Politecnico di Torino

Dipartimento di Ingegneria Gestionale e della Produzione

Laurea Magistrale in Ingegneria Gestionale

PROVA FINALE

A LASSO-integrated algorithm for endogenous competition modeling and topological segmentation in a high-dimensional short-term rental market

Relatore:
Prof. Francesco Luigi Milone

Co-Relatore:
Prof. Luigi Buzzacchi

Candidato: *Anna Lavagnino*

ANNO ACCADEMICO 2024/2025

"Se non fossi mai partita
forse avrei risparmiato
mille viaggi avanti e indietro
e pure qualche pianto.
Ma invece sono partita,
e che sorrisi che ho trovato.
Oggi so sarebbe fiera
anche me stessa del passato.

Nuovi amici, nuove cicatrici,
una valigia sempre pronta a metà,
lo spazzolino di scorta
e una fetta di torta di mamma
di tre mesi fa.
Nuovi amori, troppi nuovi errori..."

A tutti i nuovi amici e l'amore che l'università mi ha dato, agli errori fatti, a Torino, la città che mi ha reso chi sono oggi.

Indice

Elenco delle figure

Elenco delle tabelle

Elenco degli acronimi

2SRI Two-Stage Residual Insertion

ALASSO Adaptive Least Absolute Shrinkage and Selection Operator

API Application Programming Interface

CV Cross Validation

GFL Generalized Fused Lasso

HTTP HyperText Transfer Protocol

IQR Interquartile Range

KDE Kernel Density Estimation

KNN k-Nearest Neighbors

LASSO Least Absolute Shrinkage and Selection Operator

MSE Mean Squared Error

OLS Ordinary Least Squares

SLL Sistemi Locali del Lavoro

SLT Sistemi Locali del Turismo

Abstract

La crescente diffusione degli affitti brevi ha reso necessario un approccio quantitativo capace di individuare i meccanismi di formazione del prezzo e i confini effettivi della competizione locale. La tesi analizza il mercato Airbnb in Piemonte proponendo un algoritmo integrato basato su regressione LASSO per isolare la componente strutturale del prezzo e identificare, attraverso i residui, le interazioni competitive tra strutture geograficamente prossime.

A partire dalle interdipendenze residue viene costruita una rete competitiva spaziale non direzionale, sulla quale un'analisi di modularità consente di individuare cluster autocontenuti di competizione. I risultati evidenziano una marcata eterogeneità territoriale e propongono un *framework* replicabile per la definizione *data-driven* dei confini competitivi nei mercati digitali frammentati con implicazioni per le strategie di pricing e posizionamento degli host.

Capitolo 1

Introduzione

Introducendo modelli di concorrenza caratterizzati da un'elevata trasparenza informativa, le piattaforme digitali hanno modificato l'interazione tra domanda e offerta nei mercati dei servizi. Nel contesto degli affitti brevi, tali caratteristiche si traducono in una struttura competitiva nella quale la semplice prossimità geografica non costituisce condizione sufficiente per definire il perimetro del mercato. Dunque, il problema centrale affrontato nella tesi riguarda l'identificazione delle relazioni di concorrenza tra unità di offerta, ovvero le strutture, in un contesto caratterizzato da eterogeneità qualitativa, in quanto comprendere quali strutture esercitino un'influenza reciproca nella determinazione delle scelte degli utenti e delle strategie di prezzo è fondamentale per delineare confini competitivi.

La letteratura ha offerto contributi significativi sia nella comprensione dei determinanti del prezzo degli affitti brevi, sia nella comprensione delle dinamiche strategiche di interazione tra gli operatori in contesti spazialmente distribuiti. Tuttavia, i modelli di prezzo tendono a concentrarsi sulle caratteristiche osservabili dell'offerta, mentre le analisi competitive si basano frequentemente su aree territoriali definite a priori, lasciando aperta la questione di come integrare in modo coerente la dimensione strutturale e la dimensione relazionale.

La tesi si inserisce in questo contesto di ricerca proponendo lo sviluppo di un algoritmo basato su regressione LASSO, volto a isolare la componente strutturale del prezzo e a utilizzare le informazioni residue per ricostruire le interdipendenze tra strutture potenzialmente concorrenti. L'obiettivo non è soltanto stimare un modello di prezzo, ma derivare in modo endogeno la configurazione competitiva del mercato, identificando aree di competizione emergenti dai dati. In tal modo, il quadro metodologico mira a comprendere la configurazione sistemica del mercato: come si organizzano le interazioni tra strutture?

quali aggregazioni emergono spontaneamente? in che misura tali aggregazioni coincidono o divergono rispetto alle suddivisioni amministrative? L'approccio è concepito per essere replicabile e scalabile in altri contesti territoriali o mercati digitali con caratteristiche analoghe.

L'applicazione al contesto piemontese consente di identificare in modo endogeno le aree competitive effettive all'interno della zona di interesse, superando suddivisioni territoriali predefinite; proprio per la sua struttura metodologica, l'algoritmo sviluppato è progettato per essere scalabile e replicabile in altri contesti geografici o mercati digitali con caratteristiche analoghe, rendendo l'approccio trasferibile oltre il caso studio analizzato.

Capitolo 2

Presentazione del caso studio

Il seguente capitolo è dedicato alla presentazione del contesto in cui si inserisce il caso studio oggetto dell'analisi empirica. A tale scopo, si intende delineare le caratteristiche del settore e dell'ambiente competitivo entro cui la metodologia sviluppata nel Capitolo 4 verrà applicata, consentendo una corretta applicazione di quest'ultima, e una conseguente interpretazione del valore dei risultati.

2.1 Settore dell'hospitality

In primo luogo, viene analizzato il settore dell'*hospitality*, in quanto rappresenta l'industria economica di riferimento per il caso studio. In particolare, si vuole evidenziare l'evoluzione verso un modello sempre più mediato da piattaforme digitali, all'interno del quale gli affitti brevi rappresentano un caso emblematico.

Tradizionalmente, il termine *hospitality* viene definito come un atto di accoglienza volto a offrire alloggio, cibo e bevande a ospiti o visitatori, accompagnato da un atteggiamento di generosità e buona volontà. Nel contesto economico, assume il significato di "*uno scambio umano contemporaneo, volontariamente intrapreso, progettato per migliorare il benessere reciproco delle parti coinvolte attraverso la fornitura di alloggio e/o cibo e/o bevande*" (Brotherton 1999). In questa prospettiva, l'*hospitality* non si limita alla mera erogazione di un servizio, ma implica una relazione sociale basata sulla reciprocità, sulla fiducia e sull'esperienza condivisa tra *host* e *guest* (Lashley 2015).

Tale mercato è caratterizzato da un elevato numero di venditori tra loro in competizione. Quindi, sulla base della definizione di competizione proposta dal *Cambridge Dictionary*

(2025)[1], si tratta di un mercato in cui ciascun fornitore del servizio mira a vincere o ad avere maggiore successo degli altri. Tuttavia, essendoci un'elevata concentrazione di concorrenti, gli agenti che competono non reagiscono alle azioni di ciascun singolo attore (Li, Netessine e Koulayev 2018).

Tale casistica si allontana dai modelli solitamente studiati in lettura, in quanto il numero di attori è maggiore di pochi, ma non si tratta di concorrenza perfetta, in quanto, seppur in modo limitato, gli albergatori e gli host possono distinguersi.

Nel settore dell'*hospitality* tradizionale, gli elementi di differenziazione tra i *competitor* si fondano principalmente su caratteristiche verticali come la qualità dei servizi offerti, la categoria e la disponibilità di infrastrutture opzionali; tali caratteristiche rendono i servizi ordinabili lungo una dimensione di qualità condivisa dalla totalità dei consumatori. A tali aspetti si aggiungono elementi orizzontali di natura soggettiva che contribuiscono a rendere un servizio distinto dai concorrenti anche a parità di fascia di mercato, in quanto la valutazione dipende dall'eterogeneità delle preferenze individuali (Mazzeo 2002; Denizci Guillet et al. 2026).

Un ulteriore fattore cruciale è la posizione geografica, che influenza la percezione di valore per il cliente. Infatti, da un lato, l'agglomerazione con altri servizi ha effetti indiretti di incremento della domanda potenziale dovuti ad una maggiore attrattività della zona. Dall'altro lato, la distanza dai *competitor* incide sulla possibilità di applicare strategie di prezzo differenziate (Park, Kim e Frye 2022).

2.1.1 Sharing economy nell'hospitality

Le dinamiche dell'*hospitality* sono ad oggi influenzate dalla digitalizzazione dell'economia, che ha portato ad accorciare la distanza dello scambio economico tra i fornitori di servizi e i loro utenti, tramite l'introduzione di piattaforme globalmente diffuse (Hall et al. 2022). Tali evoluzioni sono studiate nella *Sharing Economy*, un modello economico basato sulla condivisione di risorse sottoutilizzate per ottenere benefici monetari e non monetari, incentrato principalmente sulle transazioni *peer-to-peer* (Mukhopadhyay e

[1] `https://dictionary.cambridge.org/dictionary/english/competition`

B.K.Mukhopadhyay 2021). Dove il concetto di ***peer-to-peer*** si riferisce ad un modello di interazione diretta tra individui, ovvero i *peer*, che agiscono contemporaneamente come fornitori e fruitori di beni o servizi, senza l'intermediazione tradizionale di un'impresa produttiva o distributiva . Nelle piattaforme digitali, tale relazione è mediata da infrastrutture tecnologiche che facilitano la fiducia, la reputazione e la transazione economica tra gli utenti.

Questo nuovo approccio ridefinisce il ruolo del consumatore, che diviene al contempo produttore e parte attiva nello scambio, generando nuove forme di valore (Hamari, Sjöklint e Ukkonen 2016).

Nello specifico, nel settore dell'*hospitality*, tale cambiamento è stato guidato da Booking.com, ed Airbnb, entrambe piattaforme digitali che hanno modificato la relazione tra gli attori nel settore del turismo (Hall et al. 2022). Dato l'interesse della tesi, il focus viene posto sulla piattaforma di Airbnb, tramite un approfondimento nella sezione successiva.

2.2 Piattaforma di studio: Airbnb

Ad oggi, Airbnb si configura come uno degli esempi più noti di piattaforme di *sharing economy* nel settore dell'*hospitality*. Il servizio consente a soggetti privati, tradizionalmente considerati meri fruitori di servizi turistici, di assumere un ruolo attivo come fornitori di alloggi, mettendo in affitto abitazioni o singole stanze. Questo modello ha dato origine ad un mercato fortemente frammentato e caratterizzato da dinamiche competitive prevalentemente localizzate, rendendo la piattaforma un contesto di analisi adatto allo studio dei meccanismi di formazione del prezzo e della competizione a livello micro-locale (Zervas, Proserpio e Byers 2017).

2.2.1 Evoluzione della piattaforma

Airbnb nasce nel 2008 a San Francisco, contesto geografico segnato da una crescente domanda di soluzioni abitative flessibili per i viaggiatori urbani. L'idea originaria fu concepita da Brian Chesky, Joe Gebbia e successivamente da Nathan Blecharczyk, i quali, per far fronte all'aumento del costo della vita, decisero di affittare spazi nel proprio appartamento a visitatori di una conferenza, offrendo materassini ad aria e colazione. Questo esperimento iniziale diede vita alla piattaforma "AirBed & Breakfast", che poi divenne semplicemente Airbnb, comunemente riconosciuta tramite il logo riportato in Figura 2.1.

Figura 2.1. Logo Airbnb

Consentendo ai privati cittadini di diventare *host*, sono state superate le barriere di accesso

delle strutture alberghiere convenzionali. Infatti, il potere della *sharing economy*, ha consentito alla piattaforma di accellerare la trasformazione digitale nel turismo, la quale è risultata in una riduzione dei costi di transazione ed un aumento della capacità ricettiva urbana, oltre a consentire una maggiore personalizzazione dell'esperienza di soggiorno (Guttentag 2015; Williamson 1985).

2.2.2 Analisi del mercato

Alla luce degli obiettivi della tesi, l'analisi del mercato della piattaforma assume un ruolo centrale per comprendere il contesto competitivo in cui operano gli *host*. In questa prospettiva, la dimensione geografica rappresenta una variabile chiave, in quanto consente di osservare come la distribuzione dell'offerta e la disponibilità dei dati varino significativamente tra paesi e aree urbane. Per tali ragioni, l'analisi delle statistiche di mercato viene impostata ponendo le aree geografiche come parametro di raggruppamento.

In questo senso, la Figura 2.2 fornisce una prima evidenza quantitativa della distribuzione globale degli annunci Airbnb, mostrando come l'Italia si collochi al quarto posto per numero di strutture disponibili in valore assoluto, nonostante abbia una popolazione inferiore rispetto a paesi quali Stati Uniti e Cina[2]. Questo dato segnala una presenza rilevante della piattaforma nel contesto italiano, suggerendo un'elevata intensità competitiva e una significativa disponibilità di osservazioni.

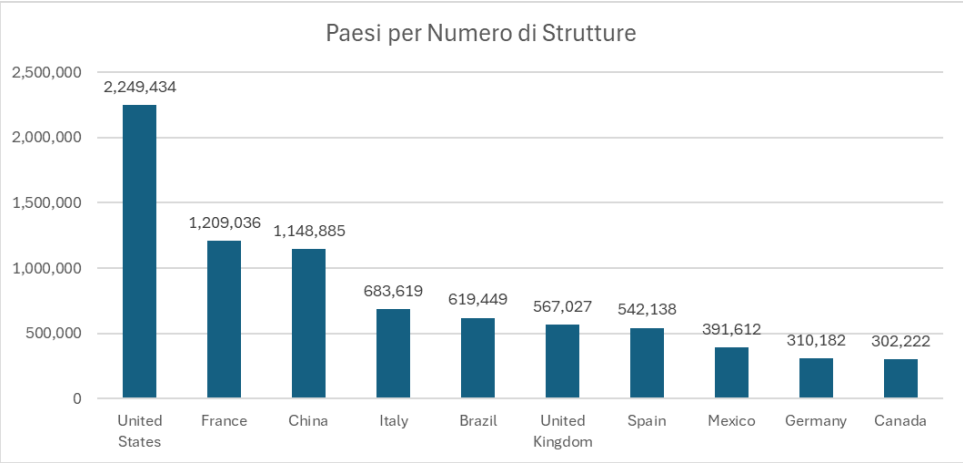


Figura 2.2. Paesi per numero di strutture Airbnb

[2]https://www.demandsage.com/airbnb-statistics/?

Ne deriva che il mercato italiano rappresenta un contesto di analisi adatto agli obiettivi della tesi, in quanto consente di disporre di un ampio dataset e di osservare in modo accurato le dinamiche di competizione e di formazione del prezzo che verranno approfondite nella sezione successiva.

Una volta riscontrata l'elevata numerosità di record disponibili a livello nazionale, l'analisi viene ulteriormente raffinata concentrando l'attenzione sulle principali città turistiche italiane[3]. Questo passaggio consente di osservare in maggiore dettaglio le differenze locali in termini di diffusione delle strutture e di performance del mercato degli affitti brevi, fornendo un contesto più mirato per la validazione dell'area di studio.

I dati riportati in Figura 6.13 offrono una panoramica comparativa delle città italiane in relazione al numero di strutture Airbnb e al tasso di occupazione medio. In tale contesto, Torino, città oggetto dell'analisi empirica, si colloca in una posizione intermedia rispetto alle principali destinazioni turistiche, registrando valori nella media sia in termini di offerta disponibile sia di tasso di occupazione. Questo risultato conferma l'adeguatezza della scelta del contesto di studio, consentendo di evitare situazioni estreme caratterizzate da outlier positivi o negativi che potrebbero compromettere la generalizzabilità dei risultati.

Figura 2.3. Città italiane per numero di strutture Airbnb e tasso di occupazione medio

[3] `https://airbtics.com`

Quindi, le caratteristiche statistiche della città di Torino precedentemente riportate pongono le basi per un'analisi territoriale robusta, in grado di fornire risultati interpretabili e potenzialmente generalizzabili, estendendoli anche alla regione di cui è capoluogo.

2.2.3 Dimensioni di differenziazione

Al fine di sviluppare un'adeguata metodologia, risulta necessario comprendere in che modo gli annunci presenti sulla piattaforma Airbnb si differenzino tra loro all'interno dello stesso contesto territoriale.

A tal proposito, la piattaforma Airbnb rende osservabili una pluralità di attributi che contribuiscono a distinguere i singoli *listing* [4], riflettendo sia caratteristiche intrinseche dell'alloggio sia elementi legati al comportamento degli *host* e all'interazione con gli utenti. Tali dimensioni di differenziazione risultano particolarmente rilevanti nel contesto collaborativo *peer-to-peer*, dove la comparabilità diretta tra offerte è facilitata dall'interfaccia della piattaforma.

Alla luce di queste considerazioni, la Tabella 2.1 sintetizza le principali variabili di differenziazione individuate in letteratura, che costituiscono la base interpretativa per l'analisi delle variabili presenti nel dataset utilizzato nel caso studio (H. Zhang, Zach e Xiang 2024a; H. Zhang, Zach e Xiang 2024b).

Tabella 2.1. Variabili di differenziazione nel contesto Airbnb

Fattore	**Descrizione**
Reputation signals	
Rating medio	Valutazione media complessiva della struttura da parte degli ospiti.
Numero di recensioni	Indicatore della popolarità e dell'esperienza percepita del *listing*.

Continua nella pagina successiva

[4] Nel contesto di Airbnb, il termine *listing* indica la scheda digitale che rappresenta l'unità di offerta sulla piattaforma (Zervas, Proserpio e Byers 2017).

Tabella 2.1. (continua)

Fattore	**Descrizione**
Superhost status	Etichetta ufficiale che segnala performance dell'*host* superiori alla media.
Tasso di risposta	Percentuale di messaggi ai quali l'host risponde rapidamente.
Strategie di prezzo	
Volatilità del prezzo	Ampiezza delle variazioni di prezzo nel tempo come risposta alla concorrenza.
Adattamento temporale	Prezzi diversificati in base a stagione, giorno della settimana o eventi.
Condizioni del mercato locale	
Numero di host attivi	Competizione diretta nel quartiere o micro-area.

Le ***strategie di prezzo*** costituiscono, se considerate come macro-categoria, un elemento centrale anche nel settore dell'*hospitality* tradizionale. Tuttavia, nel contesto Airbnb, le modalità di implementazione di tali strategie risultano profondamente trasformate, poichè consentono agli *host* di osservare in modo continuo il comportamento della domanda e il posizionamento della concorrenza, favorendo l'adozione di strategie di *pricing* dinamico (Gibbs et al. 2018). Infatti, nel contesto tradizionale, la disponibilità di informazioni sui *competitor* è limitata e spesso frammentaria, in quanto non è noto il posizionamento di prezzo della concorrenza, e sono ridotte le risorse per raccogliere informazioni consistenti sugli ospiti (Einav, Farronato e Levin 2013).

Analogamente, la differenziazione tra strutture tramite le ***caratteristiche di listing***, era un elemento di valore ancora prima dell'avvento di Booking.com ed Airbnb. Ad essersi evoluta è l'interfaccia di confronto tra le strutture, reso possibile tramite l'applicazione di filtri sulle caratteristiche d'interesse. Infatti, senza le piattaforme digitali la selezione di una o più strutture potenziali non era automatica, ma frutto di una valutazione incrociata

da parte dell'utente. Inoltre, l'avvento delle piattaforme ha esteso il concetto di struttura ricettiva, portando l'utente ad aggiungere caratteristiche d'interesse al *listing* (Guttentag 2015).

Nel complesso, l'analisi delle dimensioni di differenziazione, evidenzia come la competizione nel mercato degli affitti brevi si sviluppi lungo molteplici assi, che combinano elementi strutturali, reputazionali e strategici. La standardizzazione delle informazioni e la trasparenza introdotte dalla piattaforma amplificano il ruolo di tali fattori, rendendo il contesto particolarmente adatto allo studio delle dinamiche di prezzo e di competizione locale.

2.3 Area geografica di studio: Regione Piemonte

La presente sezione si focalizza sulla Regione Piemonte, area scelta come contesto geografico di riferimento per l'analisi empirica; essa si pone lo scopo di fornire il quadro di riferimento necessario per la successiva modellizzazione econometrica. Una maggiore enfasi verrà posta sul capoluogo, ovvero Torino, con lo scopo di estrarre in fase di analisi informazioni rilevanti e differenti rispetto al restante territorio, poichè è l'area metropolitana più densamente popolata. Inoltre, come evidenziato nell'analisi di mercato della piattaforma Airbnb, la città si colloca in una posizione intermedia rispetto alle principali città turistiche italiane sia in termini di numero di strutture disponibili sia di tasso di occupazione medio, rendendo il contesto torinese adatto allo studio delle dinamiche degli affitti brevi.

Infatti, l'area considerata, città storica e metropoli del Nord Italia, è in fase espansiva del turismo. Secondo l'Osservatorio Turistico della Regione Piemonte, nel 2024 il territorio di Torino e provincia ha registrato circa 2,89 milioni di arrivi e oltre 7,58 milioni di presenze, con una crescita rispettivamente del +5,7% e +6,9% rispetto al 2023 (Regione Piemonte 2025).

Parallelamente, anche il fenomeno degli affitti brevi ha assunto una rilevanza strutturale; Infatti in base ai dati riportati all'inizio del 2025 dal Corriere della Sera, nel comune di Torino si contano attualmente oltre 6.000 inserzioni attive su Airbnb, mentre nel solo centro storico si concentrebbe il 69% dell'offerta complessiva, con circa 4.945 alloggi su un'estensione di soli 4,5 km^2 (Corriere della Sera Torino 2025), statistiche percentuali simili sono state rilevate anche nel resto della regione.

Oltre alla crescita annuale dei flussi turistici, il calendario degli eventi svolge un ruolo determinante nella definizione dei picchi stagionali.

Infatti, esplorando la pagina dedicata agli eventi della città di Torino [5], emerge un elevato numero di manifestazioni rilevanti ospitate nell'area, tra cui il Salone Internazionale del Libro e le ATP Finals di tennis, che generano incrementi temporanei della domanda

[5] `https://www.turismotorino.org/it/esperienze/eventi`

turistica. Nel resto della regione si osservano picchi stagionali differenziati, riconducibili in inverno alla domanda generata dalla stagione sciistica nelle principali località alpine, quali Sestriere, Bardonecchia e Limone Piemonte, e in autunno agli eventi enogastronomici di rilievo, tra cui in particolare la Fiera Internazionale del Tartufo Bianco d'Alba.

In tale contesto gli affitti brevi si configurano come una componente flessibile dell'offerta ricettiva, in grado di assorbire variazioni improvvise della domanda meglio rispetto alle strutture alberghiere tradizionali (Farronato e Fradkin 2022).

Per poter sviluppare un modello flessibile, il caso studio si focalizzerà sul periodo di maggio 2024, mese privo di festività nazionali rilevanti, picchi di basso e alto livello di turismo ed eventi conclamati.

A livello comportamentale, una recente ricerca locale istituzionale segnala che quasi un turista su tre a Torino preferisce un affitto breve all'opzione alberghiera tradizionale (La Stampa 2024).

Ne risulta, che l'offerta di affitti brevi oltre a competere con quella alberghiera, esercita anche una pressione sulla struttura urbana, contribuendo a fenomeni di modifiche nei pattern di uso del suolo locale.

Perciò, in risposta all'intensificazione delle inserzioni Airbnb, in particolare nel centro storico, si è progressivamente aperto un dibattito istituzionale sulla regolamentazione del mercato degli affitti brevi nel Comune di Torino. Tale dibattito ha condotto all'incremento della tassa di soggiorno a partire da dicembre 2025, con l'obiettivo di compensare gli impatti urbani del turismo e di ristabilire condizioni di maggiore equilibrio competitivo tra affitti brevi e strutture ricettive tradizionali. Tali elementi risultano d'interesse ai fini dell'analisi empirica, in quanto il comportamento degli operatori si sviluppa all'interno di vincoli istituzionali variabili (Mentelocale Torino 2025).

Quindi, la Regione Piemonte si configura come un contesto particolarmente idoneo allo studio delle dinamiche competitive nel mercato degli affitti brevi. Torino rappresenta, all'interno di questo quadro regionale, un caso urbano paradigmatico per densità dell'offerta e concentrazione spaziale, mentre l'intero territorio piemontese completa l'analisi offrendo una varietà di contesti.

Alla luce dell'inquadramento territoriale fornito e dell'approfondimento teorico che segue, il Capitolo 4 illustrerà quindi le scelte di modellizzazione adottate per l'analisi delle dinamiche competitive nel mercato degli affitti brevi piemontese.

Capitolo 3

Stato dell'arte della letteratura

Individuati gli obiettivi della tesi e il relativo contesto, è stata svolta un'analisi dello stato dell'arte con lo scopo di comprendere quali sono gli aspetti già esplorati in letteratura che possono essere implementati nello sviluppo della metodologia presentata nel capitolo successivo.

Sulla base di quanto indicato nella sezione introduttiva, la ricerca in questione è stata guidata dagli elementi emersi dal paper di Li, Netessine e Koulayev del 2018, esplorando i concetti della letteratura che hanno condotto alle conclusioni del paper in questione, e le successive evoluzioni.

A tale fine, inizialmente è stata posta l'attenzione sulle caratteristiche e vincoli dei modelli per l'identificazione della competizione di prezzo tra strutture immobiliari, di cui si dispongono maggiori informazioni, cercando di comprendere quali conclusioni possono essere sfruttate per l'analisi del mercato degli affitti brevi, e quali invece sono da considerarsi proprie del solo mondo immobiliare. Inoltre, là dove disponibili fonti in letteratura specifiche per gli affitti brevi, sono state approfondite.

3.1 Segmentazione spaziale

Nel contesto immobiliare, la segmentazione spaziale rappresenta un presupposto teorico fondamentale per l'analisi delle dinamiche di prezzo e delle interazioni competitive. Infatti, il mercato immobiliare non è omogeneo, ma si articola in sub-mercati che riflettono preferenze locali, vincoli strutturali e caratteristiche socio-economiche; tale disomogeneità persiste anche nel contesto degli affitti brevi, ma in forma più lieve. Quindi, si vuole analizzare la teoria della segmentazione spaziale per rispondere agli interrogativi della

tesi.

3.1.1 Sistemi Locali del Turismo

Il concetto dei Sistemi Locali del Turismo (SLT), si ispira ai *Sistemi Locali del Lavoro (SLL)*, i quali si definiscono come aggregazioni di unità territoriali che identificano mercati di lavoro omogenei; ovvero aree geografiche nelle quali si realizza una sovrapposizione tra domanda ed offerta di lavoro. Si tratta quindi di zone delimitate in cui il flusso del lavoro è autocontenuto, poichè le competenze possedute ed offerte dagli individui corrispondono con quelle domandate dalle imprese (Coppola 2005).

Analogamente, la definizione dei Sistemi Locali del Turismo consente di individuare dei sottoinsiemi geografici per cui le interazioni di prezzo e la pressione competitiva risultino principalmente interne al cluster.

I cluster individuati possono essere rappresentati graficamente come in Figura 3.1, la quale riporta i gruppi del Sistema Locale del Lavoro nel 2002 in Piemonte. Una simile estensione dei cluster può essere realizzata per gli SLT; al contempo, nel contesto del settore turistico cittadino, è possibile sviluppare anche una suddivisione geografica con maggiore granularità (INSEE e IRES Piemonte 2002).

Le partizioni SLL vengono visualizzate sulla base di diversi parametri d'interesse, quali ad esempio: Densità di popolazione, Indice di vecchiaia, Tasso di occupazione, Tasso d'istruzione e Tasso d'immigrazione. Simili parametri rilevanti sono ancora da definirsi per la visualizzazione degli SLT.

Inoltre, l'IRES fornisce una visualizzazione riassuntiva dei partizionamenti, la quale viene realizzata sulla base dei valori assunti dagli indici di auto-contenimento della domanda e dell'offerta. Questi due dati confluiscono nella valutazione del numero di comuni all'interno di ciascun partizionamento per cui l'indice di centralità è 1; dove esso misura il rapporto tra la domanda di lavoro del comune rispetto all'offerta di lavoro del comune stesso. Simili considerazioni sono da valutarsi sulla base dei flussi di domanda dei turisti e di offerta delle strutture, in relazione a unità geografiche più piccole, i quartieri (INSEE e IRES Piemonte 2002).

Figura 3.1. Gruppi SLL 2002 - Piemonte

Sebbene la prospettiva dei Sistemi Locali del Turismo permetta di individuare partizionamenti territoriali autocontenuti, capaci di descrivere la distribuzione geografica delle interazioni tra domanda e offerta, essa non esaurisce la complessità dei meccanismi competitivi nei mercati turistici.

Infatti, la sola dimensione spaziale non è sufficiente a spiegare la formazione effettiva dei sub-mercati, poiché la competizione si manifesta anche in termini di percezione di equivalenza tra le strutture (Watkins 2001). Due unità possono appartenere allo stesso SLT, ma non essere in competizione diretta se non sono considerate intercambiabili dal punto di vista del consumatore. Per cogliere pienamente questa dimensione, occorre integrare la logica territoriale con una prospettiva microeconomica.

3.2 Sostituibilità tra strutture

Accanto alla logica di autocontenimento territoriale, assume quindi rilievo la dimensione della ***sostituibilità*** tra le unità offerte, la quale rappresenta il corrispettivo comportamentale della segmentazione spaziale. Quest'ultima individua i confini fisici dei mercati locali, la prima ne definisce i confini economici e percettivi, determinando l'effettiva estensione della competizione di prezzo.

Una prima definizione storica di sostituibilità vedeva i sub-mercati nel settore immobiliare come aree in cui le unità abitative sono considerate sostituibili tra loro da parte dei consumatori (Goodman e Thibodeau 1998). Essa evidenziava già al tempo un'implicazione importante, ovvero che la competizione di prezzo si sviluppa prevalentemente all'interno del sub-mercato, rispetto a mercati distinti. In tal senso, il concetto di sostituibilità guida la delimitazione endogena dei confini di mercato.

Dal punto di vista teorico, tale concetto può essere formalizzato all'interno del framework ad utilità randomica (*Random Utility Model*) introdotto da McFadden (1974).

Infatti, nel quadro della teoria microeconomica della scelta discreta, l'***utilità*** rappresenta una misura del beneficio percepito da un individuo nel consumare un determinato bene o nel scegliere una specifica alternativa. Essa costituisce una grandezza latente, non direttamente osservabile, che sintetizza il grado di soddisfazione derivante da un insieme di caratteristiche oggettive, quali ad esempio prezzo, posizione e dimensioni, e da elementi soggettivi come gusti, abitudini o percezioni individuali (Ben-Akiva e Lerman 1985).

Ogni individuo n associa a ciascuna alternativa i un livello di utilità complessiva U_{ni}, scomponibile in due componenti:

1. **Parte sistematica (V_{ni})**: determinata da variabili osservabili e rappresentativa della porzione prevedibile della scelta;
2. **Parte stocastica (ε_{ni})**: cattura gli aspetti non osservati del comportamento individuale.

Formalmente, ogni unità abitativa può essere rappresentata da un vettore di caratteristiche $\mathbf{z}_i$, mentre le preferenze individuali sono descritte da un vettore di parametri $\boldsymbol{\beta}_n$, che riflette la sensibilità del consumatore rispetto a ciascun attributo (Train 2009).

L'individuo sceglie l'alternativa che massimizza la propria utilità attesa, la quale viene calcolata come segue:

$$U_{ni} = \boldsymbol{\beta}_n \cdot \mathbf{z}_i + \varepsilon_{ni} \tag{3.1}$$

Poiché la componente casuale ε_{ni} non è osservabile, la scelta può essere descritta solo in termini probabilistici. Quindi, la probabilità di scelta per ciascun bene dipende dalla distribuzione statistica assunta per la componente stocastica.

Pertanto, due alloggi sono considerati sostituibili quando una variazione nel prezzo o in un attributo di uno di essi determina una variazione statisticamente significativa nella probabilità di scelta dell'altro (Gentzkow, Shapiro e Taddy 2019).

A fronte di quanto definito, dal punto di vista formale, la sostituibilità tra due alternative i e j può essere espressa come l'elasticità incrociata della probabilità di scelta dell'alternativa i rispetto al prezzo o un altro attributo x_j dell'alternativa j.

Dove, l'elasticità marginale misura tramite la formula riportata di seguito, la sensibilità S_{ij} della probabilità P_{ni} che l'individuo n scelga i rispetto a una variazione marginale di x_j (Belleflamme e Peitz 2015).

$$S_{ij} = \frac{\partial P_{ni}}{\partial x_j} \cdot \frac{x_j}{P_{ni}} \tag{3.2}$$

Se il valore di sensibilità è positivo, le due alternative sono *sostituti*, poiché un aumento del prezzo di j incrementa la probabilità di scelta di i, mentre al contrario se il valore è negativo, sono beni tra loro *complementari*. Infine, se $S_{ij} \approx 0$, esse appartengono a mercati distinti.

Perciò, la sostituibilità descrive il grado di correlazione comportamentale e competitiva tra alternative percepite come comparabili dal punto di vista del consumatore (Train

2009).

Nel contesto specifico del mercato degli affitti brevi, la sostituibilità può essere ulteriormente arricchita includendo dimensioni temporali come la stagionalità della domanda, oppure elementi reputazionali sulla base di quanto visto nella Tabella 2.1 nella Sezione 2.1.1 (Wang e Nicolau 2017).

Integrare questi fattori nel processo di segmentazione consente di delineare confini competitivi più accurati, in cui le unità competono effettivamente tra loro per attirare la stessa domanda marginale.

In un'ottica dinamica, la definizione di sostituibilità può essere estesa includendo esplicitamente la dimensione temporale e le caratteristiche reputazionali tipiche dei mercati digitali tramite la definizione del vettore $x_{j,t}$. Dove, esso rappresenta gli attributi dell'alternativa j al tempo t, comprendente il prezzo $p_{j,t}$, un indice reputazionale $r_{j,t}$ e parametri stagionali s_t:

Tale estensione della misura prende il nome di ***sostituibilità dinamica***, la quale può essere definita come l'elasticità incrociata della probabilità di scelta nel tempo. Analogamente a quanto visto prima viene calcolata come segue (Arcidiacono e Miller 2011):

$$S_{ij,t} = \frac{\partial P_{ni,t}}{\partial x_{j,t}} \cdot \frac{x_{j,t}}{P_{ni,t}} \tag{3.3}$$

In questo modo, la sostituibilità non è più una relazione statica fra beni, ma un concetto spazio-temporale e comportamentale, che riflette l'evoluzione delle preferenze dei consumatori nel tempo e la risposta del mercato a shock di domanda o variazioni di reputazione.

In termini empirici, ciò implica che due alloggi possono considerarsi *sostituti dinamici* se una variazione del prezzo o della reputazione di uno determina una variazione statisticamente significativa nella probabilità di scelta dell'altro nello stesso periodo o in periodi contigui (Dubé, Hitsch e Rossi 2010).

Questo approccio consente di modellare la competizione in mercati digitali altamente fluidi, in cui la disponibilità, le strategie di prezzo e la percezione reputazionale variano

continuamente, ridefinendo confini economici.

La definizione di sostituibilità è considerata ancora tutt'oggi di valore, come emerge dalle recenti ricerche di Anenberg e Ringo (2022), i quali hanno dimostrato empiricamente che gli shock di domanda tendono a propagarsi prevalentemente all'interno di segmenti omogenei per qualità e localizzazione, indicando che le dinamiche competitive sono guidate da meccanismi locali di sostituibilità.

3.3 Contiguità spaziale

L'anello di congiunzione tra la due prospettive è rappresentato dalla ***contiguità spaziale***. Infatti, essa garantisce che le interdipendenze comportamentali e le interazioni di prezzo si manifestino entro ambiti coerenti dal punto di vista territoriale, evitando che i confini econometrici della competizione si discostino da quelli geografici reali (Wu e Sharma 2012).

I metodi puramente statistici di segmentazione, basati esclusivamente su similarità nelle caratteristiche edoniche o nei livelli di prezzo, possono generare cluster privi di integrità geografica, ovvero composti da osservazioni distanti nello spazio ma artificialmente unite da analogie nei dati. Questo fenomeno, definito come "*spatial misalignment*", compromette la validità interpretativa dei sub-mercati così ottenuti, specialmente quando l'obiettivo è modellare dinamiche di concorrenza locale.

Per superare questo limite, Wu e Sharma (2012) propongono di integrare vincoli di contiguità spaziale all'interno del processo di clustering, i quali vengono formalizzati attraverso matrici di adiacenza.

Una *spatial adjacency matrix* W specifica le relazioni di vicinanza tra unità geografiche, quali quartieri, sezioni censuarie, o come nel caso oggetto di studio, strutture Airbnb. A tale scopo le matrici sono costruite così che ogni riga e colonna rappresenti un'osservazione geografica, ovvero una Struttura Airbnb. In tal modo, gli elementi della matrice w_{ij} assumono valori che descrivono la relazione tra le unità geografiche i e j. Quindi, formalmente la matrice di contiguità spaziale può essere definita come:

$$w_{ij} = \begin{cases} 1 & \text{se le unità } i \text{ e } j \text{ sono contigue} \\ 0 & \text{altrimenti} \end{cases} \tag{3.4}$$

Nel caso di unità areali la contiguità è spesso definita sulla base di confini amministrativi condivisi. Tuttavia, in presenza di unità geografiche puntuali, come nel caso delle strutture Airbnb, tale approccio risulta inapplicabile. In questi contesti, la letteratura propone

criteri di prossimità basati sulla distanza geografica, tra i più diffusi emerge l'approccio ***k-nearest neighbors (k-NN)*** .

Secondo tale criterio, a ciascuna unità geografica i vengono associate le k strutture più prossime in termini di distanza geografica. La matrice di contiguità risultante assegna quindi w_{ij} positivo se l'unità j rientra tra i k vicini più prossimi di i, e valore nullo altrimenti. Tale impostazione garantisce che ogni unità presenti un numero fisso di relazioni spaziali, evitando la presenza di osservazioni isolate e assicurando una struttura di interazione locale omogenea (LeSage e Pace 2009, Section 4.2, pp. 77–82).

Una volta definito il criterio di vicinato, la matrice di contiguità può assumere differenti forme funzionali. La forma standard è la *matrice binaria*, la quale rappresenta una semplice adiacenza topologica, per cui se due unità i e j sono vicine il valore assunto w_{ij} è pari a 1, altrimenti 0. Un'altra soluzione diffusa è la *matrice a pesi inversi*, la quale prevede che il valore di w_{ij} diminuisca all'aumentare della distanza geografica d_{ij} tra le unità geografiche.

Congiuntamente a questi due approcci, la matrice viene solitamente normalizzata in modo che la somma degli elementi di ciascuna riga sia pari a uno, ossia (Bivand, Pebesma e Gómez-Rubio 2013):

$$\sum_j w_{ij} = 1.$$

Per chiarire il concetto, di seguito è riportato un esempio di matrice binaria normalizzata per quattro alloggi Airbnb a Torino.

Esempio – Matrice di contiguità spaziale

Si considerino quattro alloggi Airbnb localizzati in quartieri contigui dell'area torinese:

ID	Quartiere	Coordinate (x, y)
1	Centro	(0, 0)
2	San Salvario	(0, 1)
3	Crocetta	(1, 0)
4	Mirafiori	(2, 0)

Le relazioni di contiguità (1 con 2 e 3; 3 con 1 e 4) danno origine alla seguente matrice binaria di adiacenza:

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Dividendo ogni riga per la somma dei suoi elementi si ottiene la matrice normalizzata per riga:

$$\mathbf{W}_{norm} = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Questa matrice descrive la rete di vicinanza tra le unità geografiche, ponendo un peso maggiore alle relazioni tra alloggi contigui. Essa costituisce il vincolo spaziale che consente di preservare la coerenza territoriale dei cluster.

Durante la procedura di segmentazione, le informazioni contenute nella matrice di contiguità vengono utilizzate per introdurre una *spatial penalty*, la quale impone un costo aggiuntivo ogni volta che due unità geograficamente adiacenti, secondo la matrice $\mathbf{W}$, vengono assegnate a cluster differenti. In questo modo, il modello è indotto a formare raggruppamenti che risultino coesi geograficamente e per caratteristiche edoniche (Wu e

Sharma 2012).

Formalmente, la penalizzazione spaziale può essere espressa come un termine aggiuntivo nella funzione obiettivo di clustering.

Esempio - Penalizzazione spaziale

Si considerino tre strutture Airbnb localizzate in aree contigue: A (Centro), B (San Salvario) e C (Crocetta). La matrice di contiguità $\mathbf{W}$ è definita come:

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Supponiamo che le strutture A e B presentino caratteristiche simili (prezzo e dimensioni), mentre C sia differente. Un algoritmo di clustering standard, privo di penalizzazione, potrebbe assegnare A e C allo stesso cluster, ignorando la distanza spaziale.
Con l'introduzione del termine di penalizzazione , il modello riceve un costo aggiuntivo se A e B, le quali sono contigue, finiscono in cluster diversi, inducendo così una soluzione più coerente con la geografia urbana.

Quindi, la matrice di adiacenza preserva l'omogeneità interna in termini di caratteristiche strutturali e di prezzo e si assicura la coesione geografica dei segmenti, in modo da rappresentare più fedelmente i mercati entro cui avviene la competizione reale.

Un'evoluzione recente di tale approccio è fornita dagli stessi autori Feng et al. (2025), che estendono il modello introducendo tecniche di penalizzazione adattiva e indicatori di autocorrelazione spaziale (vedi Sezione successiva) per calibrare dinamicamente l'intensità del vincolo di contiguità.

In questa prospettiva, la segmentazione spaziale diventa un processo ibrido, in cui la similarità edonica e la prossimità geografica sono pesate in modo ottimale per massimizzare la coerenza dei cluster, rispecchiando la reale articolazione del mercato urbano.

3.4 Tecniche di segmentazione

Alla luce di quanto emerso nella trattazione teorica nelle sezioni precedenti, risulta evidente che una segmentazione efficace del mercato non può prescindere da due elementi fondamentali:

- **Sostituibilità** percepita tra le unità ricettive, che ne determina l'appartenenza a uno stesso sub-mercato competitivo;
- **Coerenza spaziale** dei cluster, necessaria per garantire che le segmentazioni riflettano mercati reali anziché artefatti statistici.

La sezione in questione discute le principali tecniche di segmentazione adottabili, in quanto nel passaggio dalla teoria alla pratica, si rende necessario operazionalizzare tali principi all'interno di un processo di segmentazione dei sub-mercati.

In particolare, l'integrazione del concetto di prezzo edonico, inteso come funzione esplicativa del valore di mercato basata sulle caratteristiche osservabili delle strutture, assume un ruolo centrale per quantificare la sostituibilità tra le unità, divenendo anche una metrica operativa per guidare la costruzione di cluster di concorrenza.

3.4.1 Autocorrelazione spaziale

Per l'identificazione di submercati, la letteratura distingue tra approcci a priori, che si basano su suddivisioni amministrative o geografiche predefinite, quali ad esempio quartieri e CAP, e approcci data-driven, che utilizzano informazioni empiriche per determinare le frontiere effettive del mercato.

Data la definizione, se ne deriva che i metodi a priori raramente riflettono le effettive relazioni competitive tra gli attori. Al contrario, i metodi data-driven permettono una rappresentazione più fedele dei modelli di comportamento dei consumatori e delle interazioni tra venditori (Malpezzi 2002).

L'identificazione dei sub-mercati tramite tecniche data-driven deve necessariamente tenere conto della dipendenza spaziale, non solo in fase di segmentazione, ma anche nella valutazione dei modelli. In particolare, risulta cruciale progettare processi di validazione che riflettano fedelmente il pattern di ***autocorrelazione spaziale*** presente nel fenomeno analizzato, garantendo una corretta interpretazione delle dinamiche competitive locali (Brenning 2012).

Si definisce esistere un'autocorrelazione spaziale quando i valori di una variabile osservati in località geograficamente vicine tendono a essere simili, o dissimili, tra loro più di quanto ci si aspetterebbe per caso (Getis 2008). Essa può essere:

- **Positiva**: Aree vicine mostrano valori simili della variabile. Ad esempio, i prezzi immobiliari alti in un quartiere circondato da altri con prezzi alti.
- **Negativa**: Aree vicine mostrano valori molto diversi. Tale dinamica è tipica di confini netti, come zone di gentrification o periferie rispetto ai centri urbani.
- **Assente**: Non vi è un pattern spaziale evidente. I valori osservati sono distribuiti casualmente nello spazio, tale per cui la vicinanza geografica non influenza la somiglianza dei valori.

Quindi, bisogna tenere in considerazione la presenza di autocorrelazione spaziale positiva e negativa, in quanto nelle osservazioni potrebbe generare bias significativi nella fase di valutazione dei modelli predittivi.

A tale fine Wu e Sharma hanno ritenuto di valore l'introduzione di due indicatori per la valutazione dell'autocorrelazione spaziale (Feng et al. 2025):

L'***Indice di Moran*** misura la correlazione *globale* dei valori nello spazio, la quale è positiva se $I > 0$, negativa se $I < 0$, altrimenti nulla. Il calcolo dell'indicatore viene definito come segue:

$$I = \frac{N}{S_0} \cdot \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{3.5}$$

L'***Indice Geary's C*** è una misura di autocorrelazione *locale*, più sensibile alle variazioni puntuali tra unità contigue. Esso è definito come:

$$C = \frac{(N-1)}{2S_0} \cdot \frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2} \tag{3.6}$$

In questo caso, valori di C inferiori a 1 indicano autocorrelazione positiva, valori prossimi a 1 suggeriscono assenza di autocorrelazione, mentre valori superiori a 1 segnalano autocorrelazione negativa.

Nei due indicatori, date N osservazioni, sono coinvolte le variabili e parametri che seguono:

- x_i rappresenta il valore della variabile di interesse per l'unità geografica i;
- $\bar{x}$ è la media globale dei valori osservati;
- $S_0 = \sum_i \sum_j w_{ij}$ rappresenta la somma totale dei pesi.

Entrambi gli indici se utilizzati congiuntamente, forniscono due informazioni complementari. Infatti, Moran's I evidenzia la presenza di pattern globali di clustering, mentre Geary's C consente di identificare le disomogeneità locali o i confini tra sub-mercati.

Complessivamente, nella definizione dei sub-mercati solitamente risulta di valore non limitarsi alle mere applicazioni statistiche, integrando la conoscenza contestuale e le dinamiche locali del mercato (Malpezzi 2002); si deve però tenere in considerazione che gli Airbnb e hotel tendono a generare cluster distinti e non sempre sovrapponibili alle divisioni amministrative ufficiali (Gutiérrez et al. 2017).

3.4.2 Modelli edonici dei prezzi

Per tradurre la struttura competitiva dei sub-mercati in relazioni economiche osservabili, è necessario stimare in che misura ciascun attributo contribuisce al valore percepito delle unità.

In questa prospettiva, i modelli edonici dei prezzi rappresentano lo strumento analitico che permette di quantificare la sostituibilità tra strutture e di misurare l'effetto marginale di ogni caratteristica sul prezzo, costituendo così il passaggio logico dalla segmentazione spaziale alla modellizzazione econometrica.

Infatti, nel contesto dell'economia, il termine "***Edonico***" si riferisce all'utilità derivante dal consumo di beni e servizi; la sua definizione ufficiale si articola su due teorie:

- **Teoria del comportamento dei consumatori di Lancaster**: prevede che il rapporto tra il prezzo dei beni e le loro caratteristiche intrinseche sia lineare, che i prezzi impliciti siano costanti e che gli individui consumino alcuni o tutti i beni che appartengono a un gruppo in combinazioni (Chin e Chau 2003).
- **Modello di equilibrio di Rosen**: presuppone che la relazione sia non lineare, che i prezzi impliciti non siano costanti e che i consumatori scelgano e consumino ciascun bene da una gamma di beni in modo discreto per acquisire le caratteristiche preferite (Chin e Chau 2003).

Ciò nonostante, entrambi gli approcci ipotizzano che i beni e i servizi possano essere considerati come insiemi di attributi oggettivamente misurabili che influenzano l'utilità e che esista un mercato implicito in cui ogni attributo può essere valutato per dimostrare la disponibilità dei consumatori a pagare per quell'attributo (Berry e Haile 2021).

Ne consegue che i modelli di determinazione del prezzo edonistico siano ampiamente utilizzati nell'analisi del settore immobiliare e turistico, in quanto i prodotti turistici sono eterogenei e incorporano una serie di caratteristiche che forniscono valore e soddisfazione ai consumatori. In particolare, con l'avvento della *Sharing Economy* si è diffuso l'utilizzo del modello del prezzo edonistico per analizzare i fattori determinanti dei prezzi di Airbnb (Lorde, Jacob e Weekes 2019).

Oltre ai vincoli concettuali legati alla rappresentazione della concorrenza, i modelli edonici di prezzo si collocano nel quadro classico della regressione lineare, che presuppone errori indipendenti, normalmente distribuiti e a varianza costante. Nei mercati immobiliari e degli affitti brevi tali ipotesi risultano frequentemente violate. In particolare, la varianza dei residui tende ad aumentare con il livello del prezzo e con l'eterogeneità degli alloggi, generando forme marcate di ***eteroschedasticità***. In presenza di eteroschedasticità, le stime dei coefficienti rimangono imparziali, ma le varianze associate risultano distorte, con il rischio di sovra- o sottovalutare la significatività statistica degli effetti marginali (Kuminoff, Parmeter e Pope 2010).

Inoltre, il modello edonico assume implicitamente l'indipendenza delle decisioni di prezzo, trattando ciascuna osservazione come il risultato di una scelta autonoma non influenzata dalle strategie degli altri operatori. Sebbene tale assunzione semplifichi l'analisi, essa appare difficilmente sostenibile in mercati caratterizzati da elevata densità competitiva. Di conseguenza, mentre il modello edonico consente di spiegare la formazione del prezzo in funzione delle caratteristiche intrinseche dell'alloggio, esso tende a trascurare le interdipendenze esterne generate dalla concorrenza locale e dalla dinamica spaziale dei prezzi (Anselin 2008).

Per superare tali limiti, negli ultimi anni la letteratura ha progressivamente introdotto approcci capaci di combinare la dimensione edonica con quella competitiva.

3.4.3 Modelli di regressione penalizzata

Come citato in precedenza, un contributo significativo in questa direzione è fornito da Li, Netessine e Koulayev (2018), i quali propongono l'utilizzo del ***LASSO*** (*Least Absolute Shrinkage and Selection Operator*) per modellare la competizione di prezzo in mercati ad alta dimensionalità.

Dove il *LASSO*, introdotto da Tibshirani (1996), è una tecnica di regressione che minimizza la somma dei quadrati dei residui imponendo un vincolo sulla somma dei valori assoluti dei coefficienti di regressione. In tal modo, il modello individua un insieme parsimonioso di variabili rilevanti, imponendo a zero i coefficienti meno significativi e realizzando contemporaneamente stima e selezione delle variabili. Formalmente, il problema di ottimizzazione è espresso come:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{soggetto a} \quad \|\boldsymbol{\beta}\|_1 \leq t,$$

dove $\mathbf{y}$ è un vettore $n \times 1$ delle osservazioni, $\mathbf{X}$ è una matrice $n \times k$ di variabili esplicative, $\boldsymbol{\beta}$ è il vettore dei coefficienti di regressione e t è il parametro positivo di regolarizzazione che controlla il grado di sparsità del modello.

Tale problema può essere riscritto in forma equivalente come:

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right],$$

Dove λ è il moltiplicatore di Lagrange associato al vincolo l_1, il quale è un indicatore di come e quanto viene controllata la complessità del modello attraverso la penalità applicata ai coefficienti stimati (Zou e Hastie 2005). I valori ottimali di λ o t sono in genere determinati mediante tecniche di *cross-validation* [1], che bilanciano la complessità del modello e la capacità predittiva.

Prima di approfondire le proprietà del LASSO, è utile introdurre brevemente il modello di regressione Ridge, dal quale esso deriva per estensione. Entrambi appartengono alla famiglia dei metodi di regressione penalizzata, che mirano a migliorare la stabilità dei coefficienti e a ridurre la varianza delle stime in presenza di multicollinearità o di un elevato numero di variabili esplicative (Hastie, Tibshirani e Friedman 2009).

Nel caso del *Ridge Regression*, il problema di ottimizzazione prevede una penalizzazione di tipo l_2 sulla magnitudine dei coefficienti:

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right].$$

Il termine $\lambda\|\boldsymbol{\beta}\|_2^2$ impone un vincolo sulla somma dei quadrati dei coefficienti, riducendone la magnitudine complessiva e prevenendo la varianza elevata dovuta alla collinearità tra variabili. A differenza del LASSO, tuttavia, il Ridge non annulla i coefficienti meno rilevanti, ma li riduce in modo continuo, producendo una soluzione "densa" in cui tutte le variabili restano nel modello, seppur con peso ridotto (Hoerl e Kennard 1970).

Il LASSO nasce come estensione di questa logica, perciò mantiene il principio della penalizzazione, ma sostituisce la norma l_2, definita come $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ con la norma l_1, dove $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$, favorendo soluzioni sparse, ed imponendo a zero i coefficienti meno significativi (Hastie, Tibshirani e Friedman 2015).

[1] La *cross-validation* è una tecnica di valutazione statistica che consente di stimare la capacità predittiva di un modello su dati non osservati. Essa consiste nel suddividere il campione in più sottoinsiemi, utilizzando iterativamente una parte per l'addestramento e la restante per la validazione. L'approccio più comune è la *k-fold cross-validation*, in cui il campione è diviso in k parti di pari dimensione; le prestazioni medie forniscono una stima robusta dell'errore di generalizzazione (Stone 1974; Geisser 1975).

Dal punto di vista geometrico, il vincolo l_1 definisce una regione ammissibile di forma romboidale, mentre il vincolo l_2 genera una regione circolare. Tali differenze sono illustrate in Figura 3.2 .

Le curve tratteggiate rappresentano le ellissi di livello della funzione di errore quadratico $\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, la quale è una misura della capacità del modello di approssimare i dati. Ciascuna curva di livello identifica l'insieme dei punti che producono lo stesso valore di errore, perciò si tratta di combinazioni di coefficienti che garantiscono lo stesso grado di adattamento ai dati; procedendo verso ellissi più piccole ci si avvicina alla soluzione OLS, ovvero l'errore minimo (Efron et al. 2004).

Invece, le aree colorate identificano le regioni ammissibili imposte dai due vincoli di regolarizzazione, consentendo di visualizzare la distinzione concettuale tra Ridge e LASSO.

Il cerchio blu rappresenta il vincolo l_2, il quale riduce in modo uniforme la magnitudine dei coefficienti, producendo soluzioni stabili ma dense, mentre il vincolo l_1 raffigurato dal rombo rosso tende a "tagliare" gli assi del piano dei coefficienti, annullando completamente quelli meno influenti.

L'intersezione dell'ellisse di errore con il bordo del rombo determina così un sottoinsieme di coefficienti non nulli, alla base del principio di sparsità che caratterizza il LASSO (Boyd e Vandenberghe 2004).

Questa rappresentazione fornisce un'intuizione visiva del vantaggio del LASSO rispetto alle tecniche di regressione classiche e al Ridge: la capacità di effettuare simultaneamente stima e selezione delle variabili, riducendo la complessità del modello e migliorandone l'interpretabilità economica.

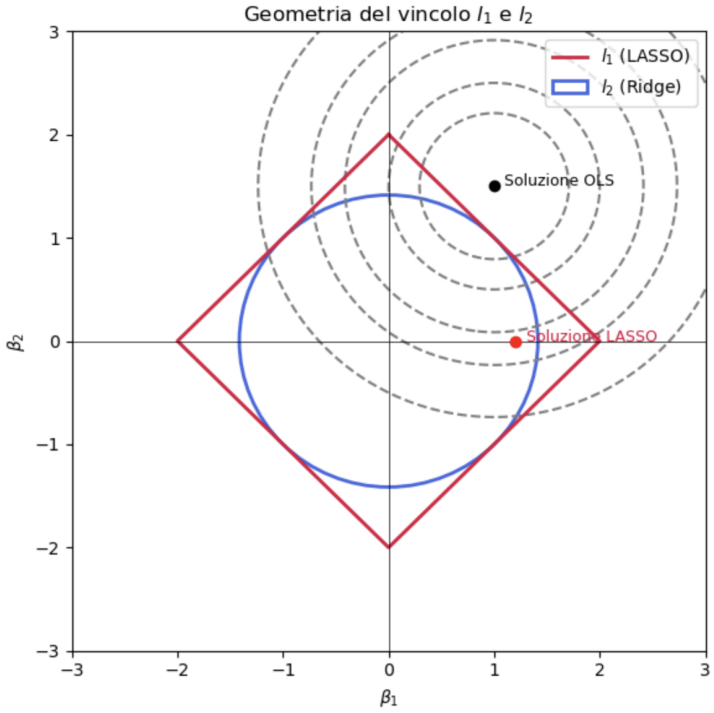


Figura 3.2. Rappresentazione geometrica del vincolo l_1 rispetto al vincolo l_2

3.4.4 Modello LASSO competitivo

In questa prospettiva, la penalizzazione assume un ruolo centrale anche nei modelli che esplicitano le interdipendenze strategiche tra operatori.

In particolare, il modello proposto da Li, Netessine e Koulayev (2018) formalizza la formazione del prezzo come il risultato di interazioni strategiche tra operatori, in cui il prezzo praticato da ciascun venditore dipende non solo dalle proprie caratteristiche, ma anche dai prezzi praticati dai concorrenti. La struttura competitiva viene quindi rappresentata mediante un sistema di equazioni simultanee, in cui la penalizzazione l_1 consente di identificare in modo endogeno le interdipendenze di prezzo rilevanti:

$$\min_{\alpha_i,\gamma_i,\beta_i,\delta_j} \frac{1}{2JT} \sum_{j,t=1}^{J,T} \left(P_{ijt} - \alpha_i - P_{-i,jt}\beta_i^T - X_{jt}\gamma_i - \delta_j E_{jt}\right)^2 + \lambda \sum_{n=1,n\neq i}^{N} |\beta_{in}|.$$

Il significato di ciascun simbolo viene riportato in Tabella 3.1.

Tabella 3.1. Variabili del LASSO competitivo

	Descrizione
P_{ijt}	Prezzo praticato dall'operatore i per l'unità j al tempo t.
α_i	Termine costante che cattura effetti fissi individuali e caratteristiche non osservate dell'operatore i.
$P_{-i,jt}$	Vettore dei prezzi praticati dagli operatori concorrenti diversi da i per l'unità j al tempo t.
β_i	Vettore dei coefficienti che misura l'intensità e la direzione della dipendenza del prezzo dell'operatore i rispetto ai prezzi dei concorrenti.
X_{jt}	Vettore delle caratteristiche strutturali e qualitativi osservabili dell'unità j al tempo t.
γ_i	Vettore dei parametri associati alle caratteristiche dell'unità j per l'operatore i.
E_{jt}	Variabile esogena che cattura shock di domanda e fattori esterni che influenzano la formazione del prezzo al tempo t.
δ_j	Parametro che misura l'effetto della variabile esogena E_{jt} sul prezzo dell'unità j.
λ	Parametro di penalizzazione che controlla l'intensità della regolarizzazione l_1 e il grado di selezione delle interazioni competitive.
J	Numero totale di unità considerate nel campione.
T	Numero di periodi temporali osservati.
n	Indice che identifica ciascun operatore concorrente rispetto all'operatore i.

Poichè, come visto nella sezione precedente, in presenza della penalizzazione l_1 , il termine associato alle interazioni competitive induce l'annullamento di una parte dei coefficienti β_{in}, selezionando endogenamente le relazioni di prezzo rilevanti. In tal modo, la matrice B dei coefficienti assume una struttura sparsa che riflette la natura locale e selettiva della competizione di prezzo: ciascun operatore reagisce solo a un numero ristretto di rivali, coerentemente con l'evidenza empirica dei mercati frammentati.

In un contesto di competizione strategica, il prezzo praticato da ciascun operatore e quello dei rivali si influenzano reciprocamente, rendendo il vettore $P_{-i,jt}$ intrinsecamente endogeno. Di conseguenza, una stima diretta dell'equazione di prezzo, anche in presenza di una corretta selezione delle relazioni competitive, può condurre a coefficienti distorti. È quindi importante distinguere il ruolo svolto dalla penalizzazione da quello relativo all'identificazione causale.

Infatti, la regolarizzazione l_1 agisce efficacemente sul piano strutturale, riducendo la dimensionalità del problema e isolando un sottoinsieme parsimonioso di interazioni di prez-

zo; tuttavia, essa non è sufficiente, da sola, a garantire la consistenza delle stime in presenza di endogeneità.

Alla luce di queste considerazioni, si rende necessario integrare la procedura di selezione penalizzata con una strategia di correzione dell'endogeneità coerente con la struttura del modello. In questo contesto, Li, Netessine e Koulayev (2018) adottano un approccio di tipo ***Two-Stage Residual Insertion*** (2SRI), che consente di correggere la simultaneità dei prezzi mantenendo invariata la formulazione dell'equazione strutturale e preservando il ruolo selettivo della penalizzazione (Terza, Basu e Rathouz 2008).

Tale metodologia permette di separare il problema della selezione competitiva da quello dell'identificazione causale, correggendo la correlazione tra il termine di errore e le variabili esplicative (Belloni, Chernozhukov e Hansen 2014).

Operativamente, nella prima fase viene stimato un modello ausiliario in cui i prezzi potenzialmente endogeni vengono regressi sulle variabili strumentali, ricavate da sorgenti esterne come dati di ricerca online.

Nella seconda fase, i residui di tale regressione vengono inseriti come ulteriori regressori nel modello principale, da cui il nome *Residual Insertion*.

Questa procedura consente di ottenere stime consistenti anche in presenza di endogeneità, preservando la capacità selettiva del LASSO.

Un'ulteriore estensione proposta dagli autori è l'***Adaptive LASSO*** (ALASSO), una versione pesata del LASSO che assegna un diverso grado di penalizzazione a ciascun coefficiente in base alla sua rilevanza stimata in un modello preliminare. L'idea alla base di tale approccio è che i coefficienti che risultano più elevati in valore assoluto in una stima iniziale contengano maggiore informazione e debbano quindi essere penalizzati in misura minore, mentre quelli di entità ridotta siano soggetti a una penalizzazione più severa.

Formalmente, a ciascun coefficiente β_j viene associato un peso w_j, definito come

$$w_j = \frac{1}{\left|\hat{\beta}_j^{(0)}\right|^{\gamma}}, \qquad \gamma > 0,$$

dove $\hat{\beta}_j^{(0)}$ rappresenta la stima preliminare del coefficiente j-esimo, ottenuta tramite un modello iniziale, mentre il parametro γ controlla l'intensità con cui la penalizzazione viene differenziata tra le variabili. In tal modo, i coefficienti più informativi ricevono una penalizzazione più debole e vengono preservati nel modello finale, mentre quelli marginali vengono spinti più facilmente verso zero (Takada e Fujisawa 2024).

In particolare, il meccanismo di stima preliminare può provenire da qualsiasi stima "*ragionevole, consistente o zero-consistente*" dei coefficienti, quindi tipicamente viene ottenuta tramite OLS o Ridge regression, ma è aperta la possibilità di usare stime derivate da sondaggi o indagini esterne purché soddisfino certe proprietà statistiche (Huang, Ma e C.-H. Zhang 2006).

Questa procedura adattiva permette di ridurre il bias di stima introdotto dalla penalizzazione l_1 standard e di migliorare la probabilità di identificare correttamente le variabili realmente influenti. Secondo i risultati sperimentali, tale approccio consente di individuare correttamente fino al 97% dei concorrenti rilevanti, confermando l'efficacia della regolarizzazione adattiva in contesti di alta dimensionalità.

La transizione dal modello edonico al LASSO competitivo segna quindi un cambiamento concettuale cruciale spostando l'attenzione dalla valutazione del contributo intrinseco delle caratteristiche al riconoscimento delle interdipendenze strategiche tra operatori.

3.4.5 Evoluzioni del LASSO competitivo

Un ulteriore avanzamento metodologico è rappresentato dal contributo di Inoue et al. (2018), che introduce il ***Generalized Fused Lasso*** (GFL) per l'identificazione endogena di segmentazioni geografiche nel mercato degli affitti. L'elemento distintivo di tale approccio risiede nell'integrazione esplicita della dimensione spaziale all'interno della funzione obiettivo.

Rispetto ai modelli penalizzati standard, il GFL introduce un termine addizionale di penalizzazione sulle differenze tra coefficienti associati a unità territoriali contigue, formalizzato tramite l'insieme E delle coppie geografiche adiacenti. In questo modo, la contiguità spaziale entra direttamente nel processo di stima, superando l'ipotesi di in-

dipendenza spaziale dei parametri e consentendo di modellare relazioni di vicinato in maniera endogena.

Il termine di seguito costituisce la principale innovazione del modello:

$$\sum_{(m,n)\in E} |\beta_m - \beta_n|$$

Esso agisce come un meccanismo di *fusion* che incoraggia l'uguaglianza dei coefficienti tra aree confinanti quando le differenze non sono supportate dai dati. Ne deriva una segmentazione geografica data-driven, in cui le regioni vengono aggregate sulla base di similarità economiche piuttosto che secondo confini amministrativi predefiniti.

Un ulteriore elemento introdotto è il parametro γ, che regola il peso relativo tra la penalizzazione di tipo LASSO e quella di *fusion*. Tale parametro consente di controllare il compromesso tra sparsità globale dei coefficienti e omogeneità locale nello spazio, rendendo il modello flessibile rispetto al grado di eterogeneità territoriale effettivamente presente nei dati.

Nel complesso, il GFL consente di integrare selezione dei parametri e regolarizzazione spaziale all'interno di un unico framework, risultando particolarmente adatto all'analisi di mercati immobiliari e degli affitti brevi.

Negli ultimi anni, la letteratura ha ampliato in modo sostanziale le applicazioni del LASSO e delle sue varianti spaziali, con l'obiettivo di gestire in maniera più robusta la dipendenza geografica e la complessità strutturale dei mercati locali.

Un primo avanzamento è rappresentato dal ***Moran's I Lasso*** proposto da Barde, Cherodian e Tchuente (2023), che integra direttamente l'indice di autocorrelazione spaziale di Moran, discusso nel Paragrafo 3.4.1, all'interno della procedura di selezione delle variabili.

Perciò, invece di basarsi sulla cross-validation tradizionale, il parametro di penalizzazione viene calibrato in funzione del livello residuo di correlazione spaziale. In questo modo, il modello è in grado di selezionare automaticamente solo quelle componenti spaziali che riducono in modo significativo la dipendenza geografica, migliorando la consistenza

statistica del modello in presenza di dati cross-section correlati. Tale approccio permette di filtrare le componenti spaziali endogene senza dover specificare ex ante la struttura del modello.

Un secondo contributo rilevante è fornito dal lavoro di Sakai, Tsuchida e Yadohisa (2024), che introduce il modello di ***Bayesian Geographically Weighted Sparse Regression*** (BGWSR), un'estensione bayesiana della *Geographically Weighted Regression* (GWR). In quest'ultima, i coefficienti del modello non sono assunti costanti sull'intero territorio, ma variano nello spazio, riflettendo eterogeneità locali nelle relazioni economiche. Formalmente, per una localizzazione $\mathbf{s}_i$, il modello può essere espresso come segue (Brunsdon, Fotheringham e Charlton 1996):

$$y_i = \beta_0(\mathbf{s}_i) + \sum_{k=1}^{K} \beta_k(\mathbf{s}_i)\, x_{ik} + \varepsilon_i,$$

dove i parametri $\beta_k(\mathbf{s}_i)$ sono stimati mediante una regressione locale che assegna pesi alle osservazioni in funzione della distanza geografica dal punto $\mathbf{s}_i$.

Sebbene tale approccio consenta di catturare la non stazionarietà spaziale, può risultare instabile in presenza di forti variazioni nella densità delle osservazioni, producendo stime eccessivamente frammentate nelle aree scarsamente popolate (Fotheringham, Yang e Kang 2023). Il modello BGWSR supera questi limiti combinando la ponderazione geografica della GWR con una regolarizzazione strutturata di tipo *Fused Lasso* sui coefficienti locali, introducendo una penalizzazione sulle differenze tra parametri associati a localizzazioni adiacenti:

$$\sum_{(m,n)\in E} |\beta(\mathbf{s}_m) - \beta(\mathbf{s}_n)|,$$

dove E rappresenta l'insieme delle coppie spazialmente contigue.

Il BGWSR, da un lato, consente di stimare coefficienti spazialmente variabili evitando una frammentazione eccessiva dello spazio; dall'altro, grazie alla formulazione bayesiana e alla penalizzazione di tipo *fusion*, riduce l'incertezza predittiva nelle aree con minore densità di dati. Questo rende l'approccio particolarmente adatto allo studio di mercati urbani disomogenei.

Parallelamente, Ohishi, Ando e Konno (2024) propongono un'estensione del *Generalized Fused Lasso* per dati raggruppati e modelli lineari generalizzati. L'approccio introduce strutture multiple di adiacenza, permettendo di rappresentare non solo la prossimità spaziale ma anche relazioni economiche o sociali, quali ad esempio reti di offerta o appartenenza a piattaforme comuni. Questo sviluppo risulta particolarmente rilevante per i mercati digitali multilocali, nei quali la competizione non segue esclusivamente la contiguità fisica, ma si distribuisce lungo reti di similarità funzionale o reputazionale.

Infine, Abella et al. (2025) applicano metodologie di segmentazione basate su network e clustering a grandi dataset di annunci immobiliari online, mostrando come la struttura competitiva possa emergere spontaneamente dai pattern di comportamento degli utenti e dalle interazioni tra domanda e offerta. Pur non adottando esplicitamente una penalizzazione LASSO, tali approcci condividono la stessa logica di individuazione endogena dei confini di mercato, basata su metriche di prossimità comportamentale e spaziale.

In sintesi, queste linee di ricerca convergono verso una visione più adattiva della segmentazione dei mercati spaziali. L'evoluzione metodologica dal modello edonico al LASSO competitivo, fino alle versioni spaziali e bayesiane del Fused Lasso, delinea una traiettoria di crescente realismo empirico e coerenza econometrica, che apre nuove prospettive per l'analisi dei mercati turistici urbani.

La Tabella 3.2 sintetizza i principali contributi teorici discussi nella presente sezione, evidenziando il percorso evolutivo che conduce dal modello edonico classico alle formulazioni penalizzate di tipo LASSO e, infine, alle loro estensioni spaziali.

L'obiettivo è fornire una visione d'insieme delle differenze strutturali tra i modelli in termini di: (i) obiettivo economico e interpretazione del prezzo, (ii) trattamento della dimensionalità e della collinearità tra variabili, (iii) capacità di modellare la competizione e la dipendenza spaziale.

Tabella 3.2. Confronto tra modelli di regressione edonica, penalizzata e spaziale

	Modello Edonico	**LASSO / ALASSO**	**(Bayesian) Fused LASSO**
Obiettivo	Spiegare la variazione dei prezzi in funzione delle caratteristiche dell'unità abitativa.	Identificare relazioni di interdipendenza e competizione tra operatori o unità di offerta.	Rilevare segmentazioni geografiche omogenee integrando vincoli di contiguità spaziale.
Struttura	Regressione lineare classica con coefficienti globali.	Regressione penalizzata con vincolo l_1 che induce sparsità.	Regressione con doppia penalizzazione: sparsità (l_1) e omogeneità tra coefficienti contigui (*fusion*).
Dimensione spaziale	Assente o trattata ex post.	Indiretta, attraverso legami competitivi tra unità vicine.	Esplicita, tramite vincoli di contiguità o dipendenza spaziale.
Selezione delle variabili	Manuale o basata su test statistici.	Automatica tramite regolarizzazione (λ).	Automatica e adattiva, con regolarizzazione spaziale o bayesiana.
Interpretabilità	Elevata ma sensibile alla multicollinearità.	Buona, con evidenza delle relazioni rilevanti.	Alta, grazie a cluster interpretabili e continui.
Limiti	Non considera interazioni o dipendenze spaziali.	Può trascurare la coerenza geografica dei legami competitivi.	Richiede elevato sforzo computazionale e calibrazione dei vincoli.

Nel complesso, la letteratura analizzata evidenzia come la segmentazione spaziale non rappresenti un passaggio preliminare o meramente descrittivo, ma costituisca una componente strutturale dei modelli di regressione applicati ai mercati immobiliari e degli affitti brevi. I contributi più recenti superano infatti l'approccio tradizionale basato su sub-mercati definiti ex ante, mostrando come le relazioni economiche possano variare nello spazio in modo continuo o discreto e debbano essere identificate endogenamente all'interno del processo di stima.

In questa prospettiva, la segmentazione spaziale e la regressione cessano di essere strumenti separati: i confini dei sub-mercati emergono direttamente dalle proprietà dei parametri stimati, dando luogo a quella che la letteratura definisce come una struttura a *spatial regimes* (Anselin 2008). Tali regimi non sono imposti a priori, ma risultano dall'interazione tra eterogeneità locale, contiguità geografica e meccanismi di regolarizzazione,

consentendo una rappresentazione più realistica delle dinamiche competitive e di prezzo (Fotheringham, Yang e Kang 2023).

Questo quadro metodologico suggerisce quindi che l'analisi dei prezzi nei mercati urbani complessi richieda modelli capaci di integrare simultaneamente eterogeneità spaziale, selezione strutturale e coerenza territoriale. Il capitolo successivo si inserisce in tale filone, adottando un approccio empirico che sfrutta queste intuizioni per identificare sub-mercati e relazioni competitive in modo *data-driven*.

Capitolo 4

Metodologia

In questo capitolo si vuole strutturare in modo puntuale la metodologia da applicare nel Capitolo 5. A tale fine, si deve prevedere una fase preliminare che mira a comprendere il dataset a disposizione in modo tale da preparare i dati alla metodologia adeguata. Quest'ultima verrà sviluppata e descritta nella seconda parte del capitolo.

4.1 Fonte dei dati

Tale analisi si basa sui dati a disposizione nella regione Piemonte per il mese di Maggio 2024. Il periodo temporale di estrazione dati è stato scelto evitando periodi di bassa ed alta stagione, ed eventuali mesi caratterizzati da eventi di aggregazione. I dati sono popolati da più di 19 mila differenti strutture, e quasi 600 mila annunci.

Le strutture in questione presentano una marcata predominanza delle intere abitazioni o appartamenti, che costituiscono l'79,98% dell'offerta complessiva. Seguono le stanze private, con una quota pari al 18,89%, mentre le stanze in hotel e le stanze condivise rappresentano rispettivamente solo circa il 0,6% e lo 0,5% del totale (Vedi Figura 4.1).

Questa distribuzione suggerisce che, nel mercato piemontese la piattaforma Airbnb è utilizzata in modo prevalente per la locazione di intere unità abitative, piuttosto che per la condivisione di spazi domestici. Tale configurazione riflette una progressiva professionalizzazione dell'offerta, poiché la disponibilità di interi appartamenti indica un uso sistematico della piattaforma a imprenditoriali, piuttosto che occasionale.

Il prezzo medio per notte risulta pari a 30,01 €, valore coerente con la fascia medio-bassa del mercato considerato.

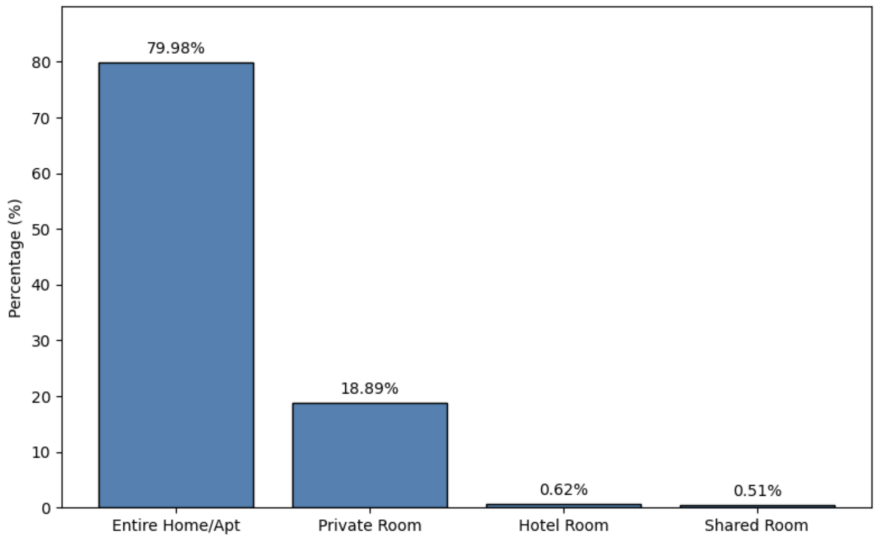


Figura 4.1. Distribuzione percentuale per tipologia di alloggio

La distribuzione degli host evidenzia una struttura fortemente sbilanciata: solo il 23,41% gestisce più di una struttura, mentre appena il 2,19% supera le cinque. Tuttavia, questi ultimi controllano circa il 18,63% dell'offerta complessiva.

Tali dati confermano la presenza di una minoranza di operatori professionali che, pur numericamente ridotta, concentra una quota significativa degli alloggi, delineando un mercato in progressiva transizione da modello *peer-to-peer* a gestione semi-imprenditoriale.

4.2 Gestione dei dati

I dati raccolti da qualsiasi fonte possono persistere incompleti, rumorosi e incoerenti, causando problemi nell'analisi dei dati, e rendendo necessario correggerli preventivamente (Maharana, Mondal e Nemade 2022). Per tale ragione, dopo aver descritto la fonte dei dati nella Sezione 4.1, devono essere valutati tutti i possibili problemi che i dati potrebbero trascinare in fase di analisi.

A tale scopo, la gestione dei dati si baserà sulle indicazioni riportate in Figura 4.2, valutando i poteniali errori derivanti dalla raccolta dati riscontrabili nel *data frame* oggetto di studio.

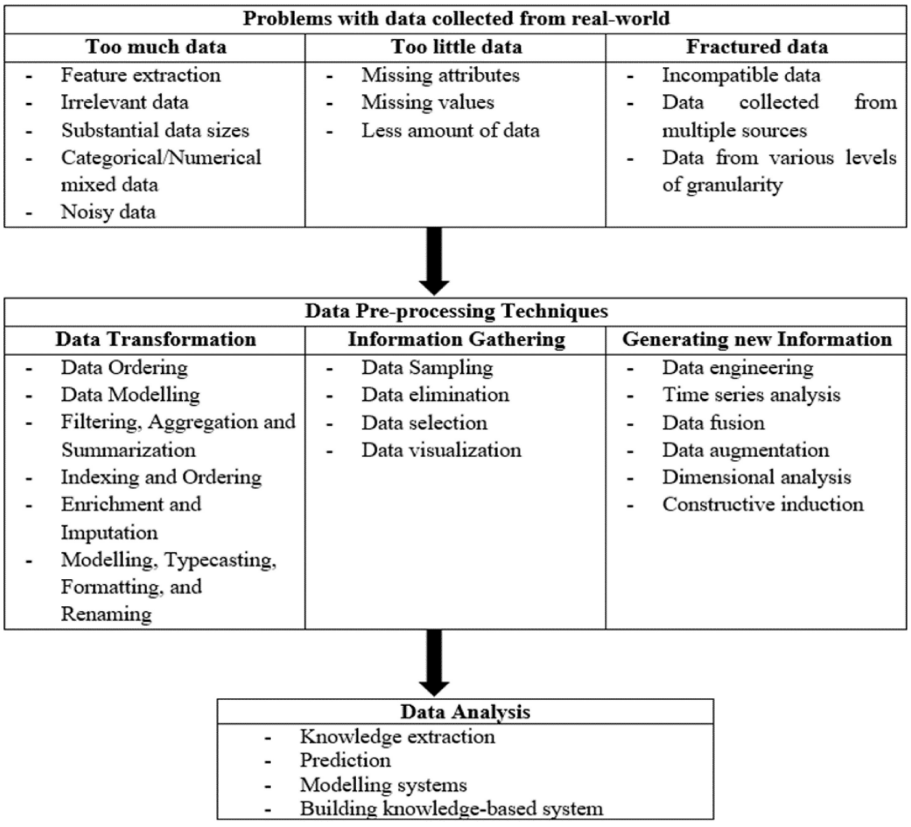


Figura 4.2. Possibili problemi dei dati raccolti

I dati di cui si dispone per il caso studio ricoprono in modo continuativo il lasso di tempo d'interesse, di conseguenza non sono da gestire le criticità legate ad un numero limitato di dati frammentati (***Too little data problems***). Inoltre, il lasso temporale scelto è

significativo, senza impattare negativamente sulle criticità legate all'elevata numerosità dei dati (***Too much data probelms***).

La base dati a disposizione è stata fornita già consistente, di conseguenza non risulta un problema l'eventuale incompatibilità dei dati o incogruenze dovute a molteplici fonti di dati (***Fractured data problems***).

Infine, per poter valutare e correggere l'eventuale presenza di dati non rilevanti o problematicità che andrebbero ad aumentare la complessità computazionale, è necessario esplorare un sample dei dati, in modo da cogliere il significato di ciascun attributo e il rispettivo tipo.

A tale scopo, sono stati stampati i primi record della base dati analizzata tramite statistiche descrittive nella sezione precedente, ed una lista della rispettiva tipologia di dato assegnata. L'esito di quest'analisi preliminare viene riportato nella Tabella di seguito.

Tabella 4.1. Tipi di dato degli attributi

Attributo	**Tipo rilevato**	**Tipo atteso**	**Descrizione**
Property ID	object	String	Identificativo univoco della struttura
Latitude	float64	Floating	Coordinata geografica (latitudine)
Longitude	float64	Floating	Coordinata geografica (longitudine)
Date	object	Data	Data di riferimento della rilevazione
Status	object	Binary	Stato della disponibilità dell'alloggio (occupato/libero)
Price (USD)	float64	Floating	Prezzo in dollari percepito dal cliente
Booked Date	object	Data	Data di prenotazione
Reservation ID	object	String	Identificativo univoco della prenotazione
Airbnb HOST ID	float64	String	Identificativo numerico dell'host
Listing Type	object	String	Tipologia di alloggio
Bedrooms	float64	Intero	Numero di camere da letto
Bathrooms	float64	Intero	Numero di bagni
Max Guests	float64	Intero	Numero massimo di ospiti ammessi
Minimum Stay	float64	Intero	Permanenza minima consentita
Number of Reviews	float64	Intero	Numero di recensioni disponibili
Number of Photos	float64	Intero	Numero di fotografie pubblicate
Airbnb Superhost	float64	Binario	Indicatore della qualifica Superhost (0/1)
Overall Rating	float64	Floating	Valutazione complessiva dell'alloggio
Amenities	object	List of String	Lista dei servizi offerti (es. Wi-Fi, cucina, parcheggio)
SLL_2011_T	int64	String	Codice territoriale SLL
DEN_SL2011	object	String	Denominazione del Sistema Locale del Lavoro
PRO_COM_T	int64	String	Codice catastale comunale
COMUNE	object	String	Nome del comune di appartenenza
MARKET	object	String	Segmento di mercato di riferimento

Nota: La colonna "Tipo rilevato" deriva dalla funzione `dtypes` del pacchetto `pandas`. Rappresentano la codifica interna utilizzata per l'elaborazione numerica e testuale. La colonna "Tipo atteso" riflette la natura semantica delle variabili nel contesto analitico e verrà utilizzata come riferimento per le fasi successive di *data cleaning* e *feature engineering*.

Una volta definite le caratteristiche degli attributi e la loro tipologia all'interno del sistema informativo, nonché identificate le principali criticità del dataset, come sintetizzato in Figura 4.2, si procede all'applicazione delle opportune tecniche di pre-processing, che verranno approfondite nella sezione successiva.

4.2.1 Fase di pre-processing

Le tecniche di *pre-processing* racchiudono la "*fase in cui i dati grezzi vengono convertiti in un formato comprensibile e valutabile dai computer e dagli algoritmi di apprendimento automatico*" (Brijith 2023). Ne consegue che dal punto di vista operativo tale frase ha lo scopo di convertire i dati grezzi in un formato comprensibile agli algoritmi utilizzati in fase di analisi e trasformare i dati in modo tale da gestire eventuali problematiche emerse nel primo punto, e allinearli con gli obiettivi dell'analisi (Kotsiantis, Kanellopoulos e Pintelas 2006).

Le tecniche di manipolazione dei dati previste in questa fase dipendono dalla classificazione dei dati sulla base di quanto segue (Brijith 2023):

- **Structured**: I dati strutturati sono presentati in modo chiaro, solitamente sotto forma di righe e colonne, seguendo uno schema ordinato che prevede una posizione specifica per ogni informazione, rendendo i dati facili da interrogare.
- **Unstructured**: La mancanza di una struttura o di un formato predeterminato ed esatto è ciò che definisce i dati non strutturati. Spesso si presentano in un formato semplice da leggere per le persone, ma difficile da comprendere per i robot senza metodi specifici, in quanto mancano di uno schema predeterminato.
- **Semi-structured**: A differenza dei dati strutturati, che rispettano un quadro rigoroso, i dati semi-strutturati sono più flessibili. Tuttavia, includono alcune componenti organizzative, come tag o chiavi, che servono a offrire una certa struttura, consentendo comunque uno schema più adattabile.

L'identificazione del tipo di dato è essenziale, in quanto, noti gli obiettivi della fase di pre-processing sopra descritti, le tecniche da applicare risultano differenti.

Nel caso studio in questione si dispone di *dati strutturati*, in quanto presentati sotto formato tabellare, con attributi ben definiti (vedi Tabella 4.1) per cui è nota la posizione.

Per questa categoria di dati, la fase di pre-processing prevede step ben definiti che devono essere seguiti, sulla base di quanto riportato in Figura 4.3.

Figura 4.3. Step nella fase di *pre-processing* per dati strutturati

Dunque, tenendo in considerazione il sample di dati stampato inizialmente, e le rispettive caratterisitche, nelle sezioni che seguono vengono riportati gli step previsti in pre-processing, in modo da discuterne lo scopo e riportare le operazioni necessarie per la base dati del caso studio. Gli step in questione consentiranno di trasfomare i dati, raccogliendo e generando da essi nuove informazioni.

Data Cleansing

La qualità dei dati è la principale problematica all'interno di ogni contesto guidato da informazioni, in quanto rappresenta il principale ostacolo all'uso efficace dei dati, conducendo a potenziali decisioni ed esiti errati (Ridzuan e Wan Zainon 2019). Per tale ragione risulta chiave lo step di pulizia dei dati, il quale prevede la pre-elaborazione dei dati grezzi estratti, tramite la rimozione di dati duplicati o ridondanti, la verifica logica dei valori delle variabili e il trattamento dei valori anomali. L'elaborazione dei dati mancanti verrà invece gestita in seguito nell'apposita sezione (Guo et al. 2023).

Inizialmente, è stata eliminata l'eventuale presenza di valori duplicati per tutti gli attributi disponibili, tramite l'apposita funzione `drop_duplicates()`.

Mentre, per quanto riguarda incongruenze di tipologia dato, tramite la funzione `dtypes` si è già verificato in Tabella 4.1 se vi è una discrepanza tra il tipo dei dati rilevato a sistema e il tipo dei dati atteso. Tale incogruenza è necessario che venga appianata per specifici attributi, in modo tale da consentire le opportune operazioni di confronto tra i record.

In particolare, nel caso degli attributi attesi con tipo "Intero", ma rilevati a sistema con

tipo ”float64”, si tratta di una differenza che non comporta problematicità in fase di analisi.

Gli attributi ”Status” e ”Airbnb Superhost” sono attesi binari, ovvero possono assumere solo due valori. Tale tipo non viene rilevato a sistema in quanto presentano valori nulli, o errori che verranno gestiti nelle sezioni successive.

Però, il tipo dei due attributi non verrà impostato come binario neppure a seguito della gestione degli errori e valori nulli, in quanto in un contesto di applicazione di algoritmi di machine learning è preferibile un tipo numerico.

Infine, si è proceduto trasformando i due attributi ”Date” e ”Booked Date” nell'adeguato tipo *DateTime* riconosciuto nell'ambiente *Jupyter Notebook*. Tale trasformazione del tipo consente eventuali confronti basati su criteri temporali e non alfanumerici.

```
Attributi = ["Date", "Booked Date"]
#Conversione in datetime
df[Attributi] = df[Attributi].apply(
    lambda col: pd.to_datetime(col, errors="coerce")
)
```

Infine, l'ultimo step ha previsto verificare che gli attributi numerici fossero in linea con le unità di misura e l'ordine di grandezza attesi. A tal proposito, è stato dimostrato che diverse istanze avevano l'*Overall Rating* moltiplicato per 10^6; sono quindi stati riscalate le recensioni basate sull'ordine di grandezza errato:

```
# Rating fuori scala
mask_scaled = df["Overall Rating"] > 1000
# Riscalati rating
df.loc[mask_scaled, "Overall Rating"] = df.loc[mask_scaled, "Overall Rating"] /
↪  1_000_000
```

Stampando nell'ambiente di programmazione una verifica dell'avvenuta trasformazione, risulta persistere una serie di valori di *Overall Rating* fuori scala, i quali sono associati ad un numero molto limitato di HOST ID che si ripetono più volte all'interno della lista. In

queste casistiche limitate, è stato reso nullo il valore registrato, in quanto errato, e verrà gestito congiuntamente ai restanti valori nulli dell'attributo.

```
# Rating anomali dopo la trasformazione
mask_out_of_range = df["Overall Rating"].notna() & (
    (df["Overall Rating"] > 100) | (df["Overall Rating"] < 0)
)
# Rating imposto nullo
df.loc[mask_out_of_range, "Overall Rating"] = np.nan
```

Lo step successivo della pulizia dei dati ha riguardato la conversione della variabile di prezzo da Dollari Statunitensi ad Euro, in quanto il caso studio riguarda una regione del territorio italiano.

Tale conversione è stata definita sulla base del cambio in vigore per la corrispondente data di ciascuna istanza secondo i dati della Banca Centrale [1]. Tali dati sono consultabili grazie all'apposita API gratuita "*Frankfurter*", la quale richiede solamente l'importazione della libreria `requests`, per poter inviare la richiesta HTTP per l'accesso.

Nel caso in cui per una specifica data non venga rilevato un tasso di cambio, o per un'istanza risulti mancare la data, allora viene utilizzato il valore medio di cambio nel periodo. Infine, se l'API non dovesse dare risultati nell'ambiente, verrà assegnato un tasso di default pari a 0,92.

```
# Definizione tasso di cambio per data
unique_dates = df["Date"].dt.date.unique()
rates = {}
for d in unique_dates:
    url = f"https://api.frankfurter.app/{d}?from=USD&to=EUR"
    r = requests.get(url)
    data = r.json()
    if "rates" not in data or "EUR" not in data["rates"]:
        print(f"Nessun tasso trovato per la data {d}. Risposta:")
        continue
```

[1] https://data.ecb.europa.eu/currency-converter

```
    rates[d] = data["rates"]["EUR"]
if len(rates) > 0:
    default_rate = np.mean(list(rates.values()))
else:
    default_rate = 0.92
df["usd_eur_rate"] = df["Date"].dt.date.map(rates)
# Valore medio per date mancanti
df["usd_eur_rate"] = df["usd_eur_rate"].fillna(default_rate)
# Conversione Definitiva
df["Price (EUR)"] = df["Price (USD)"] * df["usd_eur_rate"]
```

A seguito di tale trasformazione, al fine di attenuare i fenomeni di *eteroschedasticità*[2], si è proceduto alla rimozione dei valori estremi della variabile prezzo, applicando una troncatura della distribuzione al primo e al novantanovesimo percentile.

```
p1 = df["Price (EUR)"].quantile(0.01)
p99 = df["Price (EUR)"].quantile(0.99)
df = df[
    (df["Price (EUR)"] >= p1) &
    (df["Price (EUR)"] <= p99)
].copy()
```

Infine, poichè i dati di longitudine e di latitudine non erano correttamente salvati, è stato individuato il sistema di riferimento utilizzato, il quale è risultato essere "*EPSG:32632 (UTM/metrico)*" [3], e sono stati definiti i corretti riferimenti geografici. Inoltre, poichè in base alle istanze i dati geografici sono stati salvati con ordini di grandezza differenti, è stata creata una funzione che tramite un divisore dinamico restituisce il corretto valore di longitudine e latitudine, indipendentemente dalla scala del valore di partenza.

```
# Diagnostica iniziale CRS
for i, row in df.head(100).iterrows():
```

[2]Il termine *eteroschedasticità* indica la condizione in cui la varianza degli errori di un modello di regressione non è costante al variare dei valori della variabile indipendente. In presenza di eteroschedasticità, le stime dei coefficienti rimangono imparziali ma risultano statisticamente inefficienti.

[3]`https://3dmetrica.it/i-codici-epsg/`

```
    lat = row["Latitude"]
    lon = row["Longitude"]
    if lat > 90 or lon > 180:
        crs_guess = "EPSG:32632 (UTM/metrico)"
    else:
        crs_guess = "EPSG:4326 (geografico)"
    print(f"ID {i:>4} | Lat: {lat:<12} | Lon: {lon:<12} | {crs_guess}")
# Funzione per estrazione progressiva con divisore dinamico
def extract_progressive_dynamic(
    value,
    max_digits,
    min_digits,
    digits_before_decimal,
    valid_range
):
    try:
        if pd.isna(value):
            return np.nan
        s = str(int(abs(float(value))))
        L = len(s)
        for d in range(min(max_digits, L), min_digits - 1, -1):
            core = s[:d]
            divisor = 10 ** (d - digits_before_decimal)
            val = float(core) / divisor
            if valid_range[0] <= val <= valid_range[1]:
                return val
        return np.nan
    except Exception:
        return np.nan
# Latitudine
def fix_lat(lat):
    return extract_progressive_dynamic(
        value=lat,
        max_digits=8,
        min_digits=4,
        digits_before_decimal=2,
        valid_range=(40, 50)
```

```
    )
# Longitudine
def fix_lon(lon):
    return extract_progressive_dynamic(
        value=lon,
        max_digits=7,
        min_digits=3,
        digits_before_decimal=1,
        valid_range=(5, 10)
    )
# Applicazione al DataFrame
df["Latitude_fixed"] = df["Latitude"].apply(fix_lat)
df["Longitude_fixed"] = df["Longitude"].apply(fix_lon)
```

Si è quindi conclusa l'attività di *data cleansing*, la quale ha consentito di predisporre una base informativa analiticamente utilizzabile, in quanto l'unificazione semantica dei dati costituisce un prerequisito essenziale per le fasi successive.

Data Normalization & Standardization

In questa fase si valuta l'adeguata tecnica di ridimensionamento delle caratteristiche da applicare per trasformare l'intervallo delle caratteristiche in una scala standard.

Le principali macro categorie che si distinguono sono le tecniche di Normalizzazione, le quali mirano a ridimensionare i valori di una variabile in un intervallo prefissato, tipicamente tra 0 ed 1, e le tecniche di Standardizzazione, le quali trasformano la variabile in modo che abbia media zero e deviazione standard unitaria (Patro e Sahu 2015).

Nel contesto della stima delle regressioni penalizzate, come il LASSO, la standardizzazione delle variabili costituisce la scelta vincente per garantire la correttezza e la stabilità della stima.

Infatti, la penalizzazione l_1 tipica del LASSO agisce sulla somma dei valori assoluti dei coefficienti, imponendo una soglia comune che determina quali parametri vengano annullati. Tuttavia, se le variabili esplicative non sono espresse sulla stessa scala, il termine

di penalizzazione tenderà a favorire le variabili con valori numerici minori, poiché la loro variazione influisce meno sul valore assoluto complessivo dei coefficienti, comportando una distorsione nella selezione delle feature, non imputabile a un reale effetto economico, bensì a una differenza di unità di misura (Hastie, Tibshirani e Friedman 2009).

Quindi, si è deciso di procedere standardizzando tutte le variabili numeriche continue secondo la *trasformazione Z-score*, applicabile nell'ambiente tramite la classe `StandardScaler()`. Tale approccio centra ciascuna feature sulla propria media e la ridimensiona in base alla deviazione standard, garantendo che il termine di penalizzazione nel LASSO agisca in modo isotropo su tutte le direzioni (Mazziotta e Pareto 2020):

$$x' = \frac{x - \mu_x}{\sigma_x},$$

dove μ_x e σ_x rappresentano rispettivamente la media e la deviazione standard della variabile x.

La standardizzazione verrà applicata solamente prima della fase di implementazione del modello, e riguarderà esclusivamente le variabili numeriche continue che descrivono le caratteristiche strutturali e qualitative delle unità.

Invece, verranno escluse dalla standardizzazione le variabili categoriche e binarie, poiché già definite su una scala discreta e priva di varianza interna. Mentre, le variabili descrittive, e le coordinate geografiche, verranno trasformate in variabili derivate, risolvendo tali problematicità.

Data Transformation

La fase di *Data Transformation* viene predisposta con l'obiettivo di convertire i *raw data* in un formato e struttura consoni per eseguire l'analisi (Borrohou, Fissoune e Badir 2025). Le operazioni eseguite che comportano l'alterazione del dato sono molteplici, alcune di esse sono trattate in sezioni a parte, quali "Data Normalization & Standardization" e "Feature Engineering".

Ne risulta che le principali tecniche da discutere in questa sezione sono:

- **Categorical Encoding**: Trasforma le variabili categoriali in numeri o vettori di numeri, in quanto pochi modelli di machine learning e di regressione sono in grado di gestire direttamente variabili non numeriche (Poslavskaya e Korolev 2023).
- **Riduzione della dimensionalità**: Si pone l'obiettivo di individuare una rappresentazione dei dati con dimensione inferiore, rispetto a quanto raccolto, ma che trattenga il massimo dell'informazione dei dati originali. Infatti, disponendo di grandi masse di dati, si verifica spesso che non tutte le variabili rilevate siano informative e che possono per questo essere eliminate, sostituite o trasformate (Celon 2022).

In particolare, per poter fornire in input i dati del caso studio ad un modello di regressione, è necessario eseguire una trasformazione di encoding; a tale scopo si valuta la tecnica più adeguata tra quelle note, le quali possono essere raggrupate nei tre insiemi rappresentati nel diagramma di Venn in Figura 4.4.

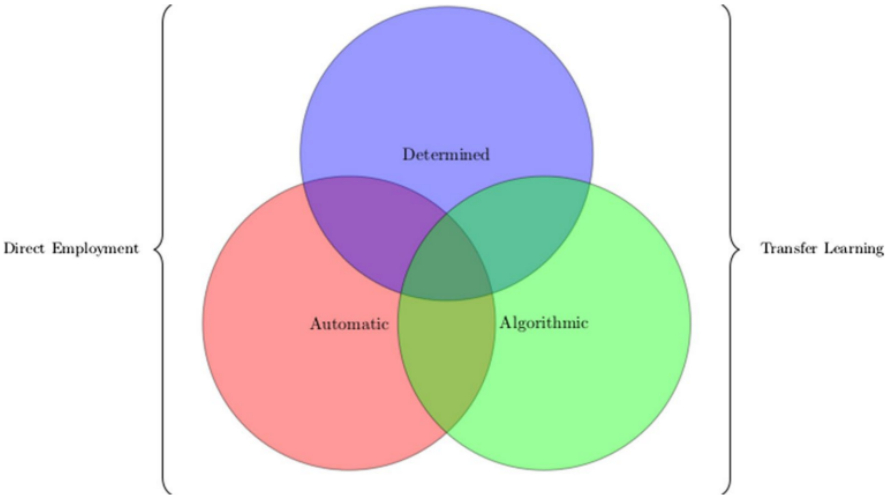


Figura 4.4. Tecniche di *encoding* delle variabili categoriche

Tali categorie si distinguono sulla base di come nasce il codice numerico ad esse assegnato e quanto dipende dai dati o dal modello. Il primo metodo applicabile è detto *determined*, tale per cui i valori di codifica rimangono fissi ogni qual volta in cui la tecnica viene utilizzata (Hancock e Khoshgoftaar 2020), a prescindere dal dataset e dal target. All'interno

del caso studio è stata applicata la tecnica di *label encoding* per l'attributo "Status", la quale ricade in questa prima categoria.

```
#Status = 1 se prenotato ("R"), 0 se disponibile ("A")
df["Status"] = df["Status"].map({"A": 0, "R": 1}).astype(int)
```

La tecnica in questione consiste nell'assegnare un numero intero a ciascuna variabile categorica, nel caso in cui si tratti di attributi binari, ovvero attributi che possono assumere solamente due valori possibili, è rispettivamente associato il valore 0 ed 1. Nel caso d'esempio si è imposto che se la struttura è prenotata (R) l'attributo Status deve essere 1, altrimenti è 0 se disponibile.

All'interno di questa prima categoria ricade anche la tecnica *one-hot encoding*, la quale è stata applicata nella variante *drop-first* per la variabile "Listening Type" relativa alla tipologia di struttura affitata. Questa tecnica prevede che dato un attributo che può assumere un insieme finito di k valori, esso venga scomposto in k attributi, uno per ciascuno dei valori possibili. Ciascun dato avrà associato un vettore di attributi la cui somma deve essere massimo 1, in quanto il valore assunto da ciascuna caratteristica sarà 1 per il campo presente nell'attributo originale, 0 per i restanti. La variante scelta prevede che vengano creati $k - 1$ attributi, detti *dummy*, tale per cui se la somma dei valori del vettore è nulla, allora il campo iniziale è da associarsi al k-esima categoria omessa, detta *baseline*. Tale alternativa consente di evitare problemi di multicollinearità perfetta, condizione in cui almeno un riga sia combinazione lineare esatta delle altre, comportando nel contesto di regressione inferenze non valida e parametri non identificati (Pillai e Mohan 2024).

Per l'attributo d'interesse *Listing Type* è stata scelta come baseline la categoria "entire home/apt", in quanto sulla base di quanto visto nel secondo capitolo in Figura 4.1 è la più rappresentata:

```
s = df["Listing Type"].astype("string").str.strip().str.title()
X_listing = pd.get_dummies(s, prefix="ListingType", drop_first=True)
df = pd.concat([df.drop(columns=["Listing Type"]), X_listing], axis=1)
```

Tale metodo si pone in contrapposizione all'obiettivo di "riduzione della dimensionalità" previsto in fase di *data transformation*. Infatti, nel pre-processing è necessario valutare le proprie necessità, prendendo le adeguate decisioni. Dunque, nel caso studio risulta più importante predisporre i dati all'applicazione della metodologia, rispetto a ridurne la dimensionalità.

Una volta conclusa una prima definizione degli step di pre-processing, è emerso che risulta migliorativo eseguire questa trasformazione dopo la gestione dei valori nulli. In quanto, pur non avendo la colonna *Listing Type* valori nulli, vi sono altri attributi con valori nulli che possono essere ricostrui tramite imputazione condizionata a partire da *Listing Type*.

Per tale ragione, se l'esecuzione della trasformazione non venisse spostata, sarebbe necessario ricostruire l'attributo d'interesse in fase di imputazione condizionata, per poi riscomporlo.

A seguire, la seconda categoria di tecniche encoding prende il nome di *algorithmic*, e prevede che il codice venga calcolato a partire dai dati con una procedura statistica. Mentre, i metodi *automatic* si distinguono in qualità di encoding apprese, in cui valori sono parametri da ottimizzare all'interno del modello. Quest'ultime due tipologie di tecniche non sono risultate necessarie per gli attributi a disposizione nel caso studio.

Missing Value Treatment

I dati mancanti possono influire in modo significativo sull'integrità e l'idoneità dei set di dati, portando a risultati statistici inaffidabili, distorsioni e decisioni errate. Inoltre, la presenza di valori mancanti nei dati introduce imprecisioni nel clustering e nella classificazione, compromettendo la validità di tali analisi (Alam et al. 2023).

Per comprendere come risulta più opportuno gestire quest'ultimi all'interno del caso studio, è necessario valutare i possibili pattern dei missing value per *structured data*, sulla base di quanto riportato in Figura 4.5 (Zhou, Aryal e Bouadjenek 2024).

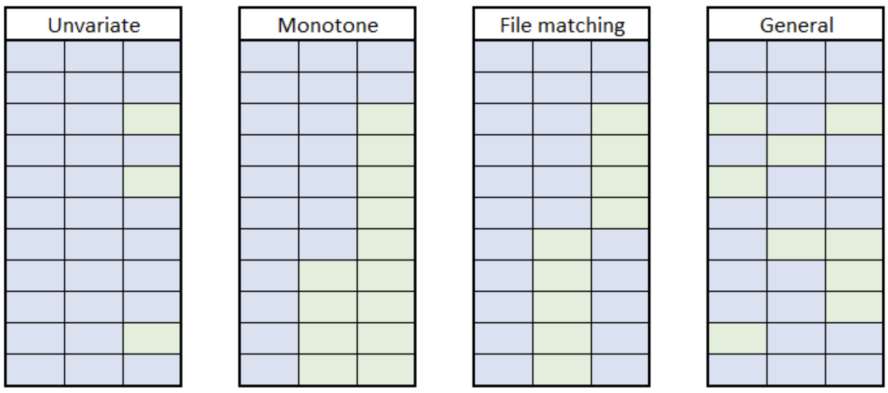


Figura 4.5. *Pattern* dei dati mancanti

Nel modello ***univariate***, i valori mancanti si verificano solo in un attributo del dati. Invece, in un modello ***monotono***, essi si verificano sistematicamente in una direzione all'interno del set di dati. Ciò implica che, una volta che un valore è mancante, anche tutti i valori successivi nel set di dati sono mancanti. Infine, il pattern ***file matching*** illustra uno schema di connessione in cui è possibile raggiungere tutti i valori non nulli con movimenti orizzontali o verticali. Tuttavia, bastano lievi variazioni per far si che questo schema diventi disconnesso.

Nel caso del data frame d'interesse, diversi attributi possono assumere valore nullo, per ragioni differenti. Quindi, lo scopo di questa sezione è identificare le possibili casistiche di record vuoti, e come gestirle sulla base del pattern presentato.

Dato tale obiettivo, il primo step ha previsto l'identificazione degli attributi con almeno un valore nullo, ed il calcolo della frequenza del valore nullo rispetto al totale dei valori, tramite lo script riportato sotto.

```
# Numero di valori nulli per colonna
null_counts = df.isnull().sum()
# Percentuale di valori nulli rispetto al totale di record
null_percentage = (df.isnull().sum() / len(df) * 100)
null_summary = pd.DataFrame({
    'Null Count': null_counts,
    'Null Percentage (%)': null_percentage
})
```

```
# Visualizzazione dei soli attributi con almeno un valore nullo
null_summary = null_summary[null_summary['Null Count'] > 0]
print(null_summary)
```

Il Figura 4.6 viene riportato l'output dello script ottenuto tramite *Jupyter Notebook*, e successivamente viene esplorato come sono stati gestiti i valori nulli:

	Null Count	Null Percentage (%)
Latitude	51	0.009703
Longitude	51	0.009703
Booked Date	372312	70.833191
Reservation ID	372312	70.833191
Bedrooms	437	0.083140
Bathrooms	162	0.030821
Amenities	1699	0.323239
Latitude_fixed	1698	0.323048
Longitude_fixed	1580	0.300599

Figura 4.6. Attributi con valori nulli

In primis, dato lo scopo della tesi, gli attributi "Latitudine", "Longitudine" e "Prezzo" sono dati che non possono essere approssimati, di conseguenza si è proceduto eliminando le intere righe di record per cui questi attributi, e i rispettivi derivati, risultavano nulli:

```
df = df.dropna(subset=["Longitude", "Latitude", "Longitude_fixed", "Latitude_fixed",
↪  "Price (USD)", "Price (EUR)"])
```

Invece, i due attributi "Booked Date" e "Reservation ID" hanno quasi il 71% di valori nulli, in quanto assumono un valore diverso da "Nan" solamente se la struttura è stata prenotata nella data d'interesse; una percentuale elevata di valori nulli non risulta quindi anomala. Però, si è deciso di procedere eliminando la Reservation ID come attributo in quanto non esprime un'informazione aggiuntiva.

Congiuntamente, è stato introdotto un attributo binario 1/0 che definisce se la struttura è stata prenotata, sulla base del valore nullo o meno della data di prenotazione, lo step in questione rientra nelle attività previste nella fase di Data Trasformation:

```
# Imporre 1 se Booked Date non è NaN, 0 altrimenti
df["Booked"] = df["Booked Date"].notna().astype(int)
# Elimina la colonna Reservation ID
df = df.drop(columns=["Reservation ID"])
```

Il numero di recensioni ha un elevato valore di missing value poichè potrebbero non essere ancora state pubblicate recensioni per quella struttura. Si è quindi deciso di imputare 0, ovvero nessuna recensione, nel caso in cui il campo risulti Nan e non vi sia un rating associato alla struttura, altrimenti 1.

```
df["Number of Reviews"] = np.where(
    df["Number of Reviews"].isna() & df["Overall Rating"].isna(),
    0,
    np.where(
        df["Number of Reviews"].isna() & df["Overall Rating"].notna(),
        1,
        df["Number of Reviews"]
    )
)
df["Number of Reviews"] = df["Number of Reviews"].astype("int32")
```

Proseguendo, sia la variabile *Bedrooms*, sia *Bathrooms* sono attributi numerici discreti correlati a *Listing Type* e *Max Guests*. Quindi, la strategia utilizzata prevede la definizione di un'imputazione condizionata per cluster logici tramite la definizione della mediana per *Listing Type* e fasce di *Max Guests*. Tale correzione deve tenere conto dei dummy creati in fase di *Data Trasformation*.

```
# Quartili Max Guests
df["MaxGuests_bin"] = pd.qcut(
    pd.to_numeric(df["Max Guests"], errors="coerce"),
    q=4,
    duplicates="drop"
)
```

```
# Funzione per imputazione condizionata
def impute_conditional_median(df, target):
    x = pd.to_numeric(df[target], errors="coerce").copy()
    med_cluster = (df.groupby(["Listing Type", "MaxGuests_bin"],
    ↪  observed=True)[target]
                    .transform("median"))
    med_type    = df.groupby("Listing Type",
    ↪  observed=True)[target].transform("median")
    med_global  = x.median()
    # Imputazione gerarchica
    m = x.isna()
    x[m] = med_cluster[m]
    m = x.isna()
    x[m] = med_type[m]
    m = x.isna()
    x[m] = med_global
    df[target] = x.round().astype("Int64")
    return df
df = impute_conditional_median(df, "Bedrooms")
df = impute_conditional_median(df, "Bathrooms")
df = df.drop(columns=["MaxGuests_bin"])
```

Quindi, sono state calcolate le mediane del numero di bagni e stanze da letto per ciascuna combinazione tra tipologia di alloggio e fasce di capienza degli ospiti ottenute mediante suddivisione in quartili.

Successivamente, tramite un approccio gerarchico, i valori mancanti dei due attributi sono stati sostituiti, in ordine, con la mediana del gruppo di appartenenza, la mediana aggregata per tipologia di alloggio, e infine, se necessario, con la mediana complessiva del dataset. Tale metodo, ispirato alla logica della ***conditional median imputation*** (Little e Rubin 2002), consente di preservare la coerenza interna dei dati, evitando distorsioni dovute a sostituzioni casuali.

Sulla base dell' output ottenuto, l'applicazione della procedura ha portato all'eliminazione completa dei valori mancanti nelle due variabili in esame.

I *missing value* per l'attributo *Overall Rating* sono strutturali, in quanto nelle circostanze in cui non vi sono reviews, risulta mancare anche la valutazione complessiva. Di conseguenza questo valore manca, sia quando è nullo il valore per l'attributo relativo al numero di reviews, sia quando esso è 0. In tal caso la soluzione non potrebbe essere imputare di default 0, in quanto risulterebbe nel confondere "nessuna valutazione" con "pessima valutazione".

Per risolvere tale vincolo, l'attributo deve essere valutato congiuntamente ai valori nulli dell'attributo Superhost, in quanto vi sono dei criteri empirici minimi di *Overall Rating* per poter essere considerati tali.

Quindi, il primo step prevede verificare per ciascun valore nullo dell'attributo "Superhost", se lo stesso *HOST ID* compare altrove con valore noto, in tal caso si impone l'utilizzo del mode definito per quell'host.

```
# Mode dell'host sui valori noti
host_mode = (df.groupby("Airbnb HOST ID")["Airbnb Superhost"]
               .apply(lambda s: s.dropna().mode().iloc[0] if not s.dropna().empty
               ↪  else np.nan))
# Imputazione del mode per host se noto
df["Airbnb Superhost"] = df["Airbnb Superhost"].fillna(df["Airbnb HOST
↪  ID"].map(host_mode))
```

Analogamente, sulla base della "Property ID", se vi sono più istanze per una struttura, ed alcune di esse hanno "Overall Rating" nullo, si impone per esse che l'attributo è uguale alla media delle valutazioni note per quella struttura.

Non si presuppone che un'istanza con la medesima Property ID debba avere il rating costante in quanto potrebbero esserci delle variazioni a seguito di nuove recensioni.

```
# Media Rating per ciascuna struttura su valori noti
property_mean_rating = (
    df.groupby("Property ID")["Overall Rating"]
      .transform("mean")
)
```

```
# Rating mancante, ma noto almeno un valore noto per la Property
mask_rating_na_and_mean_available = (
    df["Overall Rating"].isna()
    & property_mean_rating.notna()
)
df.loc[
    mask_rating_na_and_mean_available,
    "Overall Rating"
] = property_mean_rating[mask_rating_na_and_mean_available]
```

Il secondo step prevede la definizione della soglia empirica, al di sotto della quale una struttura non può essere gestita da un *Super Host*.

Infatti, come emerge dal grafico riportato in Figura 4.7, il quale mette in relazione il numero di superhost per ciascuna fascia di Overall Rating, vi è una soglia di Overall Rating al di sotto del quale il label ”Superhost” è da considerarsi outlier verso il basso.

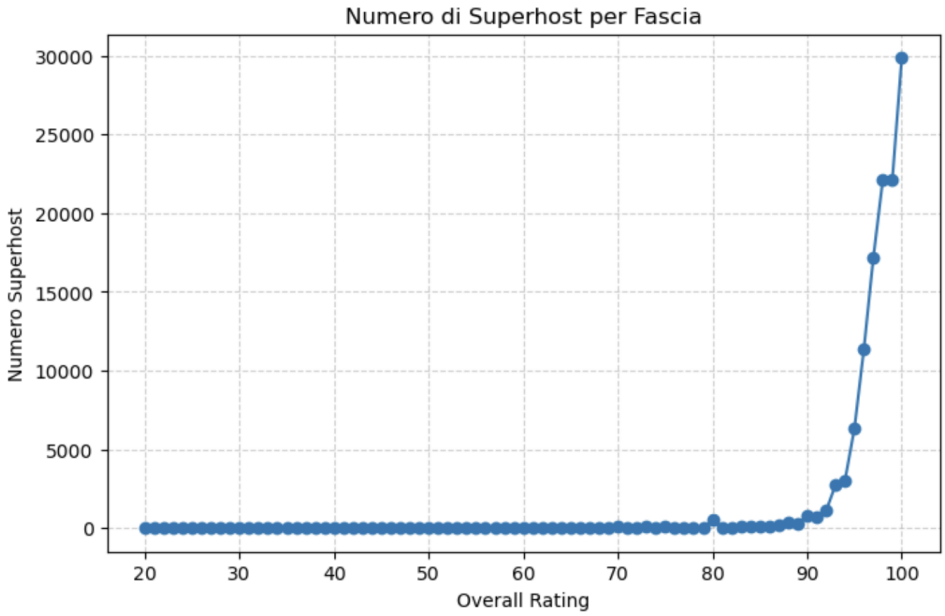


Figura 4.7. Relazione del numero di *Superhost* per ciascuna fascia di *Overall Rating*

Secondo il calcolo che segue sono stati valutati outlier i Superhost con rating inferiore a 89, in quanto la cumulata delle istanza è inferiore al 2% del totale di Superhost:

```
# Media del rating per ciascun host
host_mean_rating = df.groupby("Airbnb HOST ID")["Overall Rating"].transform("mean")
mask_rating_na = df["Overall Rating"].isna()
# Rating NaN ma altri rating noti per lo stesso host --> rating medio
df.loc[
    mask_rating_na & host_mean_rating.notna(),
    "Overall Rating"
] = host_mean_rating[mask_rating_na & host_mean_rating.notna()]
# Rating  NaN, no rating noto per altre strutture e Superhost NaN --> Superhost = 0
df.loc[
    mask_rating_na
    & host_mean_rating.isna()
    & df["Airbnb Superhost"].isna(),
    "Airbnb Superhost"
] = 0
# Soglia Rating per Superhost
superhosts = df[(df["Airbnb Superhost"] == 1) & df["Overall Rating"].notna()].copy()
rating_counts = (
    superhosts["Overall Rating"]
    .value_counts()
    .sort_index()
    .reset_index()
)
rating_counts.columns = ["OverallRating", "Count"]
rating_counts["CumulativePct"] = (
    rating_counts["Count"].cumsum() / rating_counts["Count"].sum() * 100
)
threshold_row = rating_counts[rating_counts["CumulativePct"] < 2].tail(1)
if not threshold_row.empty:
    threshold = int(threshold_row["OverallRating"].values[0])
    print(f"Soglia cumulata (<2%) = {threshold}")
else:
    threshold = 89
    print("Nessun valore di rating con cumulata < 2%. Uso soglia = 89.")
host_mean_rating_all = df.groupby("Airbnb HOST ID")["Overall
↪  Rating"].transform("mean")
```

```
# Host con rating medio </=soglia
mask_low_hosts = host_mean_rating_all.notna() & (host_mean_rating_all <= threshold)
df.loc[mask_low_hosts, "Airbnb Superhost"] = 0
mask_low_hosts_high_rating = (
    mask_low_hosts
    & df["Overall Rating"].notna()
    & (df["Overall Rating"] > threshold)
)
df.loc[mask_low_hosts_high_rating, "Overall Rating"] = \
host_mean_rating_all[mask_low_hosts_high_rating]
# Superhost NaN con rating medio > soglia
mask_high_hosts = host_mean_rating_all.notna() & (host_mean_rating_all > threshold)
mask_na_super_high_host = (
    mask_high_hosts
    & df["Airbnb Superhost"].isna()
)
df.loc[mask_na_super_high_host, "Airbnb Superhost"] = 1
mask_inconsistent_high = (
    mask_high_hosts
    & (
        (df["Airbnb Superhost"] == 0)
        | (df["Overall Rating"] < threshold)
    )
)
df.loc[mask_inconsistent_high, "Overall Rating"] =
↪  host_mean_rating_all[mask_inconsistent_high]
df.loc[mask_inconsistent_high, "Airbnb Superhost"] = 1
#Valori nulli Superhost residui
null_superhost = df["Airbnb Superhost"].isna().sum()
```

Come step successivo, si è utilizzata la soglia calcolata per eliminare le casistiche in cui l'attributo Superhost è nullo, e ridurre quelle in cui non è noto l'Overall rating.

In tale fase è necessario tenere in considerazione che lo status "Superhost" è proprio dell'host, mentre l'overall rating è caratterizzante delle strutture gestite, di conseguenza vi potrebbero essere situazioni in cui il medesimo host è associato ad Overall Rating

differenti.

Quindi, in ciascuna delle tre verifiche implementate sulla base del valore soglia calcolato, sono state integrate le adeguate considerazioni per evitare inconsistenze sul valore di *Superhost* assegnato ad un host presente in più istanze dei dati.

La prima verifica preliminare mira a ridurre il numero di casistiche in cui l'*Overall Rating* è nullo, rendendo più consistenti le verifiche successive. Perciò, nel caso in cui il valore di rating sia nullo per una struttura, ma l'host associato abbia altre strutture, si impone la media di rating di quest'ultime. Inoltre, se anche l'attributo Superhost è nullo e non vi sono rating per quell'host, allora viene imputato conservativamente che non si tratta di un Superhost.

La seconda verifica prevede di imporre, sovrascrivendo anche i valori dell'attributo già noti, che l'host non venga considerato "Super" nel caso in cui il rating medio delle strutture da lui gestite sia inferiore o uguale alla soglia. Dopo tale modifica, se una delle strutture singolarmente prese risulta avere un rating superiore alla soglia, allora tale valore è da sostituire con la media dell'host. Quest'operazione consente di rimuovere sia gli outlier, sia una quota di valori nulli per l'attributo "Superhost".

Infine, se l'attributo *Superhost* è nullo, ma le sue strutture hanno mediamente un rating maggiore della soglia, si impone l'attributo "Super" attivo. In tale verifica, se una delle istanze associate all'host ha l'attributo Superhost = 0 e/o l'overall rating inferiore alla soglia, è da reiterare la condizione, sostituendo il rating dell'istanza con il rating medio dell'host.

Conclusa questa fase, si è verificato che non vi fossero host con l'attributo "Super" inconsistente:

```
# Numero di valori distinti di Superhost per ciascun host
superhost_nunique = (
    df.groupby("Airbnb HOST ID")["Airbnb Superhost"]
      .nunique(dropna=False)
)
#Host per cui Superhost NON è sempre uguale
inconsistent_hosts = superhost_nunique[superhost_nunique > 1].index
```

```
print(f"Numero di HOST ID con Superhost incoerente: {len(inconsistent_hosts)}")
df_inconsistent = df[df["Airbnb HOST ID"].isin(inconsistent_hosts)].copy()
df_inconsistent = df_inconsistent.sort_values(
    by=["Airbnb HOST ID", "Airbnb Superhost", "Overall Rating"]
)
print(df_inconsistent[["Airbnb HOST ID", "Airbnb Superhost", "Overall
↪  Rating"]].head(50))
```

L'output dello script in questione ha restituito quanto segue:

```
Numero di HOST ID con Superhost incoerente: 0
Empty DataFrame
```

Quindi, l'approccio implementato all'interno delle tre condizioni è risultato valido nel gestire questa potenziale problematicità.

Successivamente, sono stati gestiti i restanti valori mancanti di *Overall Rating*, associando ad essi la media della rispettiva classe di Superhost:

```
# Calcolo delle medie per ciascuna categoria di Superhost
mean_super = df.loc[df["Airbnb Superhost"] == 1, "Overall Rating"].mean(skipna=True)
mean_non_super = df.loc[df["Airbnb Superhost"] == 0, "Overall
↪  Rating"].mean(skipna=True)
mask_rating_na   = df["Overall Rating"].isna()
mask_super_known = df["Airbnb Superhost"].isin([0, 1])  # Superhost noto
# Imputazione condizionata
df.loc[mask_rating_na & (df["Airbnb Superhost"] == 1), "Overall Rating"] = mean_super
df.loc[mask_rating_na & (df["Airbnb Superhost"] == 0), "Overall Rating"] =
↪  mean_non_super
# Conversione finale
df["Overall Rating"] = df["Overall Rating"].astype(float)
```

Sulla base dell'output stampato per i controlli riepilogativi, persiste una percentuale di righe che non hanno *Overall Rating*, poichè corrispondono alle casistiche in cui non vi

sono recensioni, però non vi sono istanze per cui l'*Overall Rating* è nullo nonostante vi siano recensioni.

Infine, l'attributo "Amenities" verrà trattato nella prossima sezione, convertendo ciascuna lista di stringhe in attributi distinti binarizzati. Di conseguenza, se l'attributo risulta nullo, verrà trattato come se la struttura non disponesse di amenities.

Quindi, il trattamento dei valori mancanti si è concluso restituendo valori nulli solo per attributi portatori di un significato, consentendo una maggiore consistenza e la rappresentatività del dataset tramite la combinazione di tecniche di imputazione condizionata e di sostituzione basata su regole semantiche.

Feature Engineering

Il feature engineering rappresenta un passaggio volto a trasformare i dati grezzi in rappresentazioni più informative tramite la costruzione o la manipolazione di variabili che massimizzano la capacità esplicativa del dataset rispetto al fenomeno oggetto di analisi.

Dal punto di vista operativo, il *feature engineering* può includere diverse tipologie di operazioni:

- *Feature selection*: identificazione delle variabili più rilevanti per il modello;
- *Feature construction*: creazione di nuove variabili derivate da combinazioni, interazioni o trasformazioni di quelle esistenti.

Come evidenziato da Zheng e Casari (2018), la qualità del *feature engineering* ha spesso un impatto più determinante sulle prestazioni del modello rispetto alla scelta dell'algoritmo stesso. In questo senso, esso può essere considerato un processo di "modellazione anticipata", poiché incorpora conoscenza esperta e ipotesi sul dominio nel dataset prima della fase di apprendimento vero e proprio.

Come anticipato nella sezione precedente, nel contesto di "*feature construction*" si inserisce l'attributo *Amenities*, al quale viene applicato il metodo di ***binarizzazione multilabel***.

Tramite tale metodologia, se una variabile può assumere più etichette simultaneamente, essa viene rappresentata mediante un vettore binario di lunghezza pari al numero totale di etichette possibili, in cui ciascun elemento assume valore 1 se l'etichetta è presente e 0 altrimenti (Żak e Woźniak 2020). Tale approccio è stato utilizzato per gestire la feature in questione, poichè essa prevede come campo una lista, con lunghezza variabile, delle *amenities* disponibili nella struttura corrispondente:

```
#Lista delle amenities
features_df["Amenities"] = features_df["Amenities"].apply(
    lambda x: ast.literal_eval(x) if isinstance(x, str) and x.startswith("[") else []
)
# Amenities Binarizzate
mlb = MultiLabelBinarizer()
amenities_encoded = pd.DataFrame(mlb.fit_transform(features_df["Amenities"]),
                                 columns=mlb.classes_,
                            index=features_df.index)
features_df = pd.concat([features_df.drop(columns=["Amenities"]),
↪  amenities_encoded], axis=1)
```

Tale costruzione di attributi è necessaria in quanto gli algoritmi di regressione non accettano dati testuali, categoriali o strutturati come liste; essa permette a ciascun attributo codificato di diventare una variabile esplicativa indipendente, con un propro coefficiente associato.

Infine, poichè, com'è stato evidenziato in precedenza, non vi sono in questo caso studio problemi di numerosità dei dati, dal punto di vista della "feature selection", non sono state implementate azioni.

Imbalanced Data Treatment

Un'ultima sfida da affrontare in fase di pre-processing è la gestione dei dati sbilanciati, ovvero dati caratterizzati da una distribuzione delle classi irregolare all'interno del dataset. In tale circostanza la maggior parte dei record appartengono alla classe dominante, mentre solo una ridotta porzione di dati appartiene alle altre classi, portando l'algoritmo

ad avere un bias a favore della classe dominante (Altalhan, Algarni e Turki-Hadj Alouane 2025).

Nel dataset di cui si dispone per il caso studio la variabile target "Price" è continua, per cui non sono definite classi.

Il concetto di squilibrio può essere applicato tra le classi di variabili categoriali usate come regressori del modello. Però, non essendo le variabili in questione target di classificatori, si procede solamente utilizzandole come *dummies di controllo*, ovvero includendole nel modello per isolare meglio l'effetto delle altre variabili.

Inoltre, come mostrato in letteratura da Zhao et al. 2018, l'utilizzo della penalizzazione del modello di LASSO consente di ridurre il peso di dummies associate a poche osservazioni, mitigando il rischio di *overfitting* [4] su classi rare.

Una volta concluse tutte le fase di pre-processing sono stati salvati i dati elaborati all'interno di un nuovo file csv, che verrà utilizzato in fase di analisi tramite la rinominazione `Processed_df`.

[4] Tendenza di un modello ad adattarsi eccessivamente ai dati con cui viene allenato, catturando anche il rumore e peggiorando la capacità di generalizzazione su nuovi dati (Ghojogh e Crowley 2023).

4.3 Descrizione dei dati

La presente sezione fornisce una descrizione esplorativa del dataset utilizzato nell'analisi empirica, con l'obiettivo di sintetizzarne le principali caratteristiche statistiche e distributive, supportando le scelte decisionali in fase di sviluppo della metodologia.

Distribuzione dei prezzi

Noto il contesto generale fornito nella Sezione 4.1, il primo aspetto da analizzare, è la distribuzione dei prezzi proposti dalle strutture, in quanto l'analisi preliminare della distribuzione dei prezzi consente di ottenere una prima evidenza empirica del grado di dispersione del mercato. Inoltre, la valutazione della distribuzione dei prezzi rappresenta il punto di partenza operativo per la costruzione del modello edonico e per la successiva segmentazione spaziale.

L'analisi in questione viene riportata in Figura 4.8, la quale illustra la distribuzione dei prezzi giornalieri delle strutture Airbnb attive nella regione nel periodo d'interesse. Per motivi metodologici, l'analisi è stata limitata al 99° percentile della variabile prezzo.

Graficamente emerge una forte asimmetria positiva, infatti la maggior parte delle strutture presenta prezzi medio-bassi, concentrati in un intervallo ristretto, mentre una minoranza di alloggi di fascia alta spinge la distribuzione verso destra. Tale configurazione è coerente con la natura del mercato della sharing economy, caratterizzato da un'elevata eterogeneità dell'offerta e da una competizione concentrata sulle fasce centrali di prezzo.

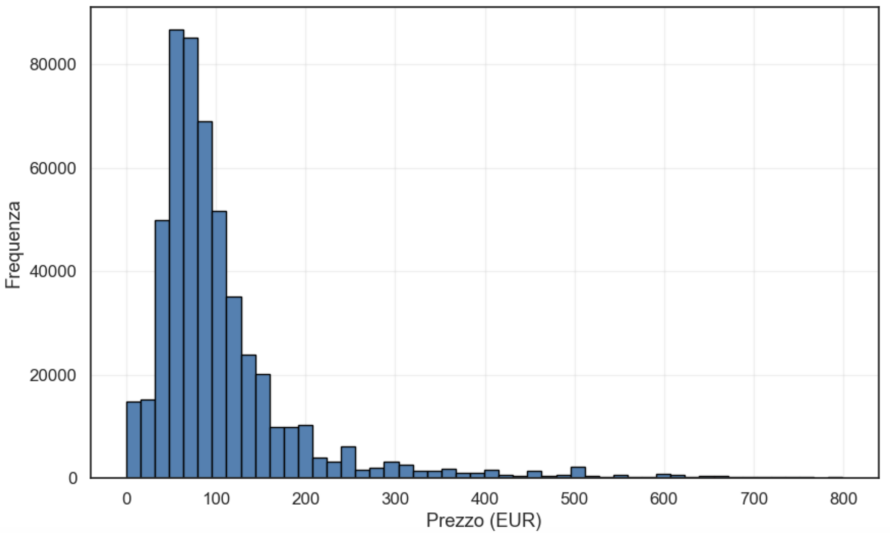


Figura 4.8. Distribuzione dei prezzi - 99º Percentile

Per approfondire la lettura dei dati, la Figura 4.9 mostra la distribuzione logaritmica dei prezzi, sempre considerando il 99° percentile. L'applicazione del logaritmo naturale ai valori di prezzo risponde a due esigenze analitiche: ridurre l'eteroschedasticità e normalizzare la scala dei valori, rendendo la distribuzione più simmetrica e comparabile con le ipotesi dei modelli econometrici successivi (Astivia e Zumbo 2019). Dopo la trasformazione, la distribuzione appare più regolare e prossima a una forma normale, confermando che, al netto delle osservazioni estreme, i prezzi degli alloggi si distribuiscono intorno a un valore centrale che rappresenta il livello medio del mercato piemontese.

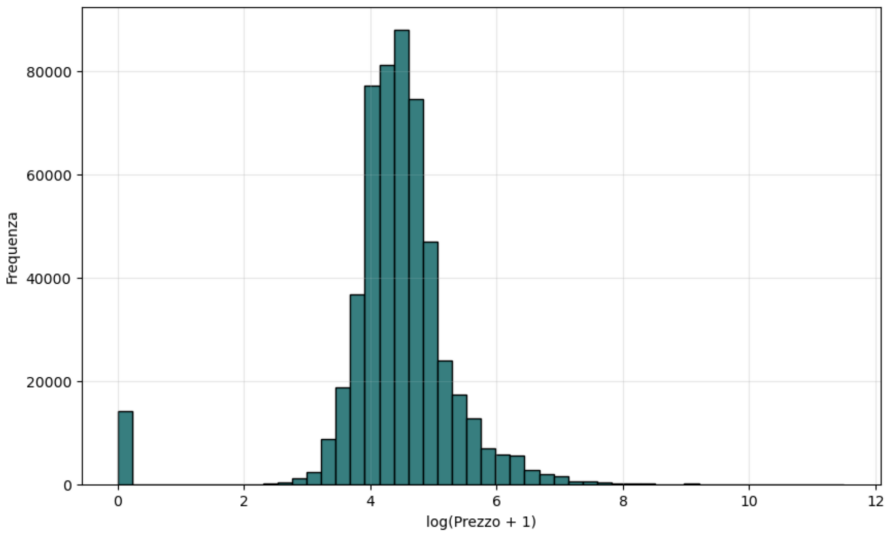


Figura 4.9. Distribuzione logaritmica dei prezzi - 99º Percentile

Distribuzione geografica delle strutture

La distribuzione geografica delle strutture Airbnb evidenzia una marcata eterogeneità spaziale dell'offerta all'interno dell'area di studio. Le strutture risultano fortemente concentrate in specifiche porzioni del territorio, mentre ampie aree presentano una densità significativamente inferiore, come mostrato dalla rappresentazione cartografica puntuale e dalla mappa di densità.

Dal punto di vista quantitativo, l'area complessivamente coperta dalle osservazioni è pari a 29.463,19 km^2, all'interno della quale si registra una densità media di 17,68 strutture per km^2. Tale valore medio, tuttavia, maschera una distribuzione fortemente non uniforme, caratterizzata da cluster locali di elevata concentrazione alternati a zone a bassa presenza di offerta.

Il centroide geografico dell'offerta, calcolato come media delle coordinate spaziali delle strutture, fornisce una sintesi della localizzazione media del fenomeno. La sua posizione, raffigurata nella Figura 4.10 risulta coerente con le aree a maggiore intensità di strutture, indicando una polarizzazione dell'offerta verso la città di Torino. Il centroide non rappresenta un punto di attrazione o di competizione, ma costituisce un riferimento spaziale utile per valutare l'asimmetria della distribuzione territoriale (Deakin, Bird e Grenfell 2002). A supporto di tale evidenza, la dispersione spaziale, misurata come deviazione standard delle distanze delle strutture dal centroide, risulta pari a 47,04 km. Questo valore segnala una diffusione geografica ampia dell'offerta attorno al centro medio, compatibile con una struttura territoriale caratterizzata da un nucleo centrale denso e da un'estensione periferica meno omogenea.

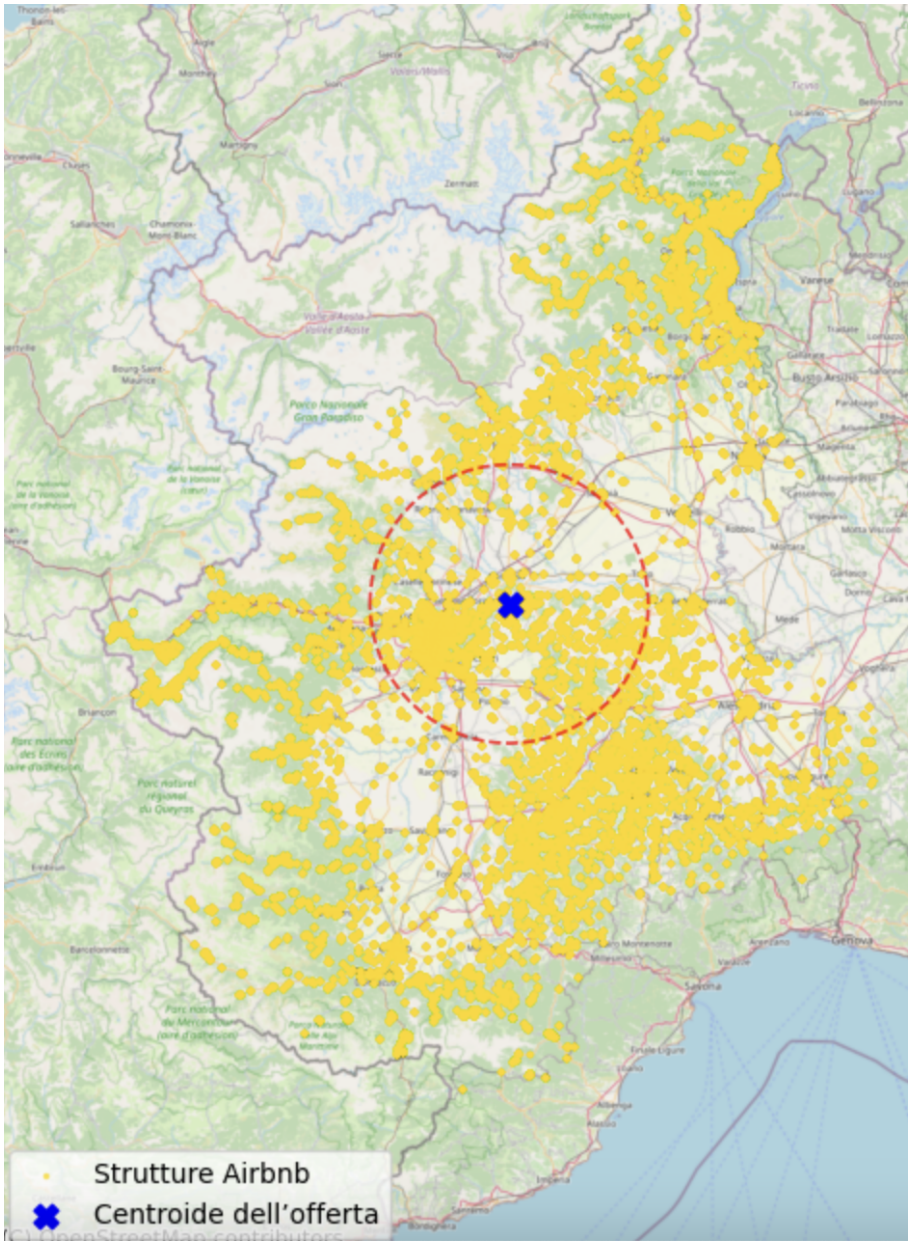


Figura 4.10. Distribuzione geografica delle strutture e centroide dell'offerta

Un'ulteriore conferma della forte concentrazione spaziale emerge dall'*indice di Gini geografico*, pari a 0,989. L'indice di Gini, originariamente sviluppato per misurare la disuguaglianza nella distribuzione dei redditi, può essere applicato alla distribuzione spaziale contando il numero di strutture all'interno di celle territoriali omogenee. Valori prossimi a zero indicano una distribuzione uniforme, mentre valori prossimi all'unità segnalano una concentrazione estrema. Il valore osservato evidenzia quindi una fortissima disuguaglian-

za spaziale dell'offerta, con una quota rilevante di strutture concentrata in un numero limitato di aree (Rey e R. J. Smith 2013). In fase di valutazione dei risultati tale valore verrà valutato anche relativamente alla distribuzione della pressione competitiva.

Nel complesso, l'evidenza descrittiva mostra che l'offerta Airbnb non è distribuita uniformemente nello spazio, ma presenta forti concentrazioni locali e un'elevata eterogeneità territoriale. Tale configurazione costituisce un presupposto empirico rilevante per l'analisi successiva, in quanto suggerisce che le interazioni tra strutture e le dinamiche di prezzo tendano a manifestarsi prevalentemente a livello locale, rendendo necessaria l'introduzione di criteri di prossimità spaziale nella modellizzazione econometrica.

4.4 Definizione della metodologia

In questa sezione viene definito il modello analitico alla base dell'analisi empirica sviluppata nel capitolo successivo. L'obiettivo è formalizzare un approccio in grado di identificare e interpretare le dinamiche di competizione di prezzo nel mercato degli affitti brevi, integrando informazioni sulle caratteristiche strutturali degli alloggi, sulla prossimità spaziale e sulla percezione dei consumatori.

L'analisi è condotta in prospettiva *cross-section*, aggregando le osservazioni giornaliere a livello di singola struttura Airbnb, e si articola in quattro fasi principali: (i) stima del modello edonico di prezzo mediante regressione penalizzata di LASSO, (ii) isolamento della componente competitiva residua, (iii) costruzione della rete di prossimità spaziale e stima dell'intensità competitiva locale, (iv) identificazione delle aree di competizione autocontenuta.

4.4.1 Definizione del modello edonico di prezzo

Si consideri un insieme di N strutture Airbnb localizzate nella regione Piemonte. Per ciascuna struttura $i = 1, \ldots, N$ si osserva un insieme di prezzi giornalieri nel periodo di riferimento, a partire dai quali viene costruita una misura aggregata del prezzo medio per notte P_i. Al fine di ridurre l'asimmetria della distribuzione e facilitare l'interpretazione dei coefficienti stimati, il prezzo viene trasformato in scala logaritmica:

$$Y_i = \log(P_i).$$

Il prezzo logaritmico viene modellato come funzione delle caratteristiche osservabili dell'alloggio, riunite nel vettore X_i, secondo una specificazione di tipo edonico:

$$Y_i = \alpha + X_i'\beta + \varepsilon_i, \tag{4.1}$$

dove α rappresenta l'intercetta, β è il vettore dei coefficienti associati alle caratteristiche dell'offerta e ε_i indica la componente del prezzo non spiegata dalle variabili osservabili

incluse nel modello.

Nel caso in esame, il vettore X_i comprende caratteristiche dimensionali dell'alloggio (`Bedrooms`, `Bathrooms`, `Max Guests`), indicatori di qualità e reputazione (`Overall Rating`, `Number of Reviews`), vincoli contrattuali (`Minimum Stay`) e una misura sintetica delle dotazioni disponibili (`amenities_count`). Tutte le variabili esplicative sono standardizzate prima della stima, al fine di garantire la confrontabilità dei coefficienti e migliorare la stabilità numerica del modello.

L'output del modello verrà utilizzato nella fase successiva per distinguere tra effetti strutturali e componenti residuali del prezzo.

4.4.2 Stima penalizzata di LASSO

La stima del modello edonico è effettuata mediante regressione LASSO, che introduce una penalizzazione ℓ_1 sui coefficienti:

$$\min_{\alpha,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (Y_i - \alpha - X_i'\beta)^2 + \lambda \sum_k |\beta_k| \right\}, \tag{4.2}$$

dove il parametro di regolazione λ deve essere selezionato tramite *cross-validation*.

Al fine di integrare una prospettiva demand-side e valutare la validità del modello, l'analisi in un secondo momento verrà estesa mediante l'utilizzo di un Adaptive LASSO, in cui la penalizzazione è ponderata da pesi specifici associati alle singole variabili, modificando l'ultimo fattore come segue: $\lambda \sum_k w_k |\beta_k|$.

I pesi w_k sono definiti inversamente proporzionali all'importanza percepita delle caratteristiche dagli utenti, rilevata tramite un questionario strutturato rivolto a 105 utenti. In particolare, maggiore è la rilevanza attribuita dall'utente su una scala da 1 a 7, ad una caratteristica nel processo di scelta della struttura, minore è la penalizzazione associata al suo coefficiente, consentendo al modello di preservarne l'effetto stimato.

Inoltre, l'impatto del valore assegnato da ciascun utente viene ulteriormente pesato sulla base della frequenza di utilizzo della piattaforma Airbnb.

4.4.3 Baseline strutturale del prezzo

Al fine di analizzare le dinamiche competitive tra le strutture, la stima del modello di prezzo viene qui utilizzata esclusivamente come *baseline strutturale*. Una volta controllato per le caratteristiche osservabili dell'offerta e per i fattori sistematici inclusi nel modello, il prezzo non costituisce più l'oggetto diretto dell'analisi.

I residui del modello, indicati con ε_i, rappresentano la componente del prezzo non spiegata dalla baseline strutturale e vengono pertanto assunti come grandezza informativa di riferimento. Tali residui sintetizzano variazioni locali del pricing non riconducibili agli attributi strutturali considerati e consentono di concentrarsi sulle interazioni tra le strutture al netto delle differenze osservabili.

Nel prosieguo dell'analisi, i residui ε_{it} sono trattati come input fissato e costituiscono la base per l'isolamento e la modellazione delle interdipendenze competitive. In particolare, la sezione successiva utilizza esclusivamente tale componente residuale per identificare e analizzare le relazioni competitive tra le strutture, mantenendo invariata la specificazione del modello di prezzo. I residui sono pertanto interpretati come misura di co-movimento condizionato dei prezzi e non in chiave causale (Li, Netessine e Koulayev 2018).

4.4.4 Isolamento della componente competitiva

Come anticipato nella sezione precedente, una volta stimato il modello edonico penalizzato, è possibile scomporre il prezzo logaritmico osservato nella componente strutturale $\widehat{Y}_i = \widehat{\alpha} + X_i'\widehat{\beta}$, spiegata dalle caratteristiche osservabili dell'alloggio, e nella componente residua $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$, che definisce lo scarto tra il prezzo logaritmico osservato per la struttura i, ed il valore stimato dal modello edonico.

Nel contesto del presente lavoro, i residui $\widehat{\varepsilon}_i$ vengono interpretati come una misura sintetica della componente di prezzo non riconducibile alle caratteristiche strutturali e qualitative osservabili dell'alloggio.

Infatti, la selezione automatica delle variabili rilevanti tramite LASSO consente di evitare che effetti competitivi vengano assorbiti spuriosamente da covariate strutturali altamente

correlate, garantendo una stima più robusta della componente strutturale del prezzo. In tal modo, i residui possono essere interpretati come deviazioni sistematiche rispetto al prezzo strutturale atteso.

Sotto questa prospettiva, un residuo positivo $\widehat{\varepsilon}_i > 0$ indica una struttura che presenta un prezzo medio superiore a quanto previsto sulla base delle sole caratteristiche osservabili, mentre un residuo negativo segnala una pressione competitiva locale più intensa.

In sintesi, l'isolamento della componente residua del prezzo consente di separare in modo sistematico la componente strutturale della formazione del prezzo dalla componente potenzialmente riconducibile a dinamiche competitive locali.

4.4.5 Set di competitor potenziali

L'analisi delle dinamiche competitive richiede anche di circoscrivere lo spazio entro il quale le interazioni tra strutture sono plausibili. In assenza di tale delimitazione, l'ipotesi di competizione globale tra tutte le unità condurrebbe a relazioni spurie.

A tal fine, per ciascuna struttura i viene definito un insieme di concorrenti potenziali ragionevoli, indicato con $\mathcal{N}(i)$.

$$\mathcal{N}(i) \subset \{1, \ldots, N\} \setminus \{i\}.$$

Tutte le analisi successive di interdipendenza competitiva sono condotte condizionatamente all'insieme $\mathcal{N}(i)$.

La costruzione di $\mathcal{N}(i)$ si basa su criteri di prossimità e similarità, volti a identificare strutture che operano in un contesto di mercato comparabile. In primo luogo, viene adottato un criterio di prossimità geografica, che riflette l'idea che la competizione tra strutture ricettive sia prevalentemente locale e che l'accessibilità spaziale costituisca una dimensione rilevante del confronto competitivo.

In secondo luogo, come discusso in letteratura, viene introdotto un criterio minimo di sostituibilità potenziale tra le strutture, volto a escludere dal set competitivo struttu-

re caratterizzate da un posizionamento dell'offerta marcatamente diverso. Tale criterio interpreta la sostituibilità in senso strutturale ed ex ante, assumendo che due alloggi possano competere solo se risultano alternative plausibili per lo stesso insieme di consumatori sulla base di vincoli fondamentali dell'offerta, quali capacità ricettiva, dotazioni essenziali e condizioni di utilizzo.

Infatti, in assenza di informazioni dirette sul comportamento di scelta, queste caratteristiche osservabili rappresentano una proxy naturale del segmento di domanda servito, consentendo di definire un primo perimetro competitivo coerente dal punto di vista economico.

Per valutare l'impatto di tale imposizione, verrà parallelamente sviluppato un approccio alternativo di natura puramente data-driven. In questo secondo caso, l'identificazione delle relazioni competitive avviene esclusivamente sulla base delle interdipendenze empiriche osservate nella componente residua del prezzo all'interno dello stesso intorno geografico, senza imporre vincoli di comparabilità dell'offerta. Tale confronto nella fase dei risultati consentirà di valutare come l'introduzione di un criterio strutturale influenzi la configurazione delle relazioni competitive e la topologia della rete, rispetto a una definizione del perimetro fondata esclusivamente sulle interdipendenze empiriche dei residui di prezzo.

Invece, la componente di prossimità geografica viene integrata mediante un approccio k-nearest neighbors applicato alle coordinate geografiche degli alloggi, utilizzando la *distanza di Haversine*, una misura della distanza geodetica tra due punti sulla superficie terrestre approssimata come una sfera (M. J. d. Smith, Goodchild e Longley 2018).

L'insieme $\mathcal{N}(i)$ è definito in modo asimmetrico, nel senso che l'appartenenza di una struttura j al set competitivo di i non implica necessariamente che i appartenga al set competitivo di j. Tale scelta riflette l'eterogeneità delle strutture analizzate e consente di catturare relazioni competitive direzionali.

Quindi, l'introduzione del set di competitors potenziali consente di ridurre la dimensionalità del problema ed imporre una struttura economica allo spazio competitivo-

4.4.6 Interazione competitiva

A partire dall'insieme dei competitori individuati, si definisce ora il livello di rappresentazione della competizione locale adottato nell'analisi. In particolare, nel presente lavoro si assume che la struttura competitiva locale possa essere rappresentata mediante un parametro di interazione medio, che sintetizza l'intensità complessiva del co-movimento dei prezzi all'interno del set di competitori rilevanti. Di conseguenza, la rete competitiva utilizzata nelle analisi successive non è stimata a livello di singole relazioni, in coerenza con l'obiettivo di identificare aree di competizione locale piuttosto che micro-relazioni bilaterali.

Quindi, sulla base dell'insieme di concorrenti potenziali identificato, viene definita la misura della pressione competitiva locale tramite la media dei residui degli alloggi appartenenti al vicinato spaziale:

$$\overline{\widehat{\varepsilon}}_{\mathcal{N}_i} = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \widehat{\varepsilon}_j.$$

L'intensità media dell'interazione competitiva viene quindi stimata mediante una regressione lineare:

$$\widehat{\varepsilon}_i = \rho \, \overline{\widehat{\varepsilon}}_{\mathcal{N}(i)} + u_i, \tag{4.3}$$

dove il coefficiente ρ misura la forza della dipendenza competitiva locale. Valori positivi indicano che le componenti di prezzo non spiegate tendono a muoversi in modo coerente tra strutture vicine, suggerendo una competizione spazialmente concentrata.

4.4.7 Costruzione della rete competitiva e delle aree di competizione

Sulla base del coefficiente ρ stimato e della struttura di contiguità definita in precedenza, viene costruita una rete competitiva non direzionale in cui i nodi rappresentano le strutture Airbnb e gli archi connettono esclusivamente coppie di strutture che risultano simultaneamente geograficamente prossime e strutturalmente sostituibili. Il peso dell'arco tra i e j è definito come:

$$w_{ij} = \frac{|\rho|}{d_{ij} + \varepsilon},$$

dove d_{ij} rappresenta la distanza geografica tra le due strutture ed ε è una costante positiva molto piccola introdotta per garantire la stabilità numerica.

La rete competitiva così costruita viene analizzata mediante algoritmi di *community detection*, i quali mirano a identificare automaticamente sottogruppi di nodi, detti comunità, che risultano più densamente connessi tra loro rispetto al resto della rete (Fortunato 2010).

In particolare, viene adottato il ***metodo di Louvain***, un algoritmo euristico per la massimizzazione della ***modularità*** [5] introdotto da Blondel et al. (2008).

Il metodo prevede che, in una prima fase, ciascun nodo venga assegnato alla propria comunità e che vengano successivamente valutati i guadagni di modularità associati allo spostamento del nodo in una comunità adiacente. Tale procedura viene iterata fino a quando non è più possibile ottenere miglioramenti della modularità. In una seconda fase, le comunità individuate vengono aggregate in nodi super-strutturali, generando una nuova rete sulla quale il processo viene reiterato. L'algoritmo converge quando la modularità non può più essere ulteriormente incrementata.

Nel contesto di questa analisi, le comunità individuate sono interpretate come aree di competizione autocontenuta, ossia sotto-mercati locali nei quali le strutture Airbnb risultano maggiormente interconnesse in termini di pressione competitiva, in coerenza con la definizione di competitori adottata a monte.

4.4.8 Risultati attesi

La metodologia proposta è finalizzata a restituire la configurazione spaziale della competizione nel mercato analizzato, tramite la costruzione di una mappa strutturata della pressione competitiva e delle sue articolazioni locali, con un occhio di riguardo per la città di Torino, e le zone rurali della regione. A tale scopo, si vuole ottenere una misura comparabile della pressione competitiva per ciascuna unità osservata, tale da consentire un ordinamento coerente lungo uno spettro di intensità concorrenziale. L'interesse principale

[5] La modularità è una funzione obiettivo che misura la qualità di una partizione della rete confrontando la densità degli archi all'interno delle comunità con quella attesa in un grafo casuale avente la stessa distribuzione dei gradi (Newman e Girvan 2004).

risiede nella forma della distribuzione aggregata: eventuali concentrazioni, discontinuità o fenomeni di polarizzazione costituirebbero evidenza di una struttura di mercato non uniforme, caratterizzata da contesti locali profondamente differenziati.

Il fulcro dell'analisi è tuttavia l'identificazione di aree di competizione. In tal senso, il risultato atteso è la delimitazione di sottosistemi spaziali coerenti sotto il profilo competitivo, all'interno dei quali l'intensità della pressione risulti relativamente omogenea rispetto al resto del territorio.

Particolare rilievo assume il confronto tra tali aree e le delimitazioni amministrative tradizionali. L'analisi dovrebbe consentire di verificare se la geografia competitiva coincida con quella istituzionale oppure se emergano configurazioni alternative. Un eventuale disallineamento rappresenterebbe un risultato di interesse, in quanto indicherebbe che le dinamiche concorrenziali si organizzano secondo logiche spaziali autonome rispetto alla suddivisione amministrativa.

Pur adottando un'impostazione comparativa in alcune specificazioni, l'obiettivo principale rimane la validazione dell'approccio basato su LASSO. In tale contesto, si intende valutare in modo mirato l'impatto che l'integrazione tramite ALASSO di una componente *demand-side* può esercitare sulla configurazione delle aree individuate. Inoltre, verrà analizzato come l'introduzione di vincoli di sostituibilità incida sulla segmentazione territoriale risultante. Tali verifiche consentono di comprendere la robustezza struttura competitiva stimata.

Capitolo 5

Analisi

Il presente capitolo è dedicato all'analisi empirica dei dati e rappresenta il cuore applicativo del lavoro di tesi. In continuità con il framework metodologico delineato nel capitolo precedente, l'obiettivo è tradurre le scelte teoriche in evidenze quantitative, valutando le dinamiche di formazione dei prezzi e i meccanismi competitivi che caratterizzano il mercato degli affitti brevi nella regione Piemonte.

5.1 Imputazioni e trasformazioni preliminari

Questa sezione prepara il terreno alle analisi successive, offrendo una base empirica solida. A tal proposito, il primo step è la creazione di attributi derivati da colonne già note all'interno del dataframe, necessaria per disporre di tutte le variabili previste dal modello teorico.

In primis, è stato definito un attributo che prende il nome di `amenities_count`, il quale tramite il conteggio del numero di servizi aggiuntivi disponibili per ciascuna struttura, consente di includere nel modello le informazioni note sulle *amenities*.

L'attributo ottenuto viene incluso all'interno del vettore `X_features` delle caratteristiche strutturali, qualitative e reputazionali dell'alloggio che determinano il valore edonico del servizio offerto. Le restanti *features* intrinseche all'alloggio, ovvero `"Bedrooms"`, `"Bathrooms"`, `"Max Guests"`, `"Minimum Stay"`, `"Number of Reviews"`, `"Number of Photos"`, `"Overall Rating"` e `"Airbnb Superhost"`, risultano già note senza richiedere trasformazioni.

Successivamente, viene definita la colonna `log_price`, la quale salva il valore della trasformazione logaritmica del prezzo Y_i corrispondente a ciascuna struttura i.

```
processed_df["log_price"] = np.log1p(processed_df[COL_P].astype(float))
```

Per concludere, al fine di ottenere una rappresentazione coerente del mercato, il *dataset* viene riorganizzato a sezione trasversale, aggregando le osservazioni giornaliere associate a ciascuna struttura. Per tale passaggio sono state rinominate le colonne che hanno richiesto un'aggregazione; si è ritenuto che il nuovo nominativo risultasse esplicativo senza necessità di riportare il blocco di codice:

```
agg_dict = {
    COL_LAT: "mean",
    COL_LON: "mean",
    COL_P: ["mean", "std"],
    "log_price": "mean",
    COL_STATUS: "mean",
    COL_SUPER: "max",
    COL_PRIVATE: "mean",
    COL_HOTEL: "mean",
    COL_SHARED: "mean",
    COL_ID: "count"
}
for c in X_features:
    if c != "Airbnb Superhost":
        agg_dict[c] = "mean"
aggregated_df = (
    processed_df
    .groupby(COL_ID)
    .agg(agg_dict)
    .reset_index(drop=False)
)
X_features_ext = X_features + [
    "n_annunci",
    "price_std"
]
```

L'approccio utilizzato consente di aggregare i dati di ciascuna struttura, preservando sia il livello medio delle variabili sia l'eterogeneità interna dell'offerta.

Le variabili in questione sono gestite come segue:

- Le coordinate geografiche vengono aggregate tramite media aritmetica, assumendo che eventuali variazioni osservate a livello giornaliero siano trascurabili, data la gestione dei relativi attributi in fase di *pre–processing*.
- La colonna relativo al prezzo viene aggregata salvando il livello medio di *pricing* della struttura, la media del `log_price` prcedentemente calcolato, e la deviazione standard del prezzo. Quest'ultimo valore è stato integrato in seguito nel *features set* esteso.
- La variabile di stato della prenotazione viene aggregata tramite media, producendo una quota di giorni prenotati rispetto al numero di annunci postati dalla struttura nel periodo.
- Se la struttura è stata gestita da un *superhost* almeno una volta, viene trattata come tale nel dataset finale.
- Le variabili *dummy* binarie costruite per definire il *Listing Type* sono mutuamente esclusive a livello di annuncio; perciò aggregandole tramite la media si ottiene la quota di offerta della struttura associata a ciascuna tipologia. Salvo eventuli rumori, ci si aspetta una media pari ad uno per una delle categorie, o pari a zero per tutte le categorie note nel caso in cui la struttura sia un "entire home/apt", poichè si tratta della *baseline* implicita.
- Il conteggio del numero di osservazioni associate alla struttura fornisce una misura della dimensione dell'offerta. L'attributo `n_annunci` creato contestualmente, è stato poi aggiunto al set esteso X delle *features*.
- Una media del numero di *amenities* offerte da ciascuna struttura viene calcolata sulla base degli annunci disponibili.
- Il valor medio di ciascun attributo in `X_features` viene calcolato e salvato.

Infine, le aggregazioni in questione vengono salvate in un nuovo *dataframe* che prende il nome di `aggregated_df`. Il *dataframe* aggregato sarà l'input di dati utilizzato nella fase successiva.

5.2 Identificazione della componente strutturale del prezzo

L'obiettivo di questa fase dell'analisi è stimare la componente strutturale del prezzo di ciascun annuncio, isolando l'effetto delle caratteristiche osservabili dell'alloggio dalla parte di prezzo potenzialmente riconducibile a dinamiche competitive locali. Tale decomposizione è necessaria per evitare che le interazioni competitive vengano catturate da attributi edonici, costruendo una misura residuale di prezzo utilizzabile nelle fasi successive. In base a quanto previsto dal modello sviluppato, la stima viene condotta mediante una regressione di LASSO.

Per garantire una penalizzazione omogenea su tutte le variabili esplicative, le covariate strutturali, ovvero le `X_features_ext`, contenute nella matrice `X1`, sono state preventivamente standardizzate e salvate in `X1s`.

Successivamente, sulla base di quanto previsto in precedenza, il modello è stato stimato utilizzando come variabile dipendente il logaritmo del prezzo `log_price`, salvato per ciascuna struttura all'interno del vettore `y1`, e come regressori le caratteristiche strutturali selezionate precedentemente standardizzate.

```
X1 = aggregated_df[X_features_ext].to_numpy(dtype=float)
y1 = aggregated_df["log_price"].to_numpy(dtype=float)
scaler = StandardScaler()
X1s = scaler.fit_transform(X1)
# Modello LASSO
lasso = Lasso(random_state=42)
param_grid = {
    "alpha": np.logspace(-4, 0, 50),
    "max_iter": [5000, 10000, 20000],
    "tol": [1e-4, 1e-3, 1e-2]
}
grid = GridSearchCV(
    estimator=lasso,
    param_grid=param_grid,
    cv=10,
    scoring="neg_mean_squared_error",
```

```
    n_jobs=-1
)
grid.fit(X1s, y1)
# Modello ottimale
best_lasso = grid.best_estimator_
aggregated_df["price_hat"] = best_lasso.predict(X1s)
aggregated_df["epsilon_hat"] = y1 - aggregated_df["price_hat"]
print(
pd.Series(best_lasso.coef_, index=X_features_ext)
    .loc[lambda x: x != 0]
)
```

Per poter realizzare empiricamente il modello nell'ambiente *Jupyter notebook* è richiesta l'importazione di due classi apparteneti alla liberia `scikit-learn`:

```
from sklearn.linear_model import Lasso
from sklearn.model_selection import GridSearchCV
```

In particolare, l'utilizzo della classe `LASSO` permette l'implementazione di una regressione lineare penalizzata con norma l_1 e richiede l'esternalizzazione della selezione degli iperparametri ad una procedura di validazione incrociata tramite `GridSearchCV`.

L'applicazione di quest'ultima classe prevede, tramite *cross-validation* (vedere la Sezione 3.4.3 per una definizione), l'identificazione dei migliori valori per gli iperparametri da fornire in input al modello LASSO, bilanciando capacità predittiva e parsimonia del modello:

- `alpha`: Parola chiave assegnata all'interno della libreria per il parametro di penalizzazione λ. Il valore minimo di λ considerato è pari a 10^{-4}, poiché per valori sufficientemente piccoli la soluzione converge alla stima OLS in presenza di regressori standardizzati, mentre riduzioni ulteriori risultano non informative (Hastie, Tibshirani e Friedman 2009). Il limite superiore pari a 1 rappresenta invece una penalizzazione sufficientemente elevata da generare modelli fortemente parsimonio-

si. L'intervallo è discretizzato mediante 50 valori equispaziati in scala logaritmica, al fine di esplorare con adeguata risoluzione l'intero percorso di regolarizzazione.

- `max_iter`: Definisce il numero massimo di iterazioni dell'algoritmo di ottimizzazione. Il valore minimo del numero di iterazioni è fissato a 5000 in quanto livelli inferiori possono risultare insufficienti a garantire la convergenza, conducendo a soluzioni premature. Il valore massimo pari a 20000 è invece introdotto come soglia prudenziale per verificare la stabilità numerica della soluzione.

- `tol`: Stabilisce la soglia di tolleranza per il criterio di arresto, favorendo la stabilità dei coefficienti prossimi allo zero.

La procedura di validazione incrociata viene effettuata a $K = 10$ $fold$, in quanto tale schema fornisce un compromesso efficace tra bias, varianza e costo computazionale nella stima dell'errore di generalizzazione ed è raccomandato in letteratura come scelta standard per la selezione del modello (Kohavi 1995). Quindi, ad ogni iterazione, il modello è stimato su nove fold e valutato sul fold rimanente, e l'errore quadratico medio di previsione viene calcolato sul sotto-campione escluso dalla stima. Il valore medio di tale errore sulle dieci iterazioni, è utilizzato come criterio per selezionare la migliore configurazione.

La stima del modello produce due output fondamentali per l'analisi empirica successiva. In primo luogo, il vettore dei coefficienti stimati (`best_lasso.coef_`) consente di identificare quali caratteristiche strutturali del *set* X esteso risultano effettivamente rilevanti nella determinazione del prezzo, imponendo a zero i coefficienti associati a variabili prive di potere esplicativo.

In secondo luogo, il modello fornisce, per ciascun annuncio, una stima del prezzo logaritmico spiegato dalle sole caratteristiche osservabili (`price_hat`) e un termine residuo (`epsilon_hat`), ottenuto come differenza tra il valore osservato `y1` e quello predetto. Tale residuo cattura sistematicamente la componente di prezzo non riconducibile agli attributi strutturali dell'alloggio.

Questa decomposizione del prezzo consente di separare l'effetto delle caratteristiche intrinseche dell'alloggio dalle interazioni tra operatori. In particolare, i residui del modello

edonico vengono successivamente utilizzati per analizzare le relazioni competitive spaziali, evitando che tali dinamiche vengano erroneamente attribuite alle variabili strutturali incluse nel modello di base.

5.3 Definizione della struttura di competizione

Una volta identificata la componente strutturale del prezzo e ottenuto il residuo del modello edonico (`epsilon_hat`), il passo successivo dell'analisi consiste nella definizione dell'interazioni competitive tra le strutture ricettive. In tale passaggio, viene affiancato un criterio minimo *ex-ante* di sostituibilità potenziale, volto a valutare l'effetto del restringimento del set competitivo alle sole strutture che risultano plausibili per lo stesso insieme di utilizzatori.

5.3.1 Contiguità spaziale

In questa fase, nota la definizione fornita in letteratura di contiguità spaziale, si mira a definire un insieme rilevante di potenziali concorrenti geograficamente prossimi.

A questo scopo, viene adottato un approccio basato sui *k-nearest neighbors*, in cui la contiguità spaziale è definita in funzione della distanza geografica tra gli alloggi. Le coordinate di ciascun annuncio, espresse in termini di latitudine e longitudine, vengono convertite in radianti e utilizzate per costruire una struttura di vicinato.

La struttura di contiguità spaziale, per poter essere costruita tramite *k-NN*, richiede l'implementazione dell'algoritmo `BallTree`[1], il cui utilizzo prevede l'importazione dell'omonima classe.

Al fine di selezionare in modo oggettivo il numero di vicini geografici K ottimale, viene inoltre utilizzata la classe `KneeLocator` [2].

Infine, per rappresentare la struttura di contiguità spaziale basata sulla *metrica di Haversine* sotto forma di rete, si rende necessaria l'importazione della libreria `NetworkX`[3].

```
import networkx as nx
from sklearn.neighbors import BallTree
from kneed import KneeLocator
```

[1] `https://github.com/tchlux/balltree`
[2] `https://github.com/arvkevi/kneed`
[3] `https://github.com/networkx/networkx`

Tramite le classi importate, una volta individuato il valore di K ottimale, ciascun annuncio viene connesso ai K annunci più vicini, generando un grafo non orientato in cui i nodi rappresentano gli alloggi e gli archi codificano relazioni di prossimità spaziale.

Gli archi del grafo sono pesati in funzione inversa della distanza geografica tra le coppie di annunci, così da riflettere un'intensità di interazione decrescente all'aumentare della distanza. In tal modo, la struttura di contiguità spaziale incorpora implicitamente un meccanismo assimilabile alla *spatial penalty* descritta in Sezione 3.3.

Al fine di garantire la stabilità numerica nella costruzione dei pesi della rete competitiva, al denominatore della distanza geografica è aggiunta la costante positiva molto piccola $\texttt{EPS} = 10^{-6}$, introdotta per evitare divisioni per zero o pesi eccessivamente elevati in presenza di strutture geograficamente quasi coincidenti.

```
coords = df[[COL_LAT, COL_LON]].to_numpy(dtype=float)
coords_rad = np.radians(coords)
props = df[COL_ID].tolist()
EPS = 1e-6
def build_graph_for_K(K, return_query=False):
    tree = BallTree(coords_rad, metric="haversine")
    distances, indices = tree.query(coords_rad, k=K + 1)
    G = nx.Graph()
    G.add_nodes_from(props)
    for i, pid in enumerate(props):
        for t, j in enumerate(indices[i][1:]):
            nid = props[j]
            d_km = distances[i, t + 1] * 6371.0
            w = 1.0 / (d_km + EPS)
            G.add_edge(pid, nid, weight=float(w))
    if return_query:
        return G, distances, indices
    else:
        return G
```

La selezione del valore di K influisce sulla densità della rete e, di conseguenza, sulla definizione delle interazioni competitive locali. Perciò, per evitare una scelta arbitraria,

K non viene fissato a priori, ma determinato in modo *data-driven*. In particolare, per una griglia discreta di valori di K, viene costruito il corrispondente grafo di prossimità e viene calcolato il numero di componenti connesse associate.

All'aumentare di K, il numero di componenti connesse tende a ridursi progressivamente, riflettendo una maggiore integrazione spaziale della rete. Il valore ottimale di K viene individuato mediante il criterio del *knee point*, applicato alla relazione tra K e il numero di componenti connesse. Tale punto identifica il valore oltre il quale incrementi ulteriori di K producono benefici marginali limitati in termini di connettività, segnando un compromesso tra rappresentatività delle interazioni locali e parsimonia della struttura di rete.

```
Ks = list(range(4, 21, 2))
rows = []
for K in Ks:
    Gk = build_graph_for_K(K)
    rows.append({
        "K": K,
        "components": nx.number_connected_components(Gk)
    })
df_K = pd.DataFrame(rows)
kl = KneeLocator(
    df_K["K"],
    df_K["components"],
    curve="convex",
    direction="decreasing"
)
K_NEIGHBORS = kl.knee
```

Una volta individuato il valore ottimale di K, la struttura di prossimità spaziale viene resa operativa estraendo, per ciascun annuncio, l'insieme dei K vicini geografici più prossimi.

```
G_final, distances, indices = build_graph_for_K(
    K_NEIGHBORS,
    return_query=True
```

```
)
neighbors = {
    props[i]: [props[j] for j in indices[i][1:]]
    for i in range(len(props))
}
```

L'output di questa fase è l'associazione ad ogni struttura `i` dell'insieme di potenziali concorrenti. Si ricorda, come già definito in fase di metodologia, che tale insieme ha la proprietà di asimmetria, ovvero la relazione di vicinanza competitiva tra due strutture non è necessariamente reciproca.

5.3.2 Sostituibilità tra le strutture

Dal punto di vista delle caratteristiche dell'offerta, sulla base di quanto affrontato nel capitolo di letteratura, è necessario che le strutture appartenenti allo stesso vicinato, per essere definite in competizione, siano tra loro sostituibili, al fine di evitare che il metodo adottato accoppi strutture palesemente eterogenee.

Per evitare tali associamenti, la sostituibilità è stata definita secondo un approccio strutturale *ex ante*, fondato sulla comparabilità delle caratteristiche osservabili dell'offerta. Tale impostazione consente di mantenere un criterio di selezione dei concorrenti economicamente interpretabile e coerente con le dimensioni di differenziazione emerse in letteratura.

Secondo quanto delinato nella metodologia, tale prospettiva verrà confrontata con un'alternativa puramente *data-driven* che prevede la modellazione delle relazioni competitive solo sulla base delle interdipendenze empiriche osservate nei residui di prezzo, all'interno dello stesso intorno geografico. Quindi, tale modello prevede l'esclusione delle operazioni riportate in questo paragrafo, proseguendo con le fasi successive.

Coerentemente con l'approccio strutturale *ex ante*, sono quindi state selezionate quattro dimensioni che definiscono il perimetro di fattibilità della struttura, mentre le restanti variabili di differenziazione vengono indirettamente valutate nel modello edonico. Le variabili considerate sono: `Max Guests`, `Bedrooms`, `Bathrooms` e `Minimum Stay`. Infatti,

una struttura con capacità ricettiva inferiore non può soddisfare una prenotazione per un gruppo numeroso, così come vincoli di soggiorno molto diversi intercettano tipicamente segmenti di domanda differenti.

La capacità ricettiva viene gestita con una tolleranza relativamente più permissiva su `Max Guests` e una tolleranza più conservativa su `Bedrooms`. Tale scelta deriva dal fatto che `Max Guests` riflette spesso una flessibilità di utilizzo, come posti letto aggiuntivi, divani letto, mentre `Bedrooms` rappresenta una configurazione più rigida dell'immobile. La variabile `Bathrooms` viene invece trattata con maggiore cautela, in quanto differenze anche contenute possono segnalare standard e target significativamente diversi. Infine, per `Minimum Stay` si utilizza una discretizzazione in fasce, scelta per rendere il confronto robusto rispetto a variazioni marginali.

```
def minstay_bucket(ms):
    if ms <= 2:
        return 0   # short stay
    elif ms <= 6:
        return 1   # mid stay
    else:
        return 2   # long stay
struct_df = aggregated_df.set_index(COL_ID)[
    ["Max Guests", "Bedrooms", "Bathrooms", "Minimum Stay"]
]
def is_substitutable(i, j, df):
    gi, gj = df.loc[i], df.loc[j]
    cond_guests = abs(gi["Max Guests"] - gj["Max Guests"]) <= 2
    cond_beds   = abs(gi["Bedrooms"]   - gj["Bedrooms"])   <= 1
    cond_baths  = abs(gi["Bathrooms"]  - gj["Bathrooms"])  <= 1
    cond_minst  = minstay_bucket(gi["Minimum Stay"]) == minstay_bucket(gj["Minimum
    ↪  Stay"])
    conditions = [cond_guests, cond_beds, cond_baths, cond_minst]
    return (sum(conditions) >= 3) and (cond_guests or cond_beds)
```

Per evitare che il filtro risulti eccessivamente rigido, la sostituibilità è implementata tramite una regola di similarità minima e non di intersezione rigida di tutte le condizioni.

Operativamente, una relazione tra i e j viene mantenuta quando almeno tre condizioni di similarità su quattro risultano soddisfatte, includendo almeno una condizione legata alla capacità ricettiva, ossia `Max Guests` o `Bedrooms`, escludendo accoppiamenti palesemente implausibili.

Quindi, tramite l'imposizione di tali condizioni, l'insieme $\mathcal{N}_i$ viene filtrato per ottenere un set finale che incorpori la sostituibilità strutturale.

```
neighbors_subst = {}
for pid, neighs in neighbors.items():
    valid = []
    for nid in neighs:
        if is_substitutable(pid, nid, struct_df):
            valid.append(nid)
    neighbors_subst[pid] = valid
```

In questo processo, si individua la presenza di una quota limitata di strutture per le quali $\mathcal{N}_i$ risulta vuoto. Tali casi corrispondono a strutture che, pur essendo localizzate nel territorio analizzato, non presentano alternative localmente comparabili secondo i vincoli considerati e vengono pertanto trattate separatamente nelle analisi basate sulla rete.

5.4 Stima della pressione competitiva locale

In questa fase viene costruita una misura sintetica della pressione competitiva locale esercitata sulle strutture. L'ipotesi di fondo è che le dinamiche concorrenziali si manifestino attraverso un co-movimento sistematico delle componenti di prezzo non spiegate, osservabile tra strutture geograficamente prossime e funzionalmente sostituibili.

In particolare, per ciascuna struttura viene calcolata una variabile aggregata definita come la media dei residui del modello edonico associati agli annunci appartenenti al suo vicinato geografico e funzionale. Tale misura sintetizza l'andamento medio delle componenti di prezzo non spiegate delle strutture comparabili localizzate nelle immediate vicinanze, fornendo una proxy della pressione competitiva esercitata a livello locale.

```
eps_map = aggregated_df.set_index(COL_ID)["epsilon_hat"]
def neigh_mean_residual_subst(pid):
    neighs = neighbors_subst.get(pid, [])
    if len(neighs) == 0:
        return np.nan
    return float(eps_map.loc[neighs].mean())
aggregated_df["neighbors_residual_mean"] =
↪  (aggregated_df[COL_ID].apply(neigh_mean_residual_subst))
df_rho = aggregated_df.dropna(subset=["neighbors_residual_mean"])
```

L'analisi procede quindi alla stima dell'intensità dell'interazione competitiva globale riconducibile a tale pressione locale. A questo scopo, la componente di prezzo non spiegata della singola struttura viene messa in relazione con il valore medio dei residui osservati tra le strutture simili e prossime, secondo la seguente relazione:

$$\widehat{\varepsilon}_i = \rho\, \overline{\widehat{\varepsilon}}_{\mathcal{N}(i)} + u_i.$$

Il parametro ρ misura il grado di dipendenza spaziale locale dei residui e può essere interpretato come una misura sintetica di autocorrelazione spaziale, concettualmente affine all'indice di Moran, pur differenziandosi da quest'ultimo per l'impostazione regres-

siva adottata. In questo contesto, ρ cattura il co-movimento medio delle componenti di prezzo non spiegate tra strutture concorrenti, condizionato alla struttura di vicinato precedentemente definita.

La stima è condotta mediante una regressione lineare con intercetta, utilizzando il metodo dei minimi quadrati ordinari e adottando errori standard robusti di tipo *heteroskedasticity-consistent* (HC1)[4], al fine di garantire affidabilità inferenziale in presenza di possibile eteroschedasticità.

```
import statsmodels.api as sm
X2 = sm.add_constant(df_rho[["neighbors_residual_mean"]])
y2 = df_rho["epsilon_hat"]
ols2 = sm.OLS(y2, X2).fit(cov_type="HC1")
rho = float(ols2.params["neighbors_residual_mean"])
```

Valori positivi e statisticamente significativi di ρ indicano che le componenti di prezzo non spiegate delle strutture vicine tendono a muoversi nella stessa direzione, suggerendo la presenza di dinamiche competitive spazialmente strutturate. Al contrario, valori prossimi allo zero sono indicativi di una debole o assente interdipendenza locale.

La relazione stimata consente di interpretare il coefficiente ρ come una misura di autocorrelazione spaziale dei residui di prezzo. In particolare, analogamente all'indice di Moran discusso in Sezione 3.4.5, ρ cattura il grado di co-movimento tra la componente di prezzo non spiegata di una struttura e quella delle strutture appartenenti al suo intorno competitivo. A differenza dell'indice di Moran, tuttavia, tale dipendenza spaziale viene qui stimata attraverso un'impostazione regressiva, che consente di quantificare direttamente l'intensità media dell'interazione competitiva locale. La stima del coefficiente ρ rappresenta pertanto un passaggio chiave dell'analisi, poiché permette di verificare empiricamente se la competizione assuma una dimensione spaziale sistematica. Essa costituisce il fonda-

[4]Gli errori standard *heteroskedasticity-consistent* (HC) sono correttivi della matrice di varianza-covarianza degli stimatori OLS che consentono di ottenere inferenza valida anche in presenza di eteroschedasticità. La versione HC1, proposta da MacKinnon e White (1985), applica un fattore di correzione per piccoli campioni alla stima robusta di White, migliorando le proprietà finite-sample degli errori standard senza modificare i coefficienti stimati.

mento quantitativo per la costruzione e l'analisi della rete competitiva presentata nella sezione successiva.

5.5 Identificazione delle aree di competizione locale

In questa fase viene costruita una rete competitiva in cui ciascun nodo rappresenta una struttura Airbnb e ciascun arco codifica una relazione di prossimità competitiva tra strutture vicine.

Gli archi della rete sono pesati in funzione inversa della distanza geografica tra coppie di strutture appartenenti al medesimo set competitivo sostituibile, e scalati dall'intensità dell'effetto competitivo globale stimato in precedenza.

```
dist_map = {}
for i, pid in enumerate(props):
    for t, nid in enumerate(neighbors[pid]):
        d = dist_km[i, t]
        dist_map[(pid, nid)] = d
        dist_map[(nid, pid)] = d
G = nx.Graph()
G.add_nodes_from(props)
for pid, neighs in neighbors_subst.items():
    for nid in neighs:
        d = dist_map.get((pid, nid))
        if d is None:
            continue
        w = abs(rho) / (d + EPS)
        G.add_edge(pid, nid, weight=float(w))
```

Sulla base della rete competitiva così costruita, l'analisi procede all'identificazione di aree di competizione locale mediante un algoritmo di *community detection*. In particolare, viene adottato il *metodo di Louvain*, che consente di individuare gruppi di strutture caratterizzati da relazioni interne più intense rispetto a quelle con il resto della rete, tramite le logiche descritte nel Capitolo 4. Tale operazione richiede l'importazione dell'apposita classe `community_louvain`.

```
import community as community_louvain
```

Il risultato del processo di partizionamento assegna a ciascuna struttura Airbnb un'etichetta di appartenenza ad un'area competitiva, definita come un sottoinsieme di strutture fortemente interconnesse dal punto di vista competitivo.

```
partition = community_louvain.best_partition(
    G,
    weight="weight",
    resolution=0.4,
    random_state=42
)
aggregated_df["competition_area"] = aggregated_df[COL_ID].map(partition)
```

Le aree individuate tramite il blocco di codice costituiscono segmenti spazialmente coerenti del mercato, all'interno dei quali le dinamiche concorrenziali risultano maggiormente concentrate.

5.6 Integrazione dell'approccio ALASSO

Parallelamente, è stata validata la componente di selezione delle variabili tramite l'intergrazione dell'approccio ALASSO, il quale prevede di includere una penalizzazione ponderata per ciascuna variabile.

A tale scopo, poichè i pesi di ciascuna variabile possono essere estratti tramite un qualsiasi ragionevole, consistente o zero-consistente approccio, in fase di metodologia è stato deciso di estrarli sulla base delle risposte collezionate per il questionario riportato nell'Annex A.

Per tale potenziamento, la fase di "imputazioni e trasformazioni preliminari" risulta invariata rispetto al modello classico LASSO. Invece, le fasi successive richiedono di inglobare i pesi al loro interno, perciò è richiesta preliminarmente la loro estrazione a partire dai dati raccolti tramite il questionario.

Perciò, a partire dai risultati in formato csv salvati all'interno della variabile `raw`, è stat creata una colonna per ciascuna domanda sottoposta agli utenti, prevedendo la gestione di errori in fase di lettura dei dati.

Oltre alla gestione degli errori non sono previste ulteriori attività di *pre-processing*, in quanto si ha diretto accesso al dataframe originale. Non vi è quindi il rischio di perdita dei dati a seguito di trasferimenti, e struttalmente non vi possono essere dati mancanti o errori in quanto le possibili risposte sono chiuse, e obbligatorie.

In fase di salvataggio delle variabili viene fatta una distinzione tra le domande che prevedono una risposta numerica da 1 a 7, e l'unica domanda che prevede una risposta standard, ma non numerica, quale la frequenza di utilizzo della piattaforma (`FREQ_COL`).

```
FREQ_COL = "Con quale frequenza utilizzi Airbnb?"
numeric_cols = raw.select_dtypes(include="number").columns.tolist()
if not numeric_cols:
    raise ValueError("Nessuna colonna numerica trovata nel questionario.")
def norm(s: str) -> str:
    return re.sub(r"\s+", " ", str(s).strip().lower())
QUESTION_TO_VAR = {
    "price_mean":  [r"prezzo", r"price", r"costo"],
```

```
    "Bedrooms":           [r"camere", r"bedroom"],
    "Bathrooms":          [r"bagni", r"bathroom"],
    "amenities_count":    [r"servizi", r"amenities"],
    "Overall Rating":     [r"valutazione", r"rating"],
    "Number of Reviews": [r"recensioni", r"reviews"],
    "ListingType":        [r"tipologia", r"listing type"],
    "Airbnb Superhost":   [r"superhost"]
}
col_map = {}
for c in numeric_cols:
    cn = norm(c)
    for var, pats in QUESTION_TO_VAR.items():
        if any(re.search(p, cn) for p in pats):
            col_map[c] = var
            break
missing = set(QUESTION_TO_VAR) - set(col_map.values())
if missing:
    raise ValueError(f"Variabili non mappate: {missing}")
survey_df = raw[[FREQ_COL] + list(col_map.keys())].rename(columns=col_map)
```

Viene quindi creato un dataframe nominato `survey_df`, all'interno del quale per ciascun utente è indicata la rispettiva frequenza di utilizzo della piattaforma e le restanti risposte numeriche. Quest'ultime sono state rinominate con i nomi standard delle variabili economiche associate.

```
survey_df = raw[[FREQ_COL] + list(col_map.keys())].rename(columns=col_map)
FREQUENCY_WEIGHT_MAP = {
    "< 1 volta / anno": 0.7,
    "1-2 volte / anno": 1.0,
    "3-5 volte / anno": 1.3,
    "> 5 volte / anno": 1.6
}
def freq_to_weight(s: str) -> float:
    s = s.replace("-", "-").lower()
    for k, w in FREQUENCY_WEIGHT_MAP.items():
        if k.replace("-", "-").lower() in s:
```

```
            return w
    return 1.0
```

Successivamente, ad ogni categoria di utilizzo viene assegnato un peso crescente, con lo scopo di dare rilevanza alle risposte degli utenti che utilizzano la piattaforma più di 3 volte l'anno, e minore importanza ai valori assegnati dagli utenti saltuari. Di conseguenza, viene salvato all'interno del dataframe un peso di rilevanza `w` a ciascun respondente sulla base dell'utilizzo, ed eliminato il valore testuale della risposta per alleggerirlo.

```
Xq = survey_df.drop(columns=[FREQ_COL]).astype(float)
importance_weighted = (Xq.mul(w, axis=0).sum() /
↪  w.sum()).sort_values(ascending=False)
```

Infine, l'output finale è la serie `importance_weighted`. Essa prevede di moltiplicare ciascuna risposta con il peso assegnato all'utente, sommare il risultato per ciascuna variabile, e normalizzare dividendo per la somma totale dei pesi.

```
gamma = 1.0
EPS_W = 1e-3
importance_safe = importance_weighted.clip(lower=EPS_W)
weights = 1.0 / (importance_safe ** gamma)
weights = weights / weights.mean()
```

Viene quindi calcolato il peso di penalizzazione `weight`, tramite una trasformazione inversa, in modo tale che per le variabili più importanti il peso sia basso, e alto per quelle meno rilevanti.

Infine, i pesi vengono riscalati in modo che il valore atteso di ciascun peso sia 1.

```
weights_full = pd.Series(1.0, index=X_features_ext)
for v in weights.index:
    if v in weights_full.index:
        weights_full[v] = weights[v]
```

Sulla base dell'output ottenuto, si crea il vettore di pesi `weights_full` per tutte le variabili. Questa fase consente di includere le variabili il cui peso non viene definito tramite il questionario, alle quali viene assegnato un peso di default pari a 1. In questo modo si ottiene un modello ibrido, per cui alcune variabili vengono penalizzate in modo adattivo, altre trattate in modo neutro.

```
W = np.diag(1.0 / weights_full.values)
X_alasso = X1s @ W
```

Gli step successivi sono analoghi a quanto visto per il modello classico, con la differenza che le covariate standardizzate salvate in `X1s` vengono moltiplicate per una matrice diagonale dei pesi (`W`).

Una volta individuata la componente strutturale del prezzo con tali considerazioni, i valori salvati di `epsilon_hat` possono essere utilizzati per l'identificazione delle aree di competizione analogamente a quanto visto nelle sezioni precedenti.

Lo scopo di tale integrazione, è validare la selezione con il solo modello LASSO, e valutare le differenze che risultano nell'applicazione di un modello adattivo contestualemtne all'isolamento della componente competitiva. Per tale ragione, nel Capitolo 6 dedicato ai risultati, verrà posta maggiore enfasi sull'eventuale divario in questa prima fase, mentre per i successivi step del modello le considerazioni risulterebbero analoghe in termini relativi, per cui l'attenzione verrà posta sulla valutazione del modello principale.

Capitolo 6

Risultati

Il capitolo in questione si pone l'obiettivo di commentare gli esiti dell'analisi e traformare gli output ottenuti in evidenze economiche interpretabili, rispondendo esplicitamente alle domande di ricerca. I valori discussi all'interno del capitolo sono stati generati dagli output numerici ottenuti dai blocchi di codice definiti in fase di analisi; eventuali calcoli statistici effettuati sui risultati, o visualizzazioni aggiuntive al supporto della discussione, sono l'esito dei blocchi di codice riportati nell'Annex B.

6.1 Risultati del modello edonico penalizzato

Questa sezione presenta i risultati della stima del modello edonico penalizzato, utilizzato come baseline strutturale per l'analisi delle dinamiche competitive. L'obiettivo è identificare le componenti del prezzo sistematicamente spiegate dalle caratteristiche osservabili dell'offerta e isolare la componente residua, che costituisce l'input informativo per l'analisi della competizione locale sviluppata nelle sezioni successive. In particolare, vengono dapprima discusse le variabili strutturali selezionate dal modello e successivamente analizzate le proprietà della componente residua del prezzo.

6.1.1 Variabili strutturali rilevanti

Le *feature* selezionate tramite LASSO e Adaptive LASSO rappresentano le dimensioni dell'offerta che contribuiscono in modo sistematico alla formazione del prezzo. L'introduzione dell'ALASSO consente di verificare la stabilità dell'isolamento della componente competitiva rispetto a una struttura di penalizzazione non uniforme, rafforzando la solidità metodologica della decomposizione prezzo–competizione adottata. L'attenzione ana-

litica si concentra pertanto sulle eventuali differenze che emergono in questa fase iniziale del modello, poiché è in tale passaggio che si definisce la separazione tra determinanti strutturali e componente residuale.

Infatti, tali variabili costituiscono la baseline strutturale del modello edonico e vengono quindi escluse dalla successiva misura della competizione: esse spiegano differenze di prezzo giustificate da caratteristiche osservabili dell'alloggio, mentre la componente residua diventa, nelle sezioni successive, il potenziale segnale competitivo.

Il confronto sistematico tra le specificazioni è riportato nella Tabella 6.1, che sintetizza le variabili selezionate e i relativi coefficienti stimati dai due modelli. Dall'evidenza empirica emerge una sostanziale coerenza nella struttura selezionata: entrambi identificano in modo stabile le principali dimensioni legate alla capacità ricettiva e alla qualità percepita dell'annuncio. In particolare, la variabile `Max Guests` si conferma il driver strutturale dominante, seguita dal numero di bagni e dal numero di fotografie, rafforzando l'interpretazione secondo cui capacità e visibilità dell'offerta rappresentano i determinanti centrali del prezzo.

Tabella 6.1. Variabili strutturali selezionate da LASSO e ALASSO

Variabile	**LASSO**	**ALASSO**
Bedrooms	0.0221	0.0303
Bathrooms	0.0692	0.0712
Max Guests	0.1372	0.1548
Number of Photos	0.0099	0.0314
Price standard deviation	-0.0189	-0.0509
Number of Reviews	–	+0.0002

L'Adaptive LASSO conferma integralmente il nucleo delle variabili selezionate dal LASSO standard, mostrando coefficienti di entità molto simile. La deviazione standard del prezzo risulta l'unica variabile con coefficiente negativo in entrambi i modelli, suggerendo che, a parità delle caratteristiche strutturali, una maggiore variabilità storica del pricing sia associata a livelli medi di prezzo inferiori. Tale evidenza è coerente con l'interpretazione della stabilità tariffaria come segnale di maggiore solidità della domanda o di posizionamento più consolidato sul mercato.

L'ALASSO seleziona inoltre una variabile aggiuntiva (`Number of Reviews`), assente nel LASSO standard. Pur presentando un coefficiente di entità contenuta, la sua inclusione risulta coerente con le evidenze emerse dal questionario, che attribuiscono alla reputazione online un ruolo rilevante nei processi decisionali degli utenti. La penalizzazione adattiva consente così di intercettare una dimensione reputazionale che, pur non configurandosi come driver strutturale dominante del prezzo, contribuisce in modo sistematico al suo posizionamento relativo.

Per quanto riguarda la selezione degli iperparametri, entrambi i modelli individuano come ottimale un numero massimo di iterazioni pari a 5000 e tolleranza 1×10^{-4}. Tale configurazione garantisce un adeguato livello di parsimonia senza compromettere la capacità del modello di assorbire le principali componenti strutturali del prezzo. Invece, il coefficiente di penalizzazione risulta pari circa a $\alpha = 0.0868$ secondo LASSO, mentre l'approccio adattivo restituisce un valore inferiore pari a $\alpha = 0.0596$; tale differenza potrebbe indicare che grazie alla ponderazione differenziata dei coefficienti, il modello necessita di una minore penalizzazione globale per ottenere una selezione efficace delle variabili, poiché la regolarizzazione risulta già strutturalmente mirata.

Nel complesso, tale confronto evidenzia una sostanziale stabilità nella struttura selezionata. Le principali dimensioni strutturali emergono in modo coerente in entrambe le specificazioni, indicando che i risultati non sono sensibili alla diversa forma di penalizzazione adottata. Infatti, l'Adaptive LASSO, pur incorporando una penalizzazione differenziata coerente con le evidenze qualitative del questionario, non modifica l'interpretazione economica del modello edonico, ma ne rafforza la robustezza e la coerenza complessiva.

6.1.2 Valutazione dell'isolamento della componente competitiva

Una volta stimata la componente strutturale del prezzo, l'attenzione si sposta sull'analisi della parte non spiegata dal modello edonico, che costituisce la base per l'identificazione del segnale competitivo. Tale passaggio è cruciale, poiché la qualità dell'isolamento della componente strutturale determina la credibilità dell'intera misura di competizione locale sviluppata nelle sezioni successive.

La Tabella 6.2 riporta le statistiche descrittive dei residui ottenuti sia dal LASSO sia dall'Adaptive LASSO. In entrambe le specificazioni, i residui risultano centrati intorno allo zero, confermando l'assenza di bias sistematico nella stima della baseline strutturale. La dispersione rimane sostanzialmente invariata tra i due modelli, segnalando che l'introduzione di una penalizzazione adattiva non altera la capacità complessiva di assorbire le principali determinanti osservabili del prezzo.

Tabella 6.2. Statistiche descrittive dei residui

Statistica	**LASSO**	**ALASSO**
Numero osservazioni	19 256	19 256
Media	-8.74×10^{-16}	-8.86×10^{-16}
Deviazione standard	0.807	0.800
Min	-5.304	-5.467
5° percentile	-0.988	-0.978
25° percentile	-0.276	-0.279
Mediana	0.054	0.053
75° percentile	0.406	0.399
95° percentile	1.056	1.041
Max	2.475	2.792

La distribuzione dei residui presenta code non trascurabili e una variabilità significativa, evidenziando che una quota rilevante della dinamica dei prezzi non è spiegata dalle sole caratteristiche strutturali. Proprio tale componente residua rappresenta, nell'impostazione adottata, il potenziale spazio entro cui si manifesta la pressione competitiva locale.

Si può quindi concludere che emerge una sostanziale sovrapponibilità delle due distribuzioni residue, indicando che l'isolamento della componente competitiva risulta robusto rispetto alla diversa struttura di penalizzazione adottata. Alla luce di tale convergenza, le analisi grafiche riportate in seguito sono sviluppate con riferimento al solo modello LASSO.

Una prima analisi viene sviluppata in Figura 6.1, la quale riporta la distribuzione dei residui all'interno di un *violin plot*[1]

[1]Il *violin plot* è una rappresentazione grafica della distribuzione di una variabile quantitativa che integra le informazioni di un box plot, ossia minimo, primo quartile (Q1), mediana, terzo quartile (Q3) e massimo, con una stima continua della densità di probabilità. La larghezza della forma in ciascun punto riflette la concentrazione delle osservazioni, consentendo di cogliere non solo la posizione e la dispersione dei dati, ma anche la loro struttura distributiva (Shishebor, Sajjadnia e Sharafi 2025).

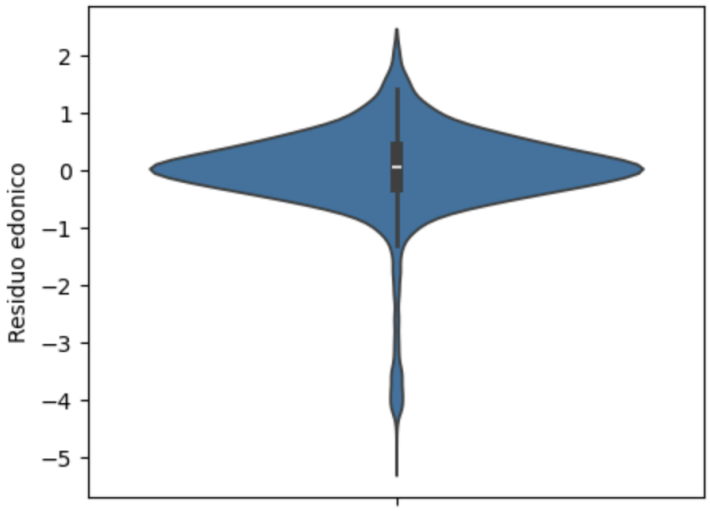
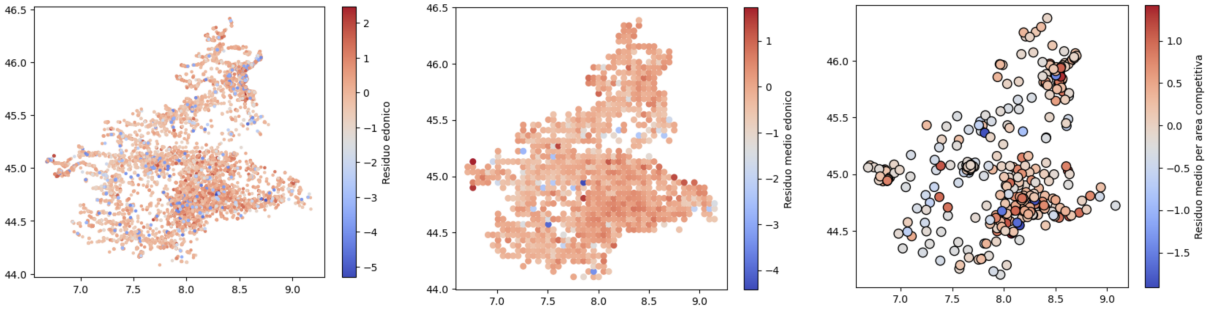


Figura 6.1. Distribuzione *violin plot* dei residui

Esso mostra una massa centrale concentrata in un intervallo relativamente contenuto, ma al contempo evidenzia la presenza di code non trascurabili. Tale configurazione suggerisce che, una volta controllato per le dimensioni e le dotazioni dell'alloggio, permangono aggiustamenti locali di prezzo che non possono essere ricondotti esclusivamente all'eterogeneità osservabile dell'offerta, configurandosi come potenziale manifestazione di pressioni competitive territoriali.

Per comprendere se tali scostamenti presentino una struttura sistematica, è quindi necessario analizzarne anche la dimensione spaziale. A complemento dell'analisi precedente, la Figura 6.2 consente infatti di valutare la distribuzione territoriale della componente residua, mostrando come gli scostamenti di prezzo non risultino distribuiti in modo puramente casuale sul territorio, passando ad una valutazione spaziale della dipendenza tra osservazioni. L'emersione di pattern locali e concentrazioni spaziali suggerisce che una parte della variabilità residua possa riflettere interazioni competitive tra strutture geograficamente prossime, piuttosto che semplice eterogeneità idiosincratica non osservata.



(a) Distribuzione residui

(b) Distribuzione media dei residui

(c) Distribuzione della media dei residui per area

Figura 6.2. Distribuzione dei residui

Queste risultanze giustificano l'utilizzo dei residui come segnale per l'analisi delle interazioni competitive locali. La successiva stima della pressione competitiva restituisce un coefficiente globale pari a $\rho \approx 0.524$, indicando un co-movimento positivo e statisticamente rilevante tra il residuo di una struttura e quello medio delle strutture vicine. Il risultato conferma che una parte della componente non spiegata dal modello edonico è condivisa a livello territoriale, coerentemente con un'interpretazione in termini di interazione competitiva.

6.2 Struttura di competizione risultante

La struttura del set competitivo associato a ciascun annuncio costituisce un elemento chiave per l'interpretazione delle dinamiche concorrenziali locali. A tale fine, l'analisi della numerosità e della variabilità del numero di concorrenti consente di caratterizzare il perimetro competitivo entro cui operano le singole strutture e di valutare l'impatto dei vincoli strutturali introdotti nella definizione dei concorrenti sulla configurazione delle interazioni locali.

Innazittuto, emerge che quando il set competitivo è definito unicamente secondo un approccio data-driven, il numero di concorrenti associato a ciascun annuncio risulta pari al parametro $K = 8$, senza variazioni tra le strutture. Questa configurazione riflette l'applicazione del criterio dei k-nearest neighbors ed implica l'adozione di un perimetro competitivo omogeneo, che delimita in modo coerente l'intorno geografico rilevante per l'analisi.

Come emerge in Tabella 6.3, l'introduzione dei vincoli strutturali di sostituibilità modifica questa configurazione, risultando nel numero medio di concorrenti per annuncio ridotto a 6.55, corrispondente a una contrazione di circa il 18% rispetto alla baseline solo data-driven.

Contestualmente, emerge una marcata eterogeneità nella numerosità del set competitivo, con una deviazione standard pari a 1.97 ed una distribuzione che varia da un minimo di 0 a un massimo di 8 concorrenti. La mediana pari a 7 indica che, per una quota significativa di strutture, solo una parte dei vicini geografici soddisfa i criteri di comparabilità dell'offerta introdotti ex ante.

In termini relativi, la riduzione del set competitivo risulta mediamente contenuta ma fortemente disomogenea tra gli annunci. Mentre per circa il 25% delle strutture l'applicazione dei vincoli di sostituibilità non comporta alcuna riduzione del numero di concorrenti, una quota non trascurabile sperimenta contrazioni molto più marcate, fino a casi estremi in cui nessun vicino geografico risulta sostituibile. In particolare, circa 455 strutture, pari a poco più del 2% del campione, rimangono prive di concorrenti sostituibili nel pro-

prio intorno geografico, suggerendo la presenza di segmenti di mercato locali scarsamente contendibili.

Tabella 6.3. Statistiche numero di concorrenti per annuncio

Statistica	**Data-Driven**	**Sostituibilità Ex-Ante**
Numero osservazioni	19 256	19 256
Media	8.00	6.55
Deviazione standard	0.00	1.97

Nel complesso, l'introduzione dei vincoli di sostituibilità comporta una riduzione del numero di concorrenti per il 56.4% degli annunci. Tale evidenza fornisce una misura della *frequenza* dell'effetto dei vincoli, indicando in quanti casi l'applicazione dei criteri di sostituibilità determina una contrazione del set competitivo. Al fine di valutare se tale riduzione costituisca un effetto sistematico dei vincoli piuttosto che un esito casuale, è stato applicato un sign test monodirezionale[2]. Il rifiuto dell'ipotesi nulla di una probabilità di riduzione pari a 0.5 ($p < 0.001$) indica che l'applicazione dei vincoli di sostituibilità determina, per una quota significativamente maggioritaria del campione, una contrazione del set competitivo.

Accanto alla frequenza, è tuttavia rilevante valutare anche l'*intensità* della riduzione del set competitivo, ossia di quanto esso si contragga nei casi in cui l'effetto è presente. A tal fine, è stata condotta un'analisi bootstrap non parametrica[3] della mediana della riduzione. I risultati mostrano un intervallo di confidenza al 95% collassato sul valore unitario ($\text{IC}_{95\%} = [1, 1]$), indicando che l'effetto tipico dei vincoli di sostituibilità si manifesta attraverso l'eliminazione di un singolo concorrente.

Considerate congiuntamente, le due evidenze suggeriscono che i vincoli di sostituibilità agiscono in modo diffuso ma di intensità contenuta: se da un lato una quota rilevante di annunci sperimenta una riduzione del proprio set competitivo, dall'altro tale riduzione

[2]Il *sign test* è un metodo statistico non parametrico utilizzato per la valutazione di differenze di mediana in dati appaiati o a un solo campione, basato sulla verifica che la probabilità di osservare deviazioni positive o negative si discosti sistematicamente dal valore di riferimento pari a 0.5, senza imporre assunzioni sulla distribuzione dei dati (Allam 2026)

[3]L'*analisi bootstrap* è una procedura di inferenza statistica basata sul ricampionamento con reinserimento dei dati osservati, utilizzata per stimare la distribuzione campionaria di uno stimatore, nel caso in esame la mediana, e costruirne intervalli di confidenza, senza assumere una forma parametrica per la distribuzione sottostante (Kostanek et al. 2024).

risulta, nella maggior parte dei casi, limitata e incrementale. Ciò conferma che i vincoli operano come un meccanismo di filtraggio selettivo delle interazioni competitive, piuttosto che come uno strumento di compressione generalizzata del contesto concorrenziale.

6.3 Evidenza della competizione spaziale

Questa sezione analizza la struttura della competizione locale risultante dalla stima delle relazioni competitive tra gli annunci. A partire dai coefficienti selezionati dal modello penalizzato, le interdipendenze competitive vengono interpretate in termini di rete, al fine di caratterizzarne la topologia, il grado di concentrazione e l'eterogeneità delle pressioni competitive a livello locale. Sulla base di quanto osservato nella sezione precedente, anche quest'analisi è condotta confrontando la configurazione con vincoli di sostituibilità con lo scenario puramente data-driven, in modo tale da valutare come le assunzioni influenzano la struttura complessiva della rete e l'intensità della competizione osservata.

6.3.1 Analisi dell'autocorrelazione dei residui

L'autocorrelazione spaziale dei residui viene analizzata confrontando il residuo di ciascun annuncio con la media dei residui degli annunci concorrenti nel relativo intorno geografico. Tale costruzione consente di verificare se la componente di prezzo non spiegata dal modello presenti una dipendenza sistematica rispetto al contesto competitivo locale, in linea con la definizione teorica riportata nella Sezione 3.4.1.

In entrambi gli scenari considerati emerge una dipendenza positiva e statisticamente significativa, indicando che la componente residua del prezzo incorpora interdipendenze locali non spiegate dalle caratteristiche strutturali dell'offerta. In particolare, valutando il metodo solo data-driven, l'intensità dell'autocorrelazione risulta pari a $\rho = 0.595$, la quale si riduce con l'introduzione dei vincoli fino a raggiungere un valore di $\rho = 0.524$. Tale riduzione suggerisce che operano una selezione delle interazioni competitive economicamente plausibili, riducendo il co-movimento indotto dall'inclusione indiscriminata di tutti i vicini geografici.

Formalmente, in entrambi i casi analizzati, il risultato si traduce in una correlazione positiva e statisticamente diversa da zero tra il residuo associato a ciascun annuncio, ε_i, e la media dei residui degli annunci concorrenti localizzati nel relativo intorno geografico, $\bar{\varepsilon}_{\mathcal{N}(i)}$, indicando la presenza di autocorrelazione spaziale positiva. Tale evidenza, coerente

con un valore positivo dell'indice di Moran applicato ai residui, suggerisce che, una volta controllato per le caratteristiche osservabili dell'offerta, permane una struttura spaziale nella componente non spiegata del prezzo, riconducibile a interdipendenze competitive locali o a fattori spaziali non osservati non catturati dal modello edonico di base.

Quindi, si può concludere che a prescindere dallo scenario la componente di prezzo contiene informazione competitiva locale rilevante, ma l'introduzione di assunzioni strutturali, isola una forma di dipendenza spaziale interpretabile dal punto di vista economico. Data tale evidenza, emerge la necessità di analizzare l'intensità della pressione competitiva locale nella prossima sezione.

6.3.2 Valutazione dell'intensità della pressione competitiva locale

In questa fase l'obiettivo è valutare come la competizione si distribuisce nello spazio, misurando a partire dalla rete competitiva stimata la pressione competitiva in termini aggregati per ciascun annuncio, così da passare da una lettura puramente relazionale a una caratterizzazione quantitativa dell'intensità competitiva locale. Dove, operativamente, la pressione competitiva locale è definita come una misura aggregata dell'influenza esercitata dai concorrenti localizzati nel vicinato rilevante di ciascun annuncio, ottenuta combinando la struttura della rete competitiva stimata con l'intensità della componente competitiva residua associata a ciascun concorrente (Pakes et al. 2021).

In termini di livello tipico, la pressione competitiva mediana risulta più elevata nello scenario privo di vincoli rispetto a quello con vincoli di sostituibilità, suggerendo, coerentemente con quanto evidenziato nella sezione precedente, che l'inclusione indiscriminata di tutti i vicini geografici tenda ad amplificare la pressione aggregata attraverso un comovimento guidato principalmente dalla prossimità spaziale, indipendentemente dalla comparabilità economica dell'offerta.

Inoltre, come mostrato nella Figura 6.3, lo scenario con vincoli presenta una maggiore dispersione relativa della pressione, coerentemente con l'aumento da 6.71 a 9.25 del rapporto tra il 90° e il 10° percentile. Tale evidenza segnala che, pur riducendosi il livello centrale della pressione, la competizione si distribuisce in modo più eterogeneo, concen-

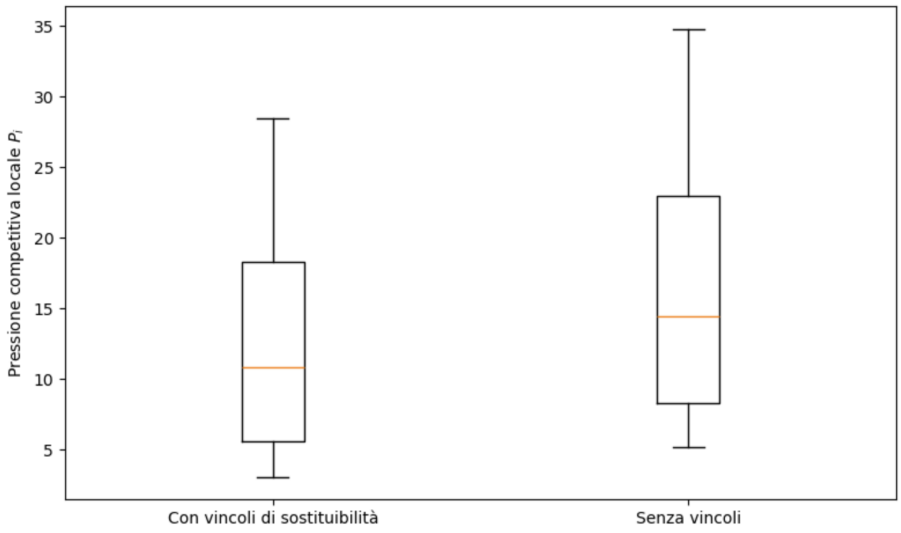


Figura 6.3. Distribuzione della pressione competitiva locale

trandosi su specifici segmenti di annunci e favorendo l'emergere di micro-contesti locali ad alta intensità competitiva. In sintesi, l'introduzione dei vincoli di sostituibilità riduce la pressione mediana ma ne accresce la segmentazione e la variabilità relativa.

La quota di pressione competitiva concentrata nel decile superiore supera il 90% per entrambi gli scenari. Tale evidenza segnala una marcata concentrazione strutturale della competizione, dominata da un numero ristretto di nodi ad alta intensità. La differenza tra i due scenari appare tuttavia contenuta, suggerendo che l'elevata concentrazione costituisca una proprietà intrinseca della rete competitiva stimata, più che una conseguenza diretta dell'introduzione dei vincoli di sostituibilità.

Nel complesso, l'analisi mostra che la pressione competitiva locale è un fenomeno strutturalmente eterogeneo, che tende a concentrarsi su specifici annunci e contesti territoriali piuttosto che distribuirsi uniformemente nello spazio. L'approccio adottato incide principalmente sulla leggibilità economica di tale pressione: mentre una definizione puramente geografica enfatizza una competizione diffusa, l'introduzione di assunzioni strutturali consente di isolare interazioni competitive più informative, riconducibili a relazioni di effettiva sostituibilità. Ne risulta una rappresentazione della competizione locale più selettiva e coerente con le dinamiche economiche sottostanti.

6.3.3 Studio dell'eterogeneità della competizione locale

La valutazione mirata dell'eterogenità della competizione locale consente di fortificare l'analisi eseguita nella sezione antecedente. Infatti, precedentemente è già emerso che a livello relativo l'introduzione di vincoli risulta in una maggiore dispersione, asimmetria e concentrazione della pressione competitiva rispetto ad il modello puramente data-driven. Quindi, in questa sezione ci si focalizzerà nello studiare questi tre elementi nella configurazione che prevede l'introduzione ex-ante della sostituibilità.

L'elevato rapporto precedentemente riportato tra il 90° e il 10° è il principale indicatore della marcata asimmetria della distribuzione e della presenza di una coda destra pronunciata.

Tale valore suggerisce infatti che, anche all'interno di un perimetro competitivo definito in modo economicamente plausibile, l'intensità della competizione locale si concentra su un sottoinsieme ristretto di annunci, mentre la maggioranza sperimenta livelli di pressione sensibilmente inferiori.

Tale evidenza è ulteriormente rafforzata dall'analisi cumulativa riportata tramite la *Lorenz curve*[4] in Figura 6.4. In particolare, la curva mira a raffigurare l'andamento della pressione competitiva cumulativa locale, al crescere del numero di annunci.

[4]La *Lorenz curve* è una rappresentazione grafica della distribuzione cumulata di una grandezza rispetto alla distribuzione cumulata delle unità ordinate in senso crescente, utilizzata per misurare il grado di concentrazione; quanto maggiore è lo scostamento dalla linea di perfetta uguaglianza, tanto più elevata risulta la concentrazione della variabile considerata (Macdonald 2017).

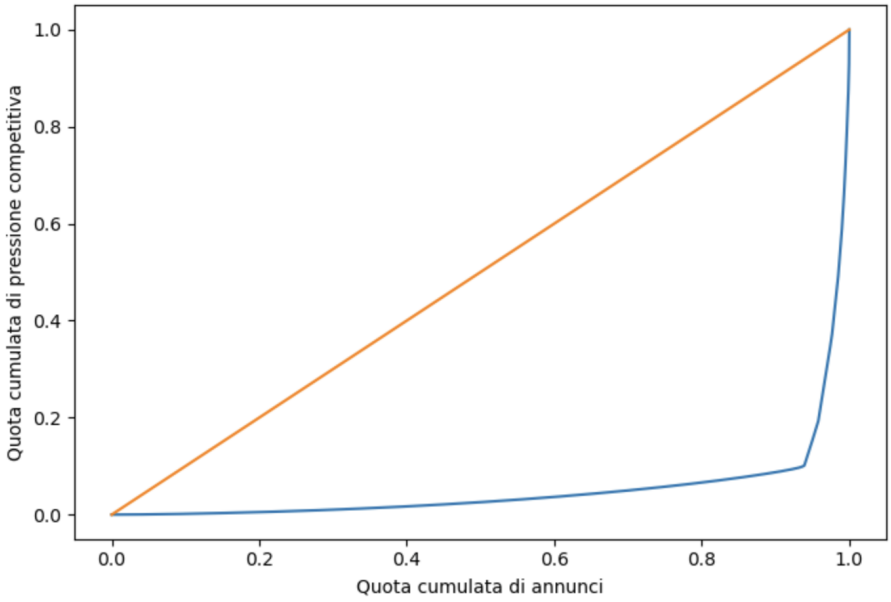


Figura 6.4. *Lorenz curve* della pressione competitiva locale

Emerge visivamente che una quota estremamente ridotta di annunci assorbe la maggior parte della pressione competitiva complessiva: il top 1% degli annunci concentra circa il 41.1% della pressione totale, mentre il top 5% ne assorbe circa l'84.7%. Tale elevato grado di diseguaglianza nella distribuzione della pressione competitiva locale è confermato dal *coefficiente di Gini*[5] pari a 0.898, rispetto ad un valore massimo di 1.

In conclusione, nel contesto piemontese, l'elevata concentrazione della pressione competitiva segnala una marcata segmentazione territoriale del mercato degli affitti brevi. L'intensità concorrenziale non risulta distribuita in modo uniforme, ma si concentra in specifiche polarità territoriali, mentre altre aree mostrano livelli di interazione competitiva significativamente più contenuti.

[5]Il *coefficiente di Gini* è una misura sintetica di concentrazione definita come il rapporto tra la differenza media (con ripetizione) osservata in una distribuzione e la differenza media massima teoricamente ottenibile nella corrispondente distribuzione massimizzante; in termini geometrici, esso coincide con il rapporto tra l'area compresa tra la linea di perfetta uguaglianza e la curva di Lorenz e l'area totale sottesa alla linea di perfetta uguaglianza (Pellegrino 2020)

6.4 Rete competitiva risultante

La presente sezione riporta i risultati relativi alla costruzione e all'analisi della rete competitiva derivata dall'intensità di interazione stimata nel modello spaziale. In un primo momento verranno discusse le proprietà topologiche della struttura, mentre successivamente verranno discusse e visualizzate le aree di competizione individuate, mettendole in relazione ai confini amministrativi corrispondenti. In queste analisi verrà accentuato il confronto tra il Comune di Torino e le aree rurali con bassa pressione competitiva.

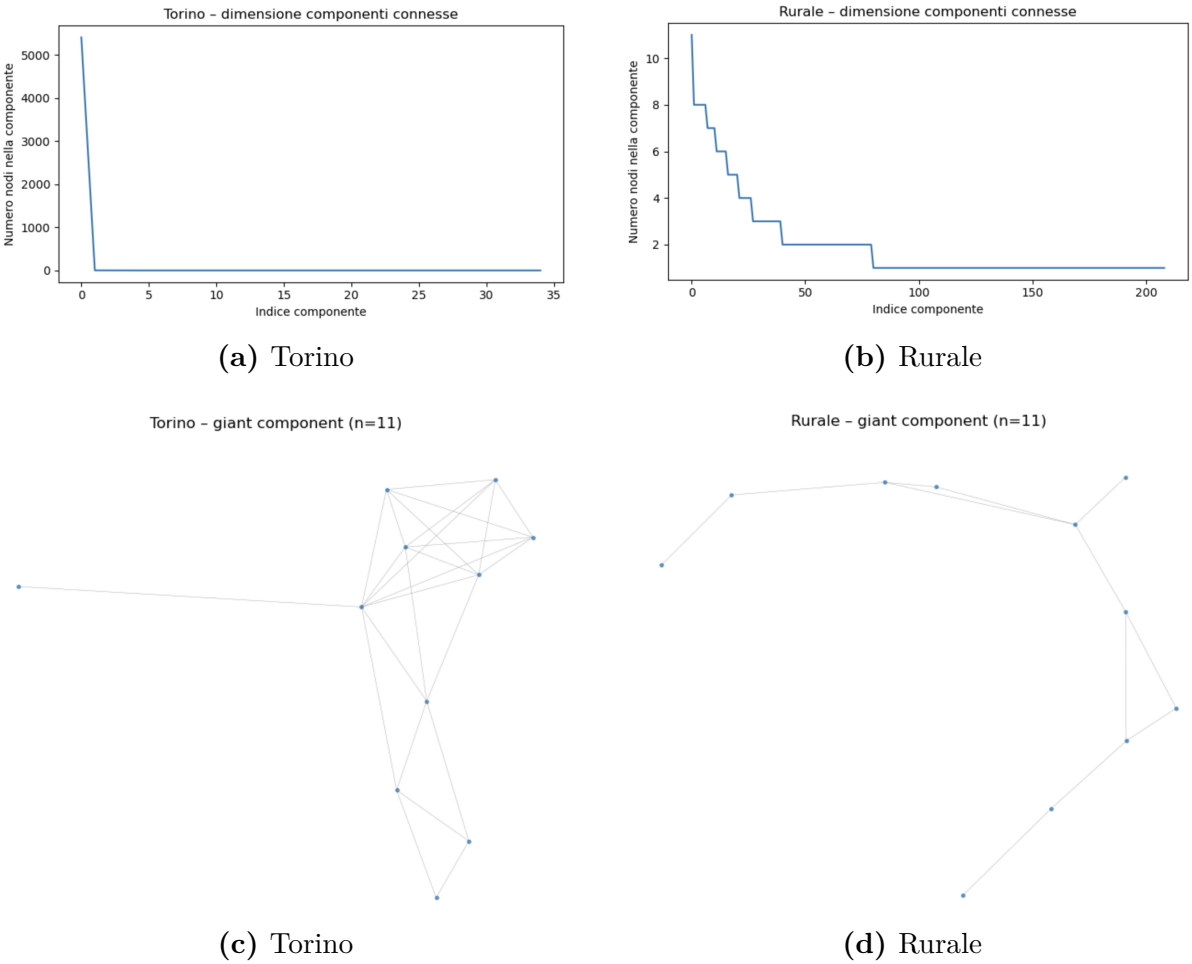
6.4.1 Struttura risultante della rete competitiva

L'analisi delle proprietà topologiche della rete competitiva consente di caratterizzare la forma empirica assunta dalla competizione nel mercato oggetto di studio. La rete è composta da 19.256 nodi, perciò viene coperto l'intero universo delle strutture attive. Però, vi sono 80.165 archi in presenza dei vincoli di sostituibilità e 98.135 in loro assenza, ciò significa che circa un 18% di contiguità geografica non si traduce in reale sostituibilità economica, rispetto alla sua definizione ex-ante. Infatti, la presenza di un arco tra due strutture indica l'esistenza di una relazione di sostituibilità competitiva economicamente rilevante, tale per cui le variazioni nei loro prezzi residui risultano interdipendenti. Inoltre, la valutazione della corrispondente densità, la quale risulta pari a 0.000432 con vincoli e 0.000529 senza vincoli, conferma che la competizione si organizza in insiemi locali di interazione.

L'effetto dei vincoli emerge in modo ancora più evidente analizzando la struttura delle comunità individuate. In assenza di vincoli di sostituibilità, la rete si articola in 256 cluster, con una dimensione media pari a 75,22 strutture e una mediana di 64, configurando una segmentazione relativamente aggregata e priva di micro-aggregati. Invece, l'introduzione dei vincoli aumenta invece il numero di cluster a 350 e riduce la dimensione media a 53,91 strutture, facendo emergere una struttura più frammentata e la presenza di sotto-mercati di dimensione molto ridotta.La competizione risulta quindi meno diffusa in senso indiscriminato e maggiormente circoscritta a nuclei relazionali coerenti.

Se si restringe l'analisi al solo Comune di Torino, già emerso nelle sezioni precedenti come l'area con la maggiore pressione competitiva in termini assoluti, con entrambi gli approcci si osserva un valore di densità circa quattro volte superiore rispetto a quello calcolato sull'intero territorio provinciale. Tale incremento segnala un livello di integrazione competitiva significativamente più elevato nel contesto urbano; a validare quanto emerso è stato quindi effettuato un confronto topologico tra il capoluogo di provincia e le aree rurali. A tale fine, sono state considerate "aree rurali", i comuni appartenenti al quartile inferiore della distribuzione della pressione competitiva totale, evitando classificazioni meramente amministrative o arbitrarie. Tale confronto viene riportato in Figura 6.5; esso è stato eseguito per il solo metodo con integrazione ex-ante della sostituibilità, poichè le considerazioni risulterebbero analoghe.

I primi due pannelli mostrano la distribuzione della dimensione delle componenti connesse ordinate per numerosità. Nel caso torinese (a), emerge chiaramente la presenza di una componente connessa dominante che assorbe la quasi totalità dei nodi, mentre le componenti residue risultano di dimensione trascurabile. Al contrario, nelle aree rurali (b) la struttura appare fortemente frammentata: non si osserva una componente prevalente di dimensione comparabile e il sistema si articola in numerose componenti di piccola taglia, segnalando una minore coesione relazionale.



(a) Torino **(b)** Rurale

(c) Torino **(d)** Rurale

Figura 6.5. Confronto topologico tra il Comune di Torino e le aree a bassa pressione competitiva.

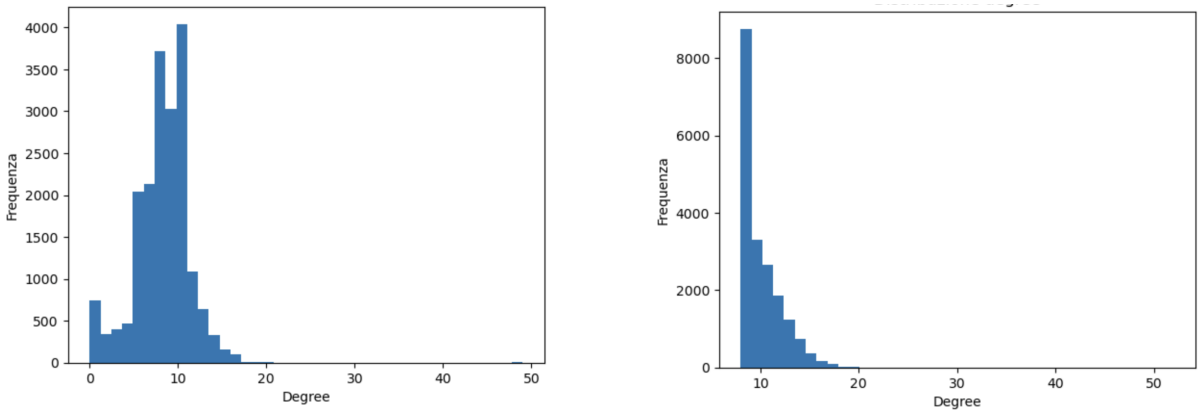
Invece, nei successivi pannelli viene approfondita la *giant component* di ciascun contesto, mantenendo costante la numerosità dei nodi al fine di garantire comparabilità strutturale. Dove con *giant component* si intende la componente connessa di dimensione massima all'interno della rete, ossia il sottoinsieme più ampio di strutture tra le quali esiste un cammino che le collega, direttamente o indirettamente, attraverso relazioni di sostituibilità competitiva (Dabrowski 2015).

Quindi, emerge che anche a parità di dimensione, la componente torinese (c) evidenzia una configurazione più densa e clusterizzata, con elevata ridondanza dei collegamenti e presenza di triangoli competitivi. La struttura rurale (d), invece, assume una forma più lineare e meno interconnessa, con minore sovrapposizione dei legami e una configurazione prossima a una struttura ad albero.

Tale rappresentazione dimostra che la maggiore pressione competitiva urbana si riflette non soltanto nell'intensità economica dei legami, ma anche nella configurazione topologica della rete, confermando che la polarizzazione territoriale della competizione è accompagnata da una differente struttura relazionale del mercato.

Come annunciato in precedenza, si è scelto di eseguire tale analisi solo per una configurazione, in quanto è stato accertato che l'introduzione dei vincoli di sostituibilità riduce il numero medio di competitor per struttura, ma non altera la natura topologica della rete. Tale evidenza è emersa studiando la distribuzione del *degree* riportata in Figura 6.6, ossia il numero di competitor diretti con cui la struttura intrattiene una relazione di sostituibilità economicamente rilevante.

Il confronto tra i due scenari evidenzia infatti una sostanziale stabilità della forma distributiva. Con vincoli di sostituibilità, il grado medio è pari a 8,33 (mediana 8), mentre in loro assenza il valore medio sale a 10,19 (mediana 10), coerentemente con l'aumento del numero complessivo di archi. Tuttavia, in entrambi i casi la distribuzione rimane compatta, priva di code pronunciate e di nodi con grado eccezionalmente elevato, dimostrando che i vincoli non alterano la natura topologica della rete.



(a) Degree con vincoli di sostituibilità **(b)** Degree senza vincoli

Figura 6.6. Distribuzione del grado della rete competitiva

Tale evidenza va letta congiuntamente ai risultati relativi alla *strength*, definita come la somma dei pesi degli archi incidenti su ciascun nodo e coincidente, nel presente contesto, con la misura di pressione competitiva locale (Fujishige 2005). A fronte di una

distribuzione del grado relativamente omogenea, la strength risulta invece fortemente concentrata.

Poiché tale concentrazione permane anche in assenza dei vincoli di sostituibilità, essa rappresenta una proprietà strutturale del sistema. Ne consegue che la polarizzazione competitiva non dipende dal numero di competitor diretti, bensì dall'intensità dei legami che gravano su un nucleo ristretto di strutture. La rete competitiva conferma quindi, in termini topologici, quanto già emerso a livello distributivo: l'asimmetria del mercato è riconducibile alla concentrazione della pressione, non alla dimensione del set competitivo locale.

6.4.2 Analisi delle aree di competizione risultanti

Per concludere la valutazione dei risultati, tale sezione si pone l'obiettivo di presentare le aree di competizione identificate con un intento interpretativo. A tale scopo, i cluster verranno messi in relazione con i confini amministrativi sottostanti, ovvero i quartieri per l'area metropolitana di Torino (fonte dati da Archivio INSPIRE della Commissione Europea[6]) ed i comuni per il resto della regione (fonte dati ISTAT[7]). Infatti, tale confronto consente di comprendere se la geografia della competizione riflette la geografia amministrativa.

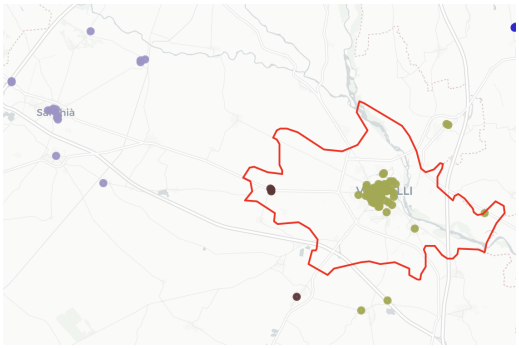
Una lettura congiunta delle mappe riportate consente di sottolineare che i cluster individuati seguono configurazioni funzionali determinate. Ne emerge un sistema competitivo policentrico, in cui la scala della concorrenza varia significativamente a seconda della densità urbana e della specializzazione turistica.

A livello aggregato, si osservano tre configurazioni principali: (i) capoluoghi o comuni a bassa vocazione turistica, in cui la competizione si sviluppa su scala intercomunale estesa; (ii) piccole destinazioni turistiche specializzate, nelle quali, nonostante l'estensione territoriale limitata, emergono configurazioni competitive frammentate; (iii) il capoluogo Torino, caratterizzato da cluster fortemente micro-localizzati.

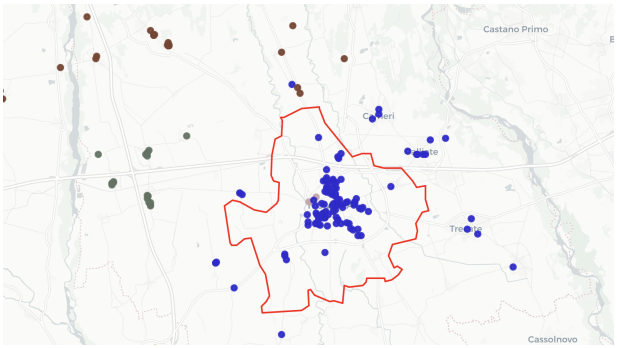
[6] `https://inspire.ec.europa.eu/metadata-codelist/OnLineDescriptionCode/accessPoint`
[7] `https://www.istat.it/notizia/Confinidelleunitàamministrativeafinistatistici`

Capoluoghi e comuni a bassa vocazione turistica

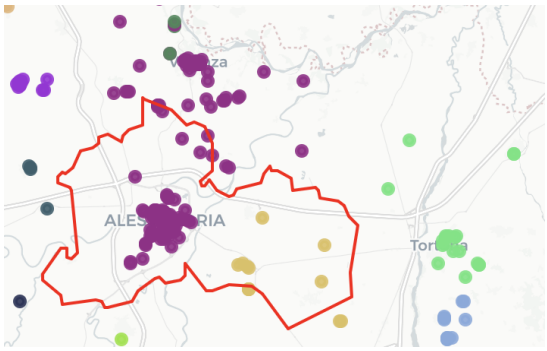
L'obiettivo di questo approfondimento è mostrare in che modo la geografia della competizione si strutturi in contesti urbani in cui la domanda di locazioni brevi è prevalentemente di natura funzionale e non esperienziale. In particolare, osservando i risultati, i cinque capoluoghi in Figura 6.8 sono stati identificati visivamente con una limitata specializzazione turistica: Vercelli, Novara, Alessandria, Cuneo e Biella.



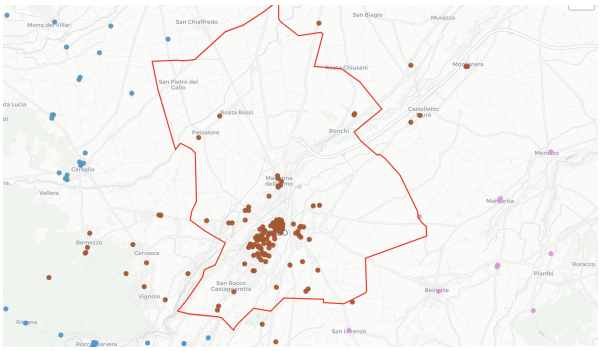
(a) Comune di Vercelli



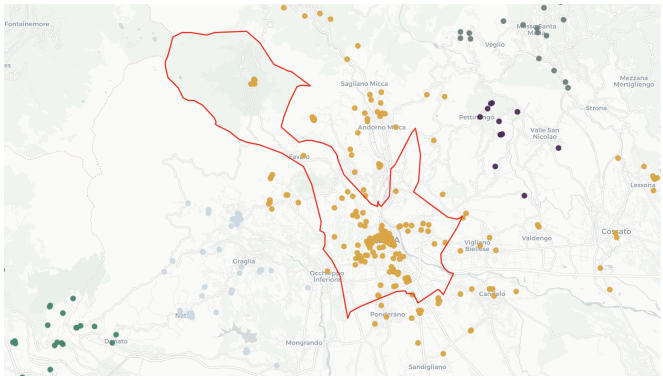
(b) Comune di Novara



(c) Comune di Alessandria



(d) Comune di Cuneo



(e) Comune di Biella

Figura 6.7. Configurazione capoluoghi a bassa vocazione turistica

Un risultato trasversale riguarda la configurazione prevalentemente monocentrica dei sistemi competitivi. In tutti i casi analizzati, le strutture si concentrano attorno al nucleo urbano centrale, senza generare una frammentazione significativa a livello infra-comunale. La scala della concorrenza appare dunque coincidere con l'intero centro urbano funzionale, piuttosto che con micro-aree interne.

Dalle immagini, emerge che per ciascuna città il cluster principale tende ad essere in larga misura circoscritto entro il territorio comunale, pur senza coincidere perfettamente con esso. A differenza delle restanti configurazioni, dove all'interno dello stesso comune coesistono numerosi micro-cluster o, al contrario, un unico sistema competitivo si estende su più comuni, in questo caso il confine amministrativo intercetta una quota significativa del mercato rilevante. Perciò, il cluster può includere aree limitrofe funzionalmente integrate, tuttavia rispetto alla maggior parte delle configurazioni il livello comunale risulta qui relativamente più affidabile come proxy territoriale.

Piccole destinazioni turistiche specializzate

In questa configurazione rientra lo studio di aree turistiche con estensione limitata, caratterizzate da una forte segmentazione. Tali zone sono state identificate visivamente, e quanto individuato è stato confermato secondo valutazione numeriche della pressione. A differenza dei capoluoghi a bassa vocazione turistica, in queste destinazioni la rete competitiva evidenzia una frammentazione infra-comunale che riflette modalità di fruizione profondamente differenti.

La prima zona distinguibile è l'area dei comprensori alpini della provincia di Torino in Figura 6.8), in cui la frammentazione intra-comunale appare coerente con un'offerta strutturata attorno alla prossimità agli impianti di risalita. In tutte le principali località è possibile distinguere almeno tre macro-tipologie di strutture: (i) strutture "ski-in/ski-out" con alta accessibilità agli impianti, (ii) strutture situate nel centro abitato o in prossimità dei principali servizi, che competono prevalentemente sulla comodità, e (iii) strutture più periferiche, spesso caratterizzate da maggiore capacità ricettiva, che competono su prezzo relativo.

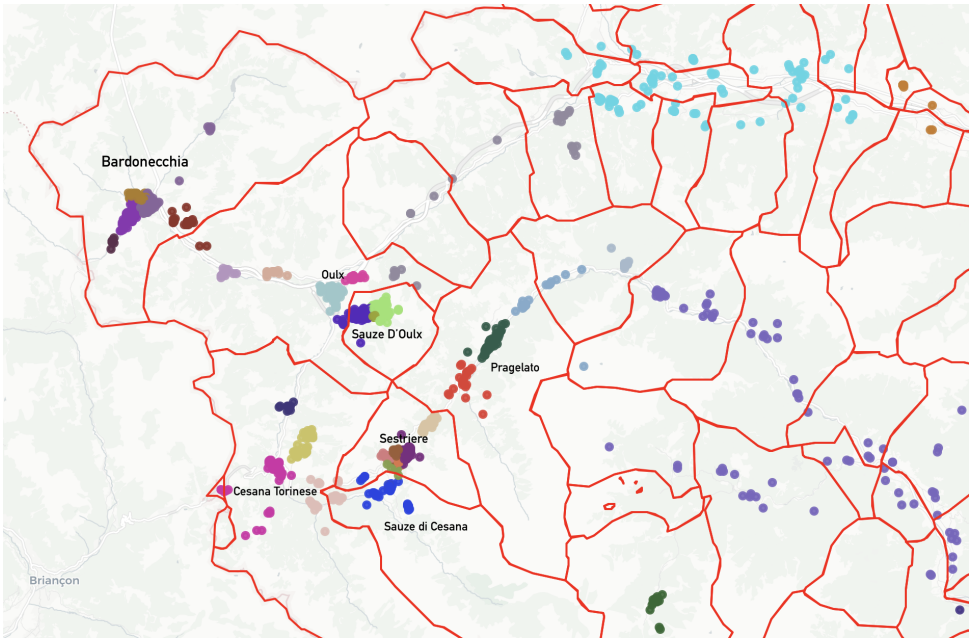


Figura 6.8. Comprensori alpini della provincia di Torino

La rete mostra come la competizione non sia isotropa rispetto al centro abitato: due strutture geograficamente vicine ma con diverso grado di accessibilità agli impianti possono appartenere a cluster distinti, suggerendo che la contiguità spaziale non coincide con la contiguità competitiva. In particolare, la pressione competitiva tende a concentrarsi nelle aree con maggiore densità di strutture ad accesso diretto alle piste, mentre le zone più distanti presentano interazioni più deboli ma relativamente omogenee al loro interno. La struttura dei cluster riflette quindi una domanda fortemente segmentata per modalità di fruizione e per sensibilità ai costi di accesso, intesi non solo in termini monetari ma anche di tempo e comodità.

Nel caso dei comprensori alpini della provincia di Cuneo in Figura 6.9, emerge una configurazione complessivamente meno compatta e più dispersa territorialmente rispetto a quella torinese. Le strutture si distribuiscono lungo direttrici vallive meno dense, generando cluster più estesi ma internamente meno intensi in termini di interconnessione. In tali contesti, la sostituibilità tra strutture sembra più elastica rispetto al prezzo e meno rigidamente ancorata alla micro-prossimità agli impianti. Ne deriva una competizione relativamente più guidata dal prezzo, in cui la localizzazione rimane rilevante ma non esclusiva.

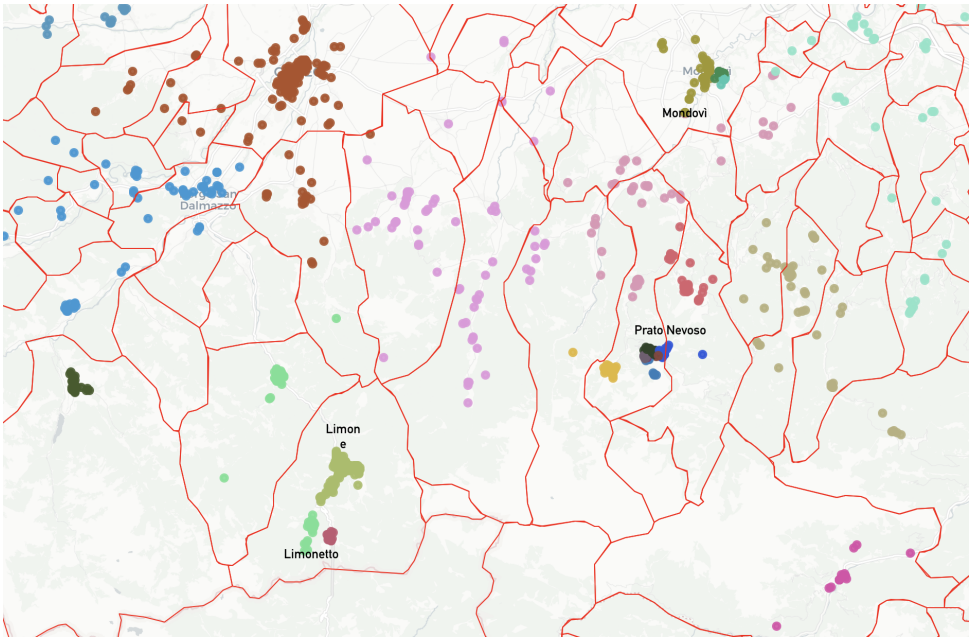


Figura 6.9. Comprensori alpini della provincia di Cuneo

All'interno di questo quadro, Prato Nevoso emerge come caso distintivo per maggiore coesione interna. Il cluster appare più compatto e autocontenuto, verosimilmente per la forte concentrazione edilizia attorno al nucleo impiantistico principale. Qui la prossimità geografica coincide in larga misura con la prossimità funzionale, rafforzando le interdipendenze competitive. Invece, al contrario di quanto ci si potrebbe aspettare, Limone Piemonte non mostra un livello di frammentazione analogo. Una possibile spiegazione risiede nella maggiore integrazione tra centro abitato, servizi e accesso agli impianti, che attenua le discontinuità nella sostituibilità tra alloggi.

Nel sistema lacustre costituito dal Lago Maggiore e dal Lago d'Orta, riportato in Figura 6.10, la logica competitiva assume una configurazione ancora più chiaramente *product-based*, in quanto il bene turistico diventa la fruizione del lago in sé. In questo caso, la segmentazione lungo l'asse costiero non è determinata soltanto dalla presenza o meno della vista lago, ma da micro-differenze di localizzazione altamente osservabili dagli utenti: esposizione panoramica, accesso diretto all'acqua, prossimità ad imbarchi, passeggiate e servizi turistici, presenza di terrazzi o spazi esterni.

L'elevata frammentazione lungo lo stesso fronte costiero indica che spostamenti di poche centinaia di metri possono implicare variazioni rilevanti nell'attrattività percepita, gene-

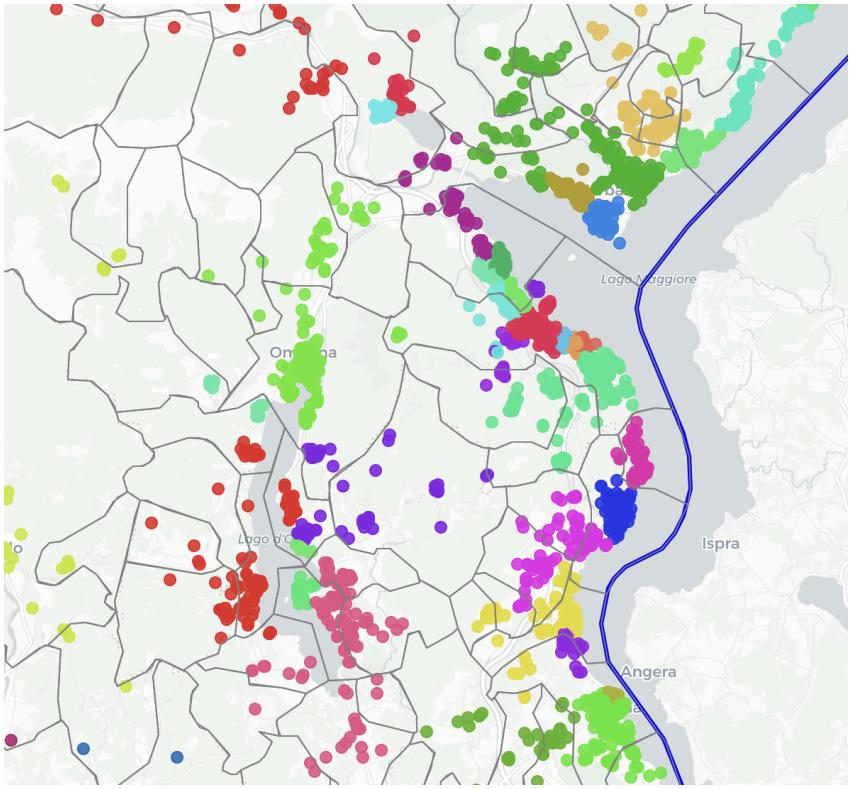


Figura 6.10. Sistema lacustre della provincia di Novara e Verbania

rando cluster distinti anche all'interno del medesimo comune. La competizione assume quindi una dimensione sovra-comunale: strutture situate in comuni diversi ma lungo uno stesso tratto funzionale di costa risultano più interconnesse tra loro rispetto ad alloggi interni allo stesso comune ma privi degli attributi core legati alla fruizione del lago. Ne emerge un chiaro disallineamento tra confini amministrativi e confini competitivi, coerente con l'idea che il mercato rilevante sia definito dalla sostituibilità effettiva e non dall'appartenenza istituzionale.

Infine, l'area di Alba rappresenta un caso qualitativamente distinto. Infatti, qui la domanda non è mono-attrattore come nel lago né rigidamente infrastrutturale come nei comprensori sciistici, ma fortemente legata a un mix tra fruizione urbana del centro storico e fruizione esperienziale diffusa sul territorio circostante, in particolare lungo itinerari enogastronomici. La rete evidenzia cluster meno polarizzati e più distribuiti, suggerendo una competizione che si sviluppa su un territorio più ampio e meno vincolato a un unico punto di attrazione. In questo caso, la sostituibilità tra strutture dipende meno dalla micro-localizzazione puntuale e più dall'inclusione in un'esperienza territoriale complessi-

va, con una componente stagionale legata a eventi che può generare picchi temporanei che non vengono catturati dalla visualizzazione statica del periodo oggetto del caso studio.

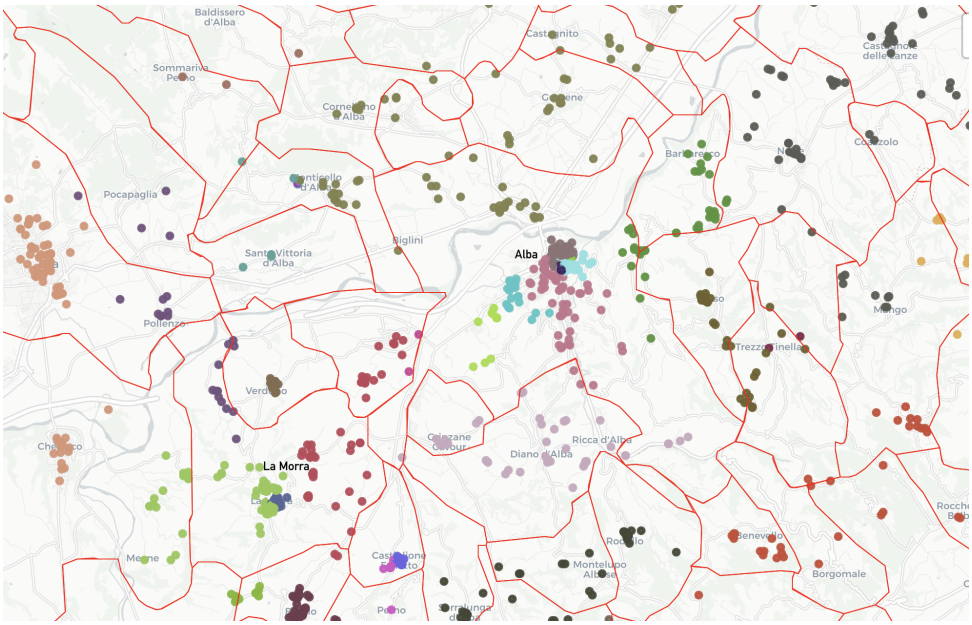


Figura 6.11. Alba e provincia

Tali configurazioni mostrano come la competizione, anche in contesti territorialmente ristretti, possa organizzarsi secondo logiche selettive, per cui strutture vicine ma differenziate per qualità o target possono non appartenere allo stesso mercato locale, mentre unità collocate in aree limitrofe ma simili per configurazione dell'offerta risultano strettamente interconnesse.

Provincia e Comune di Torino

Nel caso della Provincia di Torino, la rete competitiva evidenzia una marcata polarizzazione attorno al capoluogo, che concentra una quota dominante della pressione regionale e si configura come principale nodo di interazione tra le strutture. Tuttavia, tale centralità non implica omogeneità territoriale. Al di fuori del Comune di Torino, la competizione provinciale si articola lungo direttrici differenti: da un lato i poli turistici montani, caratterizzati da dinamiche stagionali e da cluster relativamente autocontenuti, come si è potuto osservare nel paragrafo precedente; dall'altro i comuni della cintura metropo-

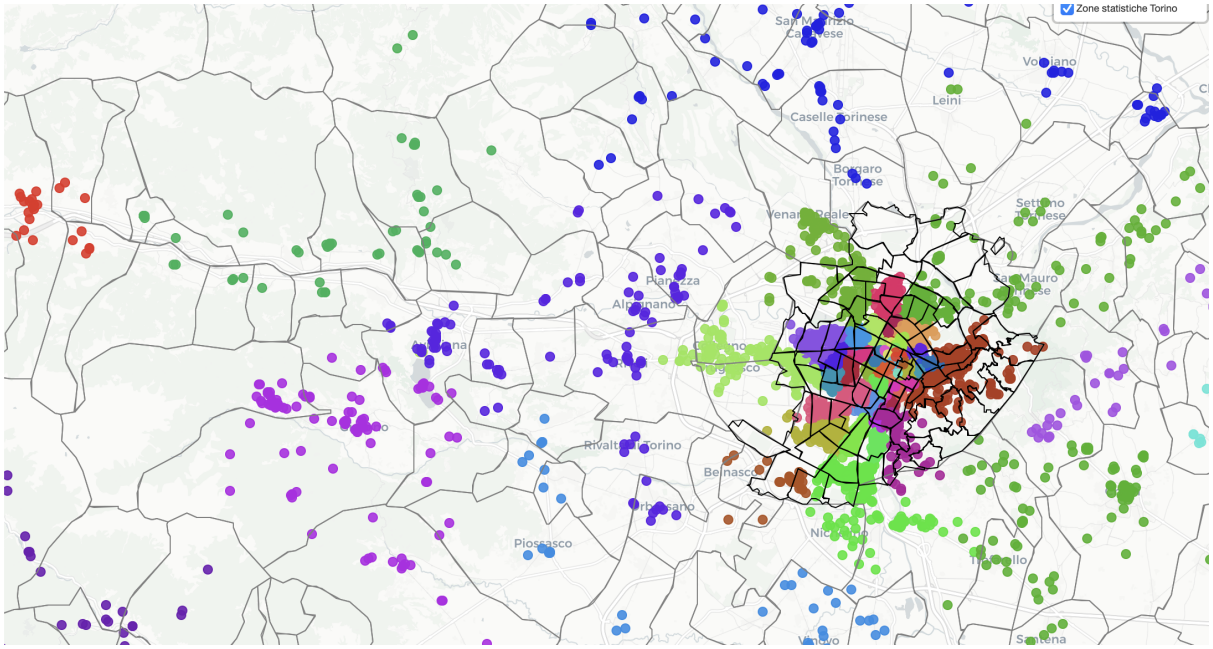


Figura 6.12. Comuni della cintura metropolitana di Torino

litana, nei quali la competizione è più debole e fortemente dipendente dalla prossimità funzionale alla città.

In questi ultimi contesti, la sostituibilità tra listing è guidata prevalentemente da prezzo residuo, dimensione dell'alloggio e accessibilità infrastrutturale, più che da una specifica attrattività turistica autonoma. Le strutture competono come alternative economiche o logistiche al soggiorno nel capoluogo, generando interazioni meno dense ma comunque riconducibili all'orbita competitiva torinese.

All'interno del Comune di Torino, la competizione assume una configurazione più stratificata. La scala urbana introduce una segmentazione significativa a livello di quartiere, con una progressiva attenuazione della pressione competitiva man mano che ci si allontana dal nucleo centrale. Nelle aree periferiche la domanda appare più funzionale, e la competizione risulta meno intensa e meno frammentata internamente.

Invece, il centro cittadino rappresenta il segmento più denso della provincia, in particolare l'area compresa tra Centro storico, San Salvario, Vanchiglia e parte di Crocetta mostra un'elevata concentrazione di strutture, restituendo un mosaico di comunità competitive che coesistono nello spazio urbano.

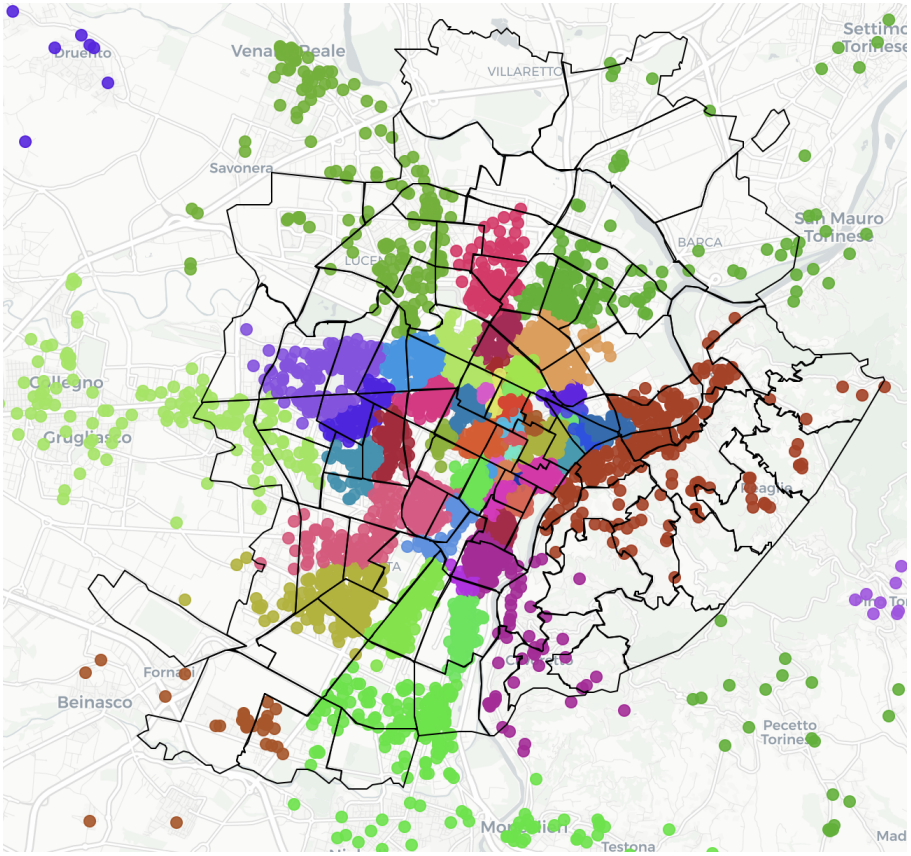


Figura 6.13. Comune di Torino

Capitolo 7

Conclusioni

Il presente lavoro ha sviluppato un framework per l'analisi della competizione endogena nel mercato degli affitti brevi. In particolare, l'algoritmo tramite un approccio integrato tra modello edonico penalizzato, isolamento della componente competitiva e rappresentazione in forma di rete pesata ha consentito l'analisi del mercato Airbnb in Piemonte, superando definizioni esogene del perimetro competitivo. I risultati evidenziano che la competizione nel mercato analizzato è locale, selettiva ed eterogenea.

La penalizzazione ℓ_1 prevista dal metodo LASSO ha consentito di stimare una matrice di interdipendenze strutturalmente sparsa, selezionando in modo endogeno, tra le possibili relazioni spaziali, soltanto quelle economicamente significative e coerenti con la sostituibilità osservata nei dati. Inoltre, sono stati inseriti vincoli *ex-ante* di sostituibilità, con l'obiettivo di evitare accoppiamenti eccessivamente eterogenei, producendo di conseguenza una delimitazione più granulare dei sotto-mercati.

La robustezza dell'approccio è stata validata integrando parallelamente una componente demand-side tramite Adaptive LASSO, i cui pesi di penalizzazione sono stati calibrati sulla base delle preferenze espresse dagli utenti. Tale integrazione ha confermato che le aree individuate riflettono anche criteri di rilevanza percepita dal lato della domanda.

La segmentazione endogena della rete competitiva restituisce sottosistemi spaziali all'interno dei quali l'intensità delle interdipendenze risulta relativamente omogenea. Tali cluster non coincidono in modo strutturato con i confini amministrativi, suggerendo che la geografia economica della competizione si organizza secondo logiche funzionali, determinate dall'interazione tra prezzi residui.

In particolare, nei contesti urbani ad alta densità emerge la presenza di macro-cluster fortemente interconnessi, mentre nelle aree periferiche la competizione assume configu-

razioni più frammentate e circoscritte. Questa eterogeneità territoriale conferma che la pressione competitiva non è distribuita uniformemente nello spazio, ma si concentra in nuclei relazionali specifici, dando luogo a sotto-mercati distinti per intensità.

Questi risultati hanno implicazioni dirette per le strategie di *pricing* della piattaforma. Infatti, la natura selettiva delle interdipendenze suggerisce che i sistemi di raccomandazione del prezzo dovrebbero basarsi su una mappatura endogena dei competitor effettivamente influenti, superando le logiche amministrative, tramite l'integrazione di misure di pressione competitiva locale nei modelli di *dynamic pricing*.

L'eterogeneità tra cluster suggerisce inoltre l'opportunità di strategie differenziate. In aree ad alta interdipendenza sono necessari strumenti di monitoraggio, mentre nelle zone periferiche, una maggiore autonomia decisionale appare compatibile con la minore intensità competitiva. Le aree di competizione individuate costituiscono dunque un'unità operativa potenzialmente più efficace rispetto ai tradizionali confini amministrativi.

Infine, l'algoritmo sviluppato può rappresentare il punto di partenza per futuri sviluppi. In tal contesto, tra le principali prospettive emerge la formalizzazione dei Sistemi Locali Turistici come evoluzione operativa delle aree di competizione individuate. Il passaggio può avvenire attraverso una fase di consolidamento dei cluster reticolari, imponendo criteri di elevata densità interna, contenimento della pressione competitiva verso l'esterno e omogeneità delle dinamiche di prezzo. In presenza di interazioni reciproche marginali e profili di domanda simili, cluster distinti possono essere aggregati in un unico sistema funzionale, indipendentemente dalla mera contiguità spaziale.

Inoltre, per la definizione degli SLT risulta necessario operare un'estensione dinamica dell'analisi, osservando l'evoluzione delle aree competitive nel tempo e distinguendo variazioni contingenti legate a stagionalità o shock regolatori.

Infine, dovrà essere testata la generalizzabilità del framework attraverso l'applicazione dell'algoritmo a contesti geografici differenti, valutando la stabilità delle aree individuate e la trasferibilità del metodo oltre il caso studio analizzato.

Annexes

A Questionario

Di seguito, viene riportato il questionario sottoposto ad 105 utenti per definire i pesi previsti dal metodo ALASSO sviluppato nella Sezione 5.6.

Il questionario è stato costruito con l'ausilio di Google Form e diffuso tramite diverse piattaforme social.

L'utente è stato profilato solamente sulla base della frequenza di utilizzo della piattaforma Airbnb, in quanto il modello non deve essere condizionato da altre caratteristiche dell'utilizzatore.

Preferenze di scelta degli alloggi Airbnb

Il questionario ha l'obiettivo di raccogliere informazioni sulle preferenze percepite dei consumatori rispetto alle principali caratteristiche degli alloggi Airbnb, al fine di integrare una prospettiva demand-side nell'analisi della competitività della mia tesi.

Per i lettori più curiosi, questa procedura incide direttamente sulla fase di **stima del modello Adaptive LASSO** attraverso la definizione dei pesi di penalizzazione associati alle singole variabili esplicative. In particolare, le importanze percepite derivate dal questionario vengono utilizzate per costruire un insieme di pesi wj che modulano l'intensità della penalizzazione ℓ1 applicata a ciascun coefficiente. Ne consegue che le caratteristiche valutate come più rilevanti dai rispondenti sono associate a pesi più bassi e risultano quindi soggette a una penalizzazione relativamente meno severa, riducendo la probabilità di essere azzerate nel processo di selezione delle variabili.

* Indica una domanda obbligatoria

1. Con quale frequenza utilizzi Airbnb? *

Contrassegna solo un ovale.

- ◯ < 1 volta / anno
- ◯ 1–2 volte / anno
- ◯ 3–5 volte / anno
- ◯ > 5 volte / anno

Istruzioni

Per ciascuna delle seguenti caratteristiche, indica **quanto è importante per te nella scelta di un alloggio Airbnb**, utilizzando una scala da **1 (per niente importante)** a **7 (estremamente importante)**.

2. Quando confronti più alloggi simili, **in che misura il prezzo finale influisce sulla tua decisione di prenotazione**?

Contrassegna solo un ovale.

	1	2	3	4	5	6	7	
Per	◯	◯	◯	◯	◯	◯	◯	Estremamente importante

3. In che misura la **disponibilità di un numero adeguato di camere da letto** incide sulla tua percezione di adeguatezza dell'alloggio rispetto alle tue esigenze?

Contrassegna solo un ovale.

	1	2	3	4	5	6	7	
Per	◯	◯	◯	◯	◯	◯	◯	Estremamente importante

4. Quanto pesa, nella tua scelta, la **presenza di un numero sufficiente di bagni** in relazione al numero di ospiti?

Contrassegna solo un ovale.

	1	2	3	4	5	6	7	
Per	◯	◯	◯	◯	◯	◯	◯	Estremamente importante

5. In che misura la **presenza di servizi e dotazioni aggiuntive** (es. cucina attrezzata, Wi-Fi, aria condizionata, lavatrice) contribuisce a rendere un alloggio preferibile rispetto ad alternative simili?

Contrassegna solo un ovale.

	1	2	3	4	5	6	7	
Per	◯	◯	◯	◯	◯	◯	◯	Estremamente importante

6. Quanto la **valutazione complessiva dell'alloggio** sulla piattaforma influenza la tua fiducia nella qualità dell'esperienza offerta?

Contrassegna solo un ovale.

	1	2	3	4	5	6	7	
Per	◯	◯	◯	◯	◯	◯	◯	Estremamente importante

7. In che misura il **numero di recensioni disponibili** influisce sulla tua percezione di affidabil dell'alloggio?

Contrassegna solo un ovale.

	1	2	3	4	5	6	7	
Per	◯	◯	◯	◯	◯	◯	◯	Estremamente importante

8. Quanto la **tipologia dell'alloggio** (intero appartamento, stanza privata, stanza condivisa, hotel) condiziona la tua decisione finale, indipendentemente dal prezzo?

Contrassegna solo un ovale.

	1	2	3	4	5	6	7	
Per	◯	◯	◯	◯	◯	◯	◯	Estremamente importante

9. In che misura il fatto che l'host sia riconosciuto come **Superhost Airbnb** influisce sulla tua propensione a prenotare l'alloggio?

Contrassegna solo un ovale.

	1	2	3	4	5	6	7	
Per	◯	◯	◯	◯	◯	◯	◯	Estremamente importante

B Visualizzazione dei risultati

In questa sezione vengo riportati i blocchi di codice utilizzati per l'analisi dei risultati. Esse prevedono sia valutazioni statistiche, sia visualizzazioni. Nel caso in cui il medesimo blocco sia stato utilizzato per la valutazione di confronto tra due metodi, quali ad esempio l'approccio LASSO rispetto alla versione adattiva, o la soluzione con i vincoli di sostituibilità, rispetto a quella data-driven, verrà riportato solamente una volta.

Gli script per l'analisi dei risultati, così come per l'implementazione del modello e la realizzazione del pre-processing, sono stati eseguiti in *Jupyter notebook*, quindi è da intendersi che le librerie e le classi tipiche di questo ambiente siano state importate.

B.1 Residui del modello edonico

Ottenimento delle statistiche e visualizzazioni per la sezione 6.1.

```
# Statistiche Descrittive
aggregated_df["epsilon_hat"].describe(
    percentiles=[0.05, 0.25, 0.5, 0.75, 0.95]
)
#Violin plot
plt.figure(figsize=(5, 4))
sns.violinplot(
    y=aggregated_df["epsilon_hat"],
    inner="box",
    cut=0
)
plt.ylabel("Residuo edonico")
plt.show()
# Distribuzione spaziale dei residui
plt.figure(figsize=(6, 5))
plt.scatter(
    aggregated_df[COL_LON],
    aggregated_df[COL_LAT],
    c=aggregated_df["epsilon_hat"],
```

```
    cmap="coolwarm",
    s=5
)
plt.colorbar(label="Residuo edonico")
plt.show()
# Pattern spaziali medi dei residui edonici
aggregated_df["lon_bin"] = pd.cut(aggregated_df[COL_LON], bins=50)
aggregated_df["lat_bin"] = pd.cut(aggregated_df[COL_LAT], bins=50)
grid_mean = (
    aggregated_df
    .groupby(["lon_bin", "lat_bin"])["epsilon_hat"]
    .mean()
    .reset_index()
)
plt.figure(figsize=(6, 5))
plt.scatter(
    grid_mean["lon_bin"].apply(lambda x: x.mid),
    grid_mean["lat_bin"].apply(lambda x: x.mid),
    c=grid_mean["epsilon_hat"],
    cmap="coolwarm",
    s=30
)
plt.colorbar(label="Residuo medio edonico")
plt.show()
```

B.2 Struttura della competizione

Ottenimento delle statistiche e visualizzazioni per la sezione 6.2.

```
rho_subst = float(ols2.params["neighbors_residual_mean"])
se_subst  = float(ols2.bse["neighbors_residual_mean"])
t_subst   = float(ols2.tvalues["neighbors_residual_mean"])
p_subst   = float(ols2.pvalues["neighbors_residual_mean"])
n_subst   = int(ols2.nobs)
#IC 95%
ci_subst = ols2.conf_int().loc["neighbors_residual_mean"]
```

B.3 Evidenze della competizione

Ottenimento delle statistiche e visualizzazioni per la sezione 6.3.

```
#Pressione Competitiva
P_subst = pd.Series(
    {
        n: sum(attr.get("weight", 0.0) for _, _, attr in G.edges(n, data=True))
        for n in G.nodes()
    },
    name="pressure_subst"
)
P_subst = P_subst.dropna()
P_subst = P_subst[P_subst > 0]
median_subst = P_subst.median()
p10_subst = P_subst.quantile(0.10)
p90_subst = P_subst.quantile(0.90)
ratio_p90_p10_subst = p90_subst / p10_subst if p10_subst > 0 else np.nan
k_subst = int(np.ceil(0.10 * len(P_subst)))
top10_share_subst = P_subst.nlargest(k_subst).sum() / P_subst.sum()
#Boxplot della distribuzione competitiva
P_subst_plot = P_subst.dropna()
P_subst_plot = P_subst_plot[P_subst_plot > 0]
data = [P_subst_plot.values]
plt.figure(figsize=(8, 5))
plt.boxplot(
    data,
    labels=labels,
    showfliers=False,
    whis=[10, 90]
)
plt.ylabel("Pressione competitiva locale $P_i$")
plt.title("Distribuzione della pressione competitiva locale")
plt.tight_layout()
plt.show()
#Eterogeneità della competizione
P = P_subst.dropna().astype(float)
```

```
stats = {
    "median": float(P.median()),
    "p10": float(P.quantile(0.10)),
    "p90": float(P.quantile(0.90)),
    "p90_p10": float(P.quantile(0.90) / P.quantile(0.10)) if P.quantile(0.10) > 0
    ↪  else np.inf,
    "p95": float(P.quantile(0.95)),
    "p99": float(P.quantile(0.99)),
}
# Concentrazione
def top_share(series, q):
    s = series.sort_values(ascending=False)
    k = max(1, int(np.ceil(q * len(s))))
    return float(s.iloc[:k].sum() / s.sum()) if s.sum() > 0 else np.nan
share_top1  = top_share(P, 0.01)
share_top5  = top_share(P, 0.05)
share_top10 = top_share(P, 0.10)
# Coefficiente di Gini
def gini(x):
    x = np.asarray(x, dtype=float)
    x = x[~np.isnan(x)]
    if np.all(x == 0):
        return 0.0
    x = np.sort(x)
    n = len(x)
    cumx = np.cumsum(x)
    return float((n + 1 - 2 * np.sum(cumx) / cumx[-1]) / n)
gini_P = gini(P.values)
# Lorenz curve
x = np.sort(P.values)
x = x[~np.isnan(x)]
cum = np.cumsum(x)
cum_share = cum / cum[-1] if cum[-1] > 0 else cum
pop_share = np.arange(1, len(x) + 1) / len(x)
plt.figure(figsize=(7,5))
plt.plot(pop_share, cum_share, label="Lorenz curve")
plt.plot([0,1], [0,1])  # linea di uguaglianza
```

```
plt.title("Lorenz curve della pressione competitiva locale")
plt.xlabel("Quota cumulata di annunci")
plt.ylabel("Quota cumulata di pressione competitiva")
plt.tight_layout()
plt.show()
```

B.4 Analisi della rete competitiva

Ottenimento delle statistiche e visualizzazioni per la sezione 6.4.1

```
N = G.number_of_nodes()
E = G.number_of_edges()
density = nx.density(G)
#Distribuzione Degree
degrees = np.array([d for n, d in G.degree()])
#Strenght
strength = np.array([
    sum(attr.get("weight", 0.0) for _, _, attr in G.edges(n, data=True))
    for n in G.nodes()
])
def gini(x):
    x = np.sort(np.asarray(x))
    n = len(x)
    cumx = np.cumsum(x)
    return (n + 1 - 2 * np.sum(cumx) / cumx[-1]) / n
#Densità
density_torino = nx.density(G_torino)
#Dimensioni Connesse
def component_size_distribution(G_sub, title):
    sizes = sorted([len(c) for c in nx.connected_components(G_sub)], reverse=True)
    plt.figure(figsize=(7,4))
    plt.plot(sizes)
    plt.title(title)
    plt.xlabel("Indice componente")
    plt.ylabel("Numero nodi nella componente")
    plt.tight_layout()
```

```
    plt.show()
    return sizes
# Struttura core-periphery
largest_cc = max(nx.connected_components(G), key=len)
G_gc = G.subgraph(largest_cc).copy()
G_gc.remove_edges_from(nx.selfloop_edges(G_gc))
core_numbers = nx.core_number(G_gc)
threshold = np.percentile(list(core_numbers.values()), 90)
core_nodes = [n for n in G_gc.nodes() if core_numbers[n] >= threshold]
periphery_nodes = [n for n in G_gc.nodes() if n not in core_nodes]
G_clean = nx.Graph()
for u, v, d in G_gc.edges(data=True):
    if u in core_nodes or v in core_nodes:
        G_clean.add_edge(u, v, weight=d.get("weight", 1))
pos = nx.spring_layout(G_clean, k=0.15, iterations=200, seed=42)
strength = dict(G_gc.degree(weight="weight"))
max_s = max(strength[n] for n in G_clean.nodes())
node_sizes = [
    1500 * (strength[n] / max_s) if n in core_nodes else 15
    for n in G_clean.nodes()
]
plt.figure(figsize=(9,9))
nx.draw_networkx_edges(
    G_clean, pos,
    alpha=0.04,
    width=0.4,
    edge_color="black"
)
nx.draw_networkx_nodes(
    G_clean, pos,
    node_size=node_sizes,
    node_color="black",
    alpha=0.85
)
plt.title("Struttura core-periphery della rete competitiva")
plt.axis("off")
plt.tight_layout()
```

```
plt.savefig("core_periphery.png", dpi=300, bbox_inches="tight")
plt.show()
```

B.5 Visualizzazione della aree competitive

Ottenimento delle statistiche e visualizzazioni per la sezione 6.4.2.

```
import folium
import colorsys
#Confini Comunali Regione Piemonte
shp_path = "/Users/apple/Documents/Tesi/Com01012024_g/Com01012024_g_WGS84.shp"
comuni_gdf = gpd.read_file(shp_path)
if comuni_gdf.crs is None:
    comuni_gdf.set_crs(epsg=32632, inplace=True)
comuni_gdf = comuni_gdf.to_crs(epsg=4326)
comuni_piemonte = comuni_gdf[
    comuni_gdf["COD_REG"] == 1
].copy()
regione_boundary = gpd.GeoDataFrame(
    geometry=[comuni_piemonte.unary_union],
    crs="EPSG:4326"
)
#Confini amministrativi Comune Torino
file_path_quartieri = "/Users/apple/Documents/Tesi/zone_statistiche.csv"
quartieri_df = pd.read_csv(
    file_path_quartieri,
    encoding="latin-1",
    sep=";"
)
quartieri_df["geometry"] = gpd.GeoSeries.from_wkt(
    quartieri_df["WKT_GEOM"]
)
quartieri_gdf = gpd.GeoDataFrame(
    quartieri_df,
    geometry="geometry",
    crs="EPSG:3003"
```

```
).to_crs(epsg=4326)
#Aree di competizione
valid_df = aggregated_df[
    aggregated_df["competition_area"] != -1
].copy()
areas = sorted(valid_df["competition_area"].unique())
def generate_large_palette(n):
    colors = []
    hue_steps = 120
    saturations = [0.85, 0.65]
    values = [0.9, 0.7]
    for v in values:
        for s in saturations:
            for h in np.linspace(0, 1, hue_steps, endpoint=False):
                r, g, b = colorsys.hsv_to_rgb(h, s, v)
        colors.append(mcolors.to_hex((r, g, b)))
    return colors[:n]
distinct_colors = generate_large_palette(n_areas)
area_color = {
    area: distinct_colors[i]
    for i, area in enumerate(areas)
}
#Mappa
m = folium.Map(
    location=[45.07, 7.68],
    tiles="CartoDB positron",
    control_scale=True
)
for row in valid_df.itertuples(index=False):
    folium.CircleMarker(
        location=[getattr(row, COL_LAT), getattr(row, COL_LON)],
        radius=4,
    color=area_color[row.competition_area],
        fill=True, fill_color=area_color[row.competition_area],
        fill_opacity=0.9,
        weight=1,
        tooltip=f"Area: {row.competition_area}"
```

```
    ).add_to(m)
folium.GeoJson(
    regione_boundary,
    name="Confine Regione Piemonte",
    style_function=lambda feature: {
        "fill": False,
        "color": "blue",
        "weight": 3
    }
).add_to(m)
folium.GeoJson(
    comuni_piemonte,
    name="Comuni Piemonte",
    style_function=lambda feature: {
        "fill": False,
        "color": "gray",
        "weight": 1
    }
).add_to(m)
folium.GeoJson(
    quartieri_gdf,
    name="Zone statistiche Torino",
    style_function=lambda feature: {
        "fill": False,
        "color": "black",
        "weight": 1
    }
).add_to(m)
minx, miny, maxx, maxy = regione_boundary.total_bounds
buffer = 0.1
m.fit_bounds([
    [miny - buffer, minx - buffer],
    [maxy + buffer, maxx + buffer]
])
m
```

Bibliografia

Abella, Andrea et al. (2025). «Exploring the Spatial Segmentation of Housing Markets from Online Listings». In: *EPJ Data Science* 14.1, p. 55. DOI: 10.1140/epjds/s13688-025-00551-z.

Alam, Shafiq et al. (2023). «An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity». In: *Decision Analytics Journal* 9, p. 100341. DOI: 10.1016/j.dajour.2023.100341.

Allam, Ishraga (gen. 2026). *THE SIGN TEST IN STATISTICS: THEORY AND A C-BASED COMPUTATIONAL IMPLEMENTATION*. DOI: 10.13140/RG.2.2.33456.01284.

Altalhan, Manahel, Abdulmohsen Algarni e Monia Turki-Hadj Alouane (2025). «Imbalanced Data Problem in Machine Learning: A Review». In: *IEEE Access* 13, pp. 13686–13699. DOI: 10.1109/ACCESS.2025.3531662. URL: https://ieeexplore.ieee.org/abstract/document/10845793.

Anenberg, Elliot e Daniel Ringo (lug. 2022). «The Propagation of Demand Shocks through Housing Markets». In: *American Economic Journal: Macroeconomics* 14, pp. 481–507. DOI: 10.1257/mac.20200037.

Anselin, Luc (2008). «Spatial Econometrics». In: *The New Palgrave Dictionary of Economics*.

Arcidiacono, Peter e Robert A. Miller (2011). «Conditional Choice Probability Estimation of Dynamic Discrete Choice Models». In: *Annual Review of Economics* 3, pp. 385–407. DOI: https://doi.org/10.3982/ECTA7743.

Astivia, Oscar L. Olvera e Bruno D. Zumbo (2019). «Heteroskedasticity in Multiple Regression Analysis: What it is, How to Detect it and How to Solve it with Applications in R and SPSS». In: *Practical Assessment, Research & Evaluation* 24.1. A peer-reviewed electronic journal. University of Massachusetts Amherst, pp. 1–19. ISSN: 1531-7714. DOI: 10.7275/q5xr-fr95. URL: https://openpublishing.library.umass.edu/pare/article/id/1590/.

Barde, Sylvain, Rowan Cherodian e Guy Tchuente (2023). «Moran's I Lasso for Models with Spatially Correlated Data». In: *arXiv preprint arXiv:2310.02773*. DOI: 10.48550/arXiv.2310.02773.

Belleflamme, Paul e Martin Peitz (2015). *Industrial Organization: Markets and Strategies*. Cambridge: Cambridge University Press. ISBN: 9781107032010. URL: https://www.researchgate.net/publication/299400200_Industrial_Organization_Markets_and_Strategies_2nd_Edition.

Belloni, Alexandre, Victor Chernozhukov e Christian Hansen (2014). «Inference on Treatment Effects after Selection among High-Dimensional Controls». In: *Review of Economic Studies* 81.2, pp. 608–650. DOI: 10.1093/restud/rdt044.

Ben-Akiva, Moshe e Steven R. Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press. DOI: https://doi.org/10.1057/jors.1987.63.

Berry, Steven e Philip Haile (gen. 2021). «Foundations of demand estimation». In: *Handbook of Industrial Organization* 4, pp. 1–62. DOI: 10.1016/bs.hesind.2021.11.001.

Bivand, Roger S., Edzer Pebesma e Virgilio Gómez-Rubio (2013). *Applied Spatial Data Analysis with R*. New York: Springer. ISBN: 9781461476184. DOI: https://doi.org/10.1007/978-1-4614-7618-4.

Blondel, Vincent D. et al. (2008). «Fast unfolding of communities in large networks». In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. DOI: https://doi.org/10.48550/arXiv.0803.0476.

Borrohou, Sanae, Rachida Fissoune e Hassan Badir (2025). «Critical Role of Data Transformation in Preprocessing: Methods, Algorithms, and Challenges». In: *Model and Data Engineering (MEDI 2024)*. Springer, pp. 108–122. DOI: https://doi.org/10.1007/978-3-031-87719-3_9.

Boyd, Stephen e Lieven Vandenberghe (2004). *Convex Optimization*. Cambridge: Cambridge University Press. ISBN: 9780521833783. URL: https://www3.diism.unisi.it/~control/seminars/boyd/book/bv_cvxbook_draft.pdf.

Brenning, Alexander (2012). «Spatial cross-validation and bootstrap for the assessment of prediction rules in spatial data». In: *International Journal of Geographical Information Science* 26.2, pp. 255–272. URL: https://ieeexplore.ieee.org/document/6352393.

Brijith, Arya (ott. 2023). «Data Preprocessing for Machine Learning». In: p. 2023.

Brotherton, Bob (1999). «Towards a definitive view of the nature of hospitality and hospitality management». In: *International Journal of Contemporary Hospitality Management* 11.4, pp. 165–173.

Brunsdon, Chris, A. Stewart Fotheringham e Martin Charlton (1996). «Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity». In: *Geographical Analysis* 28.4, pp. 281–298. DOI: 10.1111/j.1538-4632.1996.tb00936.x.

Celon, Augusto (2022). «Tecniche per la riduzione della dimensionalità nei big data: un confronto tra diversi approcci». Corso di Laurea Triennale in Statistica per le Tecnologie e le Scienze. Bachelor's thesis. Padova: Università degli Studi di Padova.

Chin, T. L. e K. W. Chau (2003). «A critical review of literature on the hedonic price model». In: *International Journal for Housing and Its Applications* 27.2, pp. 145–165.

Coppola, Gianluigi (2005). «Sistemi Locali del Lavoro e Sistemi Territoriali di Sviluppo». In: *Economia e Territorio in Italia*. FrancoAngeli.

Corriere della Sera Torino (2025). *Affitti brevi, storica apertura di Federalberghi Torino agli host di Airbnb e Booking*. URL: https://torino.corriere.it/notizie/cronaca/25_gennaio_22/affitti-brevi-storica-apertura-di-federalberghi-torino-agli-host-di-airbnb-booking-e-c-supportare-chi-ospita-nella-legalita-7339e02d-bb63-4220-b208-7f2bf3e1bxlk.shtml.

Dabrowski, Christopher (2015). «Catastrophic event phenomena in communication networks: A survey». In: *Computer Science Review* 18, pp. 1–25. DOI: 10.1016/j.cosrev.2015.10.001.

Deakin, Rod, S. Bird e R. Grenfell (dic. 2002). «The Centroid? Where would you like it to be be?» In: *Cartography* 31, pp. 153–167. DOI: 10.1080/00690805.2002.9714213.

Denizci Guillet, Basak et al. (2026). «Attribute-based pricing in hotels: Analyzing industry, customer, and behavioral insights». In: *International Journal of Hospitality Management* 133, p. 104445. DOI: 10.1016/j.ijhm.2025.104445.

Dubé, Jean-Pierre, Günter J. Hitsch e Peter E. Rossi (2010). «State Dependence and Alternative Explanations for Consumer Inertia». In: *Journal of Marketing Research* 47.3, pp. 417–432. DOI: https://doi.org/10.1111/j.1756-2171.2010.00106.x.

Efron, Bradley et al. (2004). «Least Angle Regression». In: *Annals of Statistics* 32.2, pp. 407–499. DOI: 10.1214/009053604000000067.

Einav, Liran, Chiara Farronato e Jonathan Levin (2013). «Sales Mechanisms in Online Markets». In: *National Bureau of Economic Research* 104.2, pp. 472–476. URL: https://www.nber.org/system/files/working_papers/w19021/w19021.pdf.

Farronato, Chiara e Andrey Fradkin (2022). «The Welfare Effects of Peer Entry: The Case of Airbnb and the Accommodation Industry». In: *American Economic Review* 112.12, pp. 3950–3988. DOI: 10.1257/aer.20180260.

Feng, Yunjie et al. (2025). «Spatial Reconfiguration of Housing Price Patterns and Submarkets in Shanghai Before and After COVID-19». In: *Land* 14.10, p. 2008. DOI: 10.3390/land14102008.

Fortunato, Santo (2010). «Community detection in graphs». In: *Physics Reports* 486.3–5, pp. 75–174. DOI: 10.1016/j.physrep.2009.11.002.

Fotheringham, A. Stewart, Wenbiao Yang e Wei Kang (2023). «Multiscale Geographically Weighted Regression (MGWR)». In: *Annals of the American Association of Geographers*.

Fujishige, Satoru (2005). «Submodular Analysis». In: *Submodular Functions and Optimization*. Vol. 58. Annals of Discrete Mathematics. Elsevier, pp. 199–251. DOI: 10.1016/S0167-5060(05)80006-8.

Geisser, Seymour (1975). «The Predictive Sample Reuse Method with Applications». In: *Journal of the American Statistical Association* 70.350, pp. 320–328. DOI: 10.1080/01621459.1975.10479865.

Gentzkow, Matthew, Jesse M. Shapiro e Matt Taddy (2019). «Measuring Group Differences in High-Dimensional Choices». In: *Econometrica* 87.3, pp. 859–898. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16566.

Getis, Arthur (2008). «A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective». In: *Geographical Analysis* 40.3, pp. 297–309. DOI: 10.1111/j.1538-4632.2008.00727.x.

Ghojogh, Benyamin e Mark Crowley (2023). *The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial*. arXiv: 1905.12787 [stat.ML]. URL: https://arxiv.org/abs/1905.12787.

Gibbs, Christopher et al. (2018). «Pricing in the Sharing Economy: A Hedonic Pricing Model Applied to Airbnb Listings». In: *Journal of Travel & Tourism Marketing* 35.1, pp. 46–56. DOI: 10.1080/10548408.2017.1308292.

Goodman, Allen C. e Thomas G. Thibodeau (1998). «Housing Market Segmentation». In: *Journal of Housing Economics* 7.2. Defines housing submarkets based on substitutability of dwellings, pp. 121–143. DOI: 10.1006/jhec.1998.0229.

Guo, Manping et al. (21 set. 2023). «Normal Workflow and Key Strategies for Data Cleaning Toward Real-World Data: Viewpoint». In: *Interactive Journal of Medical Research* 12. A cura di Amaryllis Mavragani. Reviewed by Rohan Alexander and Hyo Jung Kim, e44310. DOI: 10.2196/44310. URL: https://doi.org/10.2196/44310.

Gutiérrez, Javier et al. (2017). «The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona». In: *Journal of Travel Research* 56.5, pp. 612–631. DOI: 10.1177/0047287517700310.

Guttentag, Daniel (2015). «Airbnb: Disruptive Innovation and the Rise of an Informal Tourism Accommodation Sector». In: *Current Issues in Tourism* 18.12, pp. 1192–1217. DOI: 10.1080/13683500.2013.827159.

Hall, C. Michael et al. (2022). «Airbnb and the sharing economy». In: *Current Issues in Tourism* 25.19, pp. 3057–3067. DOI: 10.1080/13683500.2022.2122418. URL: https://doi.org/10.1080/13683500.2022.2122418.

Hamari, Juho, Mimmi Sjöklint e Antti Ukkonen (2016). «The sharing economy: Why people participate in collaborative consumption». In: *Journal of the Association for Information Science and Technology* 67.9, pp. 2047–2059.

Hancock, Jeffrey T. e Taghi M. Khoshgoftaar (2020). «Survey on categorical data for neural networks». In: *Journal of Big Data* 7.1, p. 28. DOI: 10.1186/s40537-020-00305-w. URL: https://doi.org/10.1186/s40537-020-00305-w.

Hastie, Trevor, Robert Tibshirani e Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer. DOI: 10.1007/978-0-387-84858-7.

Hastie, Trevor, Robert Tibshirani e Jerome Friedman (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton: CRC Press. ISBN: 9781498712163.

Hoerl, Arthur E. e Robert W. Kennard (1970). «Ridge Regression: Biased Estimation for Nonorthogonal Problems». In: *Technometrics* 12.1, pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

Huang, Jian, Shuangge Ma e Cun-Hui Zhang (2006). «Adaptive Lasso for Sparse High-Dimensional Regression Models». In: *Statistica Sinica* 18, pp. 1603–1618. URL: https://www.researchgate.net/publication/228385136_Adaptive_LASSO_for_sparse_high-dimensional_regression.

Inoue, Ryo et al. (2018). «Identification of Geographical Segmentation of the Rental Apartment Market in the Tokyo Metropolitan Area». In: *International Conference on Geographic Information Science (GIScience)*. DOI: 10.4230/LIPIcs.GIScience.2018.32.

INSEE e IRES Piemonte (2002). *Partizioni Territoriali e Sistemi Locali del Lavoro*. Istituto Nazionale di Statistica e Istituto di Ricerche Economico Sociali del Piemonte.

Kohavi, Ron (1995). «A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection». In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1137–1143.

Kostanek, Joanna et al. (2024). «Bootstrap Method as a Tool for Analyzing Data with Atypical Distributions Deviating from Parametric Assumptions: Critique and Effectiveness Evaluation». In: *Data* 9.8, p. 95. DOI: 10.3390/data9080095. URL: https://doi.org/10.3390/data9080095.

Kotsiantis, Sotiris, Dimitris Kanellopoulos e P. E. Pintelas (gen. 2006). «Data Preprocessing for Supervised Learning». In: *International Journal of Computer Science* 1.2. University of Patras.

Kuminoff, Nicolai V., Christopher F. Parmeter e Jaren C. Pope (2010). «Which Hedonic Models Can We Trust to Recover the Marginal Willingness to Pay for Environmental Amenities?» In: *Journal of Environmental Economics and Management* 60.3, pp. 145–160. DOI: 10.1016/j.jeem.2010.06.001.

La Stampa (2024). *Boom degli affitti brevi: a Torino un turista su tre rinuncia all'hotel*. URL: https://www.lastampa.it/torino/2024/02/21/news/aibnb_boom_affitti_brevi_un_turista_su_tre_rinuncia_hotel-14087200/.

Lashley, Conrad (2015). *Hospitality: A Social Lens*. Oxford: Elsevier. ISBN: 9780081007227.

LeSage, James P. e R. Kelley Pace (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: CRC Press.

Li, Jun, Serguei Netessine e Sergei Koulayev (2018). «Price to Compete... with Many: How to Identify Price Competition in High-Dimensional Space». In: *Ross School of Business, University of Michigan; The Wharton School, University of Pennsylvania; Consumer Financial Protection Bureau*. Contact: junwli@umich.edu. URL: https://orcid.org/0000-0002-9237-9147.

Little, Roderick J. A. e Donald B. Rubin (2002). *Statistical Analysis with Missing Data*. 2ª ed. Hoboken, NJ: Wiley-Interscience. ISBN: 978-0-471-18386-0.

Lorde, T., J. Jacob e Q. Weekes (2019). «Price-setting behavior in a tourism sharing economy accommodation market: A hedonic price analysis of AirBnB hosts in the Caribbean». In: *Tourism Management Perspectives* 30, pp. 251–261. DOI: 10.1016/j.tmp.2019.03.006.

Macdonald, Blair (mag. 2017). *Demonstrating Lorenz Curve Distribution and Increasing Gini Coefficient with the Iterating (Koch Snowflake) Fractal Attractor.*

MacKinnon, James G. e Halbert White (1985). «Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties». In: *Journal of Econometrics* 29.3, pp. 305–325. DOI: 10.1016/0304-4076(85)90158-7.

Maharana, Kiran, Surajit Mondal e Bhushankumar Nemade (2022). «A review: Data pre-processing and data augmentation techniques». In: *Global Transitions Proceedings* 3, pp. 91–99. DOI: 10.1016/j.gltp.2022.04.020.

Malpezzi, Stephen (2002). «The Submarket Structure of Housing Markets: Theory and Evidence». In: *Housing Studies* 17.1, pp. 69–88.

Mazzeo, Michael J. (2002). «Product Choice and Oligopoly Market Structure». In: *Journal of Economics & Management Strategy* 11.2, pp. 221–257. URL: https://www.kellogg.northwestern.edu/faculty/mazzeo/htm/rje.sum02.mazzeo.proof.pdf.

Mazziotta, Matteo e Adriano Pareto (nov. 2020). *Gli indici sintetici*. ISBN: 9788892136090.

McFadden, Daniel (1974). «Conditional Logit Analysis of Qualitative Choice Behavior». In: *Frontiers in Econometrics*. A cura di Paul Zarembka. Introduces the Random Utility Model (RUM) and the conditional logit framework. New York: Academic Press, pp. 105–142.

Mentelocale Torino (2025). *Torino aumenta la tassa di soggiorno nel 2025: nuove tariffe per hotel e affitti brevi, esenzioni e proteste del settore*. URL: https://www.mentelocale.it/torino/102579-torino-aumenta-la-tassa-di-soggiorno-nel-2025-nuove-tariffe-per-hotel-e-affitti-brevi-esenzioni-e-proteste-del-settore.htm.

Mukhopadhyay, Boidurjo Rick e Dr B.K.Mukhopadhyay (apr. 2021). *'What's Mine is Yours': The Dawn of Collaborative Consumption*. URL: https://www.researchgate.net/publication/351122120_'What's_Mine_is_Yours'_The_Dawn_of_Collaborative_Consumption.

Newman, Mark E. J. e Michelle Girvan (2004). «Finding and evaluating community structure in networks». In: *Physical Review E* 69.2, p. 026113.

Ohishi, Shintaro, Tomohiro Ando e Yoshihiro Konno (2024). «Generalized Fused Lasso for Grouped Data in Generalized Linear Models». In: *Statistics and Computing* 34.6, p. 150. DOI: 10.1007/s11222-024-10433-5.

Pakes, Ariel et al. (2021). «Moment inequalities and their application». In: *Journal of Econometrics* 222.1, pp. 48–69. URL: https://doi.org/10.3982/ECTA6865Digital%20Object%20Identifier%20(DOI).

Park, Chansoo, Young-Rae Kim e William Frye (gen. 2022). «Keeping the competition close: The impact of competitor distance in the lodging industry». In: *International Journal of Tourism Research* 24. DOI: 10.1002/jtr.2510.

Patro, S Gopal e Dr-Kishore Kumar Sahu (mar. 2015). «Normalization: A Preprocessing Stage». In: *IARJSET*. DOI: 10.17148/IARJSET.2015.2305.

Pellegrino, Simone (2020). *The Gini Coefficient: Its Origins*. Working Paper 070. Department of Economics, Social Studies, Applied Mathematics e Statistics, University of Torino.

Pillai, N. Vijayamohanan e R. Riju Mohan (mar. 2024). *Perfect Multicollinearity and Dummy Variable Trap: Explaining the Unexplained*. MPRA Paper 120376. Posted 20

Mar 2024 07:44 UTC. Author affiliation: Gulati Institute of Finance and Taxation, Trivandrum, Kerala, India. Munich Personal RePEc Archive (MPRA). URL: https://mpra.ub.uni-muenchen.de/120376/.

Poslavskaya, Ekaterina e Alexey Korolev (2023). *Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding?* arXiv: 2312.16930 [cs.LG]. URL: https://arxiv.org/abs/2312.16930.

Regione Piemonte (2025). *Nel 2024 un turismo record*. URL: https://www.regione.piemonte.it/web/pinforma/notizie/nel-2024-un-turismo-record.

Rey, Sergio J. e Richard J. Smith (2013). «A Spatial Decomposition of the Gini Coefficient». In: *Letters in Spatial and Resource Sciences* 6.2, pp. 55–70.

Ridzuan, Fakhitah e Wan Mohd Nazmee Wan Zainon (2019). «A Review on Data Cleansing Methods for Big Data». In: *Procedia Computer Science* 161. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia, pp. 731–738. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2019.11.177. URL: https://www.sciencedirect.com/science/article/pii/S1877050919318885.

Sakai, Toshiki, Jun Tsuchida e Hiroshi Yadohisa (2024). «Bayesian Geographically Weighted Regression Using Fused Lasso Prior». In: *arXiv preprint arXiv:2402.18186*. DOI: 10.48550/arXiv.2402.18186.

Shishebor, Zohreh, Zahra Sajjadnia e Maryam Sharafi (2025). «Violin Plots: An Enhanced Tool for Data Visualization in Health Studies». In: *Journal of Social Behavior and Community Health*.

Smith, Michael J. de, Michael F. Goodchild e Paul A. Longley (2018). *Geospatial Analysis: A Comprehensive Guide*. 6ª ed. The Winchelsea Press.

Stone, Mervyn (1974). «Cross-Validatory Choice and Assessment of Statistical Predictions». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 111–147. URL: https://www.jstor.org/stable/2984809.

Takada, Masaaki e Hironori Fujisawa (2024). *Adaptive Lasso, Transfer Lasso, and Beyond: An Asymptotic Perspective*. arXiv: 2308.15838 [stat.ML]. URL: https://arxiv.org/abs/2308.15838.

Terza, Joseph V., Anirban Basu e Paul J. Rathouz (2008). «Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling». In: *Journal of Health Economics* 27.3, pp. 531–543. DOI: 10.1016/j.jhealeco.2007.09.009.

Tibshirani, Robert (1996). «Regression Shrinkage and Selection via the Lasso». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Train, Kenneth E. (2009). *Discrete Choice Methods with Simulation*. 2nd. Modern textbook on discrete choice modelling and simulation methods. New York: Cambridge University Press.

Wang, Dan e Juan L. Nicolau (2017). «Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb». In: *International Journal of Hospitality Management* 62, pp. 120–131. DOI: 10.1016/j.ijhm.2016.12.007.

Watkins, Craig (2001). «The Definition and Identification of Housing Submarkets». In: *Environment and Planning A* 33.12, pp. 2235–2253. DOI: 10.1068/a34162.

Williamson, Oliver E. (1985). *The Economic Institutions of Capitalism*. New York: Free Press. ISBN: 9780029348208.

Wu, Jian e Subhash Sharma (2012). «Housing Submarket Classification: The Role of Spatial Contiguity». In: *Journal of Real Estate Research* 34.3, pp. 243–272.

Żak, Mariusz e Michał Woźniak (mag. 2020). «Performance Analysis of Binarization Strategies for Multi-class Imbalanced Data Classification». In: *Computational Science – ICCS 2020*. Vol. 12140. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 141–155. DOI: 10.1007/978-3-030-50423-6_11.

Zervas, Georgios, Davide Proserpio e John W. Byers (2017). «The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry». In: *Journal of Marketing Research* 54.5, pp. 687–705. DOI: 10.1509/jmr.15.0204.

Zhang, Huihui, Florian J. Zach e Zheng Xiang (2024a). «Multi-level differentiation of short-term rental properties: A deep learning-based analysis of aesthetic design». In: *Tourism Management* 100, p. 104832. DOI: 10.1016/j.tourman.2023.104832.

Zhang, Huihui, Florian J. Zach e Zheng Xiang (2024b). «Optimal distinctiveness of short-term rental property design». In: *International Journal of Hospitality Management* 117, p. 103737. DOI: 10.1016/j.ijhm.2024.103737.

Zhao, Yuwei et al. (mag. 2018). «Improving Generalization Based on l1-Norm Regularization for EEG-Based Motor Imagery Classification». In: *Frontiers in Neuroscience* 12, p. 272. DOI: 10.3389/fnins.2018.00272.

Zheng, Alice e Amanda Casari (2018). «Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists». In: *O'Reilly Media*.

Zhou, Youran, Sunil Aryal e Mohamed Reda Bouadjenek (apr. 2024). «A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms». In: *arXiv:2404.04905v1 [stat.ME]*. License: CC BY 4.0. URL: https://arxiv.org/abs/2404.04905v1.

Zou, Hui e Trevor Hastie (2005). «Regularization and Variable Selection via the Elastic Net». In: *Journal of the Royal Statistical Society: Series B* 67.2, pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.