secondo la correlazione spaziale.

#### 3.1.2 Distribuzioni delle variabili casuali

Pre-moltiplicando per  $H^{-1}$  ambo i termini dell'equazione 3.1, si ottiene:

$$H^{-1} \cdot \mathbf{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t.$$

$\boldsymbol{\epsilon}_t$  è un vettore di variabili aleatorie distribuite normalmente con media  $\boldsymbol{\mu}_{\epsilon} \in \mathbb{R}^N$  nulla e varianza  $\Sigma_{\epsilon} \in \mathbb{R}^{N \times N}$ , una matrice diagonale costruita utilizzando  $\sigma_{\epsilon}(l)$ , ossia la varianza funzionale. Di conseguenza:

$$H^{-1} \cdot \mathbf{y}_t \sim \mathcal{N}_N \{ \boldsymbol{\mu}_t, \Sigma_{\epsilon} \}.$$

Un discorso simile si può fare per l'equazione 3.4;  $\boldsymbol{\eta}_t$  è anch'esso un vettore di variabili casuali distribuite normalmente con media  $\boldsymbol{\mu}_{\eta} \in \mathbb{R}^{n \cdot n_z}$  nulla, tuttavia la sua matrice di varianze e covarianze  $\Sigma_{\eta} \in \mathbb{R}^{(n \cdot n_z) \times (n \cdot n_z)}$  è diagonale a blocchi, costruita utilizzando  $\Gamma_{\eta} \in \mathbb{R}^{n \times n}$ , ovvero la matrice di correlazione spaziale. Pertanto:

$$\mathbf{z}_t \sim \mathcal{N}_{n \cdot n_z} \left\{ \tilde{G} \cdot \mathbf{z}_{t-1}, \Gamma_{\eta} \right\}.$$

#### 3.1.3 Derivazione della funzione di verosimiglianza

Siano  $X$ ,  $Y$  e  $Z$  le matrici contenenti rispettivamente i valori assunti dalle covariate, dall'uscita e dalle componenti latenti. Per definizione, la funzione di verosimiglianza ha la seguente espressione:

$$L(\boldsymbol{\theta}; Y, Z, X) = L(\boldsymbol{\theta}; X) \cdot L(\boldsymbol{\theta}; Z|X) \cdot L(\boldsymbol{\theta}; Y|Z, X). \quad (3.5)$$

### 3.1 La funzione di verosimiglianza

La quantità  $L(\boldsymbol{\theta}; X)$  è unitaria perché le covariate sono deterministiche, mentre  $L(\boldsymbol{\theta}; Z|X) = L(\boldsymbol{\theta}; Z)$  poiché  $X$  e  $Z$  sono indipendenti tra loro. Quindi, l'equazione 3.5 diventa:

$$L(\boldsymbol{\theta}; Y, Z, X) = L(\boldsymbol{\theta}; Z) \cdot L(\boldsymbol{\theta}; Y|Z, X). \quad (3.6)$$

Nello specifico:

$$\begin{aligned} L(\boldsymbol{\theta}; Y|Z, X) &= \prod_{t=1}^T \left( |\Sigma_\epsilon|^{\frac{1}{2}} \cdot (2\pi)^{\frac{n \cdot q}{2}} \right)^{-1} \cdot e^{-\frac{1}{2} (\mathbf{H}^{-1} \mathbf{y}_t - \boldsymbol{\mu}_t)^\top \Sigma_\epsilon^{-1} (\mathbf{H}^{-1} \mathbf{y}_t - \boldsymbol{\mu}_t)} \\ &= \left( |\Sigma_\epsilon|^{\frac{1}{2}} \cdot (2\pi)^{\frac{n \cdot q}{2}} \right)^{-T} \cdot \prod_{t=1}^T e^{-\frac{1}{2} (\mathbf{H}^{-1} \mathbf{y}_t - \boldsymbol{\mu}_t)^\top \Sigma_\epsilon^{-1} (\mathbf{H}^{-1} \mathbf{y}_t - \boldsymbol{\mu}_t)}; \end{aligned} \quad (3.7)$$

e

$$\begin{aligned} L(\boldsymbol{\theta}; Z) &= L(\boldsymbol{\theta}; \mathbf{z}_0) \cdot L(\boldsymbol{\theta}; \mathbf{z}_1|\mathbf{z}_0) \cdots L(\boldsymbol{\theta}; \mathbf{z}_T|\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{T-1}) \\ &= L(\boldsymbol{\theta}; \mathbf{z}_0) \cdot L(\boldsymbol{\theta}; \mathbf{z}_1|\mathbf{z}_0) \cdots L(\boldsymbol{\theta}; \mathbf{z}_T|\mathbf{z}_{T-1}) \\ &= L(\boldsymbol{\theta}; \mathbf{z}_0) \cdot \prod_{t=1}^T L(\boldsymbol{\theta}; \mathbf{z}_t|\mathbf{z}_{t-1}); \end{aligned} \quad (3.8)$$

poiché l'autocorrelazione del processo gaussiano markoviano è a un passo ( $\tau = 1$ ). Da segnalare, inoltre, che l'operatore  $|\cdot|$  indica il determinante. Aumentando il livello di dettaglio si ottiene:

$$L(\boldsymbol{\theta}; \mathbf{z}_0) = \left( |\Sigma_0|^{\frac{1}{2}} \cdot (2\pi)^{\frac{n \cdot n_z}{2}} \right)^{-1} \cdot e^{-\frac{1}{2} (\mathbf{z}_0 - \boldsymbol{\mu}_0)^\top \Sigma_0^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_0)}; \quad (3.9)$$

e

$$\begin{aligned} \prod_{t=1}^T L(\boldsymbol{\theta}; \mathbf{z}_t|\mathbf{z}_{t-1}) &= \prod_{t=1}^T \left( |\Sigma_\eta|^{\frac{1}{2}} \cdot (2\pi)^{\frac{n \cdot n_z}{2}} \right)^{-1} \cdot e^{-\frac{1}{2} (\mathbf{z}_t - \tilde{\mathbf{G}} \mathbf{z}_{t-1})^\top \Sigma_\eta^{-1} (\mathbf{z}_t - \tilde{\mathbf{G}} \mathbf{z}_{t-1})} \\ &= \left( |\Sigma_\eta|^{\frac{1}{2}} \cdot (2\pi)^{\frac{n \cdot n_z}{2}} \right)^{-T} \cdot \prod_{t=1}^T e^{-\frac{1}{2} (\mathbf{z}_t - \tilde{\mathbf{G}} \mathbf{z}_{t-1})^\top \Sigma_\eta^{-1} (\mathbf{z}_t - \tilde{\mathbf{G}} \mathbf{z}_{t-1})}. \end{aligned} \quad (3.10)$$

Per rimuovere i prodotti e per predisporre la funzione di verosimiglianza alla minimizzazione, all'equazione 3.6 viene applicato il logaritmo naturale negato, ossia:

$$-2 \ln L(\boldsymbol{\theta}; Y, Z, X) = -2 \ln L(\boldsymbol{\theta}; Z) - 2 \ln L(\boldsymbol{\theta}; Y|Z, X); \quad (3.11)$$

### 3 Stima EM del modello fp-HDGM

con:

$$-2 \ln L(\boldsymbol{\theta}; Y|Z, X) = T \ln |\Sigma_\epsilon| + \sum_{t=1}^T + (H^{-1}\mathbf{y}_t - \boldsymbol{\mu}_t)^\top \Sigma_\epsilon^{-1} (H^{-1}\mathbf{y}_t - \boldsymbol{\mu}_t); \quad (3.12)$$

$$-2 \ln L(\boldsymbol{\theta}; \mathbf{z}_0) = T \ln |\Sigma_0| + (\mathbf{z}_0 - \boldsymbol{\mu}_0)^\top \Sigma_0^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_0); \quad (3.13)$$

e

$$-2 \ln \prod_{t=1}^T L(\boldsymbol{\theta}; \mathbf{z}_t | \mathbf{z}_{t-1}) = T \ln |\Sigma_\eta| + \sum_{t=1}^T (\mathbf{z}_t - \tilde{G}\mathbf{z}_{t-1})^\top \Sigma_\eta^{-1} (\mathbf{z}_t - \tilde{G}\mathbf{z}_{t-1}). \quad (3.14)$$

I termini  $T(n \cdot q) \ln 2\pi$ ,  $(n \cdot n_z) \ln 2\pi$  e  $T(n \cdot n_z) \ln 2\pi$  sono stati rimossi poiché sono costanti, quindi non influenzano la ricerca di  $\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} -2 \ln L(\boldsymbol{\theta}; Y, Z, X)$ .

Riassumendo ed esplicitando i singoli termini di  $\boldsymbol{\theta}$ , si ottiene:

$$\begin{aligned} -2 \ln L(\boldsymbol{\theta}; Y, Z, X) &= T \ln |\Sigma_\epsilon(\mathbf{c}_\epsilon)| \\ &+ \sum_{t=1}^T (H^{-1}(\rho) \cdot \mathbf{y}_t - \boldsymbol{\mu}_t(\mathbf{c}_\beta))^\top \Sigma_\epsilon^{-1} (H^{-1}(\rho) \cdot \mathbf{y}_t - \boldsymbol{\mu}_t(\mathbf{c}_\beta)) \\ &+ \ln |\Sigma_0| \\ &+ (\mathbf{z}_0 - \boldsymbol{\mu}_0)^\top \Sigma_0^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_0) \\ &+ T \ln |\Sigma_\eta(\mathbf{v}, \boldsymbol{\lambda})| \\ &+ \sum_{t=1}^T (\mathbf{z}_t - \tilde{G}(\mathbf{g}) \cdot \mathbf{z}_{t-1})^\top \Sigma_\eta^{-1}(\mathbf{v}, \boldsymbol{\lambda}) (\mathbf{z}_t - \tilde{G}(\mathbf{g}) \cdot \mathbf{z}_{t-1}). \end{aligned} \quad (3.15)$$

Il secondo addendo è quello che dev'essere minimizzato per determinare  $\rho$ .

## 3.2 Il valore atteso condizionato

Il valore atteso condizionato  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$  è l'oggetto della massimizzazione<sup>2</sup> compiuta dal passo M per determinare  $\boldsymbol{\theta}_n$ , ossia la stima all'iterazione  $n$  dei parametri da identificare. Il ricorso all'algoritmo EM è necessario poiché nel modello compaiono delle variabili latenti  $\mathbf{z}$  che devono essere ricostruite, di conseguenza non si può minimizzare direttamente la funzione 3.15.

---

<sup>2</sup>minimizzazione se viene applicato il logaritmo naturale negato alla funzione di verosimiglianza.

### 3.2.1 Derivazione di $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$

L'espressione del valore atteso condizionato è la seguente (Finazzi et al., 2013):

$$\begin{aligned}
 Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n) &= E_{\boldsymbol{\theta}_n} \left\{ T \ln |\Sigma_0| + \text{tr} \left( \Sigma_0^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_0) (\mathbf{z}_0 - \boldsymbol{\mu}_0)^\top \right) \middle| Y^{(1)} \right\} \\
 &\quad + E_{\boldsymbol{\theta}_n} \left\{ T \ln |\Sigma_\eta| + \text{tr} \left( \Sigma_\eta^{-1} \sum_{t=1}^T (\mathbf{z}_t - \tilde{G}\mathbf{z}_{t-1}) (\mathbf{z}_t - \tilde{G}\mathbf{z}_{t-1})^\top \right) \middle| Y^{(1)} \right\} \\
 &\quad + E_{\boldsymbol{\theta}_n} \left\{ T \ln |\Sigma_\epsilon| + \text{tr} \left( \Sigma_\epsilon^{-1} \sum_{t=1}^T \mathbf{e}_t \cdot \mathbf{e}_t^\top \right) \middle| Y^{(1)} \right\}; \\
 \end{aligned} \tag{3.16}$$

dove  $\mathbf{e}_t = H^{-1}(\rho) \cdot \mathbf{y}_t - \boldsymbol{\mu}_t$  e  $Y^{(1)}$  rappresenta le osservazioni disponibili<sup>3</sup>. Il terzo addendo dell'equazione 3.16 viene rinominato in  $m(\rho)$ ; nello specifico:

$$\begin{aligned}
 m(\rho) &= E_{\boldsymbol{\theta}_n} \left\{ T \ln |\Sigma_\epsilon| \middle| Y^{(1)} \right\} + E_{\boldsymbol{\theta}_n} \left\{ \text{tr} \left( \Sigma_\epsilon^{-1} \sum_{t=1}^T \mathbf{e}_t \cdot \mathbf{e}_t^\top \right) \middle| Y^{(1)} \right\} \\
 &= T \ln |\Sigma_\epsilon| + \text{tr} \left( E_{\boldsymbol{\theta}_n} \left\{ \Sigma_\epsilon^{-1} \sum_{t=1}^T \mathbf{e}_t \cdot \mathbf{e}_t^\top \middle| Y^{(1)} \right\} \right) \\
 &= T \ln |\Sigma_\epsilon| + \text{tr} \left( \Sigma_\epsilon^{-1} \sum_{t=1}^T E_{\boldsymbol{\theta}_n} \left\{ \mathbf{e}_t \cdot \mathbf{e}_t^\top \middle| Y^{(1)} \right\} \right). \\
 \end{aligned} \tag{3.17}$$

La quantità  $m(\rho)$  del valore atteso condizionato  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$  è l'unica che dipende da  $\rho$ , l'unica che dev'essere ottimizzata per stimare il nuovo parametro.

### 3.2.2 Espressione della matrice $\Omega_t$

Sia  $E_{\boldsymbol{\theta}_n} \left\{ \mathbf{e}_t \cdot \mathbf{e}_t^\top \middle| Y^{(1)} \right\} = \Omega_t \in \mathbb{R}^{(n \cdot q) \times (n \cdot q)}$ , allora (Shumway et al., 2000):

$$\begin{aligned}
 \Omega_t &= \begin{bmatrix} H^{(1)-1} \cdot \mathbf{y}_t^{(1)} - X_t^{(1)} \cdot \Phi_{\beta,t} \cdot \mathbf{c}_\beta - \Phi_{z,t} \cdot \mathbf{z}_t \\ R_{21,t} \cdot R_{11,t}^{-1} \cdot \mathbf{e}_t^{(1)} \end{bmatrix}_{(n \cdot q) \times 1} \cdot \begin{bmatrix} \mathbf{e}_t \\ R_{21,t} \cdot R_{11,t}^{-1} \cdot \mathbf{e}_t^{(1)} \end{bmatrix}_{(n \cdot q) \times 1}^\top \\
 &\quad + \begin{bmatrix} \Phi_{z,t}^{(1)} \\ R_{21,t} \cdot R_{11,t}^{-1} \cdot \Phi_{z,t}^{(1)} \end{bmatrix}_{(n \cdot q) \times (n \cdot n_z)} \cdot \Sigma_{\eta,t}^{(1)} \cdot \begin{bmatrix} \Phi_{z,t}^{(1)} \\ R_{21,t} \cdot R_{11,t}^{-1} \cdot \Phi_{z,t}^{(1)} \end{bmatrix}_{(n \cdot n_z) \times (n \cdot q)}^\top \\
 &\quad + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & R_{22,t} - R_{21,t} \cdot R_{11,t}^{-1} \cdot R_{11,t} \end{bmatrix}_{(n \cdot q) \times (n \cdot q)}. \\
 \end{aligned} \tag{3.18}$$

---

<sup>3</sup>si ricorda che l'algoritmo EM è in grado stimare  $\mathbf{y}_t$  nei punti e negli istanti in cui essa è ignota.

### 3 Stima EM del modello fp-HDGM

$$R_t = \Sigma_{\epsilon,t} = \begin{bmatrix} R_{11,t} & R_{12,t} \\ R_{21,t} & R_{22,t} \end{bmatrix}_{(n \cdot q) \times (n \cdot q)}$$

è la matrice di varianze e covarianze relativa al rumore sull'uscita;  $R_{11,t}$  rappresenta la componente di  $\Sigma_{\epsilon,t}$  relativa ai siti spaziali presso i quali si hanno delle osservazioni, mentre  $R_{22,t}$  quelli per i quali si hanno dati mancanti. Si assume che  $\epsilon_t$  sia indipendente e identicamente distribuito, ovvero  $R_{12,t} = R_{21,t} = \mathbf{0} \forall t$ ; quindi l'espressione 3.18 si semplifica e diventa:

$$\begin{aligned} \Omega_t &= \begin{bmatrix} \mathbf{e}_t^{(1)} \cdot \mathbf{e}_t^{(1)\top} + \Phi_{z,t}^{(1)} \cdot \Sigma_{\eta,t}^\top \cdot \Phi_{z,t}^{(1)\top} & \mathbf{0} \\ \mathbf{0} & R_{22} \end{bmatrix}_{(n \cdot q) \times (n \cdot q)} \\ &= \begin{bmatrix} \Omega_t^{(1)} & \mathbf{0} \\ \mathbf{0} & R_{22} \end{bmatrix}_{(n \cdot q) \times (n \cdot q)}. \end{aligned} \quad (3.19)$$

Da notare che sia  $R_{11}$  sia  $R_{22}$  non variano nel tempo  $t$  poiché si assume che  $\epsilon_t$  sia eteroschedastico solo in  $\mathcal{L}$  (assenza di autocorrelazione in  $t$ ).

# 4 Caso di studio

Dopo aver descritto il modello fp-HDGM da un punto di vista matematico e metodologico, in questo capitolo esso viene applicato al fenomeno del bike sharing. L’obiettivo iniziale del caso di studio è descrivere, sia nel tempo sia nello spazio, il numero di ritiri di biciclette presso una serie di stazioni (o punti di ritiro) dislocati nel quartiere Jersey della città di New York (Figura 4.1). Dopodiché, si vuole costruire una mappa di potenziale di mercato spaziale per capire dove conviene aprire una nuova stazione al fine di aumentare il numero di ritiri e il bacino d’utenza del servizio, tenendo conto dell’iterazione concorrenziale esistente tra le stazioni.

## 4.1 Stato dell’arte

Nelle città di tutto il mondo l’inquinamento atmosferico rappresenta una sfida sempre più urgente e complessa. Samantha Burgess, Deputy Director del Copernicus Climate Change Service, ha affermato che il 2023 non solo è stato l’anno più caldo mai registrato, ma è anche il primo in cui tutti i giorni le temperature hanno fatto registrare valori superiori di almeno 1 °C rispetto al periodo preindustriale [6.1].

L’aumento del traffico veicolare, in particolare, contribuisce in modo significativo alla diffusione di gas nocivi (es. CO<sub>2</sub> ed NO<sub>x</sub>) e particolato nell’aria (es. PM2.5



(a)



(b)

**Figura 4.1:** punto di ritiro delle biciclette gestito da Citi Bike (a) e veduta del quartiere Jersey, New York City (b).

#### *4 Caso di studio*

e PM10), compromettendo la salute pubblica e l'ambiente. In questo contesto, il bike sharing emerge come una soluzione innovativa e sostenibile per affrontare l'inquinamento urbano. Attraverso la condivisione delle biciclette, questo sistema offre un'alternativa efficace al trasporto privato a motore, riducendo le emissioni di gas serra, i consumi energetici e migliorando la qualità dell'aria nelle città (Zhang and Mi, 2018).

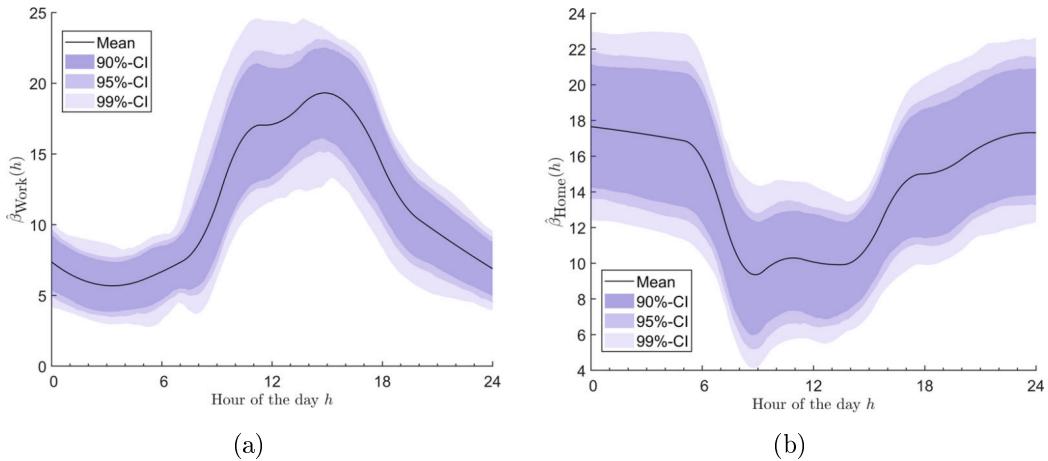
Per modellare il fenomeno del bike sharing le condizioni del tempo atmosferico risultano essere appropriate. Infatti, le variabili meteorologiche come pioggia, temperatura e vento, giocano un ruolo cruciale nel determinare la frequenza e la disponibilità delle biciclette per gli utenti, nonché la percezione stessa dell'attrattività del bike sharing come mezzo di trasporto. Inoltre, risulta interessante capire se la vicinanza di un punto di ritiro dalla più vicina fermata del treno o metropolitana incentiva l'utilizzo di quest'ultima piuttosto che della bicicletta, soprattutto in condizioni di maltempo (Gebhart and Noland, 2014).

In generale, anche gli aspetti urbanistici e demografici possono aiutare nella descrizione del fenomeno. Infatti, la densità abitativa, la disposizione delle infrastrutture ciclabili, la distribuzione della popolazione, la vicinanza ad alberghi e ai punti di interesse possono influenzare l'adozione della bicicletta a noleggio come mezzo di trasporto urbano (Li et al., 2022).

Infine, anche gli eventi straordinari possono avere un impatto significativo sul comportamento degli utenti e sulla dinamica del sistema. Uno di questi eventi straordinari è rappresentato dal lockdown imposto a causa della pandemia di COVID-19. Il lockdown ha comportato una serie di cambiamenti radicali nelle abitudini di spostamento delle persone, con effetti tangibili sull'utilizzo del bike sharing nelle metropoli di tutto il mondo. Pertanto, è essenziale comprendere come la chiusura obbligata abbia influenzato il fenomeno in oggetto, considerando sia gli aspetti legati alla riduzione del trasporto pubblico che quelli associati alla promozione di modalità di spostamento individuali e sicure (Jiao et al., 2022).

In questo contesto, l'impiego di un modello spazio-temporale funzionale si presenta come una soluzione promettente per catturare la dinamica complessa che governa il noleggio e scambio di biciclette. Tale famiglia di modelli consente di considerare non solo le variazioni spaziali dell'utilizzo delle biciclette all'interno di una città, ma anche come queste variazioni si evolvono nel tempo, consentendo una comprensione più approfondita dei pattern di utilizzo e dei fattori che li influenzano. Altresì, la modellazione funzionale permette di descrivere l'evoluzione oraria del numero di ritiri durante il giorno. Infatti, quest'ultimo non rimane costante nel corso della giornata, ma presenta un andamento periodico che raggiunge i propri massimi nelle

ore di punta. I trend periodici, inoltre, sono influenzati dalla tipologia di giorno, ovvero feriale o weekend (Figura 4.2), quindi tenere in considerazione quest'aspetto nella modellazione è fondamentale (Piter et al., 2022).



**Figura 4.2:** confronto tra l'andamento orario delle spline di Fourier per i coefficienti  $\hat{\beta}_{Work}$  (a) e  $\hat{\beta}_{Home}$  (b) (Piter et al., 2022). L'analisi è stata condotta sui dati del servizio di bike sharing della città di Helsinki, Finlandia.

## 4.2 Ricerca e acquisizione dei dati

Di seguito sono riportate le sorgenti dei dati utilizzati nel caso di studio:

- **dati riguardanti il bike sharing:** provengono dal sito web Kaggle [6.1] e contengono le informazioni sul noleggio di biciclette nel 2020 dell'azienda Citi Bike a Jersey City. Citi Bike è un sistema di bike sharing pubblico di proprietà privata che serve i distretti di New York City, nello specifico Bronx, Brooklyn, Manhattan e Queens, oltre a Jersey City;
- **dati meteorologici:** contengono le serie storiche delle variabili meteorologiche più significative per la città di New York nel 2020. Il provider di questi dati è Visual Crossing [6.1], un fornitore che offre una vasta gamma di soluzioni per l'accesso ai dati meteorologici storici e in tempo reale, nonché per la generazione di previsioni meteo;
- **dati inerenti le stazioni del treno/metropolitana:** dopo aver recuperato i nomi delle stazioni ferroviarie e della metropolitana dalla mappa dei mezzi pubblici newyorkesi [6.1], le loro coordinate sono state reperite tramite Google Maps;

- **dati demografici:** il fornitore è il SEDAC [6.1] (Socioeconomic Data and Applications Center), un data center che fa parte del programma Earth Observing System Data and Information System (EOSDIS) della NASA. La missione principale dell'ente è quella di fornire accesso a dati socioeconomici e ambientali globali, nonché a strumenti e risorse per facilitare la ricerca e la comprensione dei cambiamenti ambientali e delle dinamiche sociali.
- **dati riguardanti i giorni di festività e di lockdown:** per i primi è stato fatto riferimento al sito web OfficeHolidays [6.1], mentre per i secondi a Wikipedia [6.1].

Da sottolineare, infine, che sono state processate le sole variabili d'interesse per il caso di studio; nello specifico è stato applicato un raggruppamento su base oraria, ove necessario.

## 4.3 Descrizione del dataset

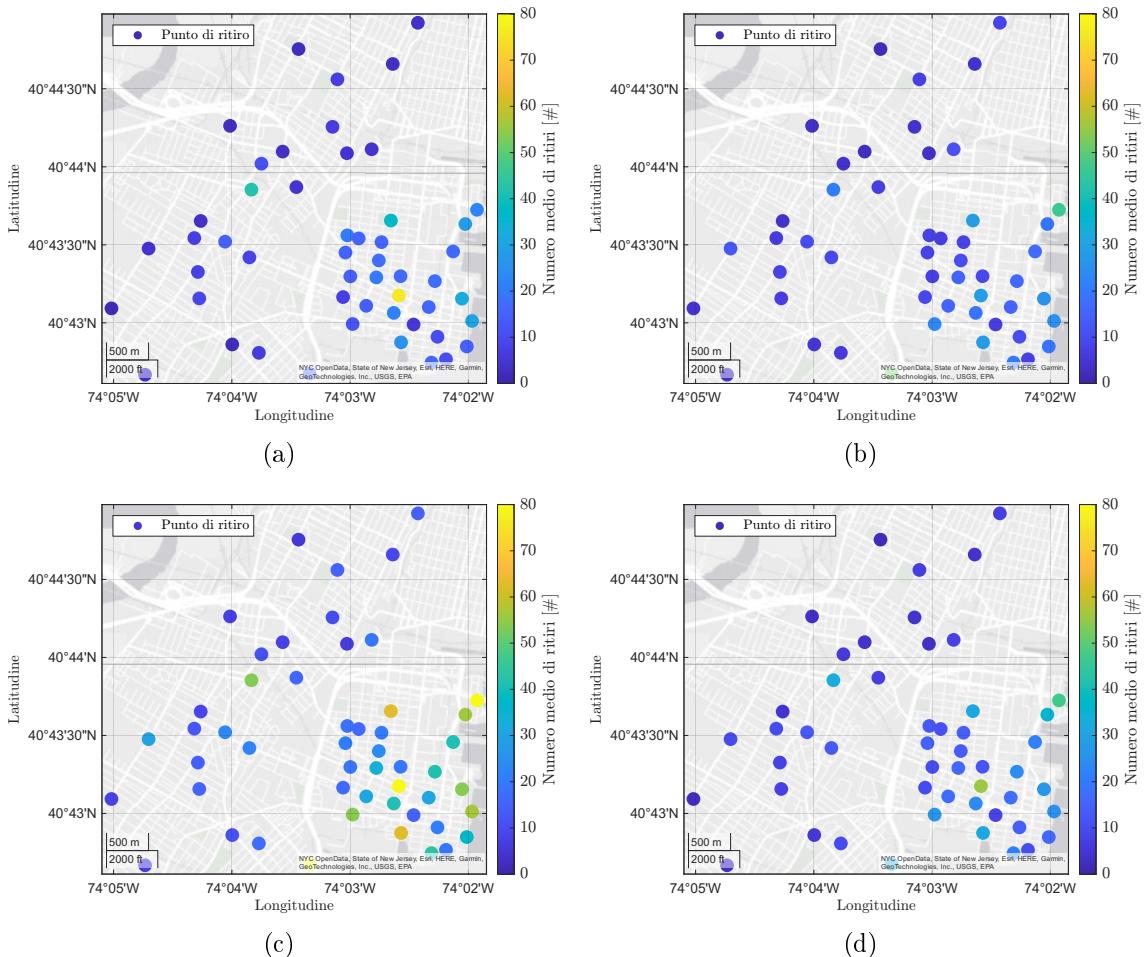
Di seguito sono riportate e descritte le variabili che compongono il dataset utilizzato per il caso di studio. Nello specifico, la variabile dipendente  $y(\mathbf{s}, l, t | \mathcal{S})$  è il numero di ritiri in una data ora  $l$ , in uno specifico giorno  $t$ , presso una determinata stazione di scambio di biciclette  $\mathbf{s}$ . Invece, le covariate possono essere suddivise in: **variabili meteorologiche**  $\mathbf{x}_{\text{meteo}}(t)$ , **variabili spaziali**  $\mathbf{x}_{\text{spazio}}(\mathbf{s})$  e **variabili dummy**  $\mathbf{x}_{\text{dummy}}(t)$ . Inoltre, per completezza, il numero di stazioni  $n$  è pari a 51, il numero massimo di osservazioni disponibili per il dominio funzionale (orario)  $q$  è 24 e l'analisi è stata condotta su dati facenti riferimento all'anno 2020, quindi  $T = 366$ . Infine, nel dataset non sono presenti dati mancati.

### 4.3.1 Variabile dipendente

In Figura 4.3 si può osservare come il numero medio di ritiri giornaliero di biciclette presso le stazioni della rete di bike sharing cambi da stagione a stagione. Questa variazione è influenzata da diversi fattori legati alle condizioni meteorologiche, agli schemi di mobilità delle persone e alle attività ricreative. In estate, quando le giornate sono più lunghe e il clima è più gradevole, si verifica un generale aumento dell'utilizzo della bicicletta (Figura 4.3(c)); le persone sono più propense a utilizzarla per spostarsi o fare gite. In autunno, invece, il numero di noleggi diminuisce leggermente poiché le giornate diventano più corte e il clima meno favorevole (Figura 4.3(d)). In inverno, il numero medio di ritiri giornaliero tende a essere il più basso dell'anno (Figura 4.3(a)); le condizioni meteorologiche avverse, come il freddo,

la pioggia e la neve, rendono meno attraente l'utilizzo della bicicletta. La primavera, infine, rispecchia la straordinarietà del 2020; il lockdown ha limitato gli spostamenti dei cittadini newyorkesi, quindi l'utilizzo della bicicletta ha subito un brusco calo. Da notare, inoltre, l'assenza di uniformità nell'utilizzo delle stazioni; quelle situate nel centro del quartiere Jersey vengono sfruttate maggiormente rispetto ai punti di scambio periferici.

Infine, l'andamento orario del numero medio di noleggi in Figura 4.4 conferma quanto detto precedentemente. Da sottolineare anche la presenza di picchi di utilizzo in concomitanza delle ore di punta.

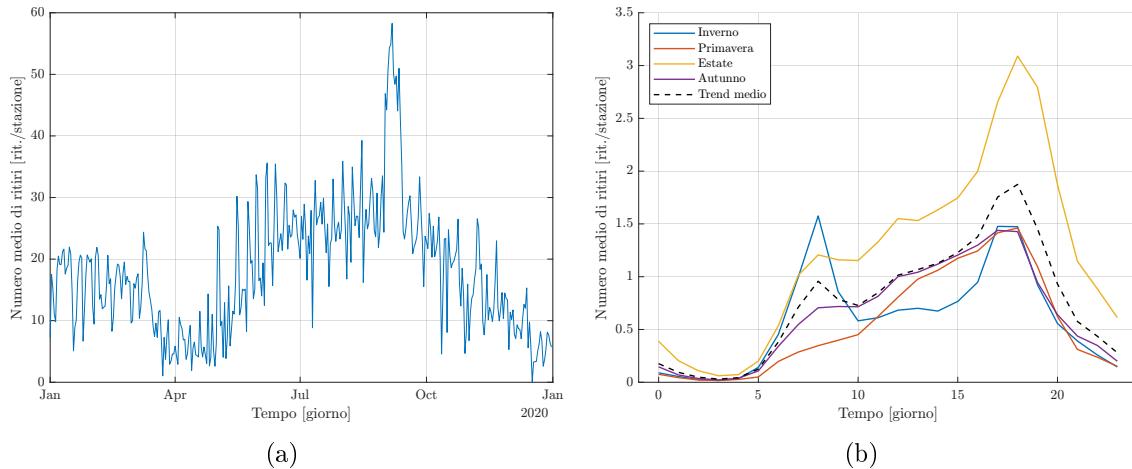


**Figura 4.3:** distribuzione spaziale del numero medio di ritiri giornaliero presso le 51 stazioni di scambio in inverno (a), in primavera (b), in estate (c) e in autunno (d), anno 2020.

### 4.3.2 Variabili meteorologiche

Le covariate meteorologiche che sono state utilizzate per l'analisi sono:

#### 4 Caso di studio



**Figura 4.4:** andamento giornaliero (a) e orario (b) del numero medio di noleggi al variare della stagione.

- la **temperatura percepita** [ $^{\circ}\text{C}$ ];
- la **piovosità** [mm];
- la **visibilità orizzontale** [km];
- la **velocità del vento** [km/h];
- la **copertura nuvolosa** [%].

Da sottolineare che le variabili sopracitate sono spazio-invarianti e orarie, ovvero sono conseguenza della media eseguita sui campioni rilevati nell'ora da una stazione meteorologica di New York. Inoltre, nella Tabella 4.1 sono riportate le principali statistiche delle covariate in questione; degne di nota la scarsa variabilità della visibilità orizzontale e l'elevata dispersione della copertura nuvolosa nel 2020.

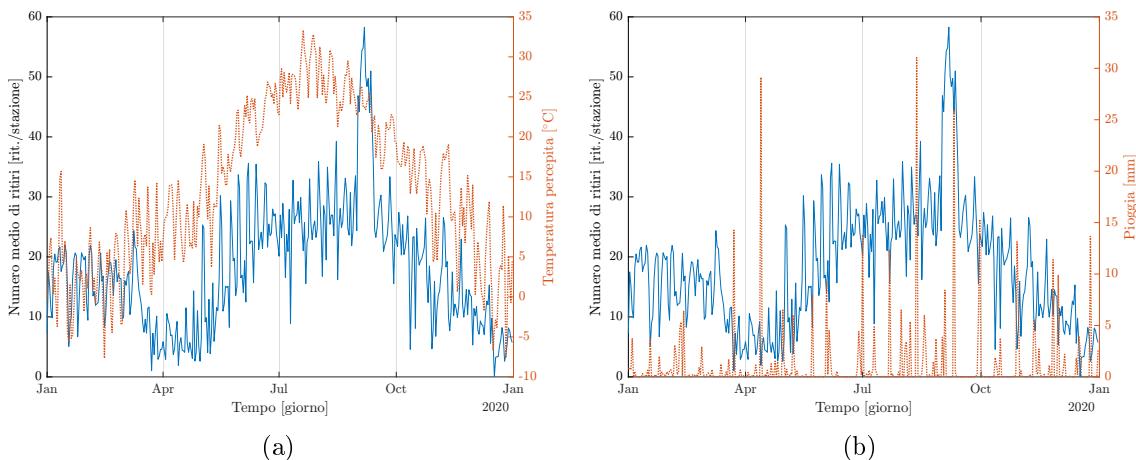
Variabile	UM	Min.	Media	Med.	Mas.	Dev.	Simm.	Curt.
<b>Temp. percep.</b>	$^{\circ}\text{C}$	-7.66	13.80	13.93	33.33	9.80	-0.04	1.96
<b>Piovosità</b>	mm	0	1.11	0	31.03	3.44	5.42	38.86
<b>Visibilità</b>	km	6.48	15.28	15.99	16	1.54	-2.99	12.71
<b>Vel. del vento</b>	km/h	4.09	10.96	9.81	27.95	4.49	1.34	4.91
<b>Copert. nuvol.</b>	%	0.08	39.45	37.29	100	29.71	0.36	1.96

**Tabella 4.1:** statistiche principali riguardanti le variabili meteorologiche.

Infine, è possibile osservare in Figura 4.5(a) la correlazione positiva esistente tra il numero medio di ritiri giornaliero e la temperatura percepita, mentre in Figu-

ra 4.5(b) l'impatto negativo nei confronti dello stesso da parte della piovosità; nei giorni uggiosi i cittadini newyorkesi prediligono mezzi di trasporto alternativi alla bicicletta.

Per gli istogrammi, i box-plot e ulteriori grafici si rimanda all'appendice.



**Figura 4.5:** confronto tra il numero medio di noleggi giornaliero, la temperatura percepita (a) e la piovosità (b).

### 4.3.3 Variabili spaziali

In questa categoria ricadono:

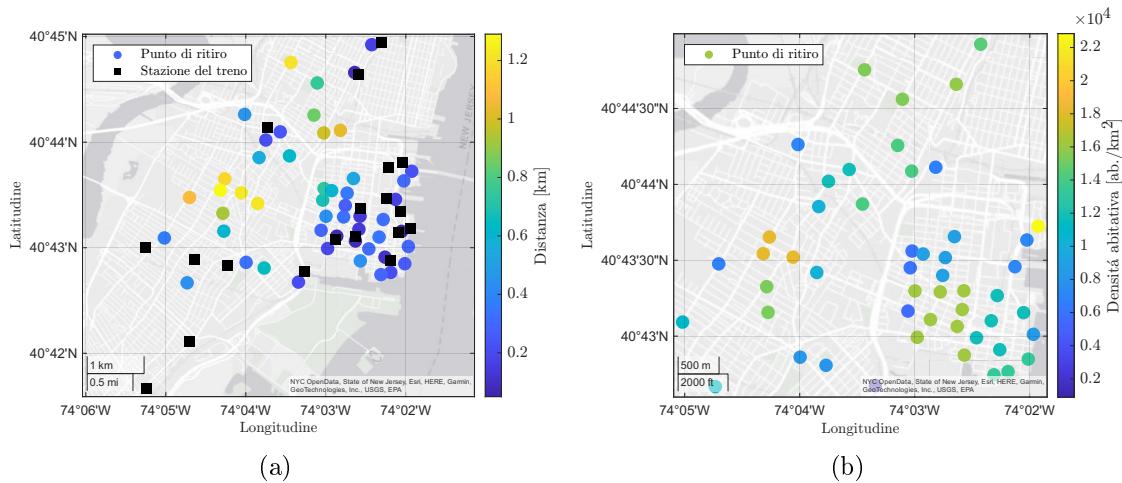
- la **distanza dalla stazione ferroviaria<sup>1</sup>** più vicina [km];
- la **densità demografica** nella zona alla quale appartiene il punto di ritiro [ab./km<sup>2</sup>].

Da evidenziare che queste covariate sono spazio-varianti e tempo-invarianti.

In Figura 4.6(a) è rappresentata la distanza di ciascun punto di interscambio dalla stazione ferroviaria più vicina. Si osserva una densa concentrazione sia di stazioni ferroviarie che di punti di ritiro delle biciclette nel centro del quartiere, densità che diminuisce man mano che ci si sposta verso le zone limitrofe. Riguardo alla densità demografica, come evidenziato nella Figura 4.6(b), si può notare una distribuzione omogenea su Jersey City. Questa osservazione è supportata dalla ridotta deviazione standard (4278 ab./km<sup>2</sup>) della covariata in oggetto, Tabella 4.2.

<sup>1</sup>o fermata della metropolitana poiché Jersey City è servita da entrambi i mezzi di trasporto pubblico.

## 4 Caso di studio



**Figura 4.6:** mappe delle distanze dei punti di interscambio dalla stazione ferroviaria più vicina (a) e della densità abitativa nei pressi dei punti di ritiro (b).

Variabile	UM	Min.	Media	Med.	Mas.	Dev.	Simm.	Curt.
<b>Dist. stazione</b>	km	0.05	0.49	0.35	1.29	0.36	0.89	2.64
<b>Dens. dem.</b>	$\frac{\text{ab.}}{\text{km}^2}$	897	12 215	11 843	22 834	4278	-0.15	2.80

**Tabella 4.2:** statistiche principali riguardanti le variabili spaziali.

### 4.3.4 Variabili dummy

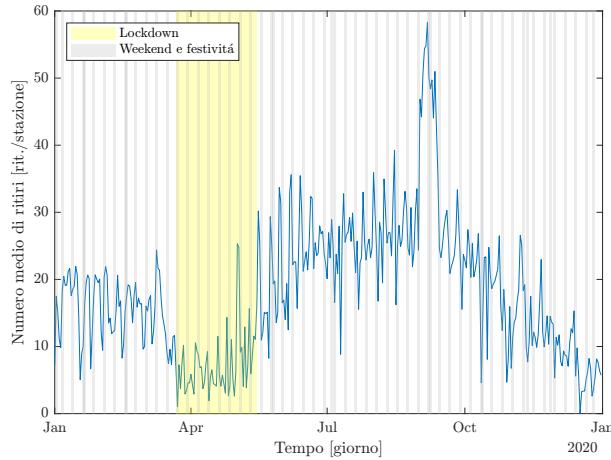
Alla luce dei risultati ottenuti in Piter et al. (2022), non solo le condizioni meteo-ologiche e le covariate spaziali influenzano la domanda giornaliera di noleggi, ma anche altri fattori possono fornire il proprio contributo. In particolare:

- il **lockdown** dovuto alla pandemia di COVID-19;
- la successiva **euforia** dovuta al ritorno alla vita quotidiana dopo mesi trascorsi in isolamento;
- i **weekend** e le **festività** federali<sup>2</sup>.

Questi eventi sono stati modellati utilizzando 3 distinte variabili categoriche binarie (covariate dummy).

Infine, la Figura 4.7 conferma che, durante la primavera, il numero ridotto di noleggi coincide con il periodo di lockdown, mentre nei weekend si osserva una generale riduzione dell'utilizzo della bicicletta, a dimostrazione che il servizio di bike sharing viene essenzialmente sfruttato nei giorni feriali, probabilmente per i consueti spostamenti lavorativi.

<sup>2</sup>giorno di festa ufficialmente riconosciuto a livello federale dal governo degli Stati Uniti.



**Figura 4.7:** confronto tra il numero medio di prelievi giornaliero, il periodo di lockdown, i weekend e le festività federali.

## 4.4 Metodologia

L'iter seguito per costruire, stimare e validare il modello fp-HDGM sul dataset del caso di studio è il seguente:

1. **cross-validation per determinare il valore del parametro  $\rho$ .** Attualmente la modifica dell'implementazione dell'algoritmo EM in D-STEM per consentire la stima anche di  $\rho$  non è stata ancora conclusa, quindi è stata impiegata la cross-validation per determinare il suo valore. Nello specifico, dopo aver suddiviso il dataset in  $k = 5$  gruppi di stazioni di interscambio, iterativamente  $k - 1$  sono stati utilizzati per stimare il modello, mentre il  $k$ -esimo per validarla. Inoltre, per ridurre sia l'onere computazionale sia il numero di osservazioni nulle, sono stati considerati soltanto i dati dei mesi di giugno e luglio, quando l'utilizzo del bike sharing è più frequente. Infatti, limitando l'analisi a questi periodi, si minimizza il numero di ore in cui non sono state prelevate biciclette; ciò consente di cross-validare il modello su un dataset più informativo. Infine, come metrica è stato impiegato il *Mean Squared Error* (*MSE*) così definito:

$$MSE = \frac{1}{P} \sum_{i=1}^{k=5} \frac{1}{\text{card}(\mathcal{S}_{val})_i} \sum_{\mathbf{s} \in \mathcal{S}_{val}} \sum_{l=1}^T \sum_{t \in \mathcal{L}} (y(\mathbf{s}, l, t) - \hat{y}(\mathbf{s}, l, t))^2; \quad (4.1)$$

$$P = k \cdot T \cdot q;$$

dove  $\text{card}(\mathcal{S}_{val})_i$  rappresenta il numero di punti di ritiro utilizzati per la validazione disponibili per l' $i$ -esimo gruppo;

2. **scelta delle covariate più significative (model selection).** Dopo aver

#### 4 Caso di studio

individuato nel passo precedente il  $\rho$  ottimo, è stato stimato un nuovo modello utilizzando tutti i dati e tutte le covariate. L'obiettivo è visualizzare le spline per i parametri  $\beta$  e i relativi intervalli di confidenza al fine di escludere i regressori poco significativi; se presenti, allora vengono rimossi e viene rieseguito il passo 1. Da evidenziare che i dati sono stati standardizzati così da consentire un confronto diretto e adimensionale tra gli andamenti  $\beta(h)$  al fine di comprendere con facilità quali regressori hanno un potere esplicativo maggiore.

3. **validazione del modello finale tramite LOOCV**<sup>3</sup>. Come metrica per valutare la bontà complessiva del modello definitivo, è stato utilizzato il *Root Mean Squared Error (RMSE)*. In aggiunta, è stato calcolato anche il  $RMSE_s$  (Wang et al., 2021), ossia:

$$RMSE_s = \sqrt{\frac{1}{T \cdot q} \sum_{t=1}^T \sum_{l \in \mathcal{L}} (y(\mathbf{s}, l, t) - \hat{y}(\mathbf{s}, l, t))^2}. \quad (4.2)$$

Grazie all'approccio LOOCV, questo indice può essere calcolato per tutte le stazioni, fornendo così un'indicazione dei punti di scambio per i quali il modello mostra le prestazioni migliori.

4. **previsione spaziale utilizzando il kriging.** Dopo aver previsto il numero di ritiri orario per ogni punto spaziale dell'area urbana di Jersey City per ogni giorno dei mesi di giugno e luglio eseguendo il kriging sul modello finale, ergo il geo-potenziale condizionato  $p(\mathbf{s}, l, t)$ , è stato computato il volume di noleggi orario così definito:

$$v(\mathbf{s}, l | \mathcal{T}) = \sum_t^{\mathcal{T}} \hat{y}(\mathbf{s}, l, t), \quad \forall \mathbf{s} \in \mathcal{D}, \quad \forall l \in \mathcal{H}; \quad (4.3)$$

con  $\mathcal{T} = [152, 213] \subset [1, 366]$ , ovvero i giorni del periodo preso in esame, e  $\mathcal{D}$  rappresentante i punti spaziali del quartiere Jersey. Dopodiché, i 24 valori di  $v(\mathbf{s}, l | \mathcal{T})$  sono stati così raggruppati:

$$\bar{v}(\mathbf{s}, l | \mathcal{T})_i = \frac{1}{\text{card}(\mathcal{R}_i)} \sum_{j \in \mathcal{R}_i} v(\mathbf{s}, j | \mathcal{T}), \quad i = 1 \dots 6; \quad (4.4)$$

dove  $\mathcal{R}_i \subset \mathcal{H}$  rappresenta la  $i$ -esima delle 6 fasce orarie in cui è stata suddivisa la giornata; così facendo si perde ovviamente la dimensione oraria del volume,

---

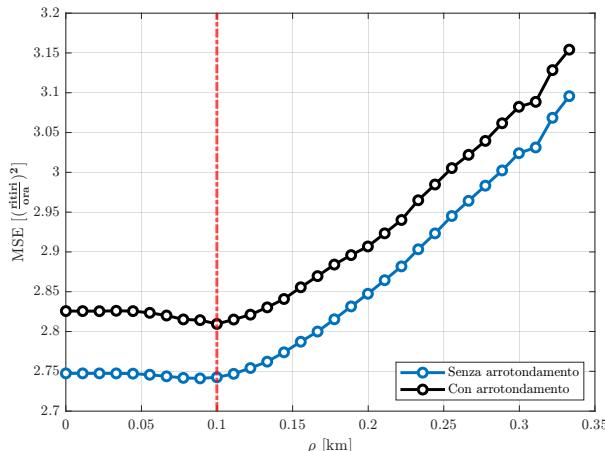
<sup>3</sup>Leave-one-out cross-validation

ciononostante si evita di riportare un numero eccessivo di mappe all'interno del capitolo. Da sottolineare, infine, che per definire la griglia di coordinate spaziali in cui fare predizione è stata scelta una risoluzione pari a 55 m.

## 4.5 Analisi dei risultati

### 4.5.1 Scelta del parametro $\rho$ tramite la cross-validation

In Figura 4.8 si nota che il  $MSE$  sia con sia senza arrotondamento<sup>4</sup> assume il valore minimo per  $\rho = 100$  m. Questo risultato è in linea con l'ordine di grandezza della media delle distanze medie tra una stazione e le 3 stazioni più vicine a essa, ossia 397 m con una deviazione standard pari a 189 m. La decisione di impiegare questa metrica per valutare la coerenza del valore di  $\rho$  ottenuto attraverso la cross-validation, anziché la media delle distanze medie tra una stazione di ritiro e tutte le altre, è motivata dalla distribuzione non omogenea dei punti di noleggio su una vasta area urbana. Questa metrica non è influenzata da tale dispersione, consentendo così una migliore descrizione dell'intensità dell'interazione tra i punti di ritiro che si trovano nelle vicinanze.



**Figura 4.8:** andamento del  $MSE$  in cross-validation ( $k = 5$ ) con e senza arrotondamento al variare del parametro  $\rho$ .

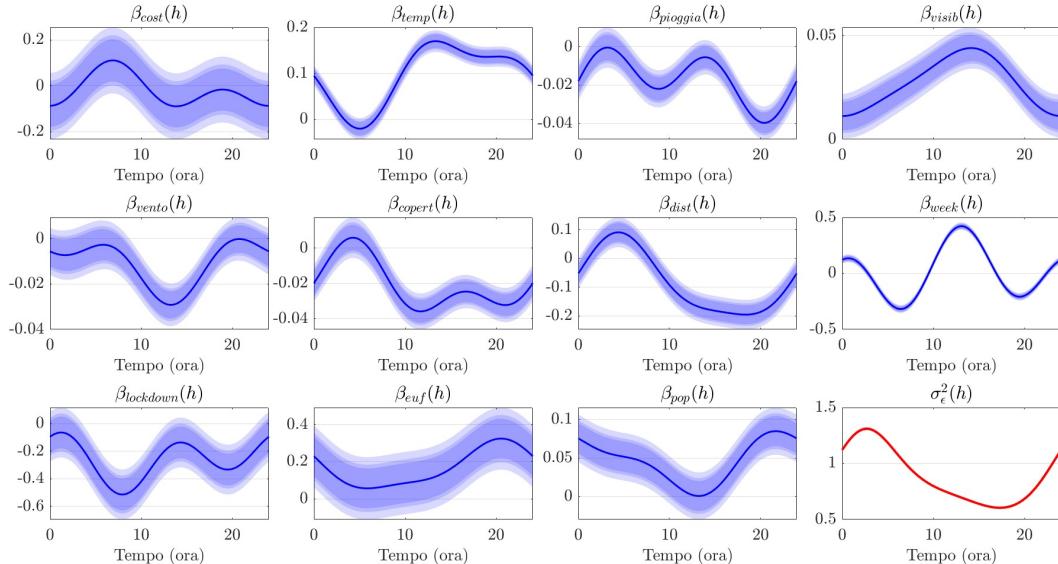
### 4.5.2 Scelta delle covariate $\beta$

Osservando gli andamenti delle spline di Fourier in Figura 4.9, si possono trarre le seguenti considerazioni:

<sup>4</sup>i valori previsti dal modello vengono arrotondati all'intero più vicino poiché il numero di noleggi è una grandezza intera.

#### 4 Caso di studio

- i dati sono stati standardizzati, pertanto è corretto che l'intercetta  $\beta_{const}$  assuma dei valori vicini a 0;
- per nessuna covariata il valore 0 permane negli intervalli di confidenza per una frazione rilevante del dominio funzionale  $\mathcal{L}$ , di conseguenza tutti i regressori sono significativi;
- la temperatura  $\beta_{temp}$ , la visibilità  $\beta_{visib}$ , l'euforia successiva al lockdown  $\beta_{euf}$  e la densità abitativa  $\beta_{pop}$  contribuiscono positivamente alla descrizione del fenomeno del bike sharing, viceversa la pioggia  $\beta_{pioggia}$ , la velocità del vento  $\beta_{vento}$ , il lockdown  $\beta_{lockdown}$  e la copertura nuvolosa  $\beta_{copert}$  negativamente;
- la distanza dalla stazione ferroviaria più vicina  $\beta_{dist}$  ha un andamento negativo durante le ore lavorative;
- la spline di  $\beta_{week}$  ha bassa incertezza ed è quella che contribuisce maggiormente alla spiegazione del numero di noleggi;
- infine, il comportamento di  $\sigma_\epsilon^2$  suggerisce una maggiore incertezza nelle ore notturne e serali.

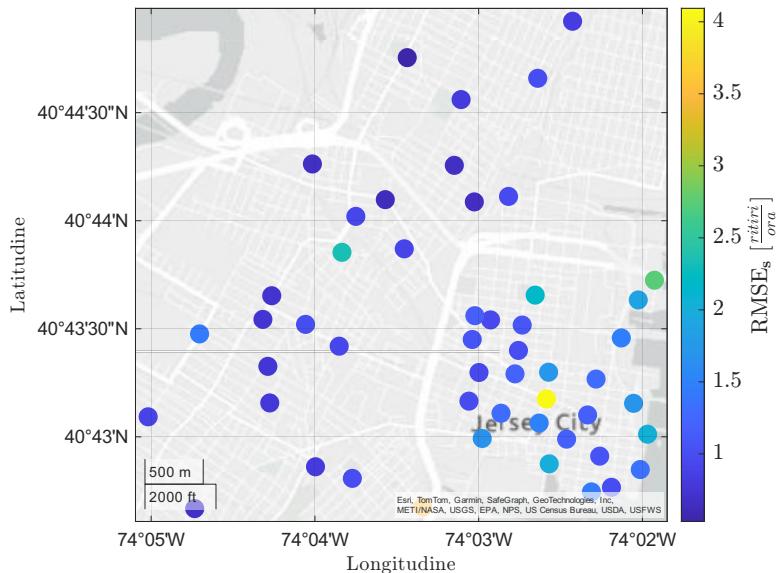


**Figura 4.9:** spline di Fourier per i coefficienti  $\beta$  e rispettivi intervalli di confidenza al 90%, 95% e 99%.

#### 4.5.3 Validazione del modello definitivo mediante la LOOCV

I punti di ritiro periferici vengono utilizzati meno rispetto a quelli situati in centro (Figura 4.3), quindi è lecito aspettarsi che il modello spazio-temporale goda di

capacità predittive migliori sui primi piuttosto che sui secondi grazie alla minor variabilità sia oraria che giornaliera del numero di noleggi. Tale assunzione viene confermata in Figura 4.10; si osserva, infatti, che il più alto  $MSE_s$  viene commesso presso i punti a elevato utilizzo, ossia quelli in centro al quartiere Jersey.



**Figura 4.10:** mappa della distribuzione del  $RMSE_s$ .

Inoltre, in Tabella 4.3 si può notare che i risultati in validazione ottenuti con l’arrotondamento sono leggermente peggiori rispetto a quelli conseguiti senza, tuttavia i primi sono più coerenti al contesto del bike sharing in quanto non avrebbe significato frazionare una bicicletta. Infine, i ridotti valori assunti dalla media e dalla deviazione standard evidenziano le buone capacità predittive del modello prima stimato e ora validato.

Variabile	UM	Min.	Media	Med.	Mas.	Dev.	Simm.	Curt.
<b>Con arrot.</b>	ritiri ora	0.53	1.29	1.01	4.09	0.71	2.10	7.90
<b>Senza arrot.</b>	ritiri ora	0.49	1.25	0.97	4.08	0.71	2.07	7.75

**Tabella 4.3:** statistiche principali riguardanti il  $RMSE_s$  con e senza arrotondamento.

#### 4.5.4 Previsione spaziale utilizzando il kriging

Nella fascia oraria che va dalle ore 00:01 alle ore 08:00, Figure 4.11(a) e 4.11(b), si nota una diminuzione significativa della domanda di ritiri; viceversa, la fascia oraria compresa tra le ore 12:01 alle ore 20:00 è quella più promettente, Figure 4.11(d)

#### 4 Caso di studio

e 4.11(e). Sempre in queste condizioni, si osserva come nei pressi delle coordinate  $(40.71^\circ\text{N}, -74.05^\circ\text{W})$  e  $(40.73^\circ\text{N}, -74.03^\circ\text{W})$  il valore di  $\bar{v}(\mathbf{s}, l|\mathcal{T})$  in quest'area appaia favorevole.

Infine, in prossimità delle stazioni di ritiro si percepisce un calo tangibile del potenziale di mercato, in particolare nella zona del centro di Jersey City, area in cui la distanza media tra i punti di scambio è confrontabile con il valore del parametro  $\rho$  stimato, ossia 100 m. Queste osservazioni sono fattori che meritano di essere presi in considerazione per ottimizzare le strategie di gestione del servizio di bike sharing.

## 4.6 Discussione sui risultati

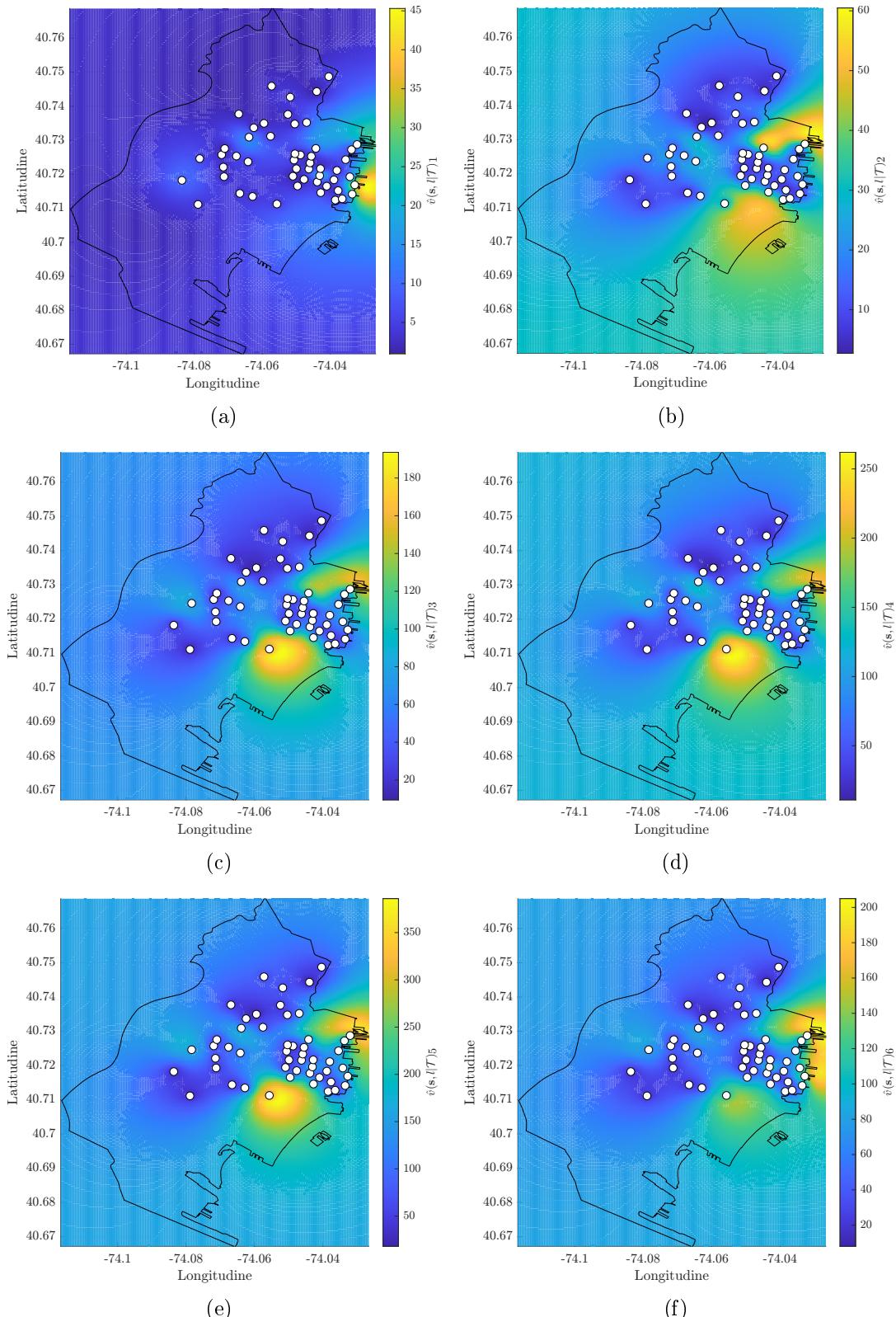
La scelta delle covariate  $\beta$ , fondata sull'analisi degli intervalli di confidenza delle spline di Fourier, mostra che i suggerimenti forniti dagli studi sopracitati sono risultati utili in quanto tutte le variabili prese in esame sono significative.

In particolare, si osserva come le covariate meteorologiche giochino un ruolo cruciale nel descrivere il fenomeno del bike sharing; esso è influenzato positivamente da condizioni favorevoli, mentre è impattato negativamente da circostanze avverse (Gebhart and Noland, 2014). Per esempio, è ragionevole supporre che un individuo sia più propenso a utilizzare la bicicletta quando la temperatura aumenta e meno nelle giornate uggiose, ovvero quando il cielo è coperto.

Il forte contributo e la scarsa incertezza della variabile  $\beta_{week}$  sottolinea che il prendere in considerazione la tipologia di giorno, feriale o meno, aumenta la capacità del modello di prevedere  $y(\mathbf{s}, l, t|\mathcal{S})$  (Piter et al., 2022). Nello specifico, la correzione apportata dalla variabile in oggetto nei weekend e nei giorni festivi fa sì che nelle ore di punta mattutine e serali il numero di ritiri previsti sia inferiore rispetto a quello nei giorni lavorativi, viceversa in quelle pomeridiane; ciò è probabilmente dovuto al cambio di abitudini dei cittadini newyorchesi quando non devono lavorare e possono quindi dedicarsi alle proprie attività ricreative.

In contrasto con i risultati di Gebhart and Noland (2014), le analisi svolte portano ad affermare che la distanza dalla stazione ferroviaria più vicina ha un impatto negativo sul numero orario di noleggi, soprattutto nelle ore lavorative. Questa discrepanza si potrebbe attribuire al fatto che nel centro del quartiere Jersey, il cuore delle attività economiche della città, si ha un'elevata concentrazione di stazioni sia ferroviarie che di noleggio. In questo contesto, l'individuo alle prese con una frenetica giornata lavorativa potrebbe preferire l'utilizzo della metropolitana alla bicicletta.

Per quanto concerne, invece, il periodo di lockdown, il trend della rispettivo  $\beta$



**Figura 4.11:** mappa del volume del numero di ritiri di biciclette previsti nei mesi di giugno e luglio, raggruppati per fasce orarie: da 00:01 a 04:00 (a), da 04:01 a 08:00 (b), da 08:01 a 12:00 (c), da 12:01 a 16:00 (d), da 16:01 a 20:00 (e) e da 20:01 a 00:00 (f).

#### 4 Caso di studio

evidenzia una generale riduzione dell'utilizzo del servizio a causa dell'impossibilità da parte dei cittadini di uscire dalle proprie abitazioni. Questo comportamento subisce un ribaltamento nei primi mesi successivi alla chiusura generale; infatti, l'andamento positivo della variabile  $\beta_{euf}$ , testimonia come il periodo post-lockdown sia stato caratterizzato da un generale incremento della voglia da parte della popolazione di svolgere attività all'aria aperta.

Altresì, la validazione tramite LOOCV conferma le buone capacità predittive del modello stimato, con delle prestazioni migliori per le stazioni periferiche rispetto a quelle centrali.

Infine, l'applicazione della previsione spaziale per determinare il volume di noleggi orario evidenzia un'interessante variazione della domanda durante il giorno, con una significativa diminuzione nelle prime ore del mattino e un picco di interesse nelle ore pomeridiane. Da evidenziare l'importanza del parametro  $\rho$ ; la sua rilevanza consente di modellare l'interazione esistente tra punti di ritiro, permettendo così di indicare due aree potenzialmente interessanti per la collocazione di una nuova stazione, rispettivamente a nord e a sud-est del centro. I denominatori comuni tra queste due zone sono:

- la vicinanza alla zona più demograficamente ed economicamente vivace;
- il sufficiente distanziamento dall'agglomerato di punti di ritiro del centro del quartiere Jersey.

Ovviamente queste conclusioni sono limitate dal fatto che il kriging è stato eseguito solo per i mesi di giugno e luglio. Per avere una visione completa, è necessario estendere lo studio anche ai restanti mesi dell'anno.

## 5 Conclusioni

Alla luce dei risultati ottenuti, il presente studio evidenzia il valore aggiunto del nuovo modello proposto nel campo della modellazione funzionale spazio-temporale come un’evoluzione rispetto al modello f-HDGM. Una delle distinzioni chiave di questo lavoro è l’introduzione del parametro di interazione spaziale  $\rho$ , il quale consente di affrontare situazioni e dinamiche in cui esiste interazione tra i punti di misura, un aspetto non considerato dal modello padre.

È importante notare che una delle limitazioni riguarda il processo di stima del nuovo parametro utilizzato nel modello proposto. Sebbene sia stata impiegata una metodologia affidabile come la cross-validation per stimare il suo valore, è necessario sottolineare che tale approccio richiede notevoli risorse computazionali; questa necessità potrebbe costituire un ostacolo pratico per alcuni utenti, specialmente in contesti nei quali le risorse computazionali sono limitate. Per superare questo ostacolo, si rende necessaria la modifica dell’algoritmo EM affinché sia in grado di minimizzare anche la funzione  $m(\rho)$ , equazione 3.17. Come anticipato nella sezione 4.4, la sua implementazione in D-STEM è già stata avviata, tuttavia richiede ancora un periodo di ricerca e sviluppo per poter rispondere ad alcuni quesiti ancora aperti. Innanzitutto, un aspetto cruciale risiede nella verifica dell’identificabilità di  $\rho$ , ossia capire se esso può essere stimato congiuntamente agli altri parametri  $\theta$ , espressione 2.13. In particolare, la stima di  $\rho$  potrebbe andare in conflitto con quella di  $\lambda$  poiché entrambi vengono identificati tramite l’ottimizzazione numerica. Se ciò dovesse concretizzarsi, allora sarà opportuno stimare uno solo dei due parametri alla volta, tenendo fisso il restante. Dopodiché, un altro fattore da attenzionare riguarda l’inizializzazione di  $\rho$ . Essendo  $m(\rho)$  una funzione non convessa, senza un’opportuna scelta del valore iniziale da assegnare al nuovo parametro si corre il rischio di identificare un minimo locale piuttosto che l’ottimo globale. Pertanto, per indirizzare l’ottimizzatore nella giusta direzione, è necessario avere un’idea a priori dell’ordine di grandezza di  $\rho$ , basandosi sulla conoscenza del dominio dello specifico caso di studio.

Infine, per quanto riguarda l’analisi svolta sul bike sharing, sarebbe interessante raffinare lo studio prendendo in esame un dataset pluriennale così da poter cattura-

## *5 Conclusioni*

re anche la componente stagionale del fenomeno. Altresì, si potrebbero prendere in considerazioni delle variazioni metodologiche, per esempio eseguire un clustering dei punti di ritiro, a monte dell’analisi, ed eseguire così la stima di un modello spazio-temporale locale per ogni cluster. Ciò consentirebbe di rilassare un’importante assunzione fatta a priori, ossia che il parametro  $\rho$  sia spazio-invariante.

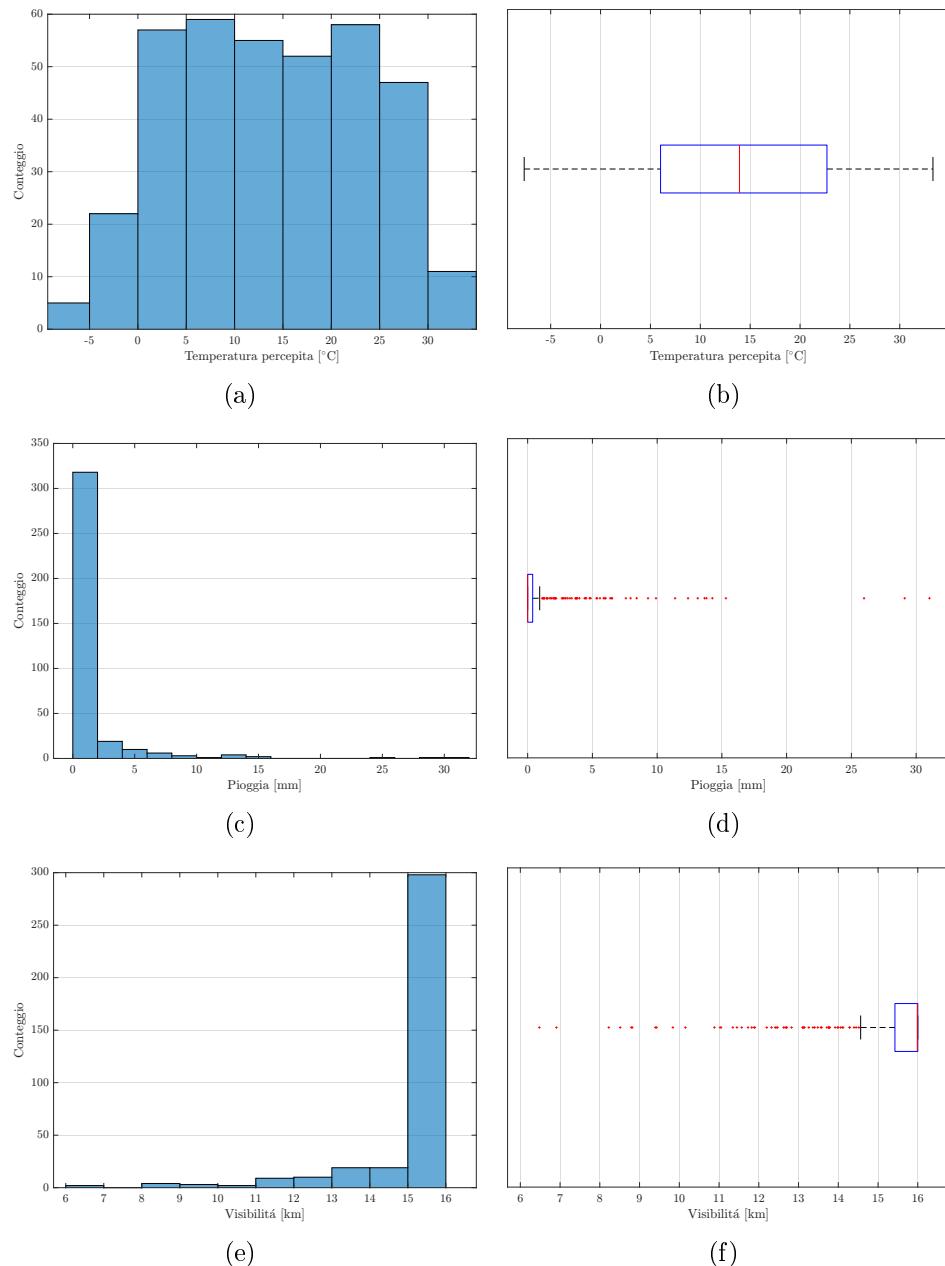
# 6 Sitografia

## 6.1 Caso di studio

- <https://climate.copernicus.eu/copernicus-2023-hottest-year-record>;
- <https://www.kaggle.com/datasets/vineethakkinapalli/citibike-bike-sharingnewyork-cityjan-to-apr-2021>;
- <https://www.visualcrossing.com/weather-api>;
- <https://stewartmader.com/subwaynynj/>;
- <https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-adjusted-to-2015-unwpp-country-totals-rev11>;
- <https://www.officeholidays.com/countries/usa/new-york/2020>;
- [https://en.wikipedia.org/wiki/COVID-19\\_pandemic\\_in\\_New\\_York\\_City](https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_City).

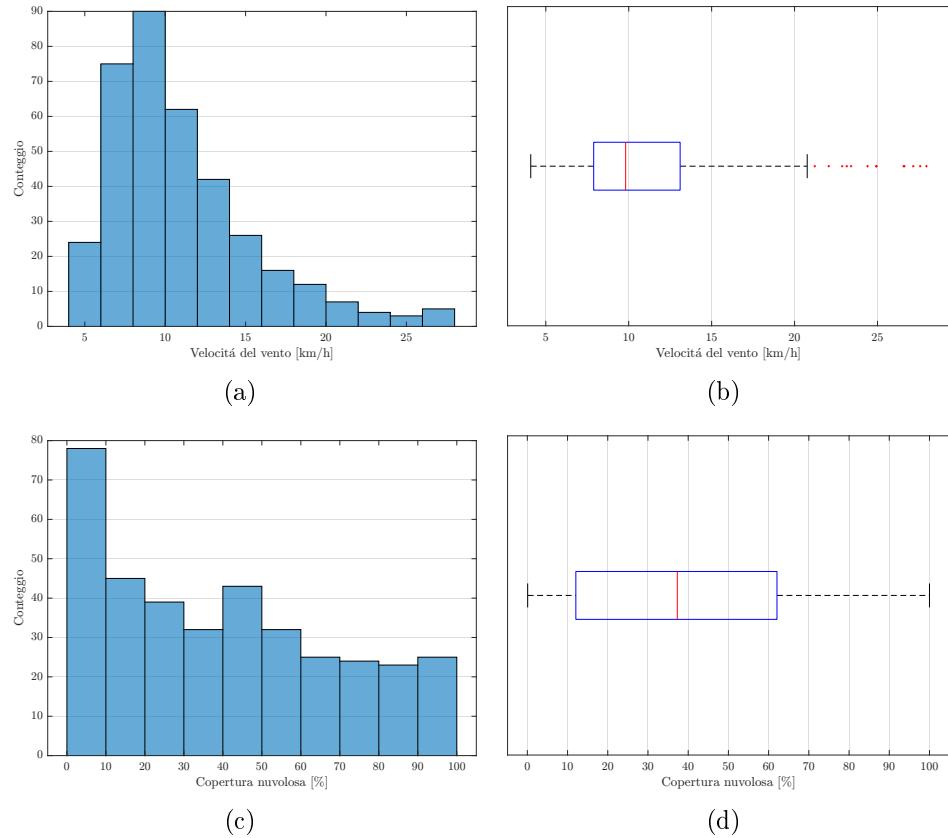


# 7 Appendice

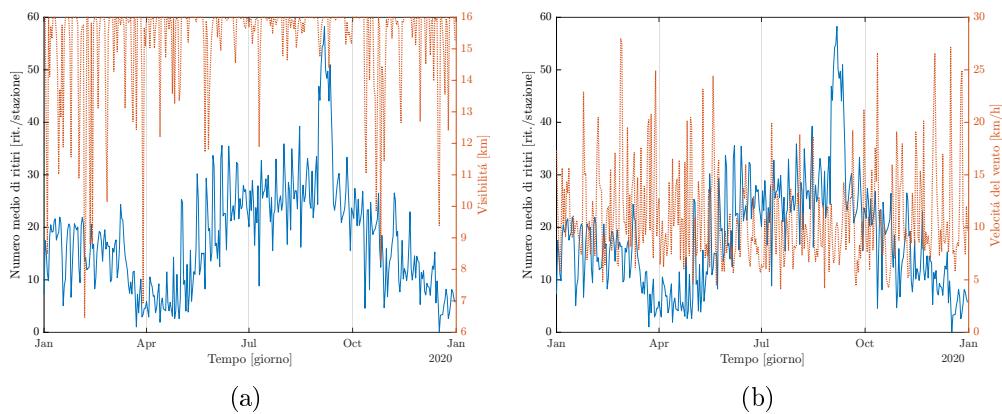


**Figura 7.1:** istogrammi e box-plot delle variabili meteorologiche, parte 1.

## 7 Appendice



**Figura 7.2:** istogrammi e box-plot delle varibili meteorologiche, parte 2.



**Figura 7.3:** confronto tra il numero medio di noleggi giornaliero, la visibilità orizzontale (a) e la velocità del vento (b).

# Bibliografia

- Borman, S. (2012). The expectation maximization algorithm: A short tutorial. 2004. *Unpublished paper available at <http://www.seanborman.com/publications>.*
- Calculli, C., Fassò, A., Finazzi, F., Pollice, A., and Turnone, A. (2015). Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. *Environmetrics*, 26(6):406–417.
- Finazzi, F. (2013). Geostatistical modeling in the presence of interaction between the measuring instruments, with an application to the estimation of spatial market potentials. *The Annals of Applied Statistics*, pages 81–101.
- Finazzi, F., Fasso, A., et al. (2013). EM estimation of a multivariate space-time data fusion model with varying coefficients. *GRASPA WORKING PAPERS*, 47.
- Gebhart, K. and Noland, R. B. (2014). The impact of weather conditions on bikeshare trips in Washington, DC. *Transportation*, 41:1205–1225.
- Jiao, J., Lee, H. K., and Choi, S. J. (2022). Impacts of COVID-19 on bike-sharing usages in Seoul, South Korea. *Cities*, 130:103849.
- Li, Z., Shang, Y., Zhao, G., and Yang, M. (2022). Exploring the multiscale relationship between the built environment and the metro-oriented dockless bike-sharing usage. *International Journal of Environmental Research and Public Health*, 19(4):2323.
- Piter, A., Otto, P., and Alkhatib, H. (2022). The Helsinki Bike-Sharing System—Insights Gained from a Spatiotemporal Functional Model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(3):1294–1318.
- Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 53(3):539–561.
- Shumway, R. H., Stoffer, D. S., and Stoffer, D. S. (2000). *Time series analysis and its applications*, volume 3. Springer.

## *Bibliografia*

- Wang, Y., Finazzi, F., and Fassò, A. (2021). D-STEM v2: A software for modelling functional spatio-temporal data. *arXiv preprint arXiv:2101.11370*.
- Zhang, Y. and Mi, Z. (2018). Environmental benefits of bike sharing: A big data-based analysis. *Applied energy*, 220:296–301.