

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

Lorenzo LEONI    Nicola ZAMBELLI

Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione

27 marzo 2024



UNIVERSITÀ  
DEGLI STUDI  
DI BERGAMO

# Indice

- 1 Introduzione**
- 2 Il Functional Potential Hidden Dynamic Geostatistical Model (fp-HDGM)**
- 3 Stima EM del modello fp-HDGM**
- 4 Modellazione di un dataset di bike sharing**
- 5 Conclusioni**

# Terminologia

- **Modello statistico funzionale:** l'oggetto della stima non sono i parametri  $\theta$  come accade per i modelli parametrici, ma una funzione  $f$  continua. Le spline sono una delle classi di funzioni più utilizzate nella Functional Data Analysis.
- **Modello spazio-temporale:** si ricorre a questa famiglia di modelli quando si vogliono cogliere le relazioni e le dinamiche che sorgono ognqualvolta le osservazioni sono influenzate sia dalla posizione spaziale sia dal tempo.
- **Interazione tra le misure:** nel caso del bike sharing, per esempio, i punti di ritiro delle biciclette possono farsi concorrenza se vengono collocati nella medesima zona.
- **Potenziale:** valore atteso di una variabile, per esempio il numero di ritiro orario presso una determinata stazione, assumendo che la misura avvenga solo in  $(s, t)$  e non negli altri punti. Il *Kriging* è una delle tecniche utilizzate per stimarlo.

# Il Functional Hidden Dynamic Geostatistical Model (f-HDGM)

$$y(\mathbf{s}, l, t) = \mathbf{x}(\mathbf{s}, l, t)^\top \cdot \boldsymbol{\beta}(l) + \Phi_z(l)^\top \cdot \mathbf{z}(\mathbf{s}, t) + \epsilon(\mathbf{s}, l, t)$$

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

dove:

- $\mathbf{s}$  è un punto spaziale,  $t$  è il tempo, mentre  $l$  è l'indice del dominio funzionale;
- $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon(l))$  con varianza  $\sigma_\epsilon(l) = \Phi_\epsilon(l)^\top \cdot \mathbf{c}_\epsilon$  funzionale;
- $\boldsymbol{\beta}(l) = [\beta_1(l) \cdots \beta_j(l) \cdots \beta_b(l)]^\top$  con  $\beta_j(l) = \Phi_\beta(l)^\top \cdot \mathbf{c}_{\beta,j}$  funzionale,  $\forall j = 1, \dots, b$ ;
- $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \Gamma(\mathbf{s}, \mathbf{s}', \lambda))$  è un processo gaussiano multivariato, con  $\Gamma(\mathbf{s}, \mathbf{s}', \lambda)$  matrice di correlazione spaziale diagonale a blocchi;
- $G$  è la matrice di transizione diagonale della dinamica markoviana su  $\mathbf{z}(\mathbf{s}, t)$ .

# La funzione di interazione spaziale

## Espressione

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

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

dove  $\mathcal{S}$  è l'insieme degli  $N$  punti spaziali nei quali è stata misurata la variabile d'interesse.

## Limiti

$$\lim_{\rho \rightarrow 0} f_\rho(\|\mathbf{s} - \mathbf{s}'\|) = 0$$

$$\lim_{\rho \rightarrow 0} h_\rho(\mathbf{s}|\mathcal{S}) = 1$$

$$\lim_{\rho \rightarrow \infty} f_\rho(\|\mathbf{s} - \mathbf{s}'\|) = 1$$

$$\lim_{\rho \rightarrow \infty} h_\rho(\mathbf{s}|\mathcal{S}) = \frac{1}{N+1}$$

## Il nuovo modello fp-HDGM

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

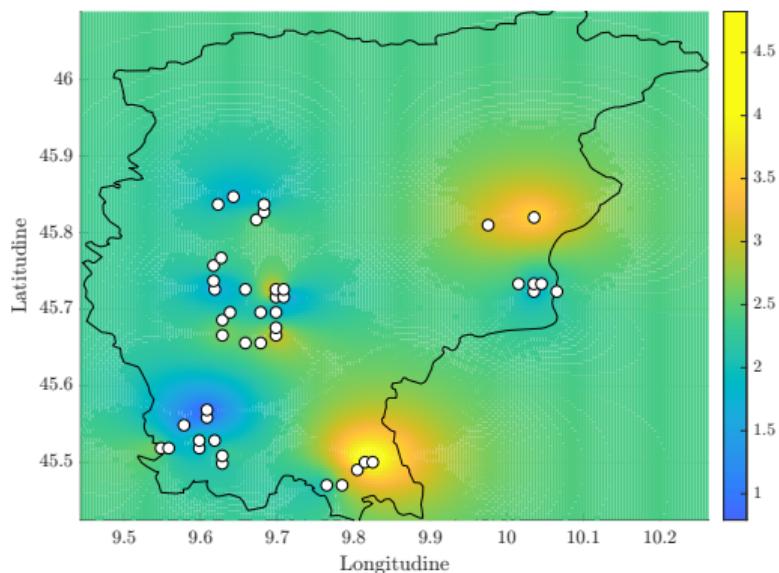
$$w(\mathbf{s}, l, t) = \mathbf{x}(\mathbf{s}, l, t)^\top \cdot \boldsymbol{\beta}(l) + \Phi_z(l)^\top \cdot \mathbf{z}(\mathbf{s}, t) + \epsilon(\mathbf{s}, l, t)$$

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

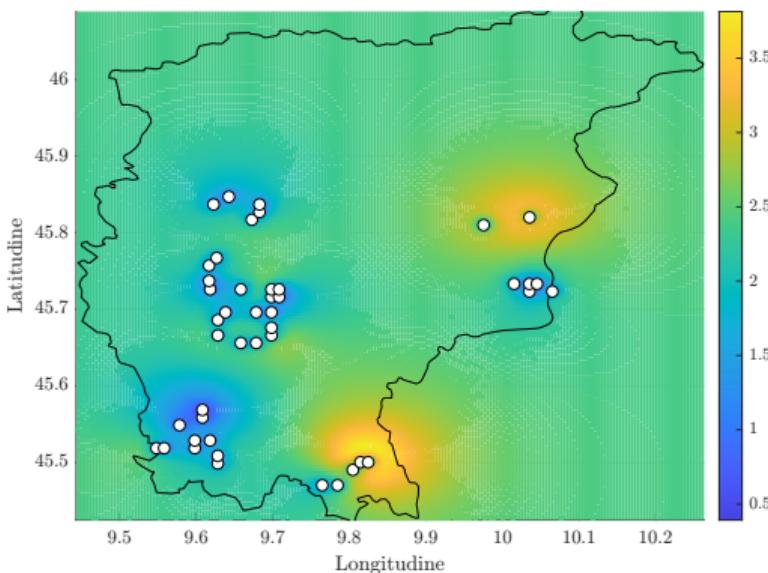
dove  $h_\rho(\mathbf{s} | \mathcal{S})$  condiziona le osservazioni  $w(\mathbf{s}, l, t)$  prive della componente interattiva e modellate utilizzando il modello f-HDGM.  $\rho$  è il nuovo parametro da stimare, il quale descrive l'intensità dell'interazione spaziale.

# Simulazione di una mappa di potenziale al variare di $\rho$

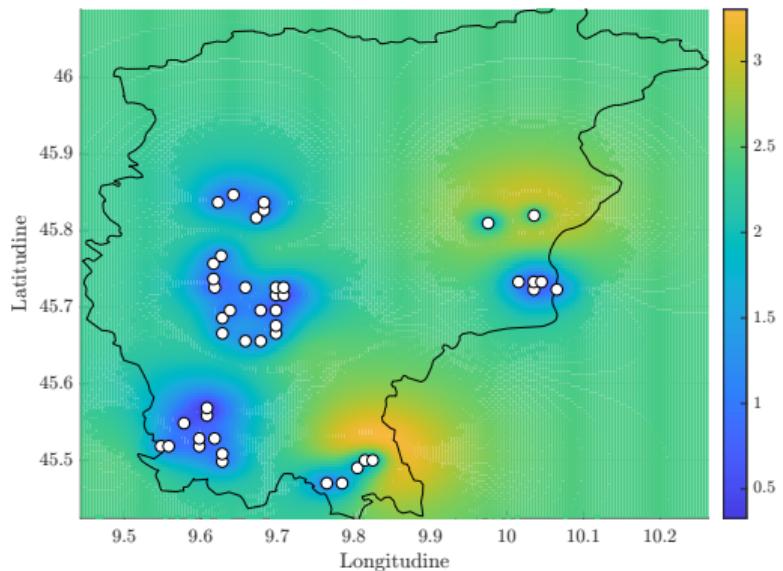
$$\rho = 0$$



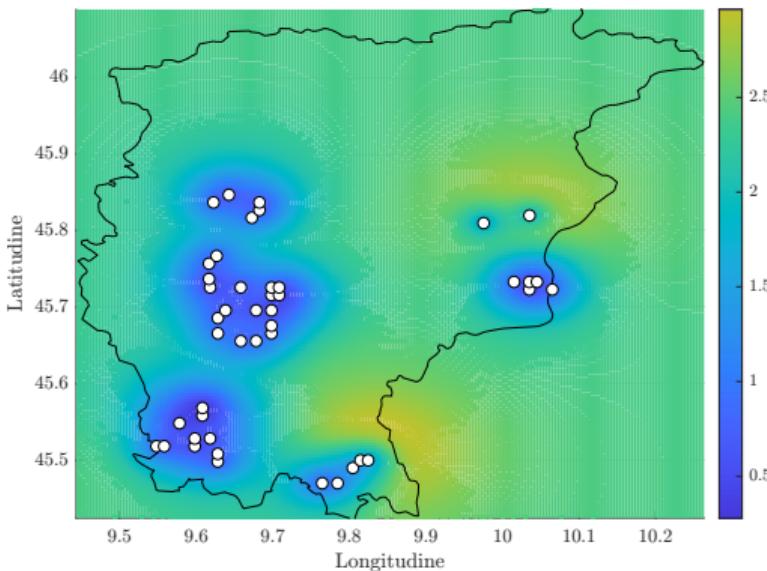
$$\rho = 0.5 \cdot 10^{-2}$$



$$\rho = 10^{-2}$$



$$\rho = 1.5 \cdot 10^{-2}$$



# Obiettivo della stima EM

Sia:

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

il vettore dei parametri.

L'obiettivo dell'**algoritmo Expectation-Maximization (EM)** è determinare, mediante un processo iterativo, la stima a massima verosimiglianza  $\theta_{MLE}$  **in presenza di dati mancanti**, in questo caso la componente latente  $\mathbf{z}(\mathbf{s}, t)$ .

# I passi dell'algoritmo

Ogni iterazione dell'algoritmo EM consiste in due passi:

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

# La funzione di verosimiglianza

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

## Descrizione del dataset

Il dataset inerente il bike sharing è così composto:

- **variabile dipendente:** l'oggetto del caso di studio consiste nel numero di ritiri orari di biciclette presso le stazioni della rete di bike sharing;
- **variabili meteorologiche:** covariate spazio-invarianti come temperatura percepita, piovosità, visibilità orizzontale, velocità del vento e copertura nuvolosa;
- **variabili spaziali:** covariate tempo-invarianti quali la distanza dalla stazione ferroviaria più vicina e la densità demografica nella zona circostante;
- **variabili dummy:** eventi come il lockdown COVID-19, il ritorno alla vita quotidiana post-lockdown, i weekend e le festività federali sono modellati come variabili categoriche binarie.

# Metodologia

- 1 **Cross-validation con  $k = 5$  per determinare il valore del parametro  $\rho$ :** al fine di determinare il  $\rho$  ottimale, come metrica è stato impiegato il MSE:

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

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

- 2 **Scelta delle covariate più significative (model selection):** l'obiettivo è visualizzare le spline per i parametri  $\beta$  e i relativi intervalli di confidenza al fine di escludere i regressori poco significativi.

- 3 Validazione del modello finale tramite LOOCV:** per valutare la bontà complessiva del modello definitivo, è stato utilizzato il RMSE:

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

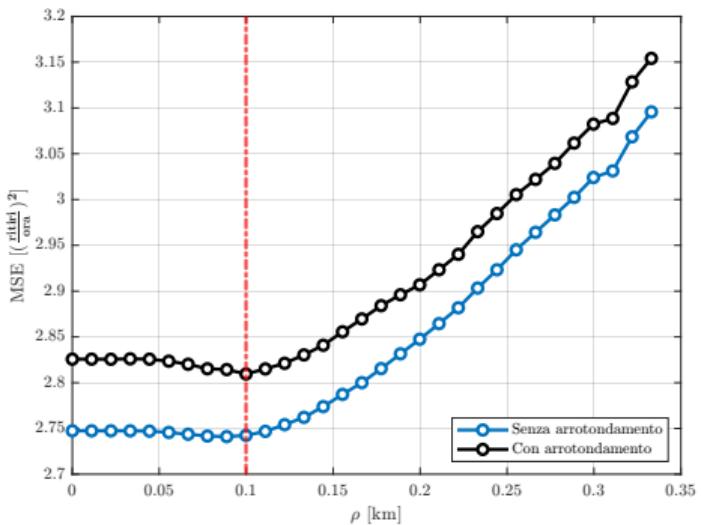
- 4 Previsione spaziale utilizzando il kriging:** è stato stimato il volume di noleggi orario così definito:

$$v(\mathbf{s}, l | \mathcal{T}) = \sum_{t \in \mathcal{T}} \hat{y}(\mathbf{s}, l, t), \quad \forall \mathbf{s} \in \mathcal{D}, \quad \forall l \in \mathcal{H}, \quad \mathcal{T} = 152, \dots, 213$$

Dopodiché, i 24 valori di  $v(\mathbf{s}, l | \mathcal{T})$  sono stati così raggruppati:

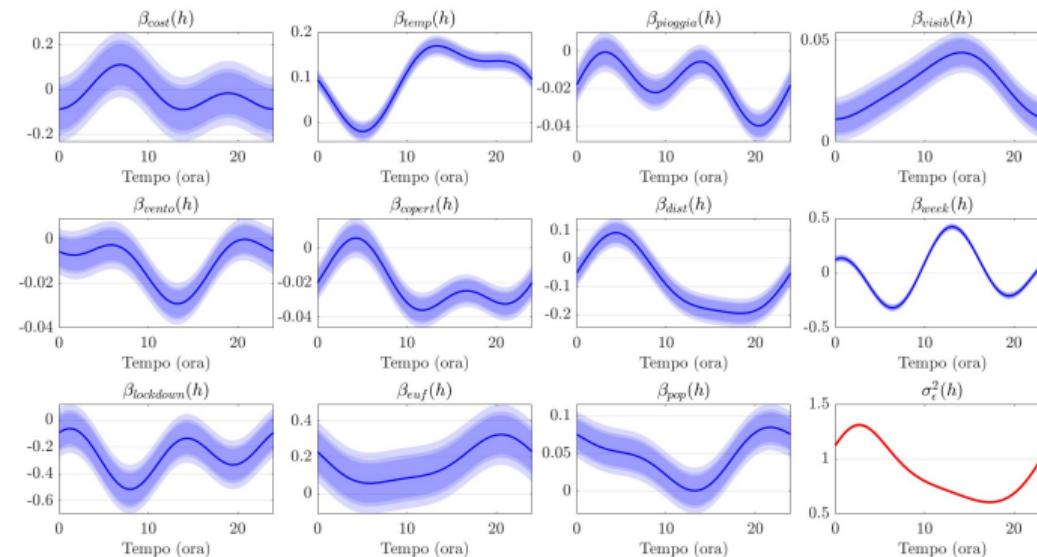
$$\bar{v}(\mathbf{s}, l | \mathcal{T})_i = \frac{1}{\text{card}(\mathcal{R}_i)} \sum_{j \in \mathcal{R}_i} v(\mathbf{s}, j | \mathcal{T}), \quad i = 1, \dots, 6$$

# Cross-validation per determinare il valore di $\rho$



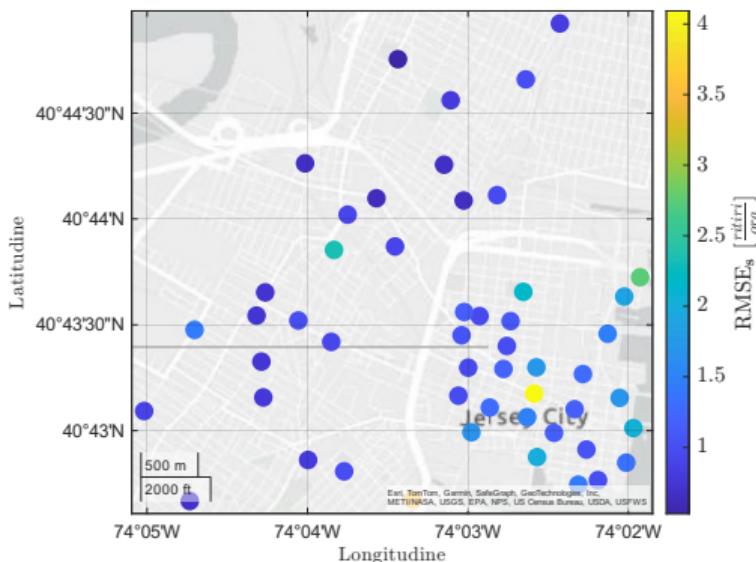
*Andamento del MSE in cross-validation con e senza arrotondamento al variare del parametro  $\rho$ .*

# Model selection



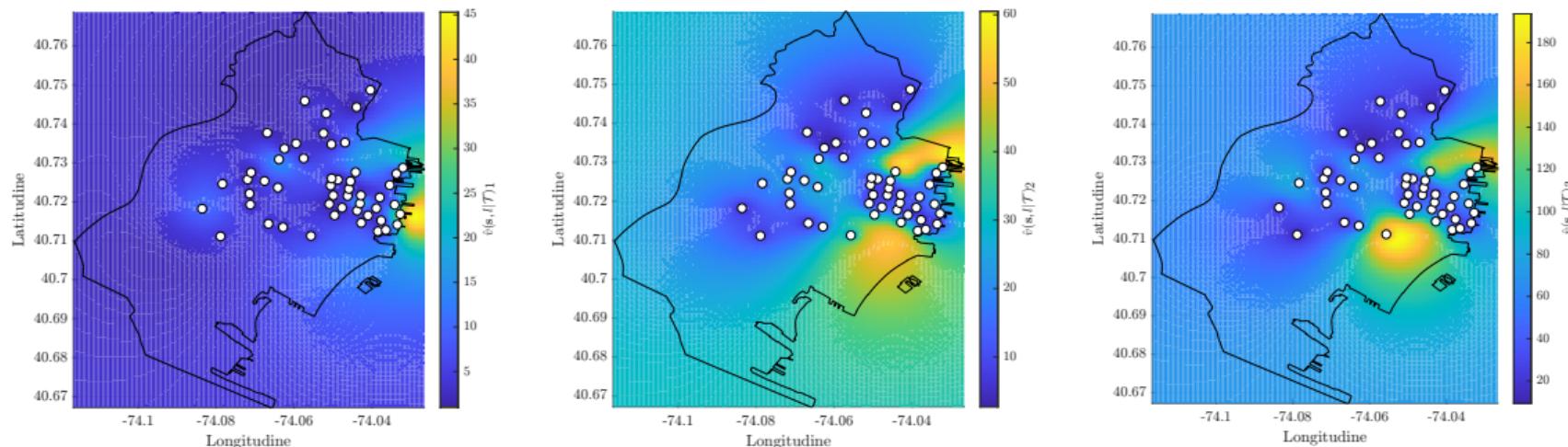
*Spline di Fourier per i coefficienti  $\beta$  e rispettivi intervalli di confidenza*

# Validazione del modello definitivo

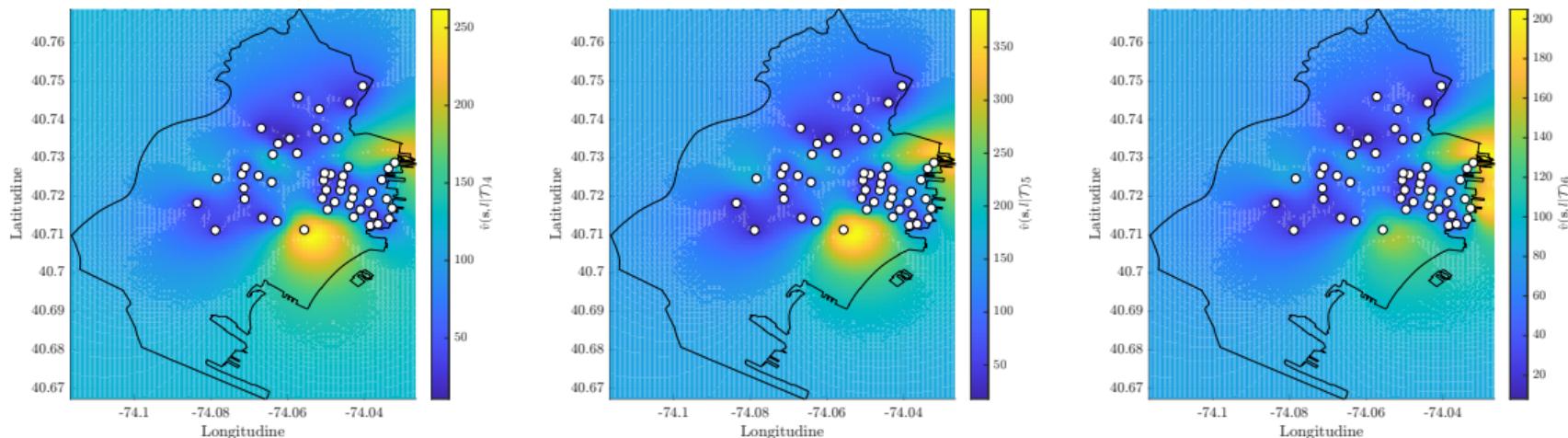


*Mappa della distribuzione del RMSE<sub>s</sub>*

# Previsione spaziale tramite Kriging



*Volume del numero di ritiri previsti, raggruppati per fasce orarie: dalle 00:01 alle 04:00 (sinistra), dalle 04:01 alle 08:00 (centro), dalle 08:01 alle 12:00 (destra).*



*Volume del numero di ritiri previsti, raggruppati per fasce orarie: dalle 12:01 alle 16:00 (sinistra), dalle 16:01 alle 20:00 (centro), dalle 20:01 alle 24:00 (destra).*

# Conclusioni

Alla luce dei risultati ottenuti:

- il presente studio evidenzia il valore aggiunto del nuovo modello proposto nel campo della modellazione funzionale spazio-temporale, distinguendosi dal modello f-HDGM grazie all'introduzione del parametro di interazione spaziale  $\rho$ ;
- una delle limitazioni riguarda il processo di stima del nuovo parametro in quanto la cross validazione richiede risorse computazionali significative;
- è necessaria la modifica dell'algoritmo EM affinché sia in grado di massimizzare la funzione di verosimiglianza  $-2 \ln L(\theta; Y, Z, X)$ , tuttavia ciò richiede ulteriori approfondimenti per risolvere questioni ancora aperte, come lo studio dell'identificabilità del modello.

Infine:

- per migliorare l'analisi sul bike sharing, si potrebbe considerare l'utilizzo di un dataset pluriennale per catturare la componente stagionale del fenomeno e l'adozione di variazioni metodologiche come il clustering dei punti di ritiro per permettere la stima di modelli spazio-temporali locali, riducendo l'assunzione che il parametro  $\rho$  sia spazio-invariante.



Grazie per l'attenzione