Facoltà di Ingegneria dell'Informazione, Informatica e Statistica

Corso di Laurea Magistrale in Scienze Statistiche

Prova Finale di Laurea Magistrale

Confronto tra Modelli Previsivi per il Mercato Immobiliare: Applicazione al caso degli Immobili di Madrid

SAPIENZA
UNIVERSITÀ DI ROMA

Relatore: Prof.ssa Cecilia Vitiello
Correlatore: Prof. Marco Alfò

Candidato: Romeo Silvestri

Anno accademico 2021 - 2022

Il Problema dei Prezzi degli Immobili

- Il mercato immobiliare è una parte vitale dell'economia di un paese che si occupa della costruzione, della gestione e della compravendita di beni immobili

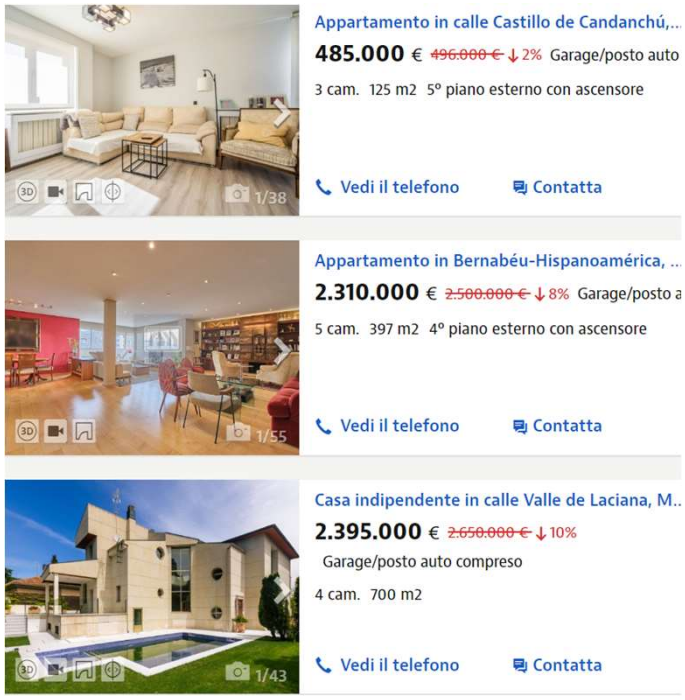
- L'obiettivo è quello di valutare e prevedere il prezzo di vendita degli immobili in un dato momento storico conoscendo alcune caratteristiche dell'abitazione

- Il lavoro si incentra sullo studio del mercato immobiliare di Madrid

Dataset degli Immobili di Madrid

Il dataset utilizzato è composto da 6287 abitazioni di Madrid, su cui sono state misurate inizialmente un totale di 43 variabili utili.

La costruzione del dataset è avvenuta ricavando le informazioni contenute nei principali portali immobiliari spagnoli nel mese di marzo 2020.



a) Sito di annunci immobiliari di Madrid

Prezzo	Superficie	Tipo	Ascensore
485.000 €	125 m²	Appartamento	Si
2.310.000 €	397 m²	Appartamento	Si
2.385.000 €	700 m²	Casa Indipendente	No

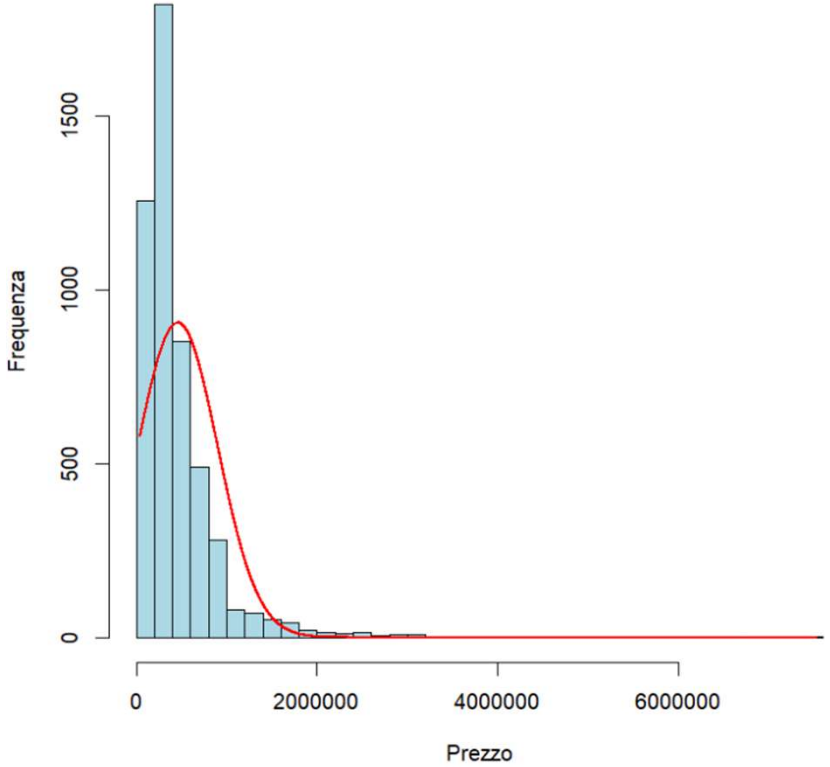
b) Subset di dati corrispondente

Analisi Esplorativa

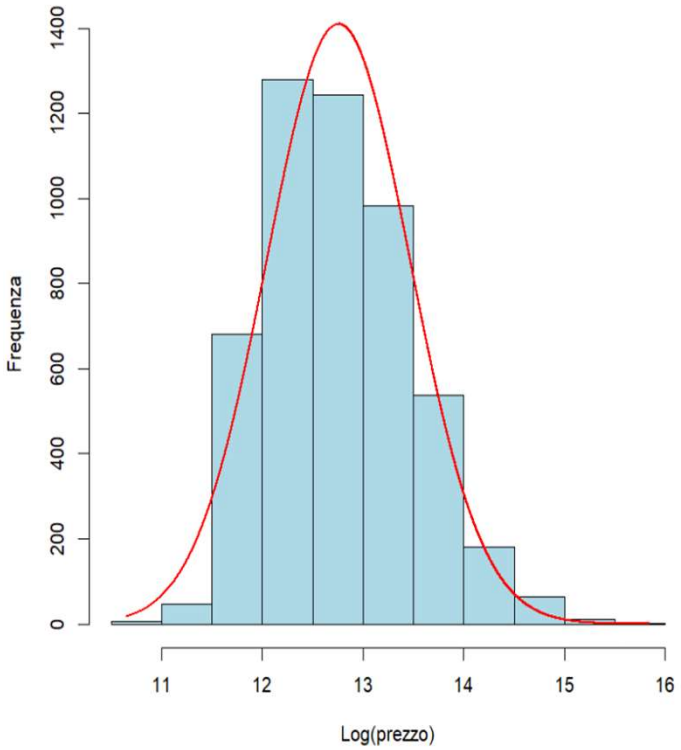
Prezzo di Vendita

Prezzo Medio: 460.000 euro

Range di Prezzo: 42.000 – 7.500.000 euro



a) Istogramma del prezzo di vendita



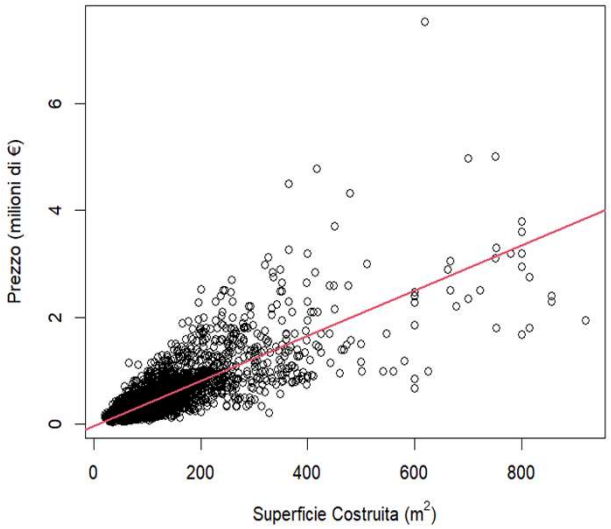
b) Istogramma del logaritmo del prezzo di vendita

Altre Variabili

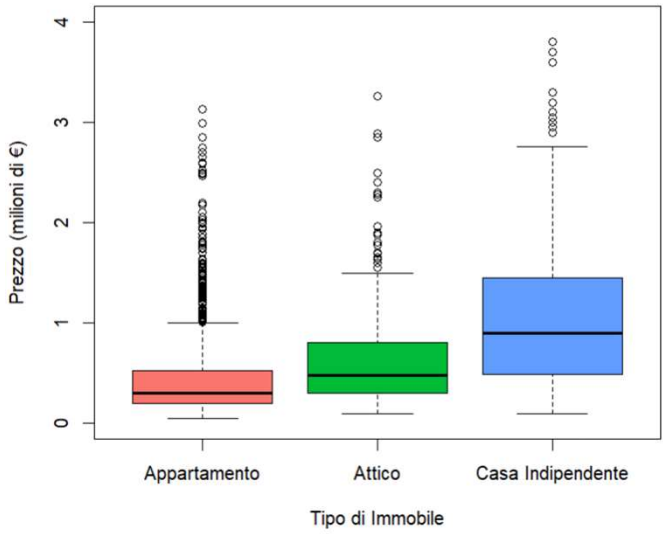
- 11 variabili numeriche
- 32 variabili categoriche:
  - caratteristiche dell'immobile
  - presenza/assenza di un attributo

Pre-Processing

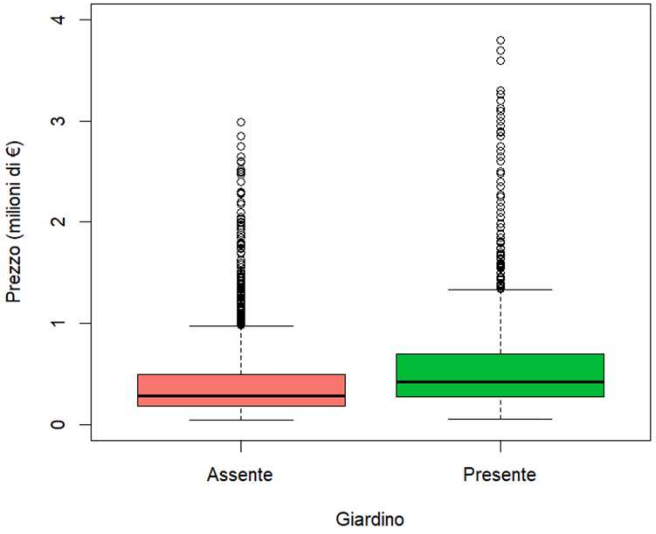
a)



b)



c)



a) Diagramma di dispersione tra la superficie costruita e il prezzo di vendita del prezzo di vendita
b) Boxplot divisi per categoria tra il tipo di immobile e il prezzo e c) tra il giardino e il prezzo

Dati Mancanti

Meccanismo di Risposta

Data la matrice dei dati $X = \{X_{oss}, X_{mis}\}$, la matrice indicatrice di risposta R che indica la presenza/assenza di un'osservazione, e i parametri ψ che descrivono le probabilità di essere mancante, allora il meccanismo di risposta è:

$$M = P(R = 0|X_{oss}, X_{mis}, \psi)$$

Classificazione di Rubin:

A. Missing not at Random (MNAR) $\quad M = P(R = 0|X_{oss}, \psi)$
B. Missing at Random (MAR) $\quad M = P(R = 0|X_{oss}, \psi)$
C. Missing Completely at Random (MCAR)

MAR o MCAR

$$P(X|X_{oss}, R = 0) = P(X|X_{oss}, R = 1)$$ Ignorabilità

Trattamento dei Dati Mancanti

I metodi per gestire i dati mancanti si possono basare sui soli dati completi, sui dati disponibili o sull'imputazione dei dati.

Per applicare al meglio i modelli previsivi su un insieme di dati omogeneo e flessibile, si decidono di utilizzare le tecniche di imputazione, come:

- Regressione
- Matching

Imputazione Singola

a. Facilità d'interpretazione
b. Adattabilità
c. Costo computazionale ridotto

Imputazione Multipla

a. Gestione dell'incertezza
b. Riduzione del bias
c. Utilizzo di tutte le informazioni

Specificazione Completamente Condizionale

Selezione del Metodo di Imputazione

Confrontare i metodi di imputazione è un problema non indifferente.
Gli usuali indicatori di errore previsivo non possono essere utilizzati.

I-Scores: assegna un punteggio ai metodi imputativi in base alla capacità di riprodurre fedelmente le distribuzioni condizionate dei dati osservati.

Metodo di Imputazione	I-Score
Random Forest	0
Albero Decisionale	-0,202
Predictive Mean Matching (PMM)	-0,405
PMM + Logistica	-0,808
Regressione Stocastica + Logistica	-0,998
Regressione Lineare + Logistica	-1,723
Moda, Media e Mediana	-2,163*

a) I-Scores per vari metodi di imputazione

Variabili	% Missing	% RMSE perso
Numero Bagni	0,1%	39,5%
Superficie Costruita	0,1%	15,1%
Nuova Costruzione	2,3%	18,6%
Ascensore	5,3%	26%
Esterno	8%	8,1%
Piano d'Ingresso	8,5%	12,1%
Classe Energetica	32%	8,2%
Riscaldamento Autonomo	48,5%	10,4%

b) Variabili con valori mancanti e relativa importanza

Componente Spaziale

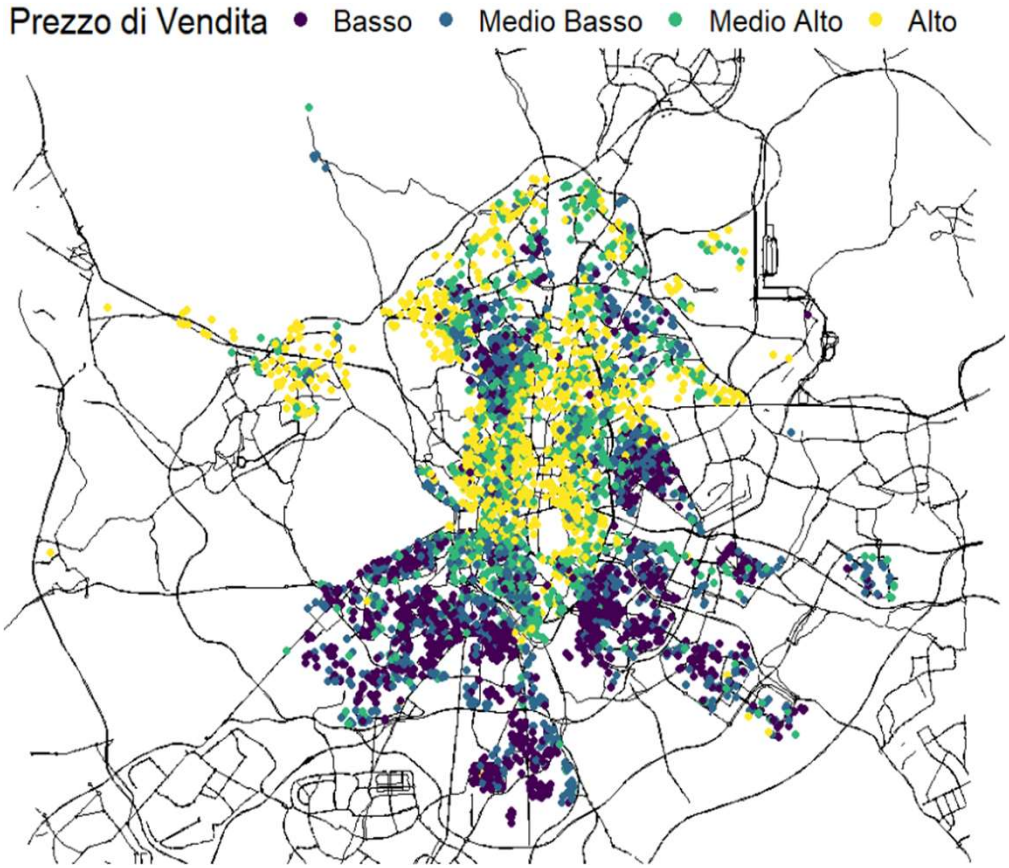


Figura: Mappa dei prezzi di vendita di Madrid

Tipi di Effetti Spaziali:

- Autocorrelazione Spaziale
- Eterogeneità Spaziale

Per elaborare le informazioni spaziali è necessario geocodificare gli indirizzi delle abitazioni (software ArcGis) e definire una misura di distanza come quella di Haversine o di Vincenty

Regressione Spaziale

Analisi Esplorativa Spaziale (ESDA)

La struttura di vicinanza stabilisce la relazione tra le osservazioni nello spazio.
Data una misura di distanza *d*, si definisce la matrice dei pesi spaziali *W*:

$$W = \left\{ w_{ij} : w_{ij} = \frac{1}{d_{ij}}, i = 1, \dots, n, j = 1, \dots, n \right\}$$ con d_{ij} distanza di Haversine tra *i* e *j*

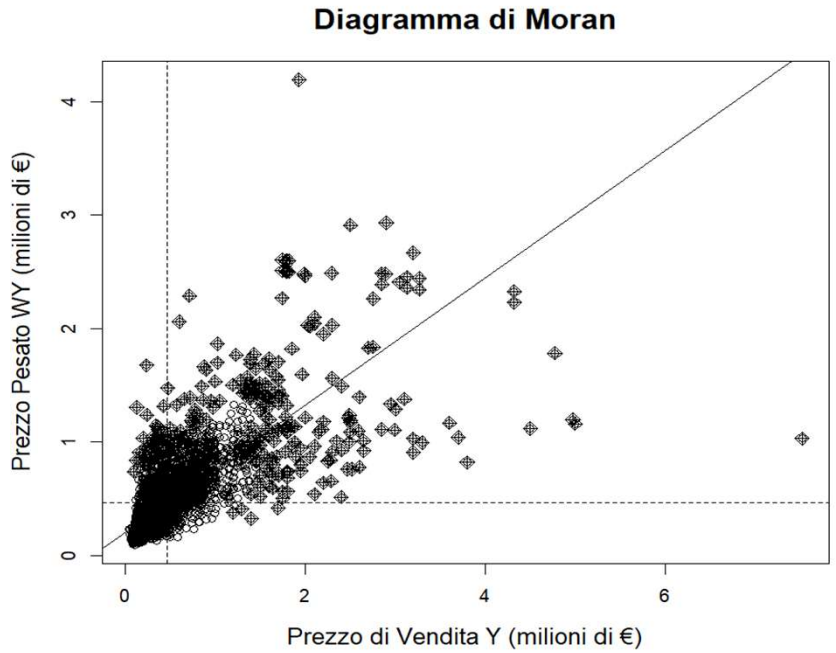


Figura: Diagramma di Moran per i Prezzi di Vendita di Madrid

Indice di Moran

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

con w_{ij} peso spaziale tra i e j,
$\bar{y}$ media della variabile risposta,
y_i valore i-esimo della risposta

Modelli Previsivi Spaziali

Una classe di modelli spaziali ampiamente utilizzata per trattare l'autocorrelazione spaziale è quella dei modelli autoregressivi simultanei (SAR).

A. Modello di Lag Spaziale $$Y = X\beta + \rho WY + \epsilon$$

B. Modello di Lag Spaziale sulle X $$Y = X\beta + WX\theta + \epsilon$$

C. Modello di Errore Spaziale $$Y = X\beta + \lambda Wu + \epsilon$$

D. Modelli Derivati

dove β è il vettore dei coefficienti di regressione, W la matrice dei pesi spaziali, ρ , θ , λ i parametri spaziali, u il vettore degli errori spaziali e ϵ il vettore degli errori indipendenti e normalmente distribuiti

Selezione del Modello Spaziale

Test di Moran: si basa sull'indice di Moran e verifica la presenza di autocorrelazione spaziale

Test dei Moltiplicatori di Lagrange: verifica la possibilità di introdurre un dato parametro statistico

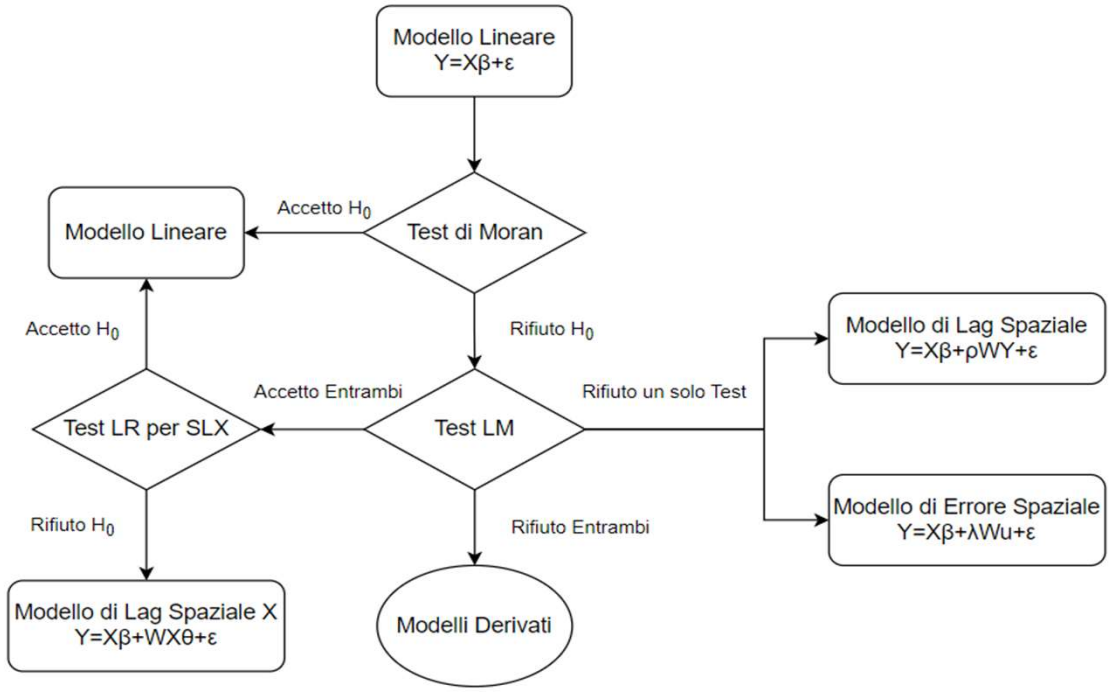


Figura: Diagramma di Flusso per la Selezione del Modello Spaziale

Selezione del Modello Spaziale per Madrid:

1) Stima del Modello Regressivo Lineare (Selezione delle Variabili)

2) Scelta della matrice dei pesi spaziali *W*

3) Test di Moran sui residui: test significativo con *p*-value < 0,001

4) Test LM per i parametri ρ e λ : test entrambi significativi

5) Test LM robusti per ρ e λ : il parametro λ è più significativo

6) Stima del Modello con Errore Spaziale (SEM)

7) Test di Moran sui residui: test non-significativo con *p*-value = 0,337

seleziono il modello SEM

Regressione Spaziale Implicita

Variabili Spaziali

Identificano un particolare sottomercato spaziale.

A. Aree Amministrative: come i quartieri o i distretti

B. Cluster Spaziali: gruppi di immobili vicini con caratteristiche interne simili

C. Cluster LISA: gruppi di immobili vicini con prezzi di vendita simili; si possono costruire a partire dall'indice di Moran locale

$$LISA_i = I_i = (y_i - \bar{y}) \sum_{j=1}^{n} w_{ij} (y_j - \bar{y})$$

con w_{ij} peso spaziale *ij* , $\bar{y}$ media della variabile risposta, y_i valore *i*-esimo della risposta

i Cluster LISA hanno mostrato capacità previsive migliori

Variabili di Distanza

Sono complementari alle variabili spaziali ed esprimono la distanza minima tra gli immobili e i luoghi d'interesse della città.

Per trovare i punti d'interesse si è utilizzato il database di OpenStreetMap, ovvero un database geografico collaborativo e libero che contiene dati spaziali dettagliati di tutto il mondo.

Sono state misurate 19 diverse variabili, raggruppabili in 6 categorie:

1) Salute/Benessere
2) Finanza
3) Educazione
4) Trasporti
5) Intrattenimento
6) Turismo

Modelli Previsivi Tradizionali

- Regressione Parametrica
- Regressione Non-Parametrica

I modelli parametrici richiedono numerose trasformazioni per ottenere risultati previsivi simili ai principali modelli non-parametrici.

La selezione dei modelli è stata eseguita tramite una procedura di 10-fold cross-validation utilizzando l'*RMSE* come indicatore di errore.

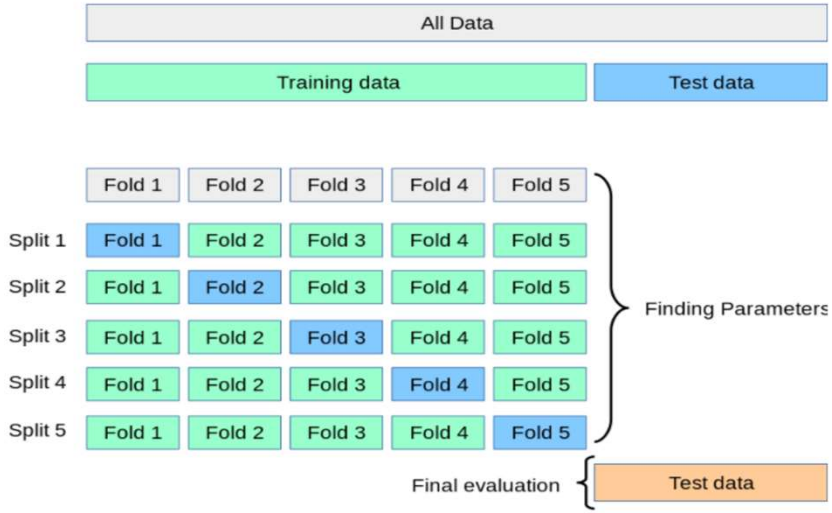
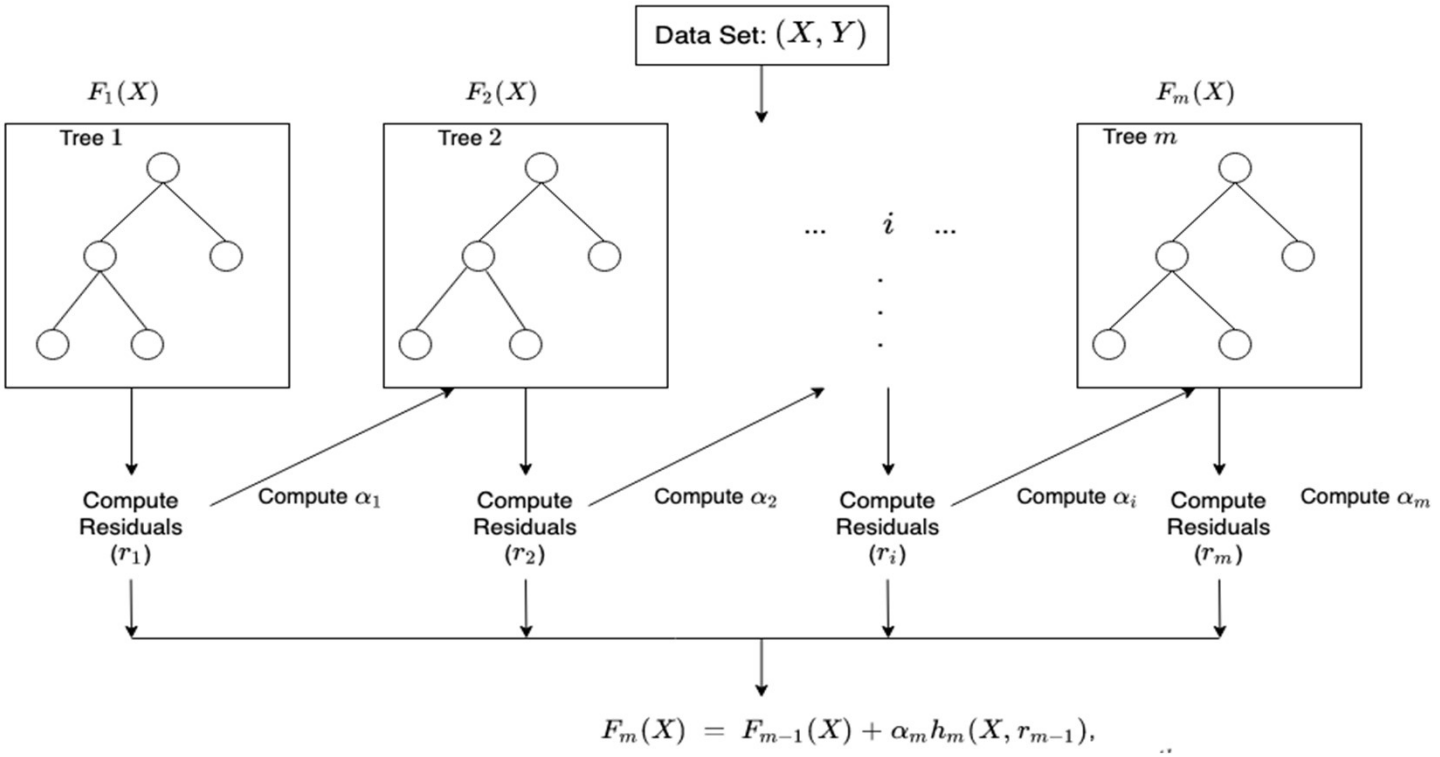


Figura: Esempio di 5-fold cross-validation

XG-Boost



L'XG-Boost regressivo è un modello che utilizza un insieme (ensemble) di alberi decisionali costruiti sequenzialmente sulla base dei residui dei precedenti.

L'algoritmo minimizza ad ogni iterazione una specifica funzione di perdita.

Vengono impiegate numerose tecniche di regolarizzazione.

Confronto tra Modelli

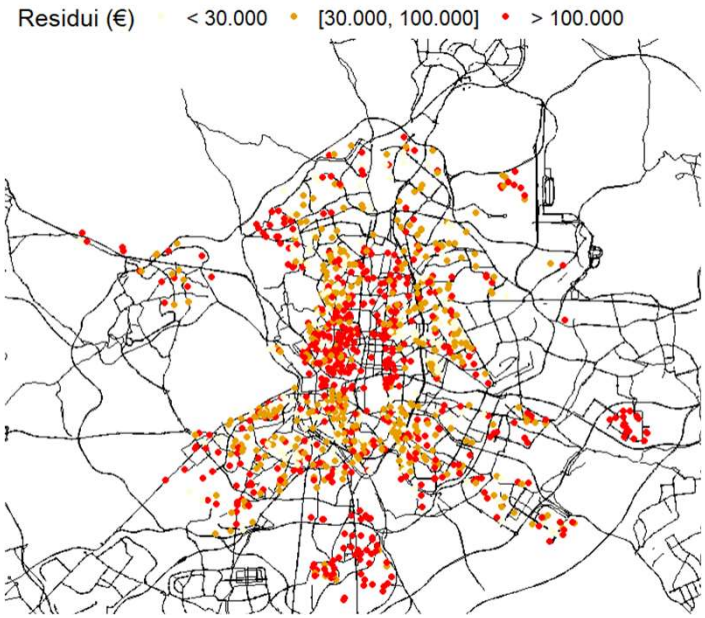
Regressione Spaziale

i. Utilizza dei parametri spaziali

ii. I Modelli SAR descrivono soltanto l'autocorrelazione spaziale

iii. Discreto livello di interpretabilità

iv. Buon adattamento ai dati, ma scarse capacità previsive



RMSE
242.000 €

a) Mappa dei residui del modello SEM

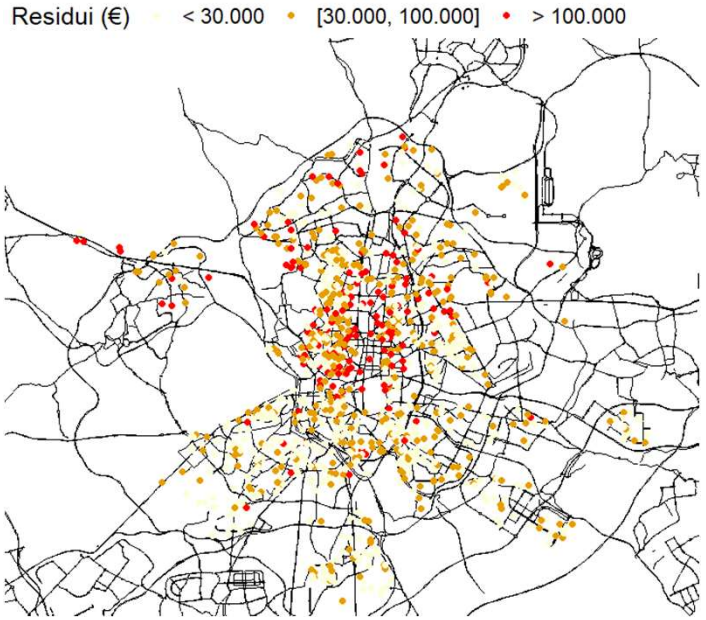
Regressione Spaziale Implicita

i. Utilizza delle variabili spaziali

ii. Possono descrivere l'autocorrelazione e l'eterogeneità spaziale

iii. L'interpretabilità dipende dal modello

iv. I modelli di ensemble di alberi hanno ottenuto i risultati previsivi migliori



RMSE
142.000 €

b) Mappa dei residui del modello XG-Boost

Conclusioni

Modello	Train Error	Test Error
XG-Boost	**48.617**	**142.482**
Random Forest	66.186	160.542
GLM (gaussiano)	154.540	165.623
KNN	165.629	200.389
Lineare Completo	202.158	208.429
Spaziale (SEM)	182.558	242.303
Lineare Ridotto	242.793	246.129

L'XG-Boost che tiene conto dei cluster di tipo LISA è il modello che prevede più accuratamente i prezzi di vendita.

Ha un *RMSE* di circa 142.000 €.

In termini di *MAE*, il modello finale sbaglia in media di 63.000 €.
Si riduce a 25.000 € per gli immobili che hanno un prezzo inferiore a 500.000 €.

In sintesi:

- Per il mercato immobiliare, la gestione dei dati mancanti è importante al fine di ottenere un dataset che permetta di applicare i modelli previsivi
- E' preferibile adottare un approccio basato sulla regressione spaziale implicita
- Si possono apportare ancora numerosi miglioramenti in termini di prestazioni

Figura: Confronto tra i Modelli Previsivi

Scuola Magistrale in Scienze Statistiche
Facoltà di Ingegneria dell'Informazione, Matematica e Statistica
Anno Accademico 2021-2022

Romeo Silvestri

Grazie per la Vostra Attenzione!

SAPIENZA
UNIVERSITÀ DI ROMA