Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche

Le regole dell'affitto perfetto: uno studio sui fattori che vanno ad influenzare i livelli dei prezzi degli affitti su Airbnb.

Relatore: Dott.ssa Mariangela Guidolin
Correlatore: Dott. Mauro Bernardi
Dipartimento di Scienze Statistiche

Laureando: Chiara Belli
Matricola: 1132486

Anno Accademico 2018/2019

Indice

Introduzione

Airbnb è un'azienda innovativa considerata da tutti come il più lampante e rappresentativo esempio di "sharing economy", in quanto consente la condivisione della propria casa, affittandola ai viaggiatori e ricavando in questo modo un reddito extra. La condivisione della casa in passato si svolgeva attraverso canali comunicazione "analogici" come passaparola, volantinaggio e annunci sui giornali e le offerte erano tipicamente rivolte al mercato del vicinato (Lehr, 2015). Airbnb ha creato un mercato globale innovativo per affittare proprietà attraverso una piattaforma online (Lehr, 2015). La piattaforma Airbnb ha globalizzato e trasformato il mercato degli affitti a breve termine, creando centinaia di migliaia di aziende di micro-ospitalità a livello globale. Chiunque può pubblicizzare e affittare la propria casa su scala mondiale, quindi secondo la filosofia propria di Airbnb qualsiasi individuo può diventare un imprenditore del settore dell'ospitalità in pochi minuti (Airbnb.com). La crescita esplosiva di Airbnb sta cambiando il concetto moderno di alloggio (Lehr, 2015), cosa che sta implicando anche problemi e trasformazioni di carattere socio-economico e legislativo mai affrontati in precedenza e di cui si farà menzione anche nelle prossime sezioni di questo capitolo. Per questa ragione il "fenomeno Airbnb" sta diventando sempre più oggetto di studio da vari punti di vista. In questo contesto deve essere collocata anche l'analisi presentata in questa tesi che ha lo scopo di comprendere quali possono essere i fattori esogeni che vanno ad influenzare i livelli degli affitti di Airbnb. In particolare, l'obiettivo primario che qui si vuole raggiungere è quello di fornire una prima risposta a due quesiti di tipo aziendale tra loro interconnessi: in primo luogo, l'individuazione dei fattori che vanno ad influenzare i livelli degli affitti di Airbnb e in seguito l'utilizzo delle informazioni ottenute nella prima fase per la previsione degli affitti. I risultati di questa analisi possono essere d'interesse per Airbnb sia per poter consigliare agli host-coloro che affittano le case o gli appartamenti-un prezzo che si avvicini il più possibile al reale valore

delle proprietà, sia per permettere all'azienda stessa di avere proiezioni più precise degli utili futuri. Si tenga presente che Airbnb come pagamento per il servizio che svolge prende il 3% dell'importo che l'host incasserà da ogni prenotazione e una commissione che va dallo 0 al 20% per sulle transazioni tra affittuari e host.

A tale scopo l'analisi condotta si concentra sul caso della città di New York ed è organizzata nel seguente modo:

- Reperimento delle informazioni sugli affitti di Airbnb. A tale scopo è stato necessario visitare il sito www.insideAirbnb.com, che contiene tutti i dati relativi alle proprietà affittate su Airbnb aggiornati ogni mese e divisi per città. In questo caso, come si vedrà in modo più dettagliato nel capitolo 3, sono quindi presi in considerazione i dati (listings) relativi alla città di New York.

- Reperimento di tutti i dati relativi a fattori socio-economici che si suppone possano andare a influenzare i livelli degli affitti. Per farlo si è visitato il sito www.opendata.cityofnewyork.us che contiene oltre 2000 dataset relativi a dati riguardanti svariati aspetti della vita cittadina di New York. Come verrà spiegato in modo più approfondito nel capitolo 3, tra questi sono stati selezionati tutti i dataset che contengono informazioni relative a fattori socio-economici e culturali che si suppone possano essere legate ai livelli degli affitti delle proprietà.

- Specificazione e stima di modelli interpretabili e adatti ai dati che sono a disposizione e alla tipologia della variabile risposta. Generalmente per tipologie di dati come i prezzi che non hanno una distribuzione normale, ma una distribuzione con code più pesanti, il modello lineare, comunemente impiegato in molti ambiti, può non essere una scelta adeguata perché il comportamento medio della variabile risposta fornisce poca o nessuna informazione sul comportamento delle code della distribuzione condizionata della variabile d'interesse. In particolare come verrà illustrato nel capitolo 4 in questa tesi si utilizzerà la regressione quantilica, ovvero un approccio regressivo che non utilizza come risposta la media condizionata della distribuzione della variabile risposta, ma i quantili condizionati.

- Utilizzo dei risultati ottenuti per implementare altri modelli con l'obiettivo di fare previsioni sul livello dell'affitto di una proprietà, date una serie di informazioni ad essa relative come per esempio il quartiere in cui si trova, il numero di stanze, etc.

Tutti questi passi verranno approfonditi nelle prossime sezioni e nei capitoli seguenti. In particolare nel capitolo 1, verrà illustrato il modello di business di Airbnb, ovvero le modalità di creazione del valore tipiche di questo business, evidenziandone sia aspetti positivi che criticità. Nel capitolo 2 si approfondiranno i criteri con cui sono state scelte le variabili da utilizzare, la costruzione del dataset finale e le analisi preliminari svolte . Nel capitolo 3 verranno descritti i modelli utilizzati, il loro funzionamento e l'interpretazione dei risultati ottenuti per quello che riguarda i fattori che influenzano il livello degli affitti di Airbnb. Nel capitolo 4 vengono illustrate alcune considerazioni conclusive unitamente a un confronto con alcuni modelli solitamente utilizzati nell'ambito dell'analisi dei dati, per approfondimenti si veda Azzalini e Scarpa (2012).

Capitolo 1

Airbnb

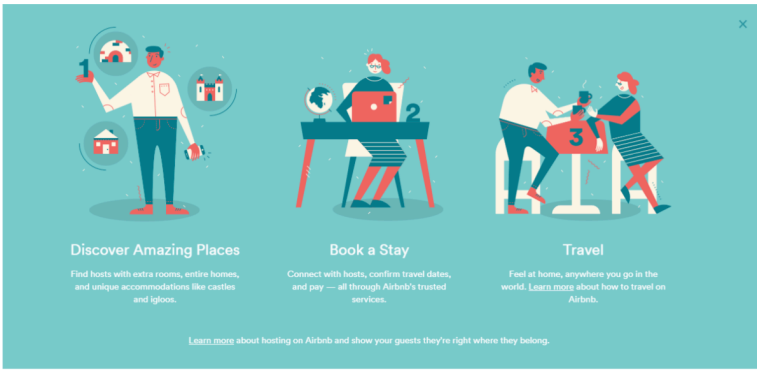
1.1 Airbnb come modello di business

Airbnb è un sito commerciale contenente annunci su case in affitto per brevi periodi creato da Brian Chesky, Joe Gebbia e Nathan Blecharczyk nel 2008. In particolare, il sito mette in contatto persone in cerca di contratti di affitto per periodi brevi con persone che hanno a disposizione uno spazio extra da affittare. Le proprietà affittabili possono essere di diverse tipologie: non soltanto stanze private e appartamenti, ma anche castelli, ville, barche, baite, case sugli alberi, igloo e isole private. Se si è interessati a prendere in affitto una proprietà bisogna compiere le seguenti azioni:

- Registrazione: ogni profilo dovrebbe includere le foto e le verifiche, soprattutto perché alcuni host lo richiedono per poter effettuare la prenotazione.

- Ricerca: in questa fase si inseriscono le date di arrivo, partenza e il numero complessivo di ospiti in modo da poter visualizzare il prezzo esatto. Per ottenere tutte le informazioni su una proprietà basta leggere le recensioni, le descrizioni, le regole della casa e quali servizi sono presenti. Per eventuali domande si può sempre contattare l'host.

- Prenotazione: in questo momento si devono seguire le istruzioni dell'host per ultimare la prenotazione. Esistono due tipi di prenotazione: la prenotazione immediata, cioè la possibilità di prenotare senza dover aspettare una conferma e la richiesta di pre-



notazione, cioè il caso in cui le prenotazioni devono essere approvate dall'host prima di diventare definitive.

Se invece si è interessati ad affittare la proprietà , dopo essersi registrati, è necessario stabilire quando si è disponibili a ospitare e il preavviso necessario prima di una prenotazione. Infine si deve decidere il prezzo con l'aiuto dei suggerimenti del sito, aggiungendo eventualmente le spese di pulizia, e fissare, eventualmente, delle regole che gli ospiti devono accettare prima di poter prenotare. Come esempio di "sharing economy" Airbnb rappresenta un vero e proprio modello di business–se non addirittura il modello di business per eccellenza–in quest'ambito. I modelli di business sono definiti come "il sistema di attività che consente

all'azienda, in accordo con i suoi partner, di creare valore e anche di appropriarsi di una quota di quel valore"(Zott e Amit, 2010). Seguendo questo approccio, gli studiosi di modelli di business sono principalmente interessati ad identificare alcune delle loro caratteristiche specifiche, come per esempio le fonti di creazione di valore (Zott e Amit, 2010), cercando di catturare attributi o caratteristiche oggettive di aziende specifiche. Tuttavia, nella letteratura sui modelli di business esiste un'altra corrente che considera principalmente il loro significato cognitivo. In questa prospettiva i modelli di business sono considerati "non solo come" fenomeni reali, "ma come strumenti cognitivi che aiutano in modo importante la comprensione dei legami causali tra gli elementi tradizionali dell'impresa e quelli esterni" (Baden-Fuller e Mangematin, 2013). Tra quelli più studiati vi sono i cosiddetti "modelli di business iconici". Per Sabatier, Mangematin e Rousselle (2010) un "modello di business iconico" è quello che è designato principalmente dal nome dell'azienda che lo ha introdotto o reso famoso, piuttosto che dalle sue caratteristiche di carattere economico. Infatti, come un "marchio iconico" (ad esempio, Coca Cola) è molto più di un semplice prodotto o servizio (Holt, 2006), così un "modello di business iconico" rappresenta molto più del mero modello di business della specifica azienda da cui prende il nome dato che si tratta di un modello di business innovativo, che stabilisce nuovi standard in termini di creazione e acquisizione di valore, oltre a diventare una fonte di ispirazione per altri attori del mercato. Ricapitolando si definisce "iconico", un modello di business innovativo che è imitato in tutti i settori ed è considerato come un prototipo esemplare per una particolare categoria di aziende.

L'innovativo modello di business di Airbnb è stato imitato in molti settori, come la cura degli animali domestici e il settore del noleggio auto, ed è considerato come l'esemplare prototipo–o almeno il migliore esempio–dei modelli di business impiegati dalla categoria emergente di aziende peer-to-peer. Airbnb è quindi diventata l'azienda iconica nell'ambito della "sharing economy". Questa idea è sostenuta da Botsman e Rogers (2011), Botsman (2015) e Gansky (2010), che definiscono il concetto di "condivisione" come il contatto e le transazioni tra i singoli consumatori. La "sharing economy" è stata anche definita da Frenken et al. (2015) come un insieme di consumatori che si concedono reciprocamente l'accesso temporaneo a beni fisici sottoutilizzati ("capacità inutilizzata"), generalmente a pagamento. Secondo Frenken et al. (2015) ci sono tre elementi che differenziano la "sharing economy" da altre forme economiche. La "sharing economy" riguarda:

- Piattaforme consumer-to-consumer, quindi non l'affitto o il leasing di un bene da un'azienda, che sono invece transazioni di tipo business-to-consumer.

- Consumatori che forniscono l'accesso temporaneo a un bene, quindi non vi è alcun trasferimento della proprietà del bene. È quindi esclusa l'economia di seconda mano, in cui ci sono vendite di prodotti e quindi passaggi di proprietà (come accade su piattaforme online tipo ebay o Facebook).

- L'uso efficiente di risorse fisiche e non scambi di servizi tra privati.

Come sostenuto in Rifkin (2014) e Olma (2014) il termine "sharing economy" designa le piattaforme digitali in quanto queste sono in grado di connettere i clienti a qualsiasi servizio o merce. La piattaforma è un collegamento – come lo sono i mercati per domanda e offerta – tra i potenziali clienti e qualsiasi bene o servizio. Tutti possono diventare fornitori di tutti i tipi di prodotti e servizi con un semplice click. In particolare, Airbnb consente alle persone di subaffittare le proprie case, come spiega Molly Turner (2013), Global Head of Civic Partnerships di Airbnb: "Il nostro modello di business si basa su persone che hanno bisogno di soldi extra, quindi affittano le loro case". I fondatori di Airbnb hanno capito che la tecnologia della piattaforma ha reso possibile la realizzazione di un business completamente nuovo che va a sfidare l'economia tradizionale del settore alberghiero. Infatti, a differenza delle catene alberghiere convenzionali, Airbnb non possiede o amministra le proprietà, ma

gestisce semplicemente la piattaforma e prende una percentuale del canone di affitto. Per far funzionare questo modello di business, Airbnb ha dovuto affrontare tre questioni chiave:

- attirare host e ospiti
- bloccare la negoziazione diretta
- stabilire la fiducia come condizione per le transazioni.

Le transazioni dirette tra le due parti vengono prevenute in quanto ciò porterebbe a un business unilaterale (Rochet e Tirole, 2004). L'azienda fa ciò direttamente attraverso un algoritmo che blocca i messaggi contenenti numeri di telefono o indirizzi e-mail, ma offre anche servizi per facilitare transazioni come il pagamento con carta di credito, strumenti di pricing e assicurazione. Tuttavia è il potere del marketing ciò che più contraddistingue Airbnb dal tradizionale mercato delle case vacanza (Guttentag, 2015). La filosofia dello "sharing" e l'immagine di una comunità calda e autentica – trasmessa principalmente nelle testimonianze video – è stata essenziale per convincere host e ospiti a entrare a far parte della rete (Stern (2010), Yannopoulou, Moufahim e Bian (2013)). Si può quindi sostenere che sono due i fattori principali che spiegano il successo di questo modello: sia cause "idealistiche", in particolare, l'autenticità del contatto nell'esperienza di alloggio, da un lato, e i benefici economici per gli ospiti, dall'altro. Airbnb è categorizzato come esempio dello "stile di vita sharing" che combina i "benefici della proprietà con oneri e costi personali ridotti, oltre a un minor impatto ambientale" (Botsman e Rogers, 2010). Inoltre il desiderio di interazione sociale è spesso visto come il principale motore della crescita del fenomeno (Gansky (2010), Ikkala e Lampinen (2015)). Per Rothkopf (2014), Twitter e Airbnb sono modi per "connettersi con gli altri in modi sia creativi che progressivi". Il CEO di Marriott Arne Sorensen ha citato il "gusto dell'autentica vita di quartiere" come motivo del successo di Airbnb (Tuttle, 2015). Tuttavia, se si passa da approcci concettuali che cercano di spiegare l'emergere della "sharing economy" agli studi sui reali motivi dei partecipanti al fenomeno, l'"idealismo" sembra giocare un ruolo secondario. Per gli ospiti Airbnb è semplicemente un'opzione a basso costo (Guttentag (2015), Liang, Choi e Joppe (2018)). Gli host di Airbnb sono guidati in primo luogo da motivi economici (Glind (2013), Stors e Kagermeier (2015)). In ogni caso questa motivazione non elimina necessariamente i vantaggi sociali o ambientali che i consumatori possono cercare nella condivisione. Quel che è certo è che Airbnb porta vantaggi sia ai proprietari sia agli affittuari: i proprietari

possono guadagnare un reddito extra affittando le loro case o stanze inutilizzate, mentre gli affittuari possono prenotare alloggi a costi inferiori.
Tuttavia, Airbnb ha sollevato diverse preoccupazioni sia per i governi sia per l'industria turistica e dell'ospitalità in generale. In primo luogo, secondo le associazioni degli albergatori americane, si configura un fenomeno di concorrenza sleale nei confronti degli alberghi operata da coloro che affittano tramite Airbnb (spesso grosse compagnie e non privati). Infatti, lo sviluppo alberghiero comporta costi iniziali molto elevati–come investimenti in immobili di pregio– l'obbligo di pagare tasse, di adottare rigorosi standard di sicurezza, oltre ad un elevato costo del lavoro (Lehr, 2015). Per contro, gli host Airbnb possono offrire prezzi competitivi perché i costi fissi sono già coperti e non c'è alcuna spesa in termini fattore lavoro, oltre al fatto che i privati non sono tenuti a fornire gli stessi standard di sicurezza degli alberghi (Oskam e Boswijk, 2016). Secondo quanto riportato da Lehr (2015) gli albergatori sostengono che tale crescita ha avuto e avrà le seguenti ripercussioni:

- Incentivazione della conversione illegale su vasta scala di unità residenziali in alloggi turistici (Lehr, 2015). La prospettiva di guadagno porterà sia compagnie sia privati ad affittare le loro proprietà ai turisti, dato che è più redditizio che stipulare contratti di affitto a lungo termine. Questo innescherà una forte penuria di alloggi affittabili a lungo termine. Il calo dell'offerta porterà quindi forti rincari sugli affitti a lungo termine e darà il via al cosiddetto fenomeno della gentrificazione, ovvero lo spostamento dei residenti dal centro della città alla periferia, come è successo nel centro di Barcellona, dove i turisti hanno preso il posto dei residenti (Oskam e Boswijk, 2016). Questo porterebbe alla distruzione del senso di comunità presente in ogni centro urbano, infatti come ha spiegato Linda Rosenthal, membro dell'Assemblea dello Stato di New York , "I quartieri (di New York) sono cambiati nelle aree ad alta densità di OVRP(Online Vacation Rental Platform). Non solo le case sono state trasformate in hotel turistici, ma la qualità della vita è cambiata nei quartieri OVRP(Online Vacation Rental Platform) ad alta densità. I genitori non sanno più chi sta camminando nei corridoi del loro condominio, cambiando il senso di comunità e di sicurezza per i loro figli. La commercializzazione di unità residenziali degrada il senso del vicinato e il senso di comunità del residente permanente". Per evitare ciò gli affitti di unità abitative per meno di 30 giorni sono stati resi illegali nella maggior parte delle città. Inoltre, le città che hanno tentato di creare ordinanze efficaci per modernizzarsi e

ospitare Airbnb, (Portland, San Francisco, Chicago e Amsterdam) hanno scoperto che le ordinanze e le leggi approvate dalle amministrazioni locali sono inapplicabili a causa della mancanza di dati su coloro che affittano le loro proprietà a breve termine.

- Evasione delle tasse di soggiorno e altre imposte, a danno di stati e amministrazioni locali, in quanto in questo modo si perdono fondi per migliorare servizi e infrastrutture. Senza considerare che le mancate entrate degli hotel significano anche tasse non pagate per il governo (tassa di occupazione transitoria, Imposta entrate lorde) (Lehr, 2015).

- Posti di lavoro nel settore alberghiero a tempo pieno sono sostituiti da impieghi part-time o lavori a cottimo (Lehr, 2015). Airbnb probabilmente sta mettendo a repentaglio anche il mercato del lavoro del settore del turismo (Lehr, 2015), almeno secondo le associazioni degli albergatori.

In realtà, in letteratura le ricerche empiriche su questo aspetto sono ancora limitate, visto che l'impatto di Airbnb è un fenomeno relativamente recente, le cui caratteristiche sono ancora da valutare (Neeser, Peitz e Stuhler, 2015). Zervas, Proserpio e Byers (2017) stimano l'impatto di Airbnb sul settore alberghiero in Texas utilizzando dati molto dettagliati sugli alberghi e i listati di Airbnb nelle più grandi città del Texas. In questo studio si nota che, dove Airbnb è penetrato di più, l'impatto sulle entrate degli hotel più vulnerabili è di circa l'8-10% in cinque anni. Zervas, Proserpio e Byers (2017) differenziano l'impatto per diversi segmenti di mercato. In particolare, emerge che gli hotel economici competono più ferocemente con questa nuova piattaforma, mentre gli hotel che si concentrano maggiormente su viaggiatori d'affari e i clienti più facoltosi raggiungono una nicchia diversa dalla tipica clientela di Airbnb. Tuttavia, il principale limite di questa ricerca è la specificità dell'industria alberghiera del Texas, dato che non è detto che gli stessi risultati si possano ripresentare in un altro luogo. In Neeser, Peitz e Stuhler (2015) viene studiato l'impatto di Airbnb in Norvegia, Finlandia e Svezia sul settore alberghiero. Il motivo principale per cui Neeser, Peitz e Stuhler (2015) sceglie di mettere insieme questi paesi è che dovrebbero essere influenzati allo stesso modo da shock comuni. Inoltre, i dati sull'industria alberghiera sono liberamente disponibili e comparabili. Questi paesi non sono stati mai considerati in questa letteratura, probabilmente perché non cercano di impedire a questa piattaforma di entrare, a differenza di altri (Neeser, Peitz e Stuhler, 2015). Infine, i tre paesi hanno visto

una penetrazione significativa di Airbnb. Al contrario di quanto rilevato in Zervas, Proserpio e Byers (2017) non si è riscontrato un impatto significativo di Airbnb sulle entrate degli hotel per camera, ma la sua presenza ha contribuito a una riduzione del prezzo medio della camera nelle zone in cui Airbnb è penetrato di più. Neeser, Peitz e Stuhler (2015) ha anche mostrato che "l'esperienza culturale" di Airbnb è relativamente più attraente per gli stranieri. Un punto di vista differente è quello assunto in Fang, Ye e Law (2016), che prende in considerazione l'impatto di Airbnb sul mercato del lavoro. Per affrontare questo problema Fang, Ye e Law (2016) hanno raccolto i dati sugli affitti in Idaho dal sito di Airbnb. Fang, Ye e Law (2016) hanno scelto questo stato in quanto il settore del turismo supera gli altri in termini di entrate (infoplease, 2015) e continua a registrare una notevole crescita (Johnson, 2014), quindi è un ottimo contesto per indagare gli effetti della "sharing economy" sull'industria turistica locale. I risultati dello studio indicano che Airbnb avvantaggia l'intera industria del turismo generando nuove posizioni lavorative in quanto il numero dei turisti aumenterebbe a causa del minor costo dell'alloggio. Tuttavia, questo effetto positivo è attenuato dal fatto che gli hotel più economici verrebbero sostituiti da Airbnb (Zervas, Proserpio e Byers, 2017). Lo studio di Fang, Ye e Law (2016) identifica sia i vantaggi sia potenziali inconvenienti della "sharing economy". A differenza degli studi precedenti che si sono concentrati solo sul lato positivo (Cervero, Golub e Nee (2007) ; Martin, Shaheen e Lidicker (2010)) o negativo (Malhotra e Van Alstyne (2014); Zervas, Proserpio e Byers (2017)). Non è ancora chiaro l'impatto complessivo che Airbnb potrà avere sul settore turistico e più in generale sull'intera società. Anche se il suo ingresso potrebbe effettivamente avvantaggiare l'intero settore del turismo poiché i visitatori che scelgono di alloggiare nelle strutture di Airbnb trascorrono più tempo nelle mete turistiche. Di conseguenza, la dimensione del mercato del settore del turismo si espanderebbe grazie all'aumento del numero di visitatori (Fang, Ye e Law, 2016). Sulla base di queste valutazioni, se Airbnb sarà un beneficio per il settore del turismo rimane una questione aperta (Fang, Ye e Law, 2016). Un aspetto ulteriore del fenomeno Airbnb, non meno importante dei precedenti, brevemente menzionato in vari studi, ma ancora non molto sviluppato è l'impatto che Airbnb ha sui mercati immobiliari. Secondo un rapporto della città di San Francisco, Airbnb potrebbe mettere circa il 40% dei potenziali affitti fuori dal mercato (Neeser, Peitz e Stuhler, 2015). Simili preoccupazioni hanno cominciato a emergere in altre grandi città del mondo come New York, Vancouver e Berlino. Si teme che Airbnb riduca

l'offerta di alloggi a lungo termine. Sebbene alcuni studi abbiano provato a verificare empiricamente queste affermazioni, o mancano i dati o l'obiettività (Neeser, Peitz e Stuhler, 2015). In ogni caso, vista la mancanza di studi riguardanti il potenziale impatto di Airbnb sul mercato immobiliare, si può affermare che la ricerca futura dovrebbe occuparsi anche di questo ambito se si vuole comprendere pienamente l'impatto delle piattaforme peer-to-peer a livello complessivo in termini storici e sociali. Ci vorranno ancora anni di ricerche e dati per poter giungere ad un livello di comprensione completo di questo fenomeno e delle sue conseguenze.

1.2 La storia di Airbnb e del suo sviluppo vista dai mass media

La storia di Airbnb inizia nel 2007 durante il ICSID'07 (IDSA World Congress, Connecting), la conferenza annuale della Industrial Design Society of America organizzata a San Francisco. Essendo la disponibilità di camere negli hotel esaurita, Brian Chesky e Joe Gebbia hanno offerto parte del loro loft come alloggio ad alcuni viaggiatori interessati a non perdersi la conferenza. Gli ospiti hanno dormito su materassi ad aria e gli è stata offerta la colazione, facendo nascere in questo modo AirBed & Breakfast. Se si considera come i media hanno descritto negli anni Airbnb è interessante notare come la sua storia possa essere divisa in tre fasi (Mikhalkina e Cabantous, 2015). Per fare questo Mikhalkina e Cabantous (2015) ha analizzato in maniera sistematica gli articoli di sei giornali molto diffusi (Wall Street Journal, il New York Times, il Financial Times, Guardian e The Times, Washington Post). In particolare utilizzando il database "Factiva" ha cercato tutti gli articoli pubblicati in lingua inglese tra il 1 gennaio 2008 e il 31 dicembre 2013 in cui appariva il termine "Airbnb" e ha individuato in questo modo 2.458 articoli di oltre 15 fonti diverse. In seguito, per rendere il campione più gestibile, ha preso in considerazione solo le fonti multimediali che hanno stampato il maggior numero di articoli su Airbnb nel periodo di studio, ovvero le fonti citate in precedenza. Stando ai risultati di questo studio il riconoscimento di Airbnb come modello di business iconico si è svolto in tre fasi.

- La prima fase (2009-2011), che corrisponde all'avviamento effettivo della società grazie all'intervento dell'incubatore Y Combinator e ad un conseguente ampliamento dell'offerta di prodotto (da semplici spazi condivisi- adatti a viaggiatori attenti al budget- si è passati a poter affittare appartamenti, intere case e qualsiasi altro tipo

di proprietà), è quella della descrizione da parte della stampa delle analogie tra il modello di Airbnb e i modelli tipici di settori di mercato o aziende già esistenti. In questo periodo di tempo, infatti, i mezzi di comunicazione di massa cercano di comprendere il funzionamento del modello di business di Airbnb e di allocarlo nell'ambito di una categorizzazione delle imprese basata sulla tipologia di prodotto o servizio offerto. In particolare, come emerge dal campione di articoli descritto in precedenza, la stampa tenta di definire Airbnb attraverso il confronto con le aziende che operano nel settore alberghiero (hotel, ostelli,etc.), dato che queste offrono la stessa tipologia dei servizi. Tuttavia, un'analisi più approfondita degli articoli che trattano questa analogia

rivela come, in molti casi, la stampa abbia sottolineato le importanti differenze che intercorrono tra Airbnb e i fornitori di alloggi tradizionali, come appare chiaro da uno stralcio di articolo del New York Times: «Dalla sua nascita nel 2008, l'azienda (...) ha ottenuto oltre due milioni di prenotazioni in tutto il mondo, ma non è un hotel. Invece, permette alle persone di affittare la loro intera casa o appartamento - o semplicemente una stanza o un letto - ad altri che trovano noioso il Marriott o vogliono vedere com' è la vita in un'area diversa da quella in cui si vive»(New York Times, 12 novembre 2011). Airbnb viene descritta come un'azienda che affitta camere ai viaggiatori- e quindi fornisce gli stessi servizi degli hotel tradizionali - ma che, in sostanza, non è un hotel. I media inoltre mettono in evidenza anche il fatto che Airbnb differisce da un hotel tradizionale in quanto - a differenza delle principali catene alberghiere globali con cui compete - «non possiede un solo letto» (New York Times, 21 luglio 2013), e i viaggiatori affittano le stanze non da compagnie, ma da persone "reali". Dato che i tentativi della stampa di comparare Airbnb ad aziende del settore alberghiero esistente non sono soddisfacenti, i media utilizzano anche altre analogie per definire Airbnb e il suo modello di business. Due di queste riguardano aziende note per i loro modelli di business originali: eBay e Couchsurfing. Per esempio Airbnb è definito come un "mercato in stile eBay ..." (Financial Times, 30 luglio 2011) o come una "versione aggiornata di Couchsurfing" (New York Times, 17 maggio 2009). Questi elementi suggeriscono che il ragionamento per analogia è il principale strumento di elaborazione delle informazioni che i media utilizzano per spiegare il funzionamento di questa azienda innovativa, cercando tuttavia di confinarla all'interno di ambiti economici preesistenti.

- La seconda fase (2011-2012) è l'elaborazione e la legittimazione del modello di business di Airbnb. In questa fase l'attenzione dei media si sposta dalle analogie con le aziende esistenti all'elaborazione di un modello di business specifico per Airbnb, dato che questa si sta affermando come un'impresa di successo. In altri termini, i media cominciano a definire Airbnb come un'azienda con una sua identità ben precisa mettendo in risalto il suo successo in ambito finanziario. In particolare, nel febbraio 2011 le notti prenotate arrivano ad un milione e il fatturato aumenta del 65% rispetto al mese precedente. In seguito ci sono le acquisizioni di Accoleo (un'azienda tedesca dello stesso settore) e Crashpadder (un sito con funzionalità analoghe ad Airbnb)

che permettono a Airbnb di inserirsi nel mercato europeo. Inoltre, nella primavera 2012 nasce l'applicazione per Android e iPhone. In effetti, i media alla luce di questi risultati commentano ampiamente il successo e le prestazioni economiche di Airbnb, fornendo anche una descrizione sempre più ricca del suo modello di business e suggerendo quindi una maggiore comprensione delle ragioni di tale successo.

- Nella terza e ultima fase (2012-2013) quello di Airbnb diviene modello di business iconico dell'economia della condivisione. In questa fase si cominciano ad osservare anche risvolti negativi di questa attività el'attenzione si sposta anche su aspetti normativi. Di qui la decisione di limitare il raggio di azione di Airbnb a Barcellona, Vancouver, New Orleans, e più recentemente anche nella città simbolo New York, dove si crede che la crisi degli alloggi che la affligge dipenda proprio dal successo di questa azienda. Infatti, a chi possiede una casa o un appartamento conviene molto di più affittarla tramite Airbnb per una o due notti alla volta piuttosto che per un anno o più, implicando una scarsità di proprietà affittabili per lunghi periodi. Questo ragionamento ha indotto il consiglio comunale newyorchese a votare una delibera che prevede che tutte le informazioni su ogni affitto vengano trasmesse ad un ufficio pubblico per verificare l'ottemperanza ad una legge vigente, ma poco applicata, per cui è vietato affittare casa per meno di 30 giorni, a meno che il proprietario non vi abiti effettivamente. Teoricamente questo provvedimento dovrebbe liberare molte proprietà.

In questa breve trattazione si è cercato di tenere conto delle molte sfaccettature del fenomeno di Airbnb, evidenziandone sia aspetti positivi che criticità. Quel che è certo è che questo tipo di business ha prodotto un notevole impatto di tipo socioeconomico, le cui implicazioni potranno essere comprese solo nel tempo. Esula dagli obiettivi di questa tesi addentrarsi in queste tematiche, ma la comprensione degli elementi che influenzano il livello degli affitti può essere utile a far luce sulla complessità del fenomeno.
Nel far questo ci si concentra sulla città di new York per due ordini di motivazioni:

- New York è rappresentativa di molte altre realtà - in particolare la suddivisione in quartieri rende conto di una realtà che può essere molto diversificata.

- disponibilità dei dati

Capitolo 2

I prezzi di Airbnb a New York: costruzione del dataset e analisi preliminari

2.1 Obiettivi dello studio

L'obiettivo principale di questo studio è comprendere quali fattori sociali, economici o ambientali vadano ad influenzare il livello degli affitti su Airbnb nella città di New York. Questo può essere interessante per Airbnb sia per poter consigliare agli host (coloro che affittano le case o gli appartamenti) un prezzo che si avvicini il più possibile al reale valore delle proprietà, sia per permettere all'azienda stessa di avere proiezioni più precise degli utili futuri. In generale si può ipotizzare che il prezzo degli affitti delle proprietà dipenda dalla presenza di servizi nelle aree circostanti ad esse, come per esempio la presenza di impianti sportivi o di qualche fermata dei mezzi pubblici nelle vicinanze. Tenendo conto che nella maggior parte dei casi gli affittuari sono turisti, un altro fattore che farà aumentare il prezzo per soggiornare nelle proprietà potrebbe essere la vicinanza o meno a luoghi di interesse turistico come per esempio la Statua della Libertà. Oltre a fattori legati ai servizi, anche la qualità della vita nel quartiere in cui la proprietà è ubicata può incidere sull'affitto, motivo per cui nell'analisi vengono presi in considerazione anche aspetti socio-economici e ambientali. Per esempio si considerano indicatori ambientali come misure della qualità dell'aria e dell'acqua potabile; oppure sociali come il livello di istruzione nel quartiere, il livello di criminalità e infine economici come il reddito medio, misurati tutti per quartiere.

2.2 Preparazione dei dati e analisi preliminari

Procedura di costruzione del dataset

Unione dei listings

Il dataset principale è quello che contiene le informazioni sulle proprietà affittate su Airbnb. Questo dataset è stato costruito a partire da 39 listing derivati da insideairbnb.com, ovvero gli scraping [1] a cadenza mensile o bimestrale del portale web Airbnb in riferimento alla città di New York, che vanno da gennaio 2015 ad agosto 2018. Ogni listato include le informazioni relative agli annunci di affitto presenti in Airbnb fino alla data di ogni scraping del sito. I listati contengono un numero di annunci che va da un minimo di 27101 fino ad un massimo 50914. Ogni riga dei listati rappresenta un annuncio. La maggior parte delle righe, tuttavia, si ritrova nei diversi listati essendo ognuno di essi una scansione nel tempo. Si procede quindi ad un'integrazione tra loro dei vari file: un'operazione necessaria per eliminare le ripetizioni delle osservazioni e quindi l'informazione ridondante. Il modo più rapido per farlo sarebbe quello di unire i vari dataset e in seguito eliminare le osservazioni che si ripetono. In questo caso, però non si può operare nel modo descritto in precedenza, in quanto i dataset contenti i 38 scraping non presentano lo stesso numero di colonne. In particolare i listati hanno un numero di colonne che va da un minimo di 52 (nel caso del primo dataset mancano tutte le variabili con informazioni specifiche su host e guest) ad un massimo di 96 (come nel caso dei listati dal ventottesimo in poi), di conseguenza non è possibile unirli direttamente. Per rendere questa operazione più efficiente, sono state messe in atto contemporaneamente l'unione dei dataset e la prima scrematura delle variabili. In particolare, si è deciso di eliminare tutte le variabili con molti dati mancanti, come per esempio la superficie dell'appartamento espressa in metri quadri, il costo della pulizia etc. Tuttavia, anche con queste eliminazioni si è notato che molte variabili che sono presenti negli ultimi listati riguardanti soprattutto le caratteristiche degli host e dei guest, mancano completamente nei primi. Quindi, dato che la finalità dell'analisi richiede che i dataset che si vogliono unire abbiano lo stesso numero di colonne e che le colonne abbiano lo stesso nome, sono stati creati dei vettori di NA (host-is-superhost, host-response-time, host-identity-verified, neighbourhood-group-cleansed, instant-bookable, cancellation-policy, require-guest-profile-

[1] metodologie che consentono di estrarre e collezionare informazioni da differenti portali Web

picture, require-guest-phone-verification, reviews-per-month) per i primi quattro listati, da inserire al posto delle variabili che mancano rispetto ai listing fatti nei periodi successivi. In seguito tutte le colonne dei primi quattro listati vengono riordinate in modo da essere messe esattamente nello stesso ordine e rinominate come quelle dei listing successivi. Dopo aver compiuto le operazioni descritte in precedenza i 38 listati sono stati uniti in un unico dataset di 1474759 di righe. Il passo successivo dell'analisi è quello di eliminare tutta l'informazione ridondante presente all'interno del dataset, dato che le informazioni relative ad una stessa proprietà possono essere presenti in più listati e quindi ripetersi più volte all'interno del dataset. Tuttavia, prima di fare questo bisogna accertarsi che la variabile "id" identifichi un appartamento e che questo sia sempre lo stesso in ognuno dei listati in cui la proprietà è presente. In altri termini bisogna assicurarsi che ognuno degli identificativi rappresenti sempre la stessa proprietà in ogni listato. Per verificare che questa condizione sia rispettata per tutti gli identificativi, innanzi tutto si ordina il dataset per identificativo. In seguito si verifica che le informazioni relative alla collocazione delle proprietà nello spazio, ovvero latitudine e longitudine, e "host-id" ovvero l'identificativo del proprietario dell'appartamento non mutino nel tempo per ogni valore di "id". Infatti, se queste variabili riferite ad una proprietà con un certo "id" restano invariate per ogni listato significa che l'identificativo designa sempre la stessa proprietà in tutti i listati. La condizione descritta sopra viene rispettata sempre, anche nei rari casi in cui latitudine e longitudine differiscono leggermente tra un listing e il successivo. Infatti dalla verifica di tutti gli attributi (come per esempio numero di stanze e numero di bagni) delle proprietà che presentavano questo problema è emerso che la proprietà a cui l'identificativo si riferiva era sempre la stessa in ogni listato. Si può quindi procedere a estrarre dal dataset di partenza per ogni "id" la riga con la data di scraping più recente, senza considerare le precedenti. In questo modo si andrà ad utilizzare l'informazione più aggiornata, eliminando tutta l'informazione ridondante che altrimenti sarebbe presente nel dataset. Si ottiene in questo modo un dataset con 151467 record e 35 variabili esplicative. Le variabili considerate sono le seguenti:

- *id*, l'identificativo che designa ogni proprietà.
- *last-scraped*, la data in cui è stato fatto l'ultimo scraping.
- *host-id*, l'identificativo che designa ogni proprietario.

- *host-reponse-time*, è un fattore che indica il tempo che l'host impiega a rispondere alle richieste di prenotazione e ha 4 modalità.

- *host-is-superhost*, è una variabile che indica se l'host a cui si riferisce è un superhost, cioè un host esperto che deve avere valutazione media complessiva pari o superiore a 4.8, calcolata in base alle recensioni di almeno il 50% dei loro ospiti Airbnb dell'anno precedente, oltre ad aver ospitato almeno 10 soggiorni, avere un tasso di risposta pari al 90% e non avere mai cancellato una prenotazione nell'anno precedente.

- *host-identity-verified* è una variabile che indica se l'identità dell'host è stata verificata o meno, attraverso un confronto tra le informazioni contenute nelle pagine online dell'host come i profili Linkedln e Facebook e le informazioni "off-line" dell'host come per esempio i documenti di identità.

- *neighbourhood-cleansed* indica il sottoquartiere in cui si trova l'appartamento.

- *neighbourhood-group-cleansed* indica il quartiere in cui si trova l'appartamento.

- *latitude* indica la latitudine di ogni appartamento.

- *longitude* indica la longitudine di ogni appartamento.

- *property-type* è un fattore che indica la tipologia di proprietà con 54 categorie.

- *room-type* è un fattore che indica la tipologia di camera con 3 categorie.

- *accommodates* indica il numero di stanze di ogni proprietà.

- *bathrooms* indica il numero di bagni di ogni proprietà.

- *bedrooms* indica il numero di camere da letto di ogni proprietà.

- *beds* indica il numero di letti di ogni proprietà.

- *bed-type* è un fattore che indica la tipologia di letto con 5 categorie.

- *price* è la variabile di interesse dello studio cioè il prezzo dell'affitto per notte risalente all'ultima data di scraping dell'annuncio disponibile .

- *guests-included* è il numero di ospiti extra inclusi nella prenotazione, in questo caso per il soggiorno si paga il prezzo per l'affitto indicato da *price* senza supplementi.
- *extra-people* è il supplemento che deve pagare l'ospite che porta una persona in più oltre agli ospiti inclusi nella prenotazione.
- *minimum-nights* numero minimo di notti da prenotare se si vuole soggiornare nella proprietà.
- *maximum-nights* è il numero massimo di notti che si possono prenotare nella proprietà.
- *number-of-reviews* è il numero di recensioni degli ospiti che hanno soggiornato nell'appartamento.
- *review-scores-rating* è il voto medio dato dagli ospiti al soggiorno nella proprietà.
- *review-scores-accuracy* è il voto medio dato dagli ospiti alla descrizione della proprietà fornita dal host.
- *review-scores-cleanliness* è il voto medio dato dagli ospiti alla pulizia nella proprietà.
- *review-scores-checkin* è il voto medio dato dagli ospiti all'accoglienza del host.
- *review-scores-communication* è il voto medio dato dagli ospiti alle capacità comunicative del host.
- *review-scores-location* è il voto medio dato dagli ospiti alla proprietà.
- *review-scores-value* è il voto medio dato dagli ospiti al prezzo di affitto.
- *instant-bookable* è una variabile che indica se c'è la possibilità di prenotare istantaneamente un soggiorno nell'appartamento.
- *cancellation-policy* è un fattore che indica la tipologia di politica di cancellazione delle prenotazioni del host.
- *require-guest-profile-picture* indica se l'host richiede che l'ospite abbia un immagine profilo nel suo account di Airbnb per accettare una prenotazione.

- *require-guest-phone-verification* indica se l'host richiede una verifica telefonica prima di accettare una prenotazione.

- *reviews-per-month* è il numero di recensioni che gli ospiti che hanno soggiornato nell'appartamento lasciano ogni mese.

Pulizia del dataset

Dopo la costruzione di questo dataset sono state effettuate alcune operazioni di pulizia necessarie per una corretta lettura dei dati. Innanzi tutto i 9 fattori che venivano letti come stringhe (char) sono stati convertiti in fattori (factor). La variabile *res-time-host* è stata ricodificata in modo che i dati mancanti venissero designati con NA invece che da N/A. Inoltre si è provveduto a ricodificare i dati mancanti in modo che fossero sempre indicati da NA, invece che da " ". Come si è osservato anche nel passo precedente è importante ricordare che i listing originali presentano un numero di variabili differente, in particolare, il listing di gennaio 2015 ne ha solo 52, cioè informazioni utili sulle caratteristiche dell'host e del guest non sono disponibili, lo stesso vale per i listing di marzo, maggio e giugno 2015 che ne hanno solo 68 (mancano le stesse variabili del listing di gennaio 2015), mentre gli altri listati ne hanno 95. In altri termini si è notato che le 13923 osservazioni da gennaio a giugno 2015 hanno molte informazioni mancanti, che non possono essere integrate tramite imputazione. Inoltre, tra i dati mancanti presenti ci sono anche i voti ai servizi offerti per tutte le proprietà di cui abbiamo traccia degli annunci, ma che non sono mai state affittate, circa 50000. Infine ci sono circa 4100 osservazioni di cui non si hanno a disposizione i nomi dei quartieri e sottoquartieri in cui si trovano oppure a cui mancano informazioni sulla proprietà, come per esempio il numero di camere o bagni. Dato che non è possibile fare imputazione di valori per tutte le informazioni mancanti si è deciso di eliminare tutte le osservazioni che presentavano questo problema, essendo comunque consapevoli del fatto che non è necessariamente vero che questa sia la soluzione più adeguata per risolvere il problema. Resta così un dataset di 77758 record e 35 variabili. Le variabili numeriche dei prezzi venivano lette come fattori (factor) quindi sono state convertite in formato numerico. Infine sono stati ricodificati i fattori con molte categorie o con categorie che contengono un numero molto basso di osservazioni. In particolare, nel caso di bed-type si uniscono le categorie "Pull-out" "Sofa", "Futon","Couch", "Airbed" in un'unica categoria "Other", che raccoglie 2187 osservazioni, mentre le rimanenti 75571 appartengono

tutte alla categoria “Real Bed”. Nel caso di property-type sono lasciate invariate le prime 3 categorie con maggiore numerosità (“Apartment”, “House”, “Loft”), che da sole raccolgono al loro interno 72859, mentre le altre 51 sono state raccolte in un’unica categoria chiamata “Other”. Per quello che riguarda cancellation-policy si uniscono “super-strict-30” e “super-strict-60” perché da sole raccolgono soltanto 57 osservazioni nella categoria “strict”.

Pulizia della variabile price

In questa fase sono stati fatti degli ulteriori aggiustamenti sui dati: in particolare osservando la distribuzione della variabile *price*, cioè il prezzo di affitto a notte, si nota che c’è moltissima variabilità, come si vede nella tabella 2.1.

Minimo	Primo Quantile	Mediana	Media	Terzo Quantile	Massimo
0.0	70.0	110.0	142.6	175.0	10000.0

Tabella 2.1: Statistiche descrittive sulla variabile “price”

Per risolvere eventuali problemi collegati a ciò, si decide innanzi tutto di eliminare le 36 osservazioni con l’affitto pari a zero. Si è anche deciso di eliminare i prezzi sopra i 4000 dollari (9 osservazioni) che avrebbero comportato una varibilità eccessiva . Inoltre un altro aspetto da considerare è il costo di portare un’ospite extra che in certi annunci è uguale o addirittura più alto del prezzo di affitto dell’appartamento stesso, anche di 10 volte nei casi più estremi. Si è deciso di non eliminare la variabile (extra-people) o queste 1248 osservazioni in quanto non è chiaro se questo sia un errore nella trascrizione dei dati o una tattica di certi host mirata a disincentivare gli ospiti dal portare persone in più oltre a quelle incluse nella prenotazione.

Minimo	Primo Quantile	Mediana	Media	Terzo Quantile	Massimo
10.0	70.0	110.0	141.9	175.0	3750.0

Tabella 2.2: Statistiche descrittive sulla variabile “price” dopo la pulizia

Pulizia degli Opendata

Le altre variabili incluse nell'analisi provengono da altri 45 dataset presi dal sito www.opendata.cityofnewyork.us (tranne poche eccezioni) e che appartengono alle seguenti categorie:

- Business: tabelle (2.3) – (2.9)
- City Government: tabelle (2.10) – (2.20)
- Education: tabellae(2.21) – (2.27)
- Environment: tabelle (2.28) – (2.30)
- Health: tabelle (2.31) – (2.34)
- Recreation: tabelle (2.35) – (2.43)
- Transportation: tabelle (2.44) – (2.47)

Le categorie considerate includono dataset che al loro interno contengono variabili legate a servizi o a fattori socio-economici che si ipotizza possano influenzare i prezzi di Airbnb. Di seguito si riportano tutti i dataset raccolti, le variabili in essi contenute con relativa descrizione, divise per tipologia.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	company.name	stringa	nome attività
2	subindustry	categoriale	tipologia di attività
3	latitude	numerica	latitudine
4	longitude	numerica	longitudine

Tabella 2.3: Database della città di New York riguardo i locali presenti in Times Square. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

Per selezionare questi dataset si è partiti dagli oltre 2000 dataset presenti nel sito www.opendata.cityofnewyork.us. In seguito sono stati eliminati i dataset che riguardavano le attività commerciali, aspetti burocratici (numero di multe fatte) dell'amministrazione

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	company.name	stringa	nome azienda
2	borough	categoriale	quartiere dell'azienda
3	latitude	numerica	latitudine
4	longitude	numerica	longitudine

Tabella 2.4: Database della città di New York riguardo le aziende presenti. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome locale
2	address	stringa	indirizzo locale
3	latitude	numerica	latitudine
4	longitude	numerica	longitudine

Tabella 2.5: Database contenente la collocazione di tutti gli Starbucks presenti negli Stati Uniti, da cui sono stati selezionati solo i locali che si trovano nella città di New York. Disponibile sul sito: `https://opendata.socrata.com`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	licence.expiration.date	stringa	nome attività commerciale
2	licence.status	stringa	active, se la licenza è attiva
3	licence.creation.date	data	data di creazione della licenza
4	industry	fattore	settore a cui appartiene l'attività commerciale
5	business.name	stringa	nome attività commerciale
6	latitude	numerica	latitudine
7	longitude	numerica	longitudine

Tabella 2.6: Database contenente la collocazione di tutte le attività commerciali della città di New York con licenze attive. Disponibile sul sito: `https://opendata.socrata.com`.

cittadina non inerenti a servizi generalmente utilizzati da tutti i cittadini o dai turisti, che sono coloro che più usufruiscono del servizio di Airbnb. Per esempio un dataset contenente l'ubicazione di tutti i retailer di tabacco e sigarette elettroniche è stato escluso dall'analisi.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome fastfood
2	longitudine	numerica	longitudine
3	latitudine	numerica	latitudine
4	neighborhood	categoriale	nome sottoquartiere del fast food
5	borough	categoriale	nome quartiere del fast food

Tabella 2.7: Database della città di New York riguardo i fast food. Disponibile sul sito: `https://overpass-turbo.eu`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome hotel
2	longitudine	numerica	longitudine
3	latitudine	numerica	latitudine
4	neighborhood	categoriale	nome sottoquartiere del hotel
5	borough	categoriale	nome quartiere del hotel

Tabella 2.8: Database della città di New York riguardo gli hotel. Disponibile sul sito: `https://overpass-turbo.eu`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome ristoranti
2	longitudine	numerica	longitudine
3	latitudine	numerica	latitudine
4	neighborhood	categoriale	nome sottoquartiere dei ristoranti
5	borough	categoriale	nome quartiere dei ristoranti

Tabella 2.9: Database della città di New York riguardo i ristoranti. Disponibile sul sito: `https://overpass-turbo.eu`.

In particolare, sono stati scelti preferibilmente dataset che contenessero l'ubicazione dei servizi o delle attrattive turistiche espressa in latitudine e longitudine. Inoltre sono stati esclusi i dataset che non riguardassero l'intera città, ma solo un singolo quartiere. Per

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	SchoolDist	numerica	numero del distretto scolastico in cui si trova il terreno
2	FireComp	stringa	la compagnia dei vigili del fuoco che ha giurisdizione su quel terreno
3	PolicePrct	numerica	distretto di polizia che ha giurisdizione sul terreno
4	HealthArea	numerica	distretto sanitario in cui si trova il terreno
5	BldgClass	categoriale	codice che indica l'uso del terreno
6	LandUse	numerica	odice dll'uso principale del terreno
7	OwnerType	categoriale	tipologia di proprietario
8	OwnerName	stringa	nome del proprietario
9	LotArea	numerica	area totale del terreno
10	BldgArea	numerica	area complessiva edificio
11	ComArea	numerica	area dell'edificio utilizzata per fini commerciali
12	ResArea	numerica	area dell'edificio utilizzata per fini residenziali
13	OfficeArea	numerica	area dell'edificio utilizzata come uffici
14	RetailArea	numerica	area dell'edificio utilizzata per il commercio al dettaglio
15	GarageArea	numerica	area dell'edificio utilizzata come garage
16	StrgeArea	numerica	area dell'edificio utilizzata come magazzino
17	FactryArea	numerica	area dell'edificio utilizzata come area industriale
18	OtherArea	numerica	area dell'edificio utilizzata per finalittà diverse da quelle elencate sopra
19	NumBldgs	numerica	numero di costruzioni sul terreno
20	NumFloors	numerica	numero di piani della costruzione principale sul terreno
21	UnitsRes	numerica	numero di edifici residenziali presenti sul terreno
22	UnitsTotal	numerica	numero di edifici presenti sul terreno
23	YearBuilt	numerica	anno di costruzione dell'edificio
24	Landmark	stringa	nome dell'edificio
25	HistDist	stringa	nome del distretto storico a cui appartiene il terreno
26	latitude	numerica	latitudine
27	longitude	numerica	longitudine

Tabella 2.10: Database contenente la collocazione di tutte le costruzioni della città di New York registrate fino all'anno 2016. Disponibile sul sito: `https://opendata.socrata.com`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome sottoquartiere
2	population	numerica	popolazione

Tabella 2.11: Database della città di New York contenente la popolazione divisa per sottoquartiere aggiornata al 2018 e intregrata per i sottoquartieri per cui questo dato mancava. Disponibile sul sito: `https://https://opendata.cityofnewyork.us/`.

quello che riguarda la scelta degli indicatori (livelli di inquinamento, livello di istruzione, dati sul traffico per esempio) che rappresentano la qualità della vita nei quartieri sono stati selezionati soltanto i dataset che non contenevano solo gli indicatori aggregati per tutta New York, ma anche questi ultimi disaggregati almeno per singolo quartiere perchè qui si ipotizza che la variabilità all'interno dei dati sia dovuta proprio alla divisione della città in

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome della struttura
2	borough	categoriale	quartiere in cui si trova la caserma
3	latitude	numerica	latitudine
4	longitude	numerica	longitudine

Tabella 2.12: Database della città di New York riguardo alle caserme dei pompieri. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome dell'ufficio
2	borough	categoriale	quartiere in cui si trova l'ufficio
3	latitude	numerica	latitudine
4	longitude	numerica	longitudine

Tabella 2.13: Database della città di New York riguardo agli uffici di collocamento. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	provider	stringa	nome associazione creditrice
2	host	stringa	nome associazione che ospita l'ente
3	borough	categoriale	quartiere in cui si trova l'istituto
4	latitude	numerica	latitudine
5	longitude	numerica	longitudine

Tabella 2.14: Database della città di New York riguardo alle istituzioni di credito ad imprese e cittadini. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

quartieri e sottoquartieri.

Per tutti i dataset sono state rimosse le variabili non utili ai fini delle analisi o con molti valori mancanti e tutte le osservazioni per cui latitudine e longitudine non sono disponibili. Nei casi in cui mancano latitudine e longitudine (tabella 2.33,tabella 2.40, avendo a

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome sito
2	type	categoriale	tipologia sito
3	borough	categoriale	quartiere in cui si trova il sito
4	latitude	numerica	latitudine
5	longitude	numerica	longitudine

Tabella 2.15: Database della città di New York riguardo ai siti di raccolta differenziata dei rifiuti. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	incident-address	stringa	indirizzo
2	borough	categoriale	quartiere in cui si trova il sito
3	latitude	numerica	latitudine
4	longitude	numerica	longitudine

Tabella 2.16: Database della città di New York riguardo le aree della città imbrattate con graffiti in cui il Department of Sanitation è intervenuto nell'anno 2018 (fino a ottobre 2018). Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome seggio
2	postcode	numerica	codice postale
3	latitude	numerica	latitudine
4	longitude	numerica	longitudine

Tabella 2.17: Database della città di New York riguardo i seggi elettorali. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

disposizione soltanto gli indirizzi, per ottenere la collocazione geografica espressa con latitudine e longitudine ci si è avvalsi di datasciencetoolkit, un sito che offre una raccolta gratuita di set di dati aperti e di strumenti open source per il data science, sfruttando le

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome
2	postcode	numerica	codice postale
3	latitude	numerica	latitudine
4	longitude	numerica	longitudine

Tabella 2.18: Database della città di New York riguardo gli uffici postali. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	CMPLNT-FR-DT	data	data esatta del reato
2	KY-CD	numerica	codice del reato
3	OFNS-DESC	categoriale	descrizione del reato corrispondente al codice
4	PD-DESC	categoriale	descrizione del reato corrispondente al codice della polizia
5	CRM-ATPT-CPTD-CD	categoriale	se completed, il reato è stato completato con successo
6	LAW-CAT-CD	categoriale	gravità del reato
7	BORO-NM	categoriale	codice del quartiereè in cui stato commesso il reato
8	Latitude	numerica	latitudine
9	Longitude	numerica	longitudine
10	PATROL-BORO	categoriale	nome della giurisdizione dove è stato commesso il reato
11	VIC-AGE-GROUP	categoriale	classe di età della vittima
12	VIC-RACE	categoriale	etnia della vittima
13	VIC-SEX	categoriale	sesso della vittima

Tabella 2.19: Database della città di New York sui crimini commessi dal 1900-2017. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	latitude	numerica	latitudine
2	longitude	numerica	longitudine
3	borough	stringa	quartiere

Tabella 2.20: Database della città di New York riguardo agli edifici dichiararati inagibili dai vigili del fuoco, aggiornato a luglio 2017 . Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

informazioni derivanti dall'ultimo censimento USA e quelle di OpenStreetMap per operare la geocodifica degli indirizzi stradali negli Stati Uniti. Purtroppo effettuando questa ope-

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome dell'università
2	ZIP	numerica	codice postale
3	latitude	numerica	latitudine
4	longitude	numerica	longitudine

Tabella 2.21: Database della città di New York riguardo ai college e alle università. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	borough	categoriale	quartiere
2	grade	categoriale	grado della scuola
3	year	numerica	anno
4	number-tested	numerica	numero di studenti a cui è stato somministrato il test
5	level1-absolute	numerica	numero di studenti che hanno ottenuto un punteggio pari a 1
6	level1%	numerica	percentuale di studenti che hanno ottenuto un punteggio pari a 1
7	level2-absolute	numerica	numero di studenti che hanno ottenuto un punteggio pari a 2
8	level2%	numerica	percentuale di studenti che hanno ottenuto un punteggio pari a 2
9	level3-absolute	numerica	numero di studenti che hanno ottenuto un punteggio pari a 3
10	level3%	numerica	percentuale di studenti che hanno ottenuto un punteggio pari a 3
11	level4-absolute	numerica	numero di studenti che hanno ottenuto un punteggio pari a 4
12	level4%	numerica	percentuale di studenti che hanno ottenuto un punteggio pari a 4
13	level3-4-absolute	numerica	numero di studenti che hanno ottenuto un punteggio pari a 3-4
14	level3-4%	numerica	percentuale di studenti che hanno ottenuto un punteggio pari a 3-4

Tabella 2.22: Database della città di New York riguardo ai risultati dell' English Language Arts Test alle scuole superiori. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

razione in alcuni casi sono state perse alcune osservazioni, perché non si è stati in grado di localizzarle. Infine, nel caso del dataset descritto in tabella (2.10) dato che c'erano le coordinate geografiche espresse nel sistema di riferimento dello stato di New York è stato necessario esprimerle nella forma di latitudine e longitudine prima di poter utilizzare la relativa informazione (per ulteriori approfondimenti si veda l'Appendice A). Nei casi in cui nei dataset siano state incluse anche variabili categoriali spesso è stato necessario ordinare i livelli. Per esempio nel dataset descritto in tabella 2.27, variabili come "weekeends" che indicano se il doposcuola è aperto anche nel finesettimana presentava i seguenti livelli : "no" "No" "yes" "Yes" "Yes"', quindi è stato necessario ricodificare i livelli in modo che alla fine se ne avessero solo due "No", "Yes". Questa operazione è stata effettuata per le altre

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	borough	categoriale	quartiere
2	grade	categoriale	grado della scuola
3	year	numerica	anno
4	number-tested	numerica	numero di studenti a cui è stato somministrato il test
5	level1-absolute	numerica	numero di studenti che hanno ottenuto un punteggio pari a 1
6	level1%	numerica	percentuale di studenti che hanno ottenuto un punteggio pari a 1
7	level2-absolute	numerica	numero di studenti che hanno ottenuto un punteggio pari a 2
8	level2%	numerica	percentuale di studenti che hanno ottenuto un punteggio pari a 2
9	level3-absolute	numerica	numero di studenti che hanno ottenuto un punteggio pari a 3
10	level3%	numerica	percentuale di studenti che hanno ottenuto un punteggio pari a 3
11	level4-absolute	numerica	numero di studenti che hanno ottenuto un punteggio pari a 4
12	level4%	numerica	percentuale di studenti che hanno ottenuto un punteggio pari a 4
13	level3-4-absolute	numerica	numero di studenti che hanno ottenuto un punteggio pari a 3-4
14	level3-4%	numerica	percentuale di studenti che hanno ottenuto un punteggio pari a 3-4

Tabella 2.23: Database della città di New York riguardo ai risultati del NY State Math Test alle scuole superiori. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	school-name	stringa	nome scuola superiore
2	neighborhood	stringa	sottoquartiere dove si trova la scuola superiore
3	total-students	numerica	numero totale di studenti iscritti
4	borough	stringa	quartiere della scuola superiore
5	latitude	numerica	latitudine
6	longitude	numerica	longitudine

Tabella 2.24: Database della città di New York riguardo le scuole superiori. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	loc-name	stringa	nome asilo
2	seats	stringa	numero totale di bambini iscritti
3	meals	numerica	numero di pasti al giorno inclusi
4	latitude	numerica	latitudine
5	longitude	numerica	longitudine

Tabella 2.25: Database della città di New York riguardo agli asili e alle scuole per l'infanzia. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	School.Name	stringa	nome della scuola
2	Latitude	numerica	latitudine
3	Longitude	numerica	longitudine
4	Zip	numerica	codice postale
5	Grades	categoriale	grado della scuola
6	Community.School	categoriale	Yes, se la scuola fa parte di un distretto scolastico
7	Economic.Need.Index	numerica	misura della povertà media degli studenti della scuola
8	School.Income.Estimate	numerica	reddito medio delle famiglie degli studenti della scuola
9	Percent.Asian	numerica	percentuale di studenti di etnia asiatica
10	Percent.Black	numerica	percentuale di studenti di etnia afroamericana
11	Percent.Hispanic	numerica	percentuale di studenti di etnia ispanica
12	Percent.Black...Hispanic	numerica	percentuale di studenti di etnia ispanica e afroamericana
13	Percent.White	numerica	percentuale di studenti di etnia caucasica
14	Student.Attendance.Rate	numerica	numero totale di giorni di lezioni svolti da tutti gli studenti
15	Percent.of.Students.Chronically.Absent	numerica	percentuale di studenti che perdono almeno il 10%delle lezioni
16	Rigorous.Instruction.Rating	categoriale	misura della qualità della preparazione fornita
17	Collaborative.Teachers.Rating	categoriale	misura della capacità degli insegnanti di migliorare l'ambiente scolastico
18	Supportive.Environment.Rating	categoriale	misura di quanto la scuola sia un ambiente adatto per crescere e imparare
19	Effective.School.Leadership.Rating	categoriale	misura della leadership della scuola nella comunità scolastica
20	Strong.Family.Community.Ties.Rating	categoriale	misura della collaborazione scuola famiglie
21	Trust.Rating	categoriale	misura della fiducia che c'è tra i membri della comunità scolastica
22	Student.Achievement.Rating	numerica	misura del livello di preparazione medio degli studenti

Tabella 2.26: Database della città di New York riguardo alle scuole presenti a Central Harlem. Disponibile sul sito: `https://kaggle.com`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	borough	categoriale	quartiere
2	ZIP	categoriale	grado della scuola
3	summer	numerica	anno
4	weekends	categoriale	se Yes ,il doposcuola è aperto anche nei weekend
5	evenings	categoriale	se Yes ,il doposcuola è aperto anche la sera
6	elementary	categoriale	se Yes ,il doposcuola è aperto agli studenti della scuola elementare
7	middle	categoriale	il doposcuola è aperto agli studenti della scuola media
8	high	categoriale	il doposcuola è aperto agli studenti della scuola superiore
9	weekly	categoriale	se Yes , il doposcuola è aperto per tutta la settimana
10	enrollment	numerica	numero iscrizioni
11	sports	categoriale	se Yes , nel doposcuola è si pratica sport
12	arts	categoriale	se Yes , nel doposcuola è si studiano le arti
13	name	stringa	nome doposcuola
14	longitude	numerica	longitudine
15r	latitude	numerica	latitudine

Tabella 2.27: Database della città di New York riguardo ai doposcuola. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

variabili categoriali presenti in tabella 2.27 e in altri dataset, per evitare che ci fossero più livelli che designassero la stessa categoria.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	created.date	data	data del reclamo
2	descriptor	categoriale	tipologia di reclamo
3	status	categoriale	stato del reclamo
4	borough	categoriale	quartiere in cui è stato fatto il reclamo
5	latitude	numerica	latitudine
6	longitude	numerica	longitudine

Tabella 2.28: Database contenente la collocazione di tutti i reclami fatti sulla qualità dell'acqua potabile fatti nella città di New York dal 2010 al 2018. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	location	stringa	luogo dell'ispezione
2	Residual-Free-Chlorine	numerica	concentrazione totale cloro misurata in mg/L
3	turbidity	numerica	misura della torbidità dell'acqua in NTU
4	Coliform-MPN-100mL	categoriale	concentrazione di coliformi su 100mL di acqua
5	escherichia coli-MPN-100mL	categoriale	concentrazione di escherichia coli su 100mL di acqua
6	latitude	numerica	latitudine
7	longitude	numerica	longitudine

Tabella 2.29: Database contenente gli esiti delle analisi della qualità dell'acqua potabile fatti nella città di New York nell'anno 2018. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	indicator	stringa	nome dell'indicatore
2	type of measurement	categoriale	tipologia di misura
3	entity-type	categoriale	tipologia di entità geografica
4	name-entity	strings	nome entità geografica
5	year-first-measurement	numerica	anno prima rilevazione
6	value	numerica	valore dell'indicatore

Tabella 2.30: Database contenente una raccolta di indicatori della qualità dell'aria della città di New York. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

Costruzione delle variabili da utilizzare nell'analisi

In tabella 2.48 viene illustrato l'intero set di variabili utilizzato nei modelli presentati nel prossimo capitolo. Si procede ora a fornirne una breve descrizione. Le variabili da 1 a 36

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome della struttura sanitaria
2	type	stringa	itipologia di struttura sanitaria
3	borough	categoriale	quartiere in cui si trova la struttura sanitaria
4	latitude	numerica	latitudine
5	longitude	numerica	longitudine

Tabella 2.31: Database della città di New York riguardo alle strutture sanitarie. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome della struttura
2	ZIP	numerica	codice postale
3	children	categoriale	se Yes, viene effettuata la vaccinazione sui bambini
4	latitude	numerica	latitudine
5	longitude	numerica	longitudine

Tabella 2.32: Database della città di New York riguardo ai centri per la vaccinazione antinfluenzale. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome della struttura
2	address	stringa	indirizzo
3	borough	categoriale	quartiere
4	ZIP	numerica	codice postale
5	lowcost	categoriale	se true, tariffa minima per gli indigenti
6	free	categoriale	se true, test gratis per gli indigenti
7	latitude	numerica	latitudine
8	longitude	numerica	longitudine

Tabella 2.33: Database della città di New York riguardo ai centri per la prevenzione dell'HIV. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

provengono dal dataset riguardante gli affitti in Airbnb e riguardano tutte le caratteristiche degli appartamenti o delle stanze affittate, degli host e dei guest. La variabile 18 price, prezzo di affitto per notte è la variabile risposta, la 36 lp è la sua trasformata logaritmica. Le variabili da 37 a 62 descrivono se un certo servizio è presente o meno nel sottoquartiere in cui

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	inspection type	categoriale	tipo ispezione
2	latitude	numerica	latitudine
3	longitude	numerica	longitudine
4	borough	categoriale	nome quartiere
5	inspection date	date	data ispezione
6	result	categoriale	risultato ispezione

Tabella 2.34: Databases della città degli interventi della derattizzazione dal 2009 al 2018. Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome della proprietà
2	location	stringa	indirizzo della proprietà
3	number of courts	numerica	numero del campo
4	accessible	categoriale	Y, se la proprietà è aperta al pubblico
5	latitude	numerica	latitudine
6	longitude	numerica	longitudine

Tabella 2.35: Databases della città di New York riguardo ai campi da basket, cricket,tennis e pallamano di proprietà del Department of Parks & Recreation di New York. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome della proprietà
2	location	stringa	indirizzo della proprietà
3	accessible	categoriale	Y, se la proprietà è aperta al pubblico
4	latitude	numerica	latitudine
5	longitude	numerica	longitudine
6	public skate admission price adult	numerica	prezzo ingresso adulti
7	public skate admission price ahild	numerica	prezzo ingresso bambini

Tabella 2.36: Databases della città di New York riguardo ai palaghiaccio di proprietà del Department of Parks & Recreation di New York. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome della proprietà
2	location	stringa	indirizzo della proprietà
3	accessible	categoriale	Y, se la proprietà è aperta al pubblico
4	latitude	numerica	latitudine
5	longitude	numerica	longitudine
6	pool type	stringa	tipologia di piscina

Tabella 2.37: Databases della città di New York riguardo alle piscine coperte o scoperte di proprietà del Department of Parks & Recreation di New York. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome della proprietà
2	location	stringa	indirizzo della proprietà
3	accessible	categoriale	Y, se la proprietà è aperta al pubblico
4	latitude	numerica	latitudine
5	longitude	numerica	longitudine
6	surf	stringa	Yes, se si può fare surf
7	barbecue allowed	stringa	Yes, se si può fare il barbecue
8	mobile charging station	stringa	Yes, se si ci sono luoghi dove caricare il cellulare

Tabella 2.38: Databases della città di New York riguardo alle spiagge di proprietà del Department of Parks & Recreation di New York. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome della proprietà
2	latitude	stringa	latitudine
3	longitude	numerica	longitudine
4	ZIP	numerica	Codice postale

Tabella 2.39: Databases della città di New York riguardo a orti botanici,musei, biblioteche, teatri e zoo. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

si trova ogni appartamento. I dataset descritti nelle tabelle 2.5, 2.10, 2.12, 2.13, 2.17, 2.18, 2.21, 2.31, 2.35, 2.36, 2.37, 2.38, 2.39, 2.40, 2.41 sono stati utilizzati per la loro costruzione.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome della proprietà
2	latitude	stringa	latitudine
3	longitude	numerica	longitudine

Tabella 2.40: Databases della città di New York riguardo a aree per i cani, aree barbecue, parchi, aree accessibili ai portatori di handicap e case storiche di proprietà del Department of Parks & Recreation di New York. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	latitude	numerica	latitudine
2	longitude	numerica	longitudine
3	type	categoriale	tipologia di wifi
4	borough	stringa	quartiere
5	activated	data	data di attivazione

Tabella 2.41: Databases della città di New York riguardo a aree con hotspot Wi-Fi . Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	latitude	numerica	latitudine
2	longitude	numerica	longitudine
3	borough	stringa	quartiere

Tabella 2.42: Databases della città di New York riguardo a aree con LinkNYC,cioè una rete di chioschi che forniscono Wi-Fi gratuito ad alta velocità, chiamate nazionali, un pulsante dedicato 911, porte di ricarica per dispositivi mobili e accesso a siti Web selezionati. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

Inoltre è stato necessario reperire l'informazione riguardante i sottoquartieri dove si trovano tutti i servizi, dato che questa è sempre mancante. Per risolvere il problema si va ad utilizzare una funzione che, a partire da latitudine e longitudine, fornisce il sottoquartiere in

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	lm-name	stringa	nome attrazione
2	boroughID	categoriale	codice quartiere
3	desig-address	stringa	indirizzo dell'attrazione
4	lm-type	data	tipo di attrazione turistica
5	latitude	numerica	latitudine
6	longitude	numerica	longitudine

Tabella 2.43: Database della città di New York riguardo le attrazioni turistiche (distretti storici e monumenti). Disponibile sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome aeroporto
2	longitude	numerica	longitudine
3	latitude	numerica	latitudine

Tabella 2.44: Database della città di New York riguardo agli aeroporti. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome fermata
2	longitude	numerica	longitudine
3	latitude	numerica	latitudine
4	line	categoriale	numero linee metro che passano per quella fermata

Tabella 2.45: Database della città di New York riguardo le entrate della metropolitana. Disponibili sul sito: `https://opendata.cityofnewyork.us/`.

cui si trova il punto designato dalle coordinate gps (per ulteriori dettagli si veda l'Appendice B). Da notare, inoltre, che per questi servizi si è preferito non calcolare la distanza minima tra ogni appartamento e il servizio in questione perché i dataset a loro riferiti hanno un numero piuttosto basso di unità statistiche (meno di 1000). Le variabili dalla 63 alla 72 sono le distanze di Haversine minime tra gli appartamenti e alcuni servizi (scuole) o attrazioni

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome fermata
2	longitude	numerica	longitudine
3	latitude	numerica	latitudine
4	neighborhood	categoriale	nome sottoquartiere fermata
5	borough	categoriale	nome quartiere fermata

Tabella 2.46: Database della città di New York riguardo le fermate degli autobus. Disponibile sul sito: `https://overpass-turbo.eu`.

	Nome della variabile	Tipo di variabile	Descrizione della variabile
1	name	stringa	nome fermata
2	longitudine	numerica	longitudine
3	latitudine	numerica	latitudine
4	neighborhood	categoriale	nome sottoquartiere della compagnia
5	borough	categoriale	nome quartiere della compagnia

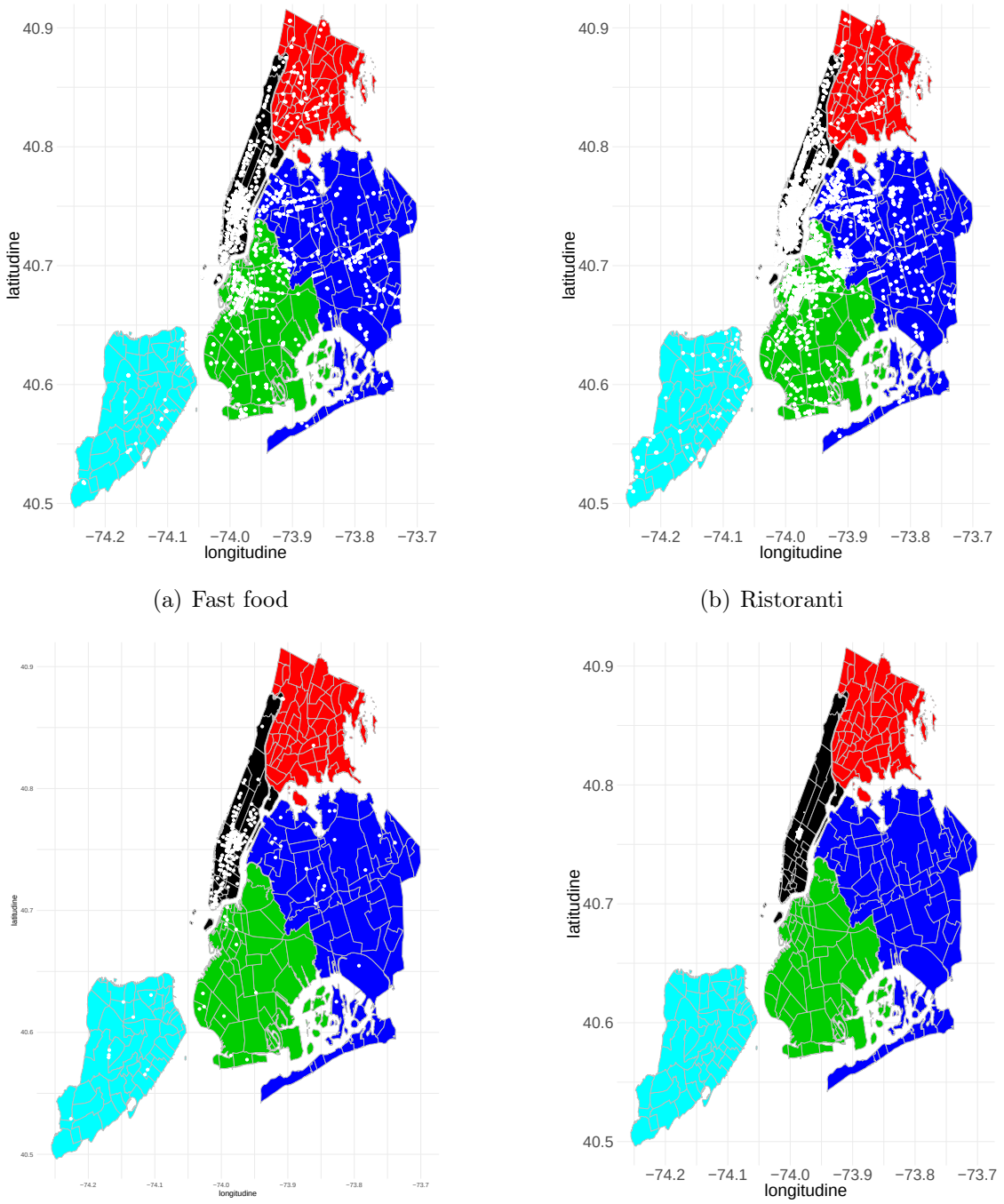
Tabella 2.47: Database della città di New York riguardo le compagnie di taxi. Disponibile sul sito: `https://overpass-turbo.eu`.

turistiche (per ulteriori dettagli su come sono state calcolate si veda l'Appendice C). I dataset da cui sono state tratte variabili utilizzate sono descritti nelle tabelle 2.7, 2.8, 2.9, 2.10, 2.43, 2.45. Le variabili da 73 a 106 sono dei tassi di criminalità per alcune tipologie di reati socialmente pericolosi, come omicidi di donne e di bambini, rapine e violenze sessuali, calcolati nel modo seguente. Innanzi tutto a partire dal dataset descritto nella tabella 2.20 è stato calcolato, per ogni tipo di crimine considerato, il numero di reati di quella categoria commessi in un raggio di 500, 1000 e 1500 metri da ogni proprietà nei 2 o 4 anni–per i reati più gravi- -precedenti la data dello scraping più recente di ogni proprietà. Per ulteriori approfondimenti sulla procedura utilizzata si veda l'appendice D. Successivamente il numero totale dei crimini in un raggio di 500, 1000 e 1500 riferito a ogni proprietà è stato diviso per la popolazione del sottoquartiere in cui la proprietà si trova, non essendo disponibile il numero esatto di abitanti nelle aree considerate per il conteggio dei crimini. I dati relativi alla popolazione per ogni sottoquartiere si trovano nel dataset

descritto nella tabella 2.11

Analisi preliminari

Dal momento che la variabile di interesse è il prezzo degli affitti per notte, si procede ad un analisi esplorativa di questa. Come si vede dalla tabella 3.2 i prezzi presentano una variabilità piuttosto ampia dal momento che si va da un minimo di 10 dollari per notte a un prezzo massimo di 3750 per notte. Per questo motivo si è scelto di utilizzare come variabile risposta non i prezzi, ma la loro trasformata logaritmica. Questo serve per renderne la distribuzione più stabile e con meno variabilità, anche per cercare di evitare che si possano ottenere nei modelli risultati non ragionevoli. Osservando la figura (2.23a) si nota che la distribuzione della trasformata logaritmica degli affitti non ha distribuzione normale, in quanto sembra avere code più pesanti soprattutto la destra, cosa che si nota anche dall'istogramma della figura 2.24a. Dal grafico 2.23b si nota che la trasformata logaritmica degli affitti ha una distribuzione diversa per ogni quartiere di New York (Manhattan, Queens, Brooklyn, Staten Island, Bronx). In generale si nota che in tutti i casi la distribuzione non è mai normale. Per i quartieri di Queens, Brooklyn, Staten Island, Bronx si ha una distribuzione ipernormale con un'asimmetria negativa. Mentre nel caso di Manhattan si ha una distribuzione iponormale con asimmetria negativa. In tutti i casi la coda destra è molto più pesante della coda sinistra come si vede dalla figura 2.23c. Manhattan e Brooklyn sono i sottoquartieri che presentano più variabilità degli affitti in quanto sono i quartieri in cui si trovano la maggior parte degli appartamenti affittati, rispettivamente 39475 e 28444. In generale il quartiere con il log-prezzo medio di affitto più alto sembra essere Manhattan. Quanto osservato in precedenza è confermato anche se si osservano gli istogrammi da figura 2.24b a 2.24f. Per quello che riguarda la tipologia di proprietà nel caso di Brooklyn e di Queens i log-prezzi degli affitti medi sembrano essere simili e il *Loft* risulta essere la categoria che presenta più variabilità. Negli altri casi ci sono tre categorie che hanno circa lo stesso log-prezzo medio e una che sembra spiccare tra le altre con un log-prezzo medio più alto, cioè il *Loft*. La categoria che risulta presentare la varibilità maggiore è *House*, tranne nel caso del Bronx e di Brooklyn. Infine per quanto riguarda la tipologia di soggiorno in tutti i sottoquartieri si può notare che la categoria col log-prezzo dell'affitto più alto è l'intero appartamento, mentre quella con il log-prezzo di affitto più basso risulta essere la stanza condivisa. Naturalmente ci si aspetta di trovare questa evidenza in quanto è logico



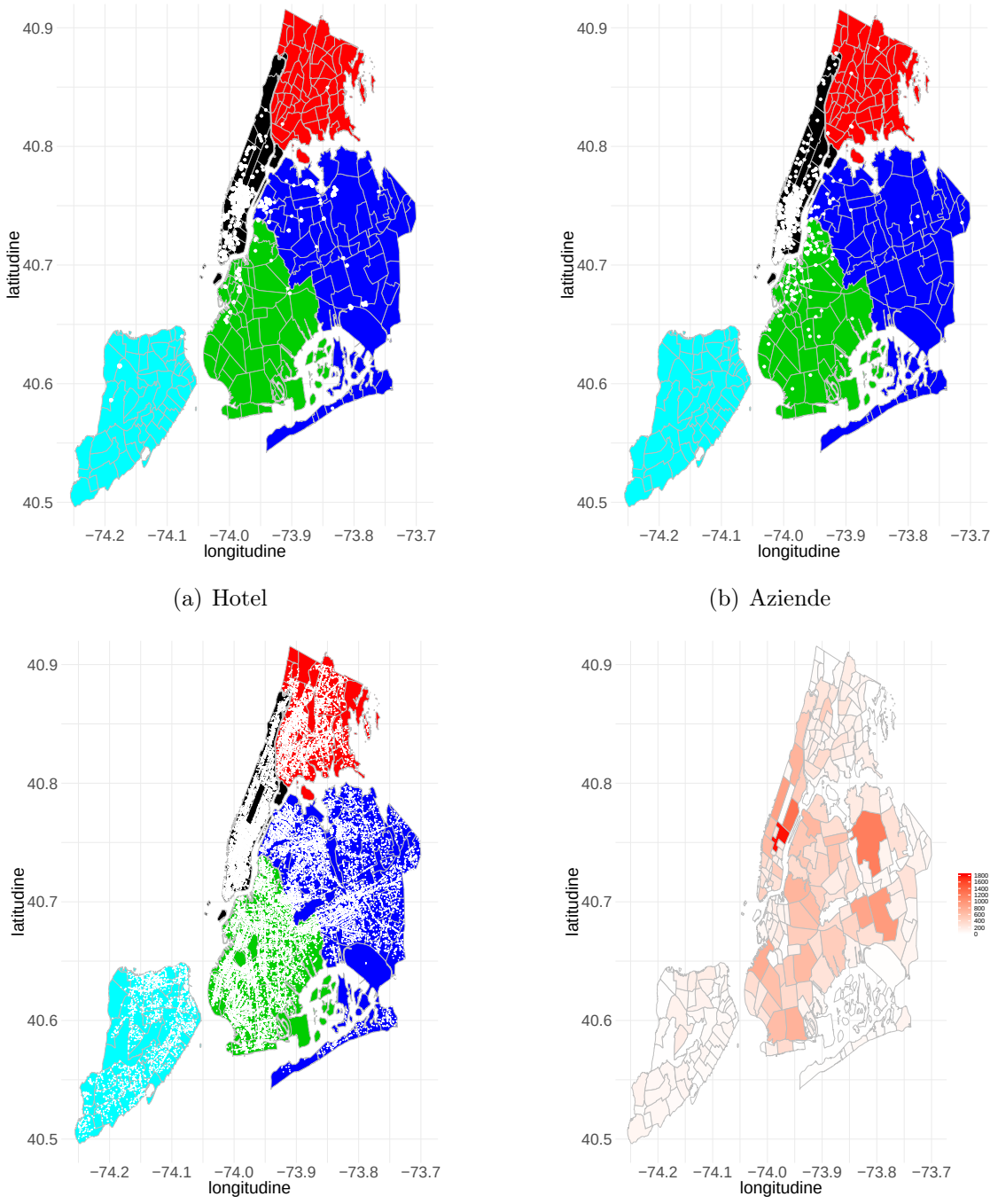
(a) Fast food

(b) Ristoranti

(c) Starbucks

(d) Locali di Time Square

Figura 2.1: Collocazione dei locali



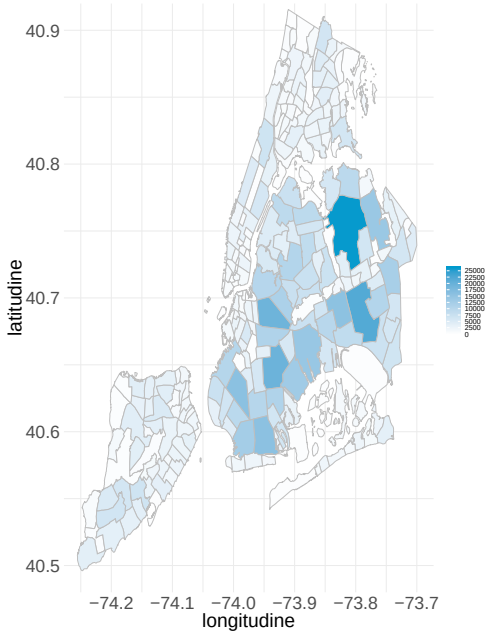
(a) Hotel

(b) Aziende

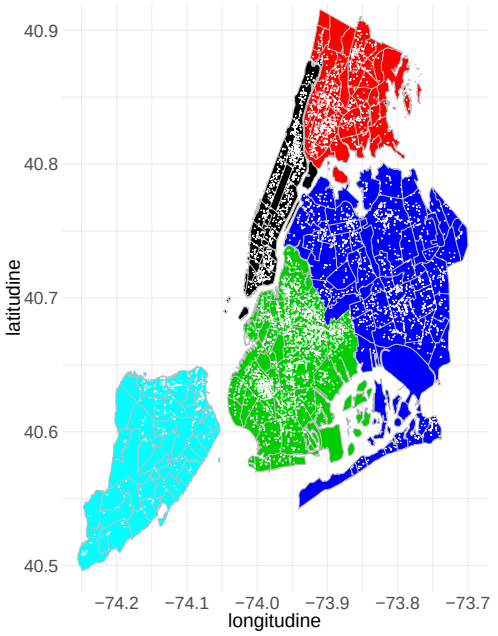
(c) Attività commerciali

(d) Heat map attività commerciali

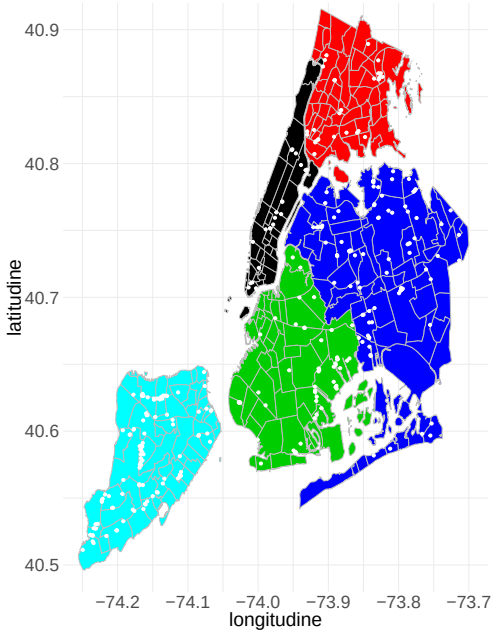
Figura 2.2: Collocazione delle attività commerciali



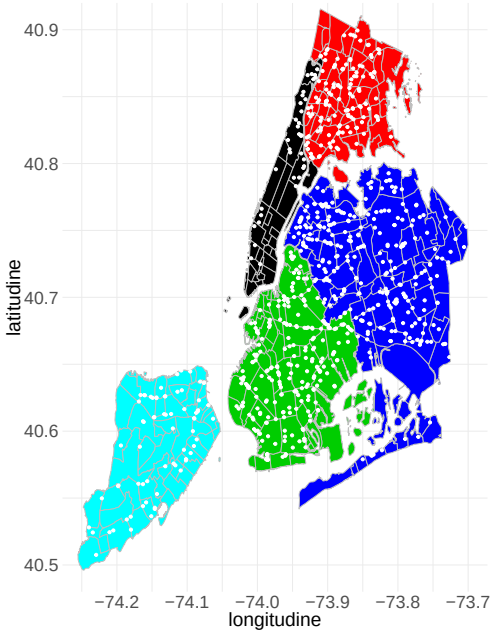
(a) Numero costruzioni in ogni sottoquartiere



(b) Luoghi di culto

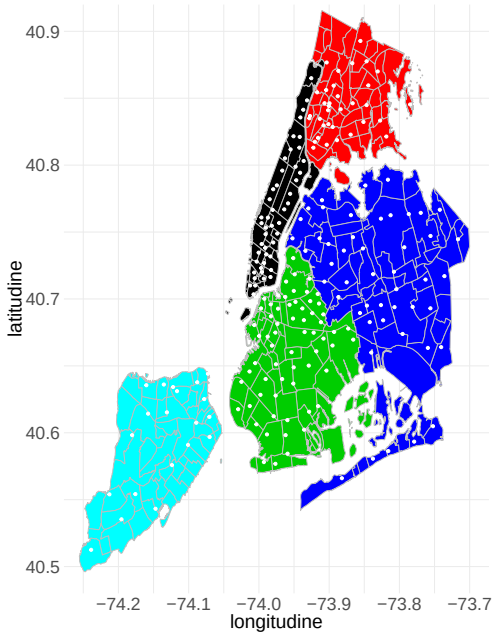


(c) Centri commerciali

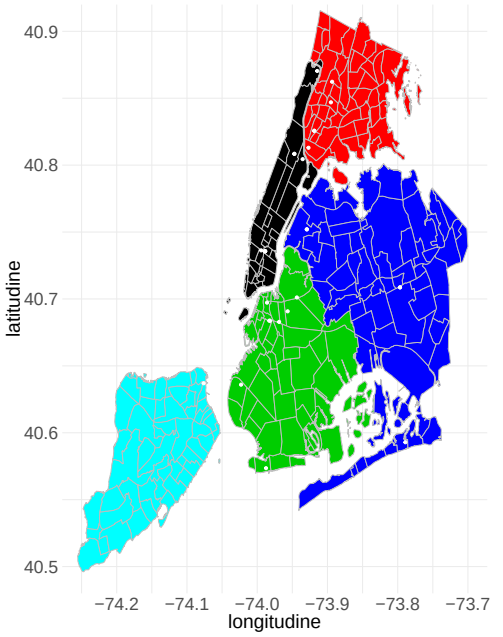


(d) Benzinai

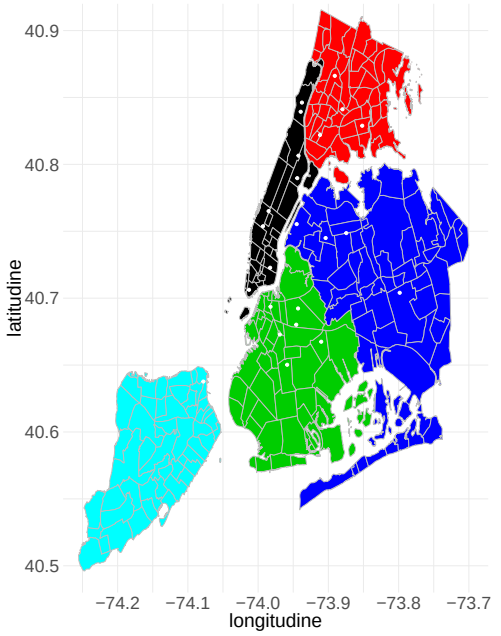
Figura 2.3: Collocazione di luoghi pubblici quali centri commerciali e luoghi di culto



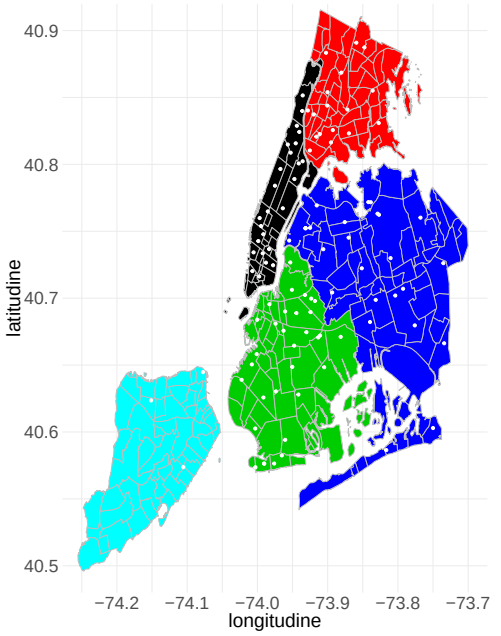
(a) Caserme dei pompieri



(b) Uffici di collocamento

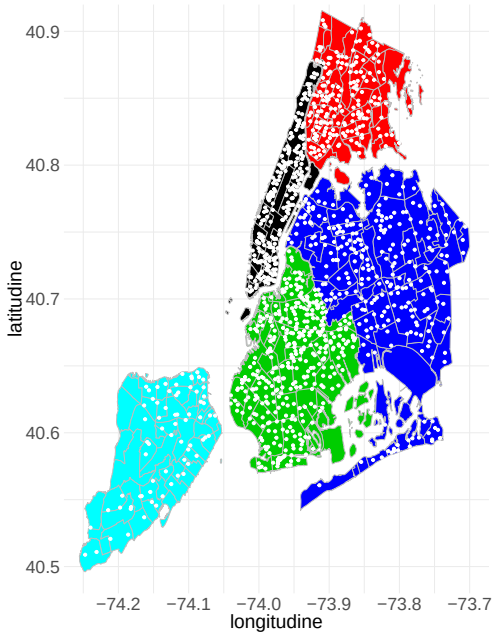


(c) Istituzioni di credito per imprese

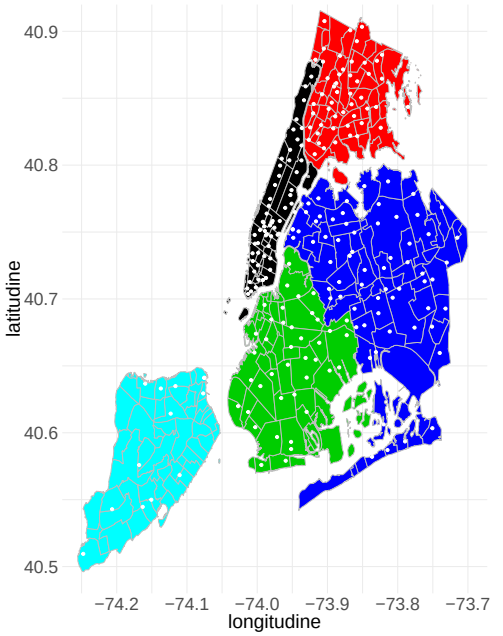


(d) Distretti di polizia

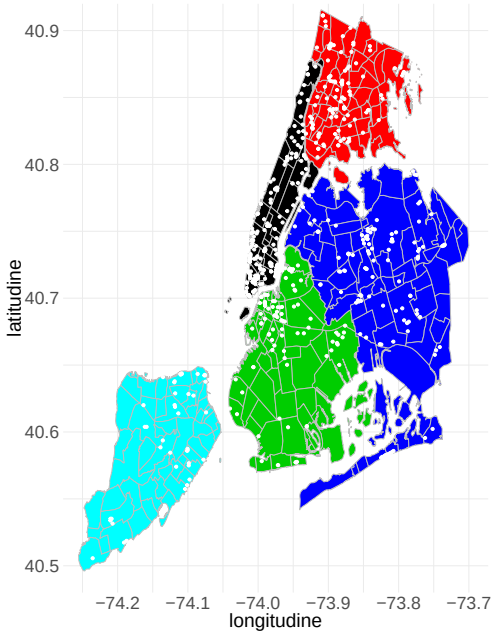
Figura 2.4: Collocazione dei servizi pubblici



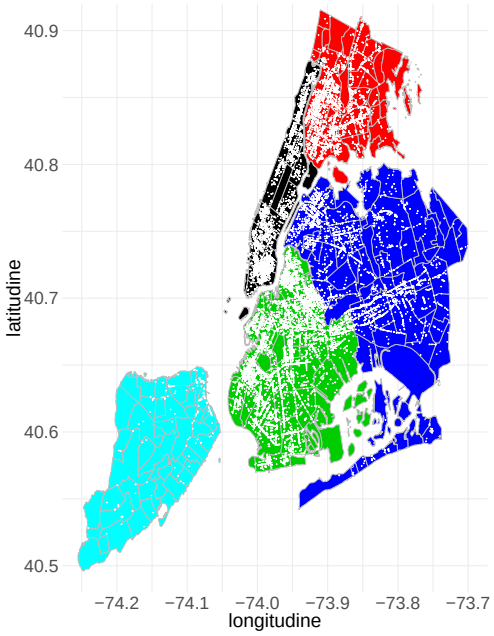
(a) Seggi elettorali



(b) Uffici postali

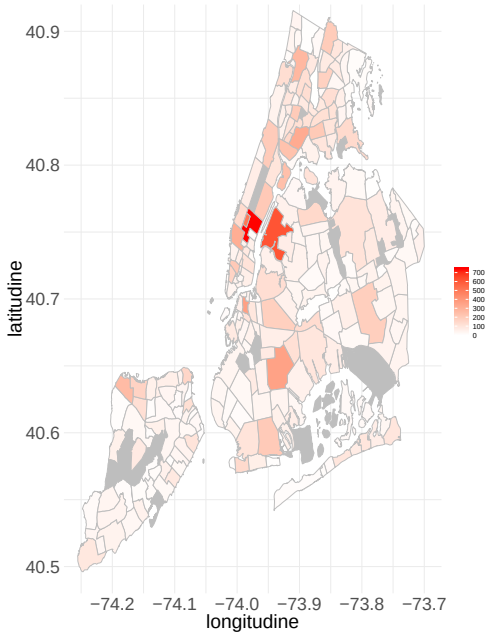


(c) Aree raccolta differenziata

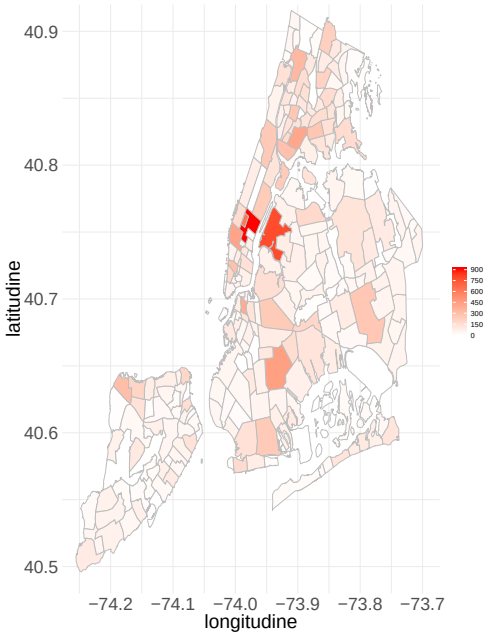


(d) Interventi del servizio di riqualificazione

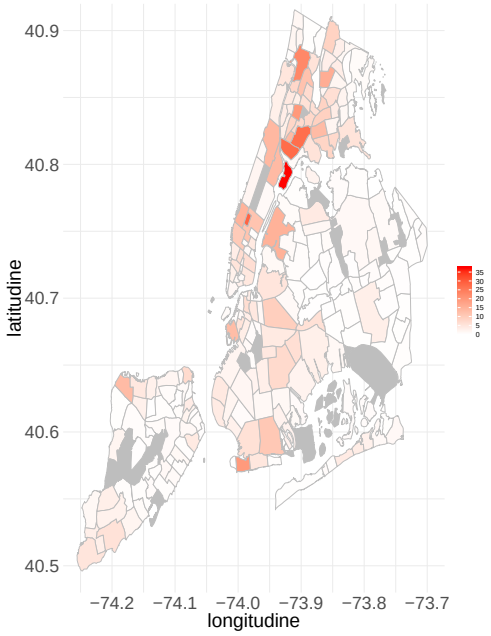
Figura 2.5: Collocazione dei servizi pubblici



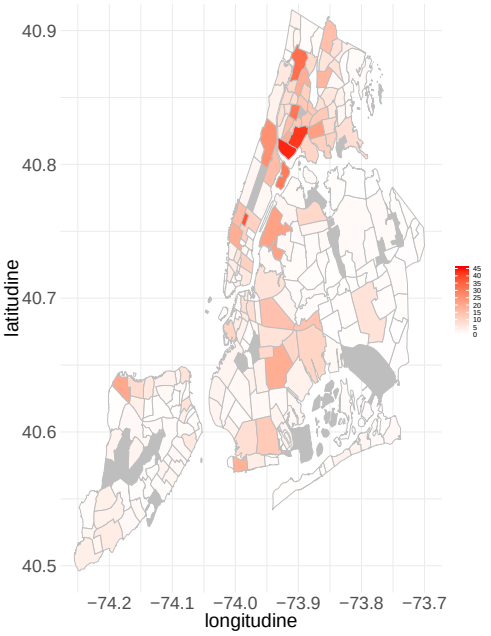
(a) Tasso criminalità 2017 (‰)



(b) Tasso criminalità 2012-2017(‰)

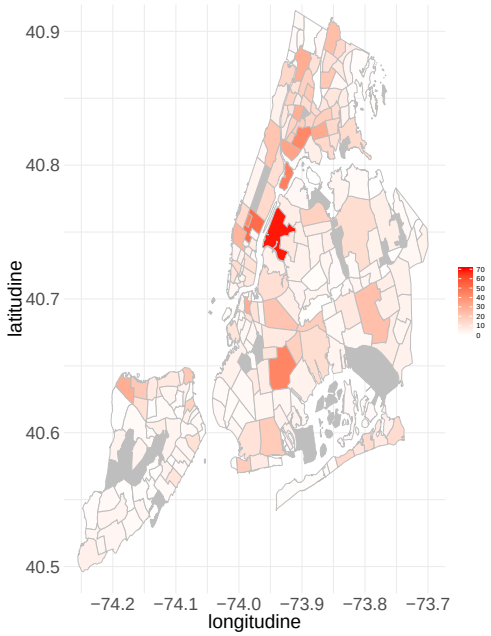


(c) Tasso reati di droga 2017 (‰)

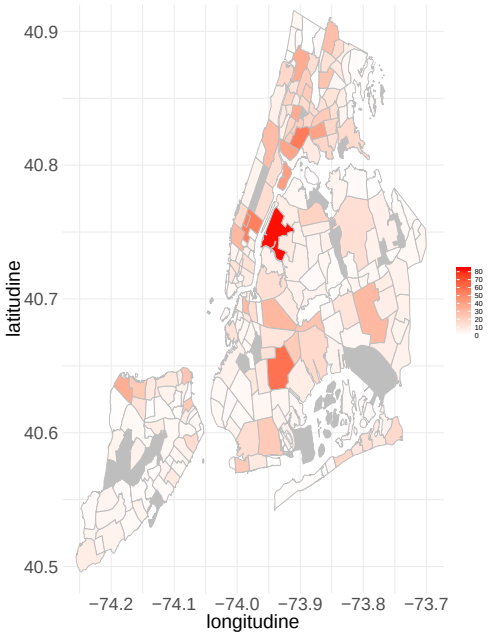


(d) Tasso reati di droga 2012-2017(‰)

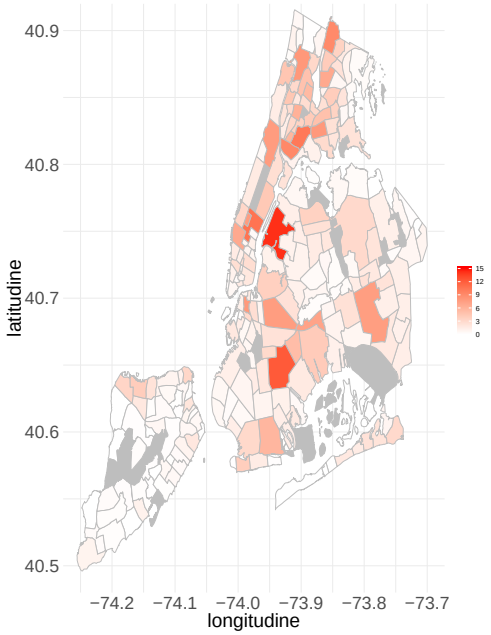
Figura 2.6: Heat map dei reati legati alla droga



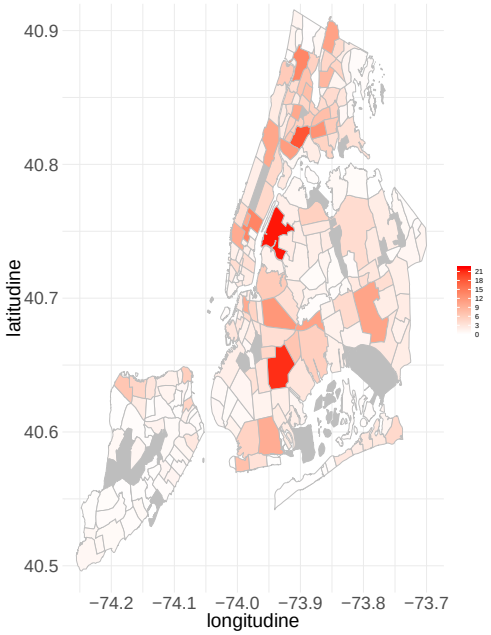
(a) Tasso aggressioni 2017 (‰)



(b) Tasso aggressioni 2012-2017(‰)

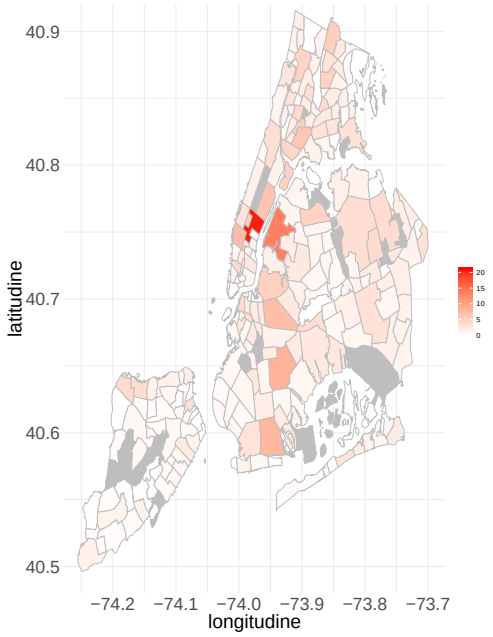


(c) Tasso rapine 2017 (‰)

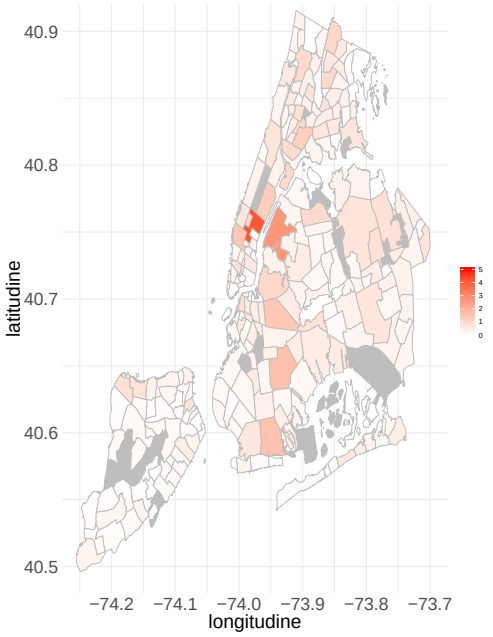


(d) Tasso rapine 2012-2017(‰)

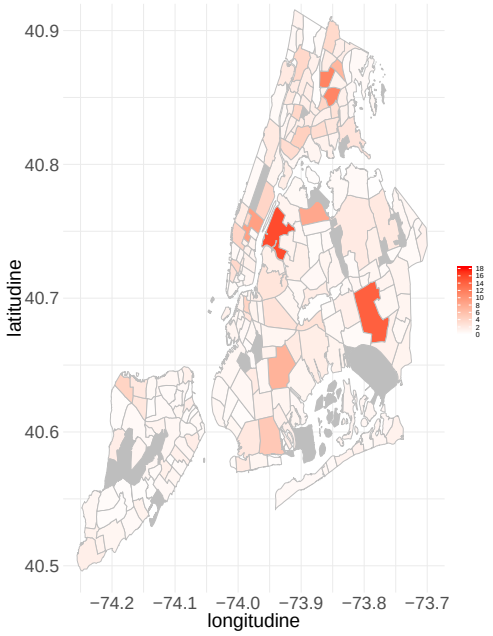
Figura 2.7: Heat map dei reati violenti (Aggressioni e Rapine)



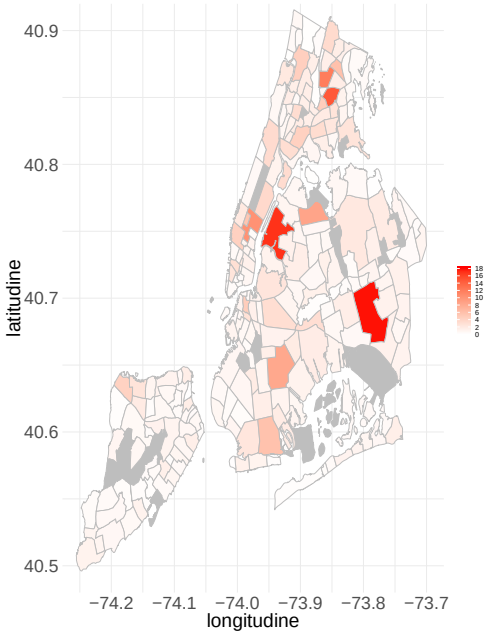
(a) Tasso furti 2017 (‰)



(b) Tasso furti 2012-2017(‰)

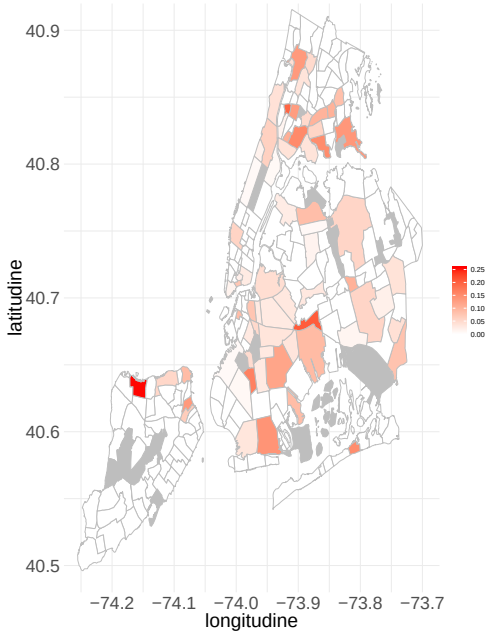


(c) Tasso multe 2017 (‰)

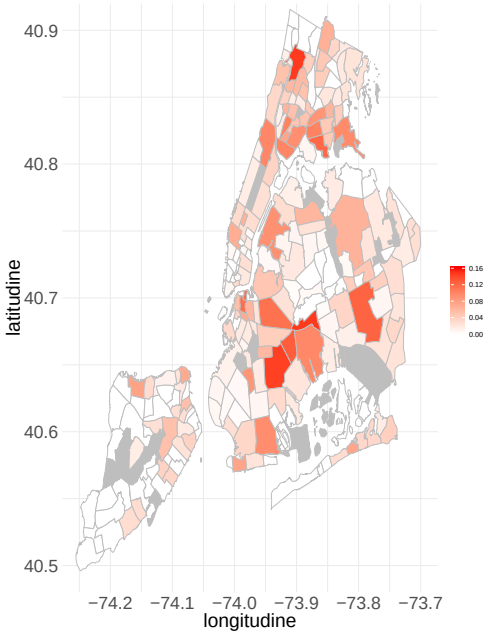


(d) Tasso multe 2012-2017(‰)

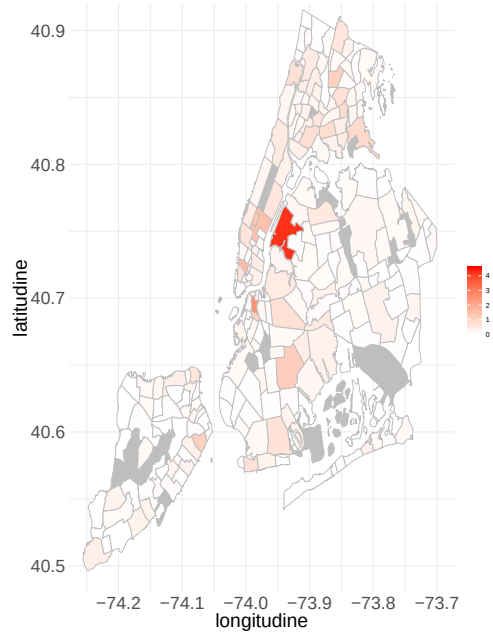
Figura 2.8: Heat map dei reati meno gravi (Furti e Multe)



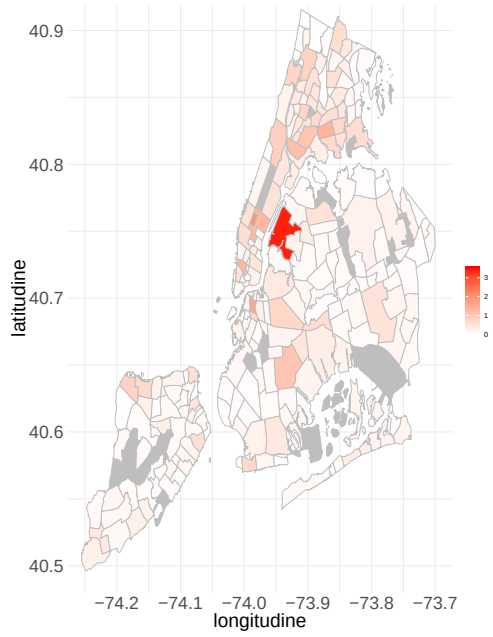
(a) Tasso rapimenti 2017 (‰)



(b) Tasso rapimenti 2012-2017(‰)

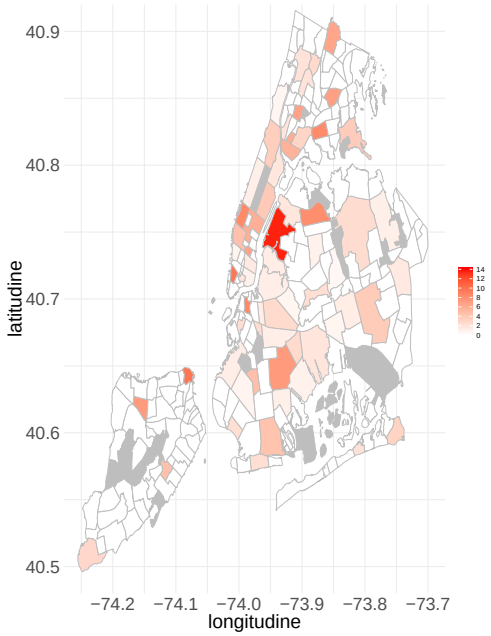


(c) Tasso reati contro la persona 2017 (‰)

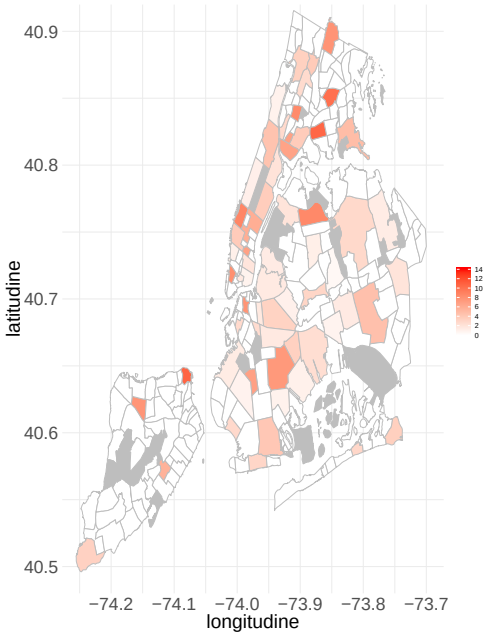


(d) Tasso reati contro la persona 2012-2017(‰)

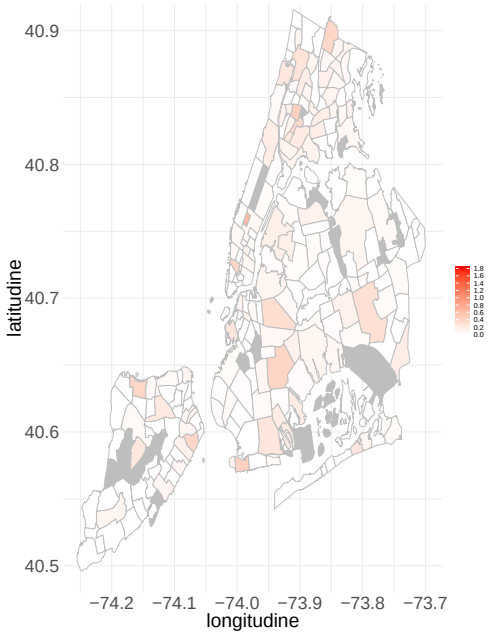
Figura 2.9: Heat map dei reati violenti (rapimenti e reati contro la persona)



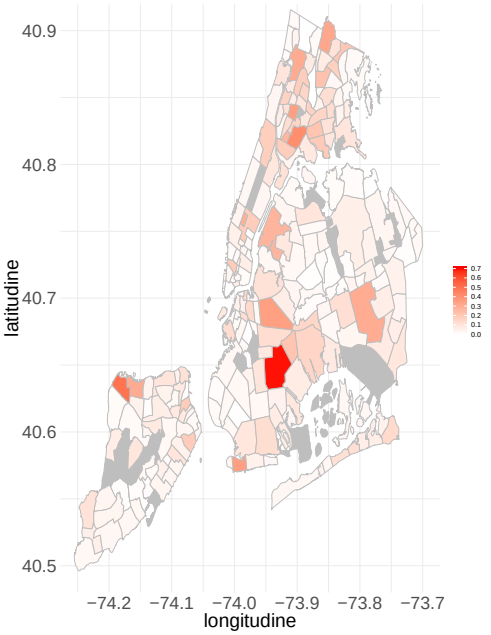
(a) Tasso stupri 2017 (‰)



(b) Tasso stupri 2012-2017(‰)

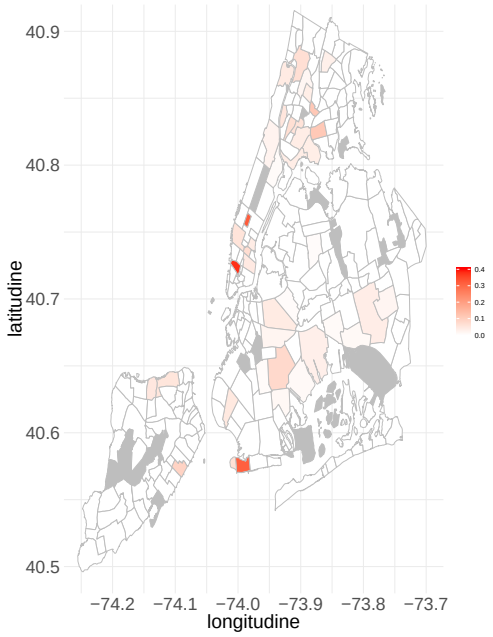


(c) Tasso omicidi 2017 (‰)

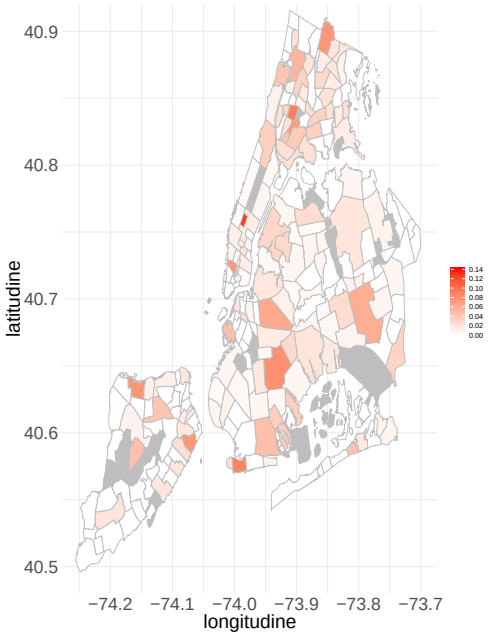


(d) Tasso omicidi 2012-2017(‰)

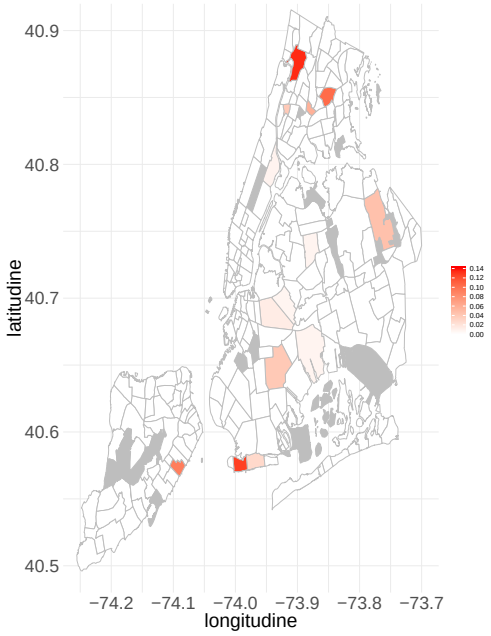
Figura 2.10: Heat map dei reati violenti (violenze sessuali e omicidi)



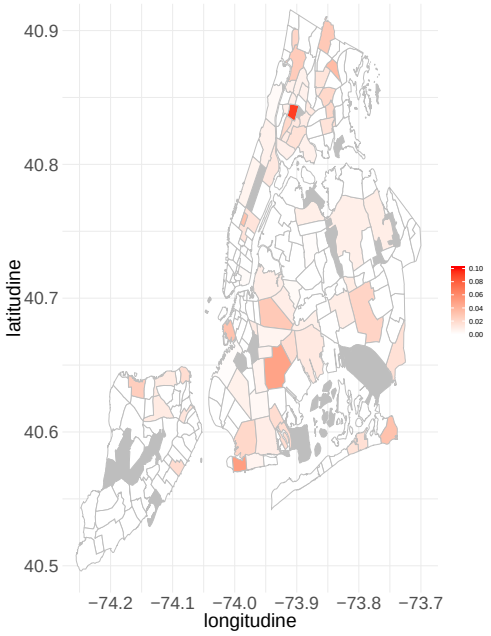
(a) Tasso omicidi donne 2017 (‰)



(b) Tasso omicidi donne 2012-2017(‰)

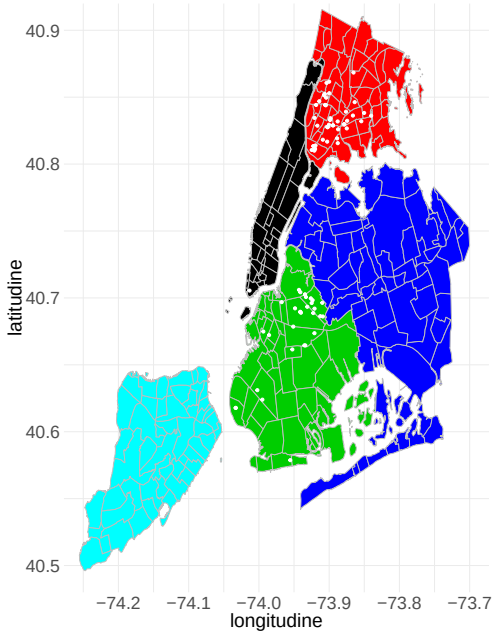


(c) Tasso omicidi bambini 2017 (‰)

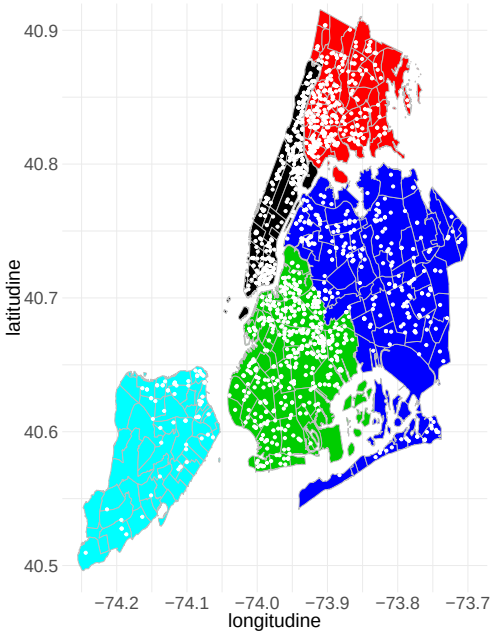


(d) Tasso omicidi bambini 2012-2017(‰)

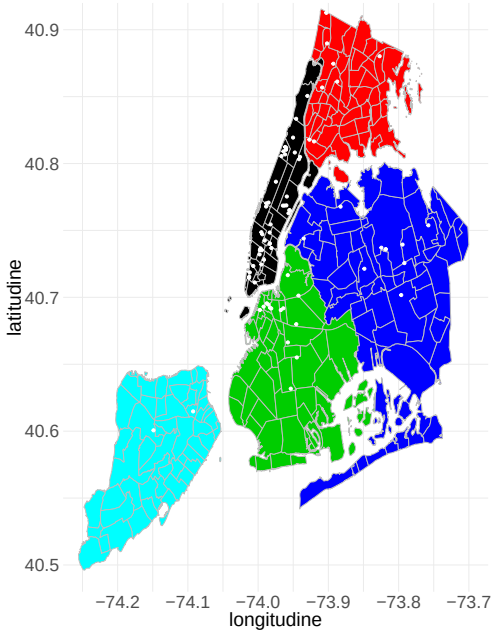
Figura 2.11: Heat map dei reati violenti (omicidi di donne e bambini)



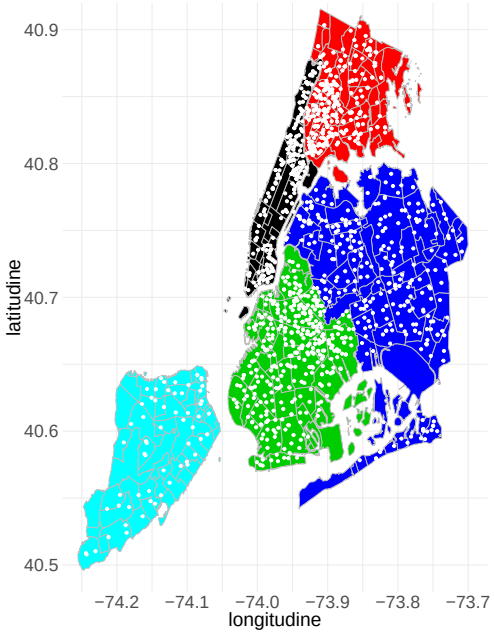
(a) Case dichiarate inagibili dai vigili del fuoco



(b) Dopo scuola

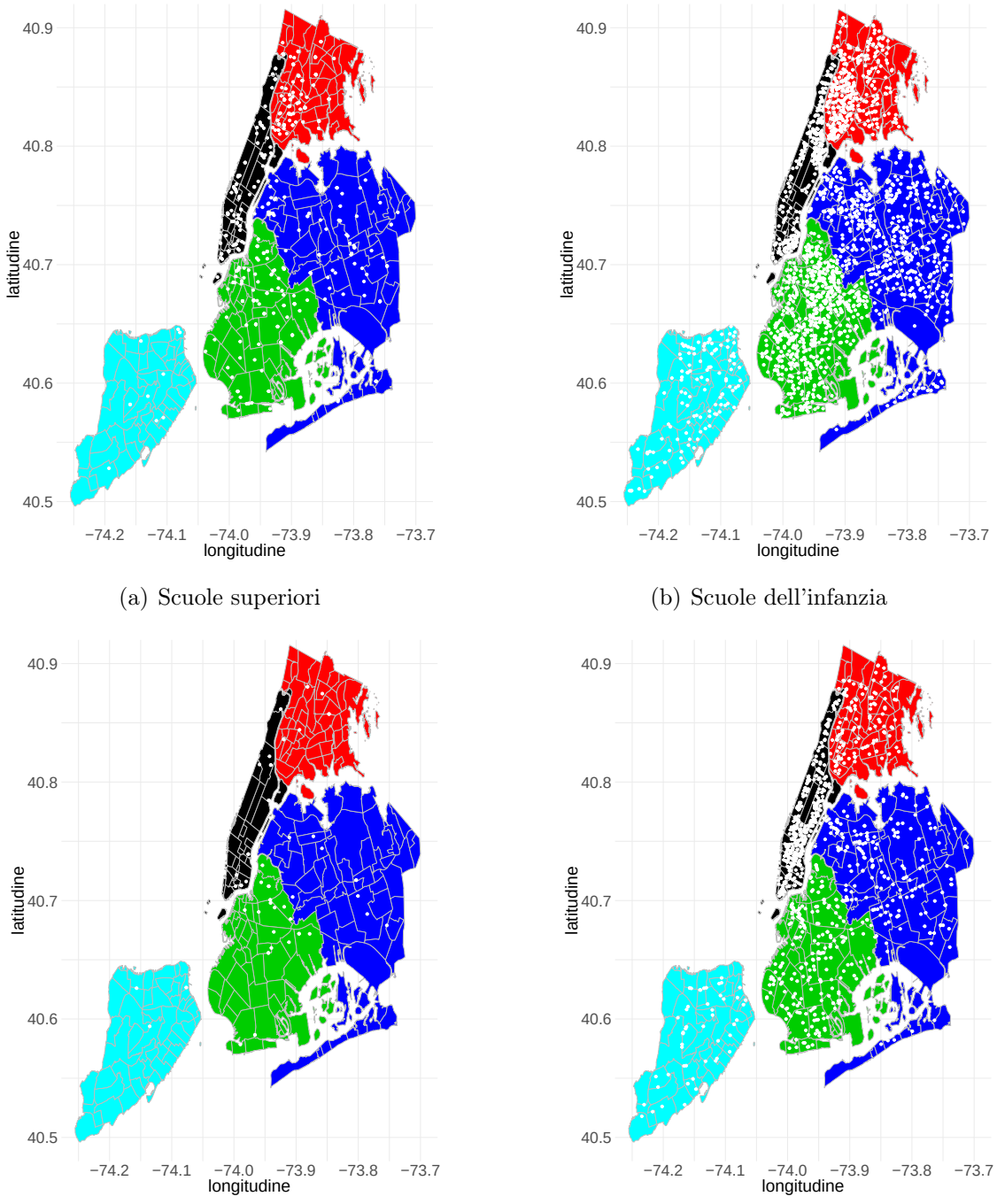


(c) Università e college



(d) Scuole di Central Harlem

Figura 2.12: Collocazione delle scuole



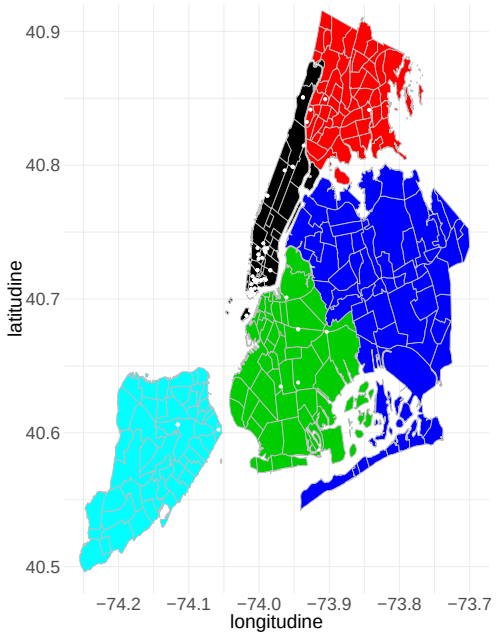
(a) Scuole superiori

(b) Scuole dell'infanzia

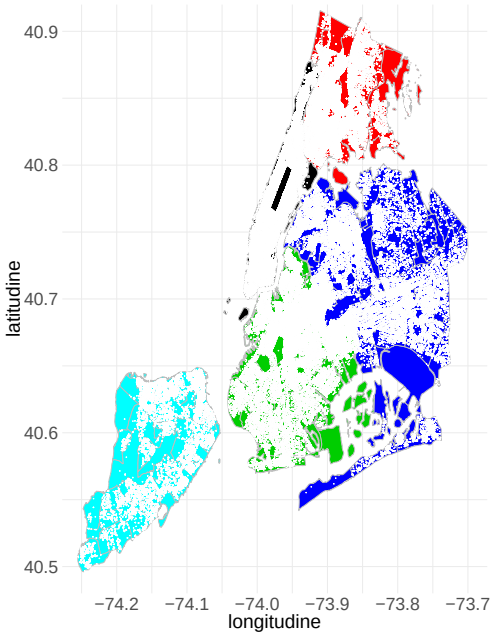
(c) Ospedali

(d) Centri vaccinazioni antinfluenzali

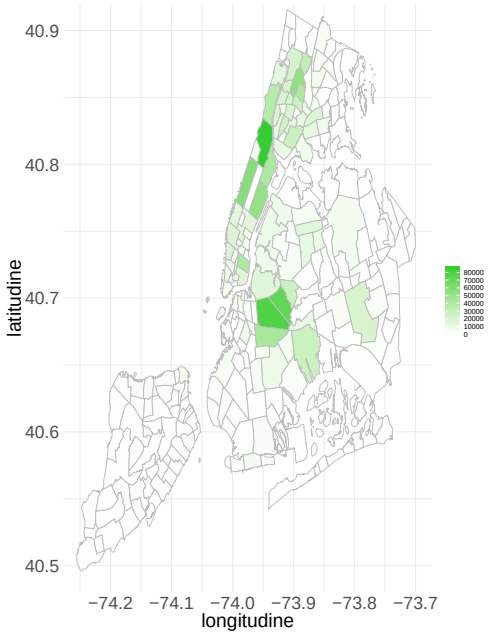
Figura 2.13: Servizi pubblici come scuole e ospedali



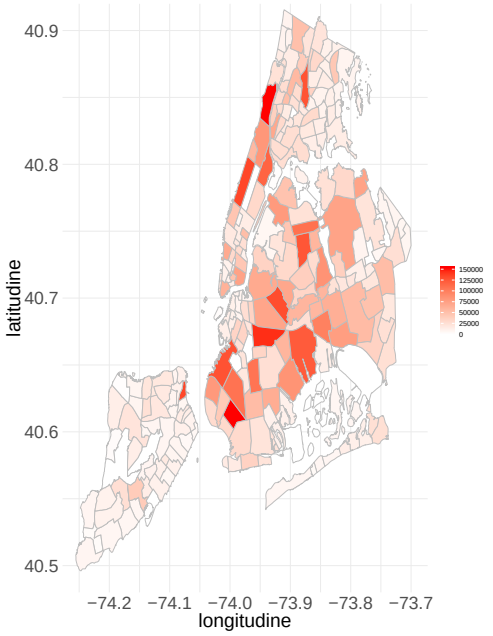
(a) Centri prevenzione HIV



(b) Interventi derattizzazione

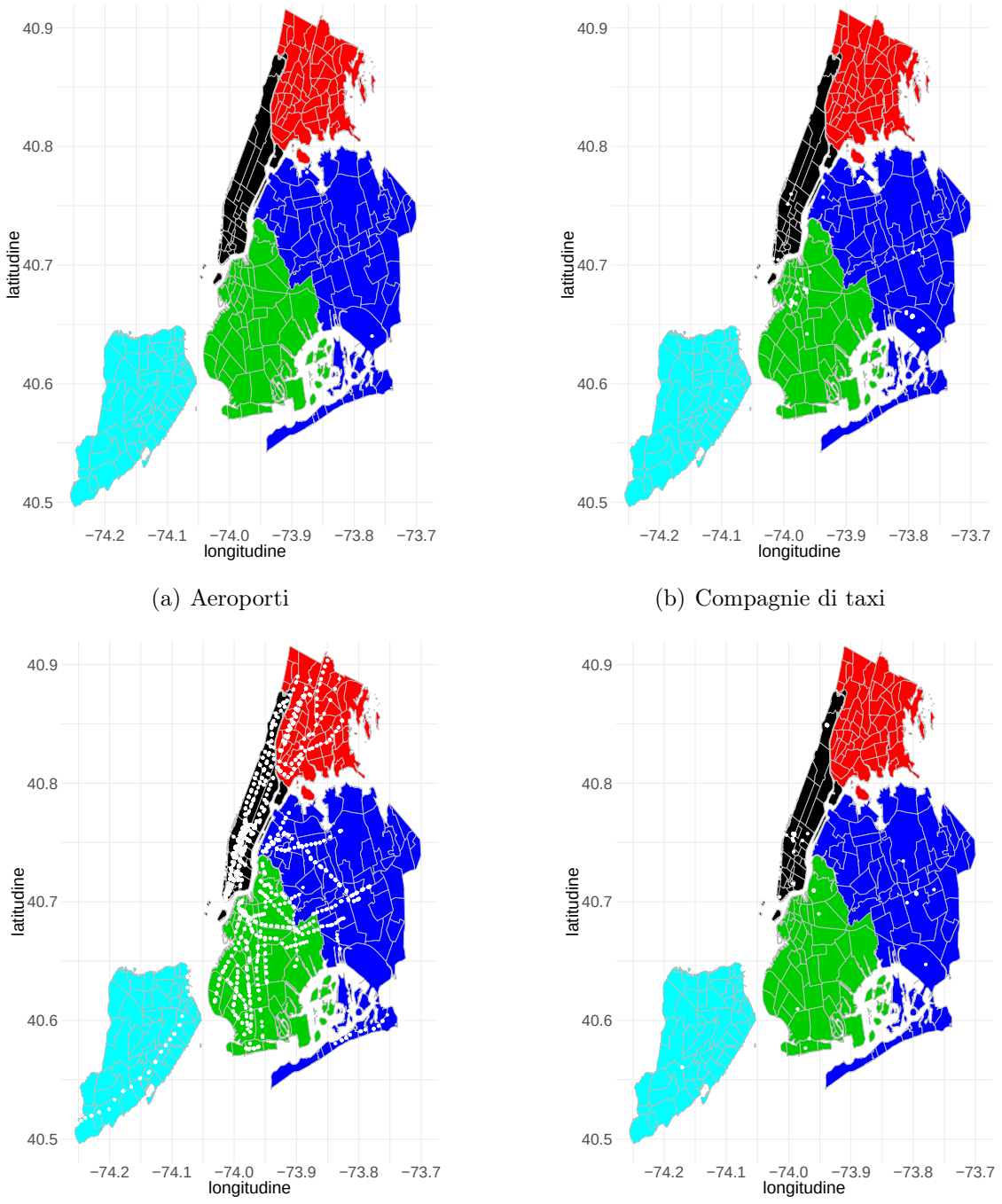


(c) Heat map interventi derattizzazione



(d) Popolazione divisa per sottoquartieri

Figura 2.14: Aree interventi derattizzazione



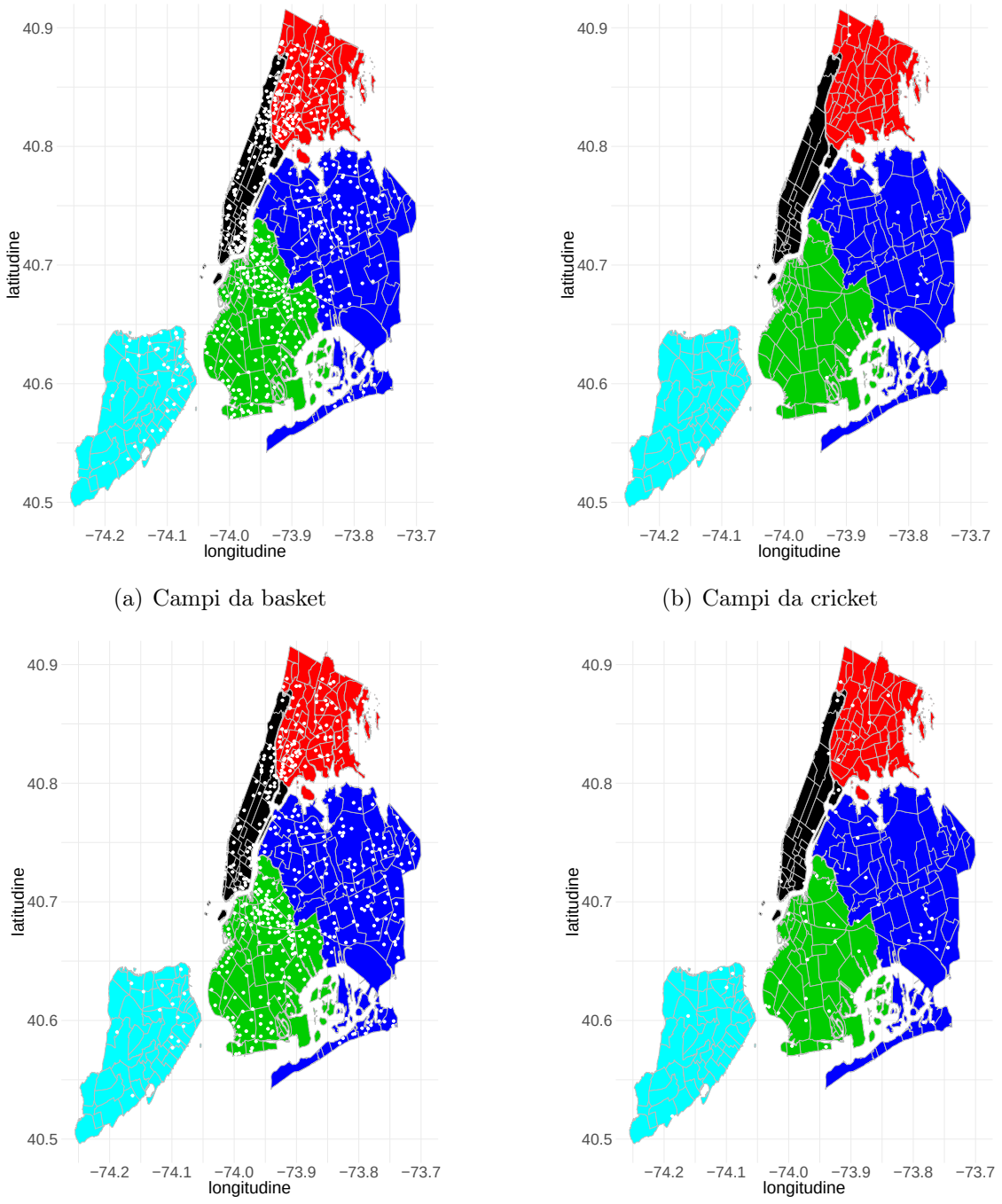
(a) Aeroporti

(b) Compagnie di taxi

(c) Entrate metropolitana

(d) Fermate autobus

Figura 2.15: Collocazione trasporti pubblici



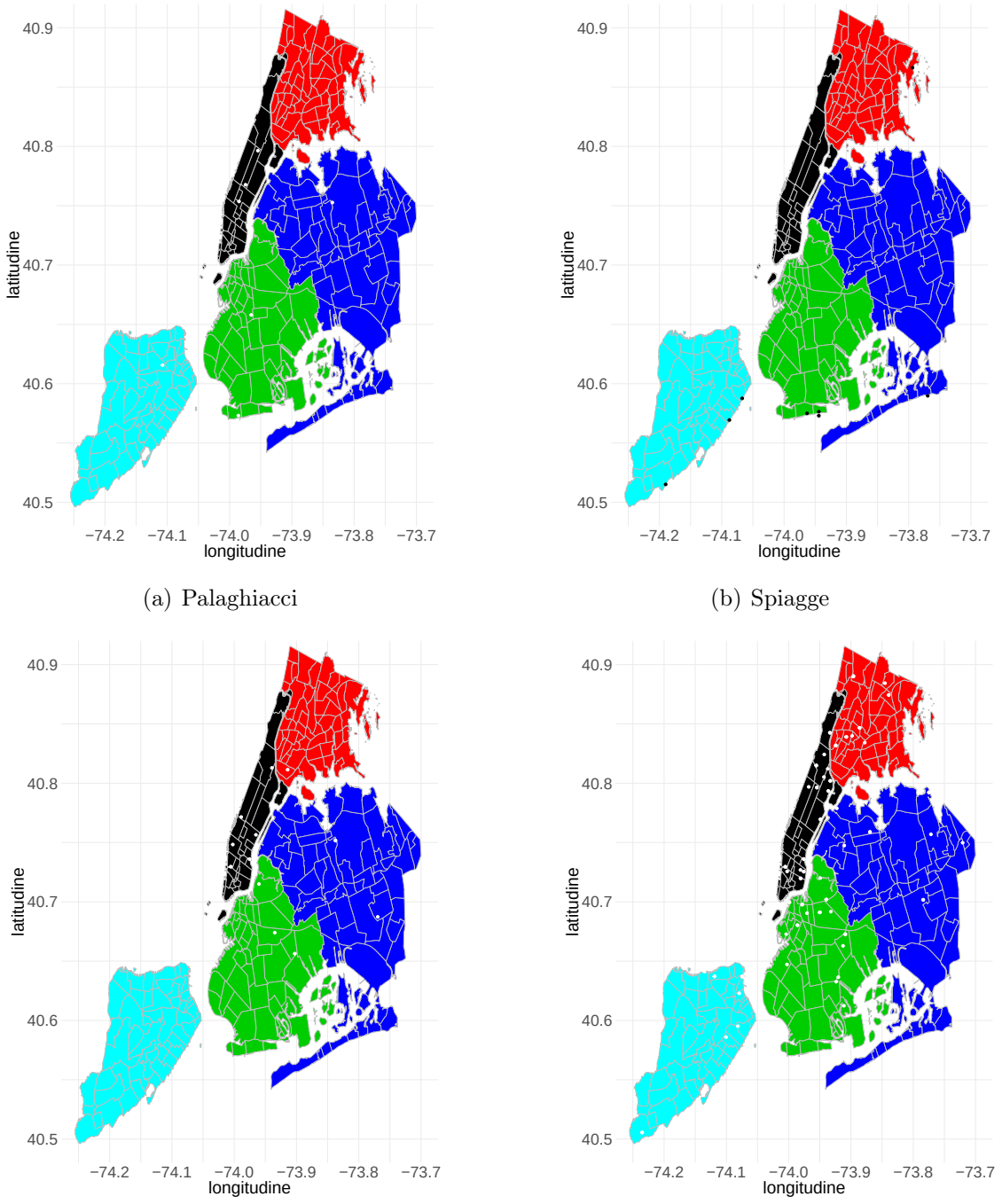
(a) Campi da basket

(b) Campi da cricket

(c) Campi da pallamano

(d) Campi da tennis

Figura 2.16: Collocazione campi da gioco



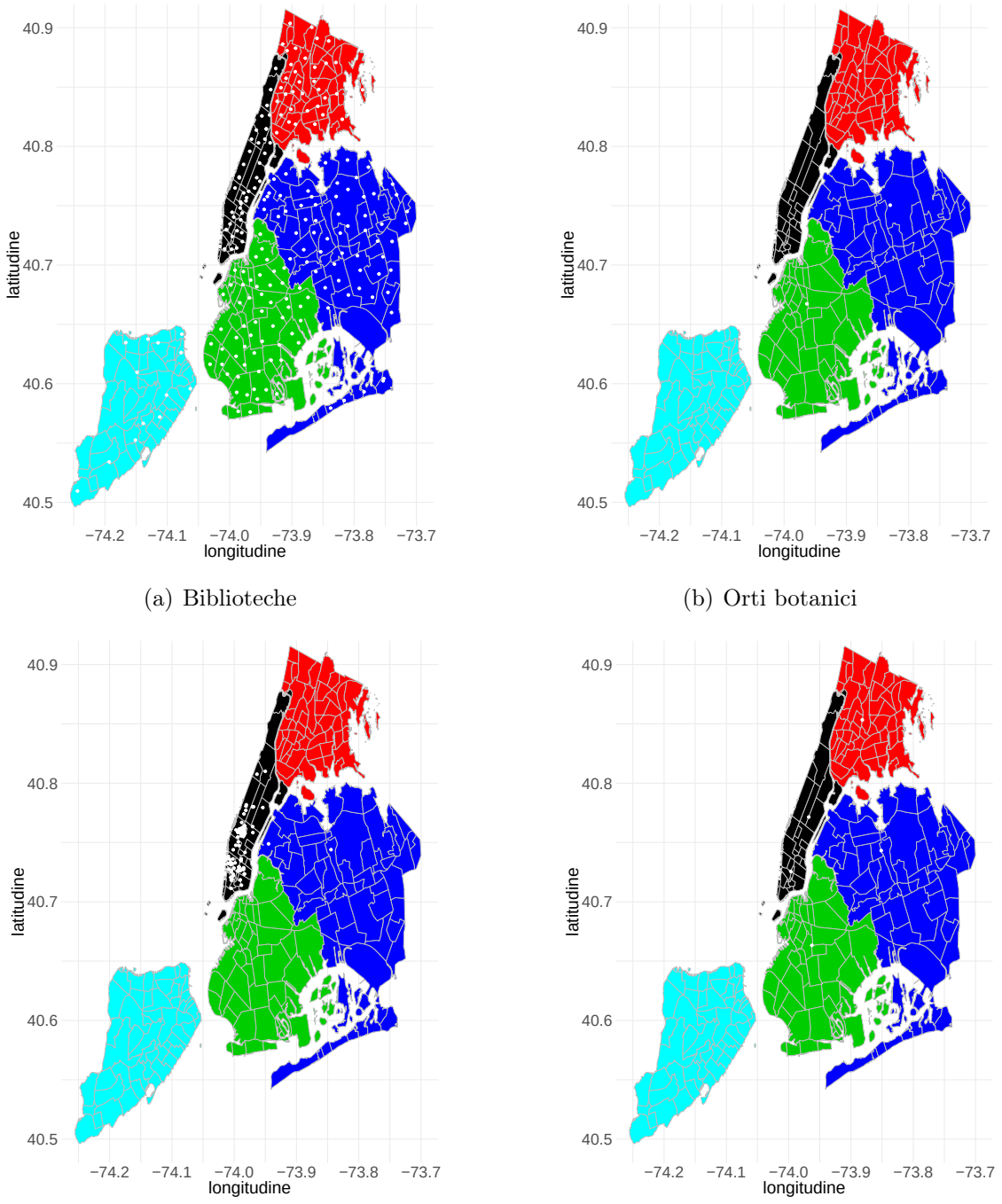
(a) Palaghiacci

(b) Spiagge

(c) Piscine coperte

(d) Piscine scoperte

Figura 2.17: Collocazione piscine e altre aree ludiche



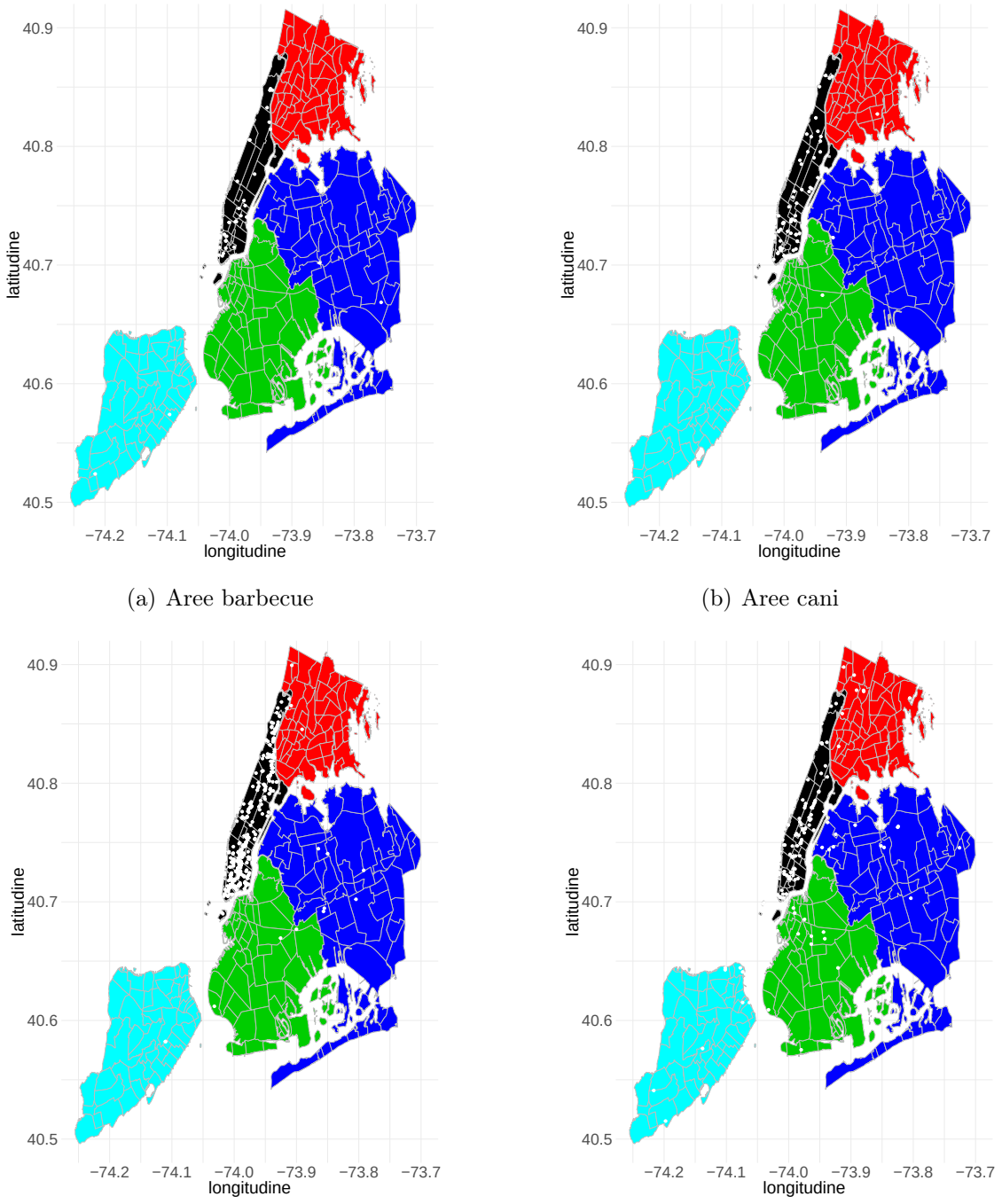
(a) Biblioteche

(b) Orti botanici

(c) Teatri

(d) Zoo

Figura 2.18: Collocazione teatri e aree culturali



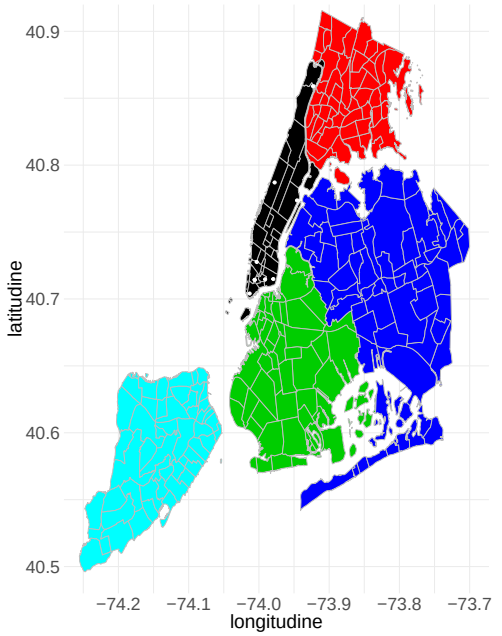
(a) Aree barbecue

(b) Aree cani

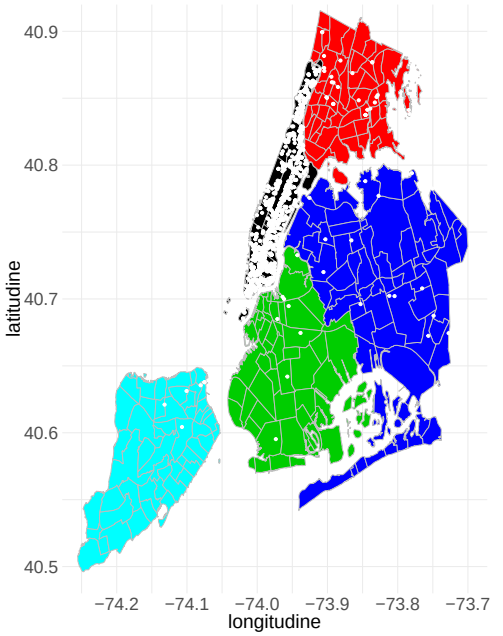
(c) Aree accessibili ai disabili

(d) Musei

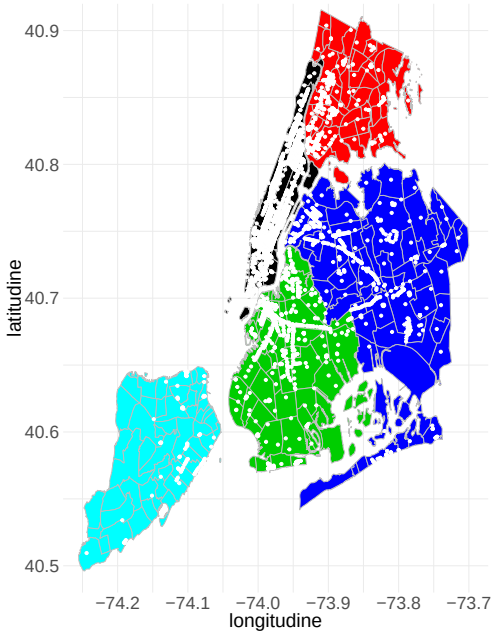
Figura 2.19: Collocazione aree all'aperto



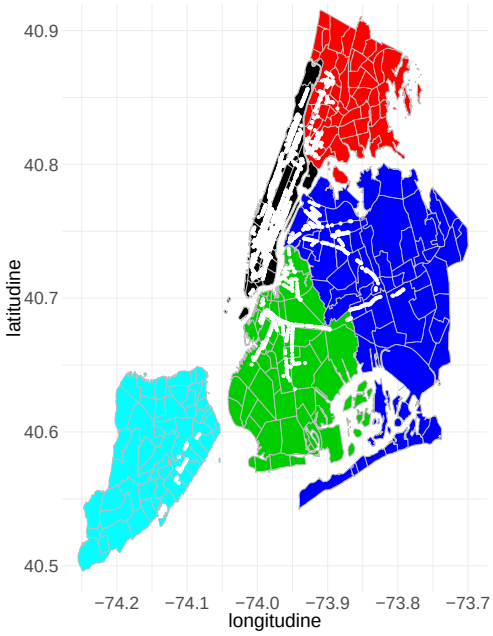
(a) Case storiche



(b) Parchi

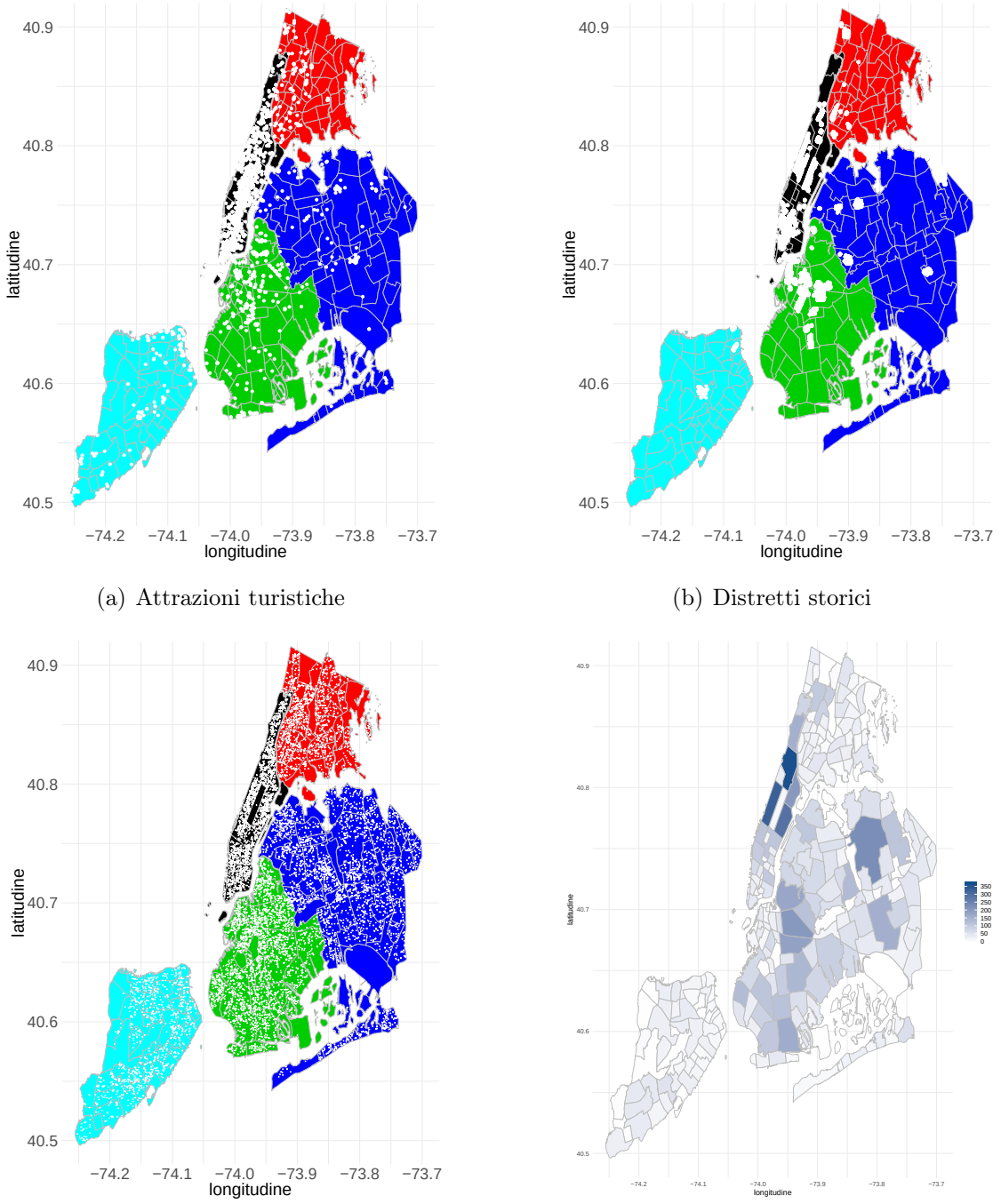


(c) Wifi



(d) Punti linkNYC

Figura 2.20: Collocazione aree con wifi libero



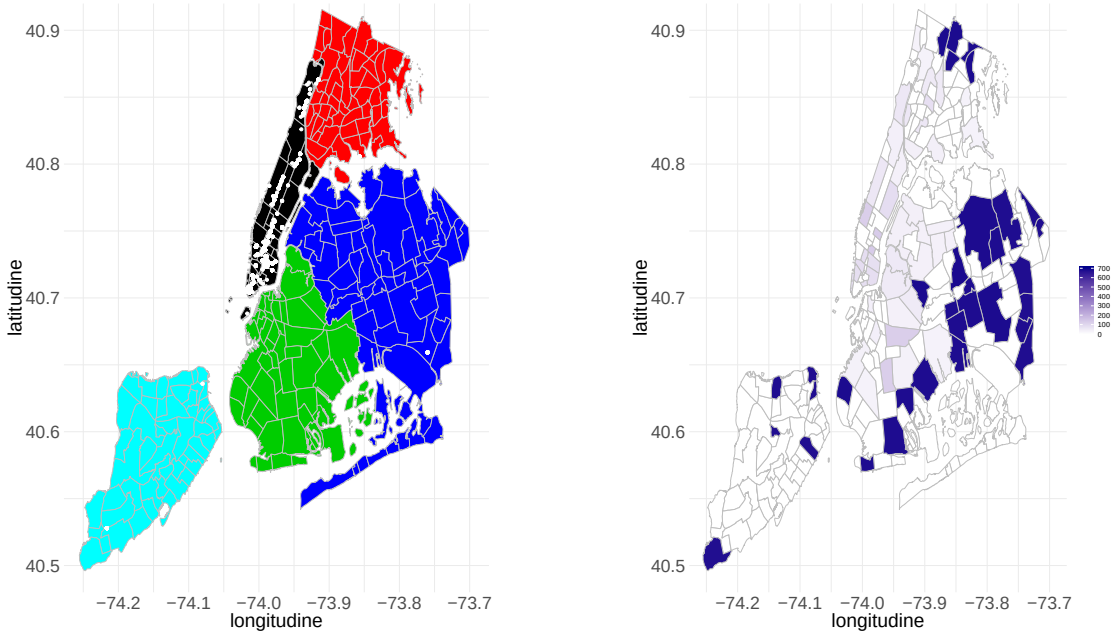
(a) Attrazioni turistiche

(b) Distretti storici

(c) Lamentele sulla qualita dell'acqua

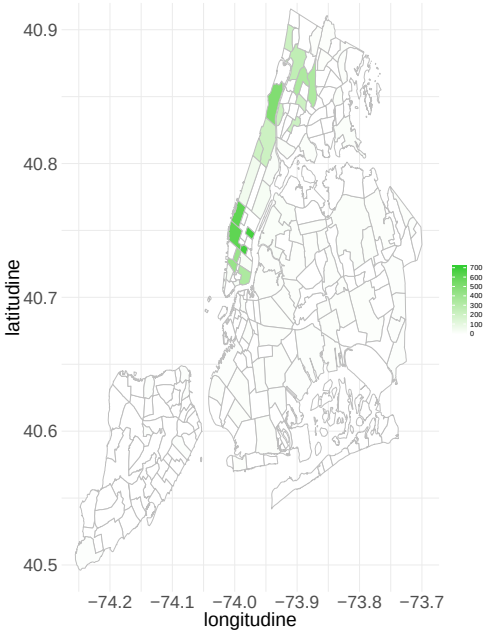
(d) Heatmap lamentele sulla qualita dell'acqua

Figura 2.21: Collocazione attrattive turistiche



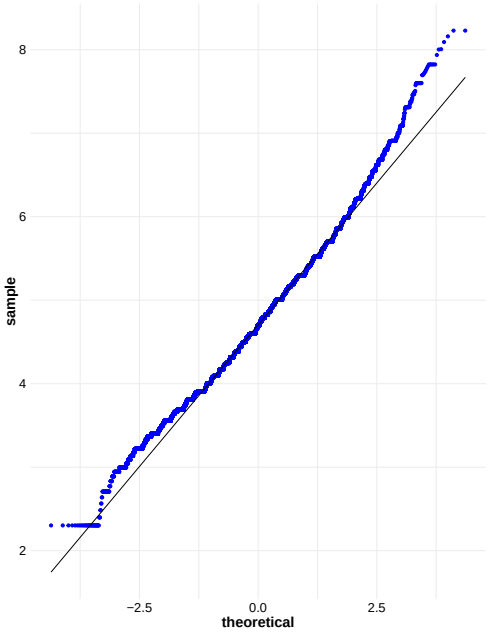
(a) Rilevazioni qualità dell'acqua potabile del 2018

(b) Livello di polveri sottili rilevato

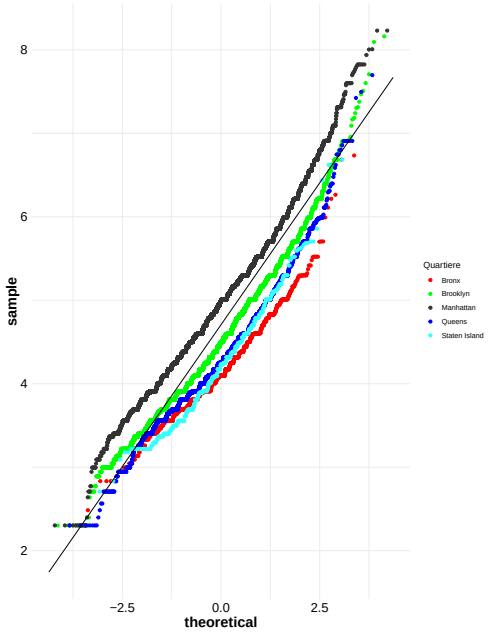


(c) Livello S02 emesso dalle caldaie

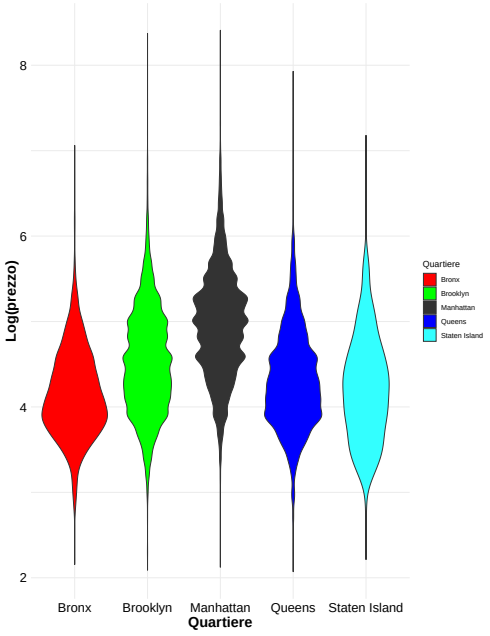
Figura 2.22: Heatmap di alcune sostanze inquinanti presenti nell'aria



(a) Qqplot log-prezzo degli affitti Airbnb



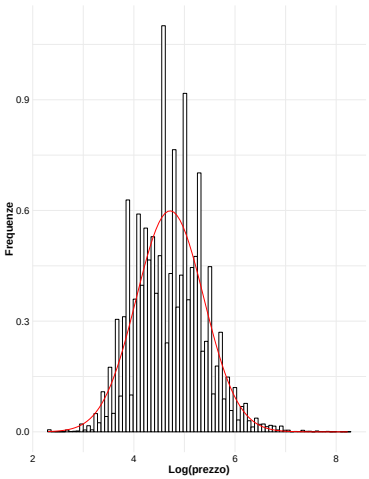
(b) Qqplot log-prezzo degli affitti Airbnb diviso per quartiere



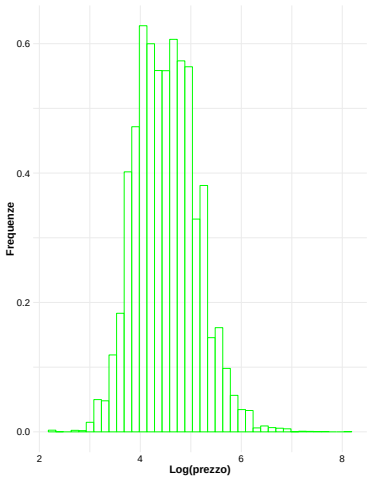
(c) Violin-plot del log-prezzo degli affitti Airbnb per quartiere

Figura 2.23: Grafici che rappresentano la distribuzione dei log-prezzi degli affitti di Airbnb

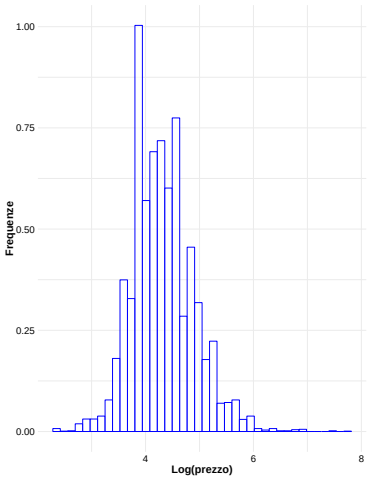
che l'affitto per notte di un appartamento sia i più alto di quello di una stanza condivisa. Generalmente la categoria che sembra avere più variabilità è *Entire home*, tranne nel caso del Bronx e di Queens, *Shared room* è la categoria che presenta maggiore variabilità.
Le figure da 2.1 a 2.22 mostrano le altre variabili presenti nel set di dati descritto nella tabella 2.48. Si è scelto di presentarle utilizzando delle mappe, invece che con tabelle delle loro statistiche descrittive o grafici di altro tipo (per esempio, istogrammi o boxplot) per avere una rappresentazione visivamente più chiara possibile e che rispettasse la loro natura di dati spaziali.



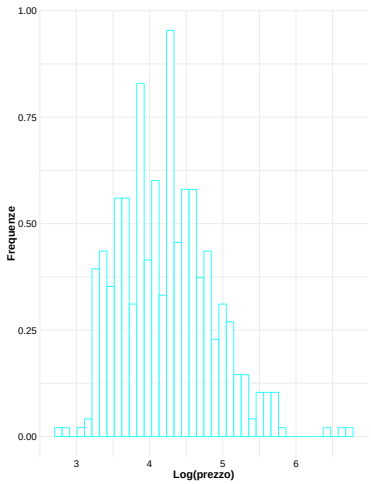
(a) Complessiva



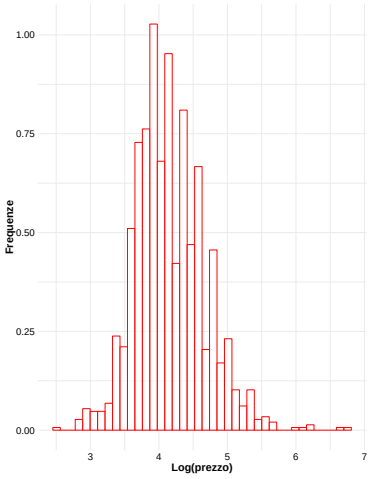
(b) Brooklyn



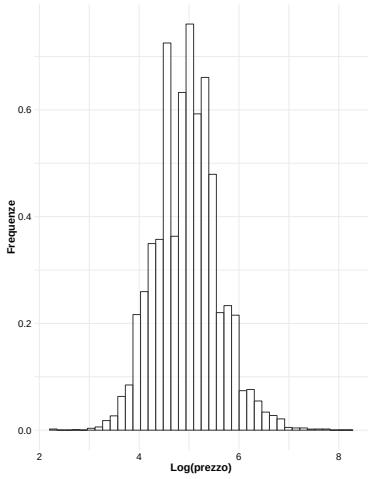
(c) Queens



(d) Staten Island

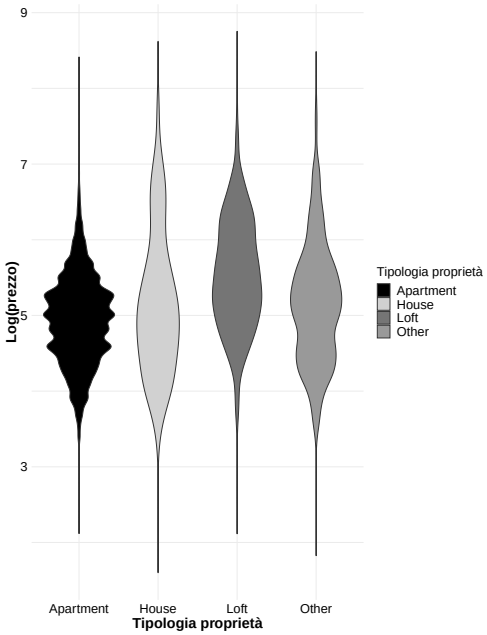


(e) Istogramma log-prezzo affitto Airbnb Bronx

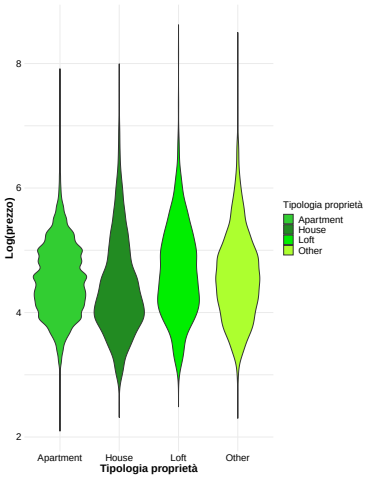


(f) Manhattan

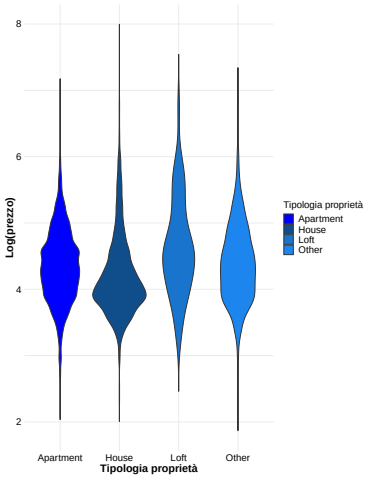
Figura 2.24: istogrammi log-prezzi Airbnb per quartiere



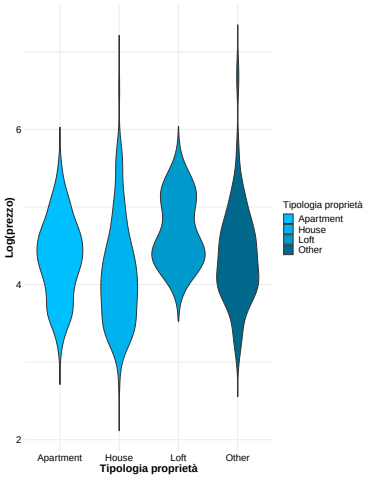
(a) Manhattan



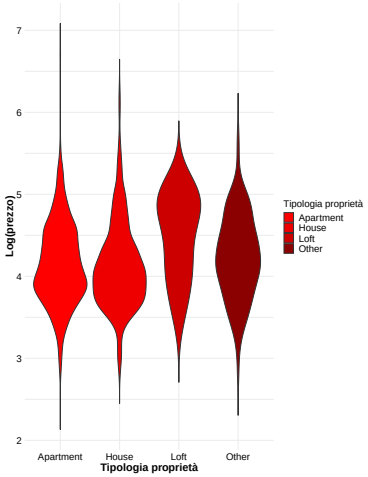
(b) Brooklyn



(c) Queens

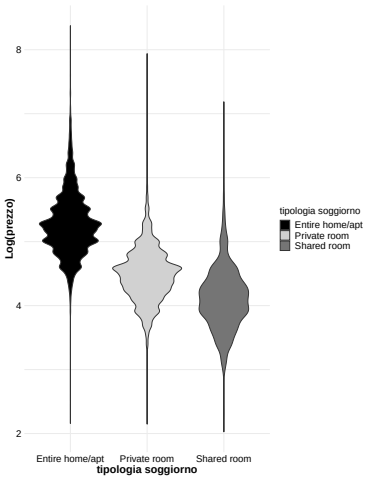


(d) Staten Island

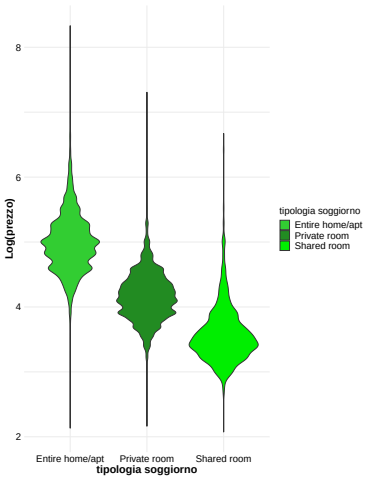


(e) Bronx

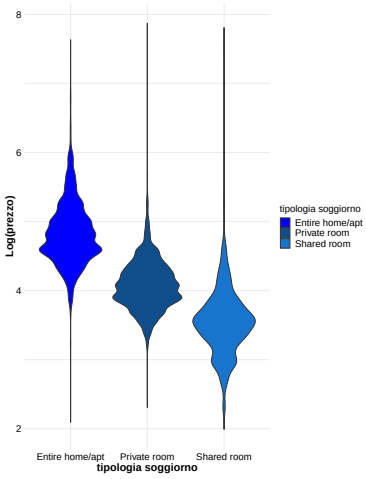
Figura 2.25: Violin-plot log-prezzo affitti per tipologia proprietà



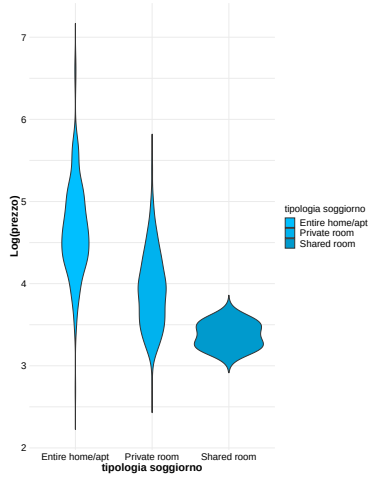
(a) Manhattan



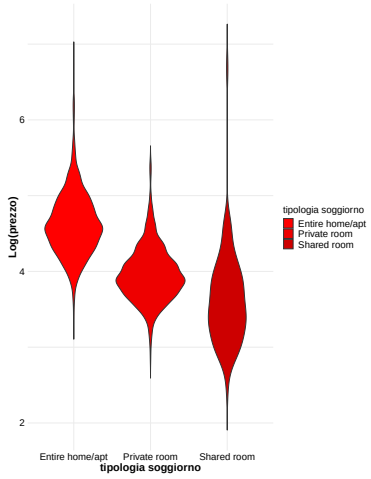
(b) Brooklyn



(c) Queens



(d) Staten Island



(e) Bronx

Figura 2.26: Violin-plot log-prezzo affitti per tipologia di soggiorno per quartiere

	Nome variabile	Tipologia variabile	Descrizione variabile
1	id	numerica	identificativo appartamento
2	last scraped	data	data dell'ultimo scraping
3	host id	numerica	identificativo dell'host
4	host response time	categoriale	tempo di risposta host
5	host is superhost	categoriale	1, se host è "super host", 0 altrimenti
6	host identity verified	categoriale	1, se l'identità dell'host è certificata, 0 altrimenti
7	neighbourhood cleansed	categoriale	sottoquartiere
8	neighbourhood group cleansed	categoriale	quartiere
9	latitude	numerica	latitudine
10	longitude	numerica	longitudine
11	property type	categoriale	tipologia di proprietà
12	room type	categoriale	tipologia di stanza
13	accommodates	numerica	numero totale di stanze
14	bathrooms	numerica	numeo totale di bagni
15	bedrooms	numerica	numero totale di camere da letto
16	beds	numerica	numero totale di letti
17	bed type	categoriale	tipo di letti
18	price	numerica	prezzo in USD per notte risalente all'ultima data di scraping dell'annuncio disponibile
19	guests included	numerica	è il numero di ospiti extra sono inclusi nella prenotazione, in questo caso per il soggiorno si paga il prezzo per l'affitto indicato da *price* senza supplementi.
20	extra people	numerica	prezzo in USD per portare un ospite in più, oltre a quelli già inclusi
21	minimum nights	numerica	lunghezza minima del soggiorno (in notti)
22	maximum nights	numerica	lunghezza massima del soggiorno (in notti)
23	number of reviews	numerica	numero di recensioni
24	review scores rating	numerica	punteggio medio recensioni
25	review scores accuracy	numerica	punteggio medio accuratezza descrizione proprietà fornita dall'host
26	review scores cleanliness	numerica	punteggio medio pulizia
27	review scores checkin	numerica	punteggio medio checkin
28	review scores communication	numerica	punteggio medio comunicazione
29	review scores location	numerica	punteggio medio location
30	review scores value	numerica	punteggio medio dato al prezzo di affitto
31	instant bookable	categoriale	1, se è possibile prenotare senza preavviso, 0 altrimenti
32	cancellation policy	categoriale	politica di cancellazione della prenotazione
33	require guest profile picture	categoriale	1, se l'host richiede che il guest abbia una foto profilo, 0 altrimenti
34	require guest phone verification	categoriale	1, se l'host richiede un recapito telefonico del guest, 0 altrimenti
35	reviews per month	numerica	numero di recensioni mensili
36	lp	numerica	logaritmo del prezzo in USD
37	dog-areas	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un'area dove portare a passeggio i cani, 0 altrimenti
38	barbecue-areas	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un'area barbecue, 0 altrimenti
39	basket	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un campo da basket, 0 altrimenti
40	cricket	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un campo da cricket, 0 altrimenti
41	beach	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno una spiaggia, 0 altrimenti
42	handball	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un campo da pallamano, 0 altrimenti
43	tennis	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un campo da tennis, 0 altrimenti
44	ice-rink	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un palaghiaccio, 0 altrimenti
45	indoor-pool	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno una piscina coperta, 0 altrimenti
46	outdoor-pool	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno una piscina scoperta, 0 altrimenti
47	disability areas	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un'area accessibili ai disabili, 0 altrimenti
48	library	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno una biblioteca, 0 altrimenti
49	museum	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un museo, 0 altrimenti
50	historical-houses	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno una casa storica, 0 altrimenti
51	botanical-garden	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un orto botanico, 0 altrimenti
52	zoo	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno uno zoo, 0 altrimenti
53	theatres	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un teatro, 0 altrimenti
54	university	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un'università, 0 altrimenti
55	hospital	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un ospedale, 0 altrimenti
56	starbucks	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un locale di Starbucks, 0 altrimenti
57	police district	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un distretto di polizia, 0 altrimenti
58	firefighter	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno una caserma dei pompieri, 0 altrimenti
59	post-office	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un ufficio postale, 0 altrimenti
60	credit	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un ufficio di credito, 0 altrimenti
61	shopping centre	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un centro commerciale, 0 altrimenti
62	job centers	categoriale	1, se nel sottoquartiere dell'appartamento c'è almeno un ufficio di collocamento, 0 altrimenti
63	d-individual	numerica	distanza tra l'appartamento e l'attrazione turistica pi'ù vicina
64	d-interior	numerica	distanza tra l'appartamento e l'attrazione turistica interna pi'ù vicina
65	d-scenic	numerica	distanza tra l'appartamento e il punto panoramico pi'ù vicino
66	d-parks	numerica	distanza tra l'appartamento e il parco più vicino
67	d-metro	numerica	distanza tra l'appartamento e la fermata della metropolitana più vicina
68	d-luoghi-culto	numerica	distanza tra l'appartamento e il luogo di culto più vicino
69	d-schools	numerica	distanza tra l'appartamento e la scuola più vicina
70	d-fast-food	numerica	distanza tra l'appartamento e il fast food più vicino
71	d-hotel	numerica	distanza tra l'appartamento e l'hotel più vicino
72	d-ristoranti	numerica	distanza tra l'appartamento e il ristorante più vicino
73	population	numerica	popolazione del sottoquartiere in cui si trova l'appartamento
74	aggressioni-500	numerica	numero di aggressioni nel raggio di 500 metri dell'appartamento pesato con population
75	aggressioni-1000	numerica	numero di aggressioni nel raggio di 1000 metri dell'appartamento pesato con population
76	aggressioni-1500	numerica	numero di aggressioni nel raggio di 1500 metri dell'appartamento pesato con population
77	droga-500	numerica	numero reati di droga nel raggio di 500 metri dell'appartamento pesato con population
78	droga-1000	numerica	numero di reati di droga nel raggio di 1000 metri dell'appartamento pesato con population
79	droga-1500	numerica	numero reati di droga nel raggio di 1500 metri dell'appartamento pesato con population
80	furti-500	numerica	numero di furti nel raggio di 500 metri dell'appartamento pesato con population
81	furti-1000	numerica	numero di furti nel raggio di 1000 metri dell'appartamento pesato con population
82	furti-1500	numerica	numero di furti nel raggio di 1500 metri dell'appartamento pesato con population
83	rapine-500	numerica	numero di rapine nel raggio di 500 metri dell'appartamento pesato con population
84	rapine-1000	numerica	numero di rapine nel raggio di 1000 metri dell'appartamento pesato con population
85	rapine-1500	numerica	numero di rapine nel raggio di 1500 metri dell'appartamento pesato con population
86	reati-contro-la-persona-500	numerica	numero di reati contro la persona nel raggio di 500 metri dell'appartamento pesato con population
87	reati-contro-la-persona-1000	numerica	numero di reati contro la persona nel raggio di 1000 metri dell'appartamento pesato con population
88	reati-contro-la-persona-1500	numerica	numero di reati contro la persona nel raggio di 1500 metri dell'appartamento pesato con population
89	violenza-sessuale-500	numerica	numero di violenze sessuali nel raggio di 500 metri dell'appartamento pesato con population
90	violenza-sessuale-1000	numerica	numero di violenze sessuali nel raggio di 1000 metri dell'appartamento pesato con population
91	violenza-sessuale-1500	numerica	numero di violenze sessuali nel raggio di 1500 metri dell'appartamento pesato con population
92	multe-500	numerica	numero di multe nel raggio di 500 metri dell'appartamento pesato con population
93	multe-1000	numerica	numero di multe nel raggio di 1000 metri dell'appartamento pesato con population
94	multe-1500	numerica	numero di multe nel raggio di 1500 metri dell'appartamento pesato con population
95	omicidi-500	numerica	numero di omicidi nel raggio di 500 metri dell'appartamento pesato con population
96	omicidi-1000	numerica	numero di omicidi nel raggio di 1000 metri dell'appartamento pesato con population
97	omicidi-1500	numerica	numero di omicidi nel raggio di 1500 metri dell'appartamento pesato con population
98	omicidi-donne-500	numerica	numero di omicidi con vittime donne nel raggio di 500 metri dell'appartamento pesato con population
99	omicidi-donne-1000	numerica	numero di omicidi con vittime donne nel raggio di 1000 metri dell'appartamento pesato con population
100	omicidi-donne-1500	numerica	numero di omicidi con vittime donne nel raggio di 1500 metri dell'appartamento pesato con population
101	omicidi-bambini-500	numerica	numero di omicidi con vittime bambini nel raggio di 500 metri dell'appartamento pesato con population
102	omicidi-bambini-1000	numerica	numero di omicidi con vittime bambini nel raggio di 1000 metri dell'appartamento pesato con population
103	omicidi-bambini-1500	numerica	numero di omicidi con vittime bambini nel raggio di 1500 metri dell'appartamento pesato con population
104	rapimenti-500	numerica	numero di rapimenti nel raggio di 500 metri dell'appartamento pesato con population
105	rapimenti-1000	numerica	numero di rapimenti nel raggio di 1000 metri dell'appartamento pesato con population
106	rapimenti-1500	numerica	numero di rapimenti nel raggio di 1500 metri dell'appartamento pesato con population

Tabella 2.48: Variabili utilizzabili per l'analisi

Capitolo 3

Fattori che influenzano i prezzi Airbnb a New York

3.1 Introduzione

Alla luce delle analisi esplorative svolte nel capitolo precedente emerge la presenza di una grande variabilità dei livelli degli affitti per notte di Airbnb, dovuta principalmente al fatto che la città di New York contiene al suo interno vari distretti molto differenti tra loro e con spiccati tratti distintivi. Un'ulteriore fonte di eterogeneità è la sua suddivisione in quartieri e sottoquartieri. Si tenga poi conto del fatto che i prezzi in generale hanno una distribuzione a code pesanti. Per poter considerare tutti questi aspetti relativi alla struttura dei dati è necessario utilizzare strumenti statistici idonei. In quest'ottica, la scelta è ricaduta sulla regressione quantilica, la cui caratteristica principale è quella di non studiare solamente la media condizionata della variabile risposta rispetto ad un set di covariate, ma di considerare anche i quantili della distribuzione condizionata, spiegandone così la variabilità in modo più adeguato. A tale scopo, prima di procedere all'utilizzo di tale metodologia per il caso di studio qui considerato, si propone una trattazione delle principali caratteristiche della regressione quantilica, unitamente ad alcuni aspetti di stima e identificazione.

La regressione quantilica

Molti metodi utilizzati in statistica possono essere visti come elaborazioni complesse del modello di regressione lineare e del metodo di stima ad esso associato: i minimi quadrati ordinari. In particolare, negli studi empirici i ricercatori sono spesso interessati ad analizzare il comportamento di una certa variabile risposta data l'informazione presente in un certo set di covariate. Solitamente in questo ambito viene specificato un modello di regressione lineare del tipo $y = X\beta + \epsilon$, con n numerosità del campione, p il numero di variabili esplicative, y vettore n×1 variabile risposta, X matrice n×p dei dati, ϵ vettore n×1 termine di errore e β vettore p×1 dei parametri da stimare. Questi ultimi in generale rappresentano gli effetti marginali di ogni variabile esplicativa sul valore atteso condizionato $E(Y|X = x)$ della variabile risposta. Tuttavia, il valore atteso condizionato esprime soltanto il comportamento medio della variabile risposta, ma fornisce poca – o nel caso peggiore – nessuna informazione sul comportamento delle code della distribuzione condizionata. Infatti come osservano Mosteller e Tukey (1977) "Quello che fa la curva di regressione è una grande sintesi dell'andamento dei valori attesi condizionati rispetto a X. Quindi per ottenere un'analisi più completa e approfondita bisognerebbe calcolare le curve di regressione corrispondenti ai vari quantili delle distribuzioni condizionate. (...) Infatti, come il solo valore atteso fornisce un quadro incompleto di una singola distribuzione, così la sola curva di regressione fornisce un'immagine incompleta di un insieme di distribuzioni condizionate". Per questa motivazione generalmente quando si analizza un singolo campione si utilizzano misure di concentrazione, boxplot, istogrammi e altre tipologie più sofisticate di analisi della densità per ottenere ulteriori informazioni sulla distribuzione e non si considera soltanto la sua media. Analogamente nell'ambito della regressione sono stati individuati strumenti e tecniche per l'analisi della distribuzione condizionata nel suo insieme, in modo da ottenerne una rappresentazione più completa possibile. Infatti, in molti campi non si è interessati a studiare il comportamento della media condizionata $E(Y|X = x)$, ma quello delle code della distribuzione della variabile risposta Y condizionata a $X \in R^p$. In particolare, lo studio delle code è utile in molte applicazioni:

- Finanza: Value-at-Risk (VaR)

- Meteorologia e agricoltura: temperatura, cambiamenti climatici, caduta delle piogge

- Economia: clustering delle abitudini dei consumatori, previsioni della forza del vento, dati sull'energia

- Industria: previsione degli eventi estremi

Spesso in questi ambiti sono utilizzati modelli che abbiano un termine di errore con distribuzione a code pesanti, ma lo strumento principe in queste situazioni è la regressione quantilica. Infatti, se si considera il problema della stima dell'intera distribuzione condizionata della variabile risposta Y rispetto al set di covariate X, i metodi di regressione quantilica ne stimano adeguatamente il comportamento a diversi livelli di confidenza (Koenker e Bassett Jr, 1978).
Nelle prossime sezioni verrà illustrato il problema di ottimizzazione da cui derivano i quantili e l'estensione che è necessario farne per implementare la regressione quantilica.

Formulazione generale del problema

In generale come una certa variabile casuale Y può essere adeguatamente caratterizzata dalla sua funzione di ripartizione (Koenker, 2005)

$$F_Y(y) = P(Y \leq y) \tag{3.1}$$

a partire da questa per ogni $0 < \tau < 1$ si ha che,

$$Q_Y(\tau) = F_Y^{-1}(y) = inf\{y : F_Y(y) \geq \tau\} \tag{3.2}$$

è chiamato il τ-esimo quantile di Y.
I quantili derivano dalla soluzione di un problema teorico di ottimizzazione, che è alla base del metodo della regressione quantilica: l'implementazione di un metodo per ricavare una

stima puntuale di una variabile casuale con una generica funzione di ripartizione $F_Y(y)$ (Koenker, 2005). È importante ricordare che non si alcun tipo di assunzione riguardo i momenti di Y.
La funzione di perdita considerata da Koenker (2005) in questo specifico contesto è

$$\rho_\tau(u) = |u|(1-\tau)\mathbb{1}_{(-\infty,0)}(u) + |u|\tau\mathbb{1}_{(0,\infty)}(u) = u(\tau - \mathbb{1}_{(-\infty,0)}(u)). \tag{3.3}$$

L'obiettivo finale è di trovare il valore $Q_Y(\tau) = argmin_{q\in R}E_{\rho_\tau}(Y-q)$ che minimizzi (3.3). Per semplificare i calcoli è utile ricordare che la (3.3) può essere riformulata nel modo seguente (Koenker, 2005):

$$(1-\tau)\int_{-\infty}^{q}(q-y)dF_Y(y) + \tau\int_{q}^{\infty}(y-q)dF_Y(y). \tag{3.4}$$

Derivando (3.4) si possono determinare le condizioni del primo ordine e ponendole poi uguali a zero si otterrà il minimo di (3.3).

$$0 = (1-\tau)\int_{-\infty}^{Q_\tau(Y)} dF_Y(y) + \tau\int_{Q_\tau(Y)}^{\infty} dF_Y(y) = (1-\tau)F_Y(q) - \tau[1-F_Y(q)] = -\tau + F_Y(q) \tag{3.5}$$

Dato che $F_Y(y)$ è monotona, ogni elemento dell'insieme $\{y : F_Y(y) = \tau\}$ minimizza la funzione di perdita. Quando la soluzione è unica $\hat{y} = F_Y^{-1}(\tau)$, altrimenti si andrà a scegliere il valore minimo nell'insieme $\{y : F_Y(y) = \tau\}$ per rispettare la convenzione secondo cui la funzione quantile empirico sia continua a sinistra (Koenker, 2005). È naturale che lo

stimatore puntuale ottimale per la funzione di perdita lineare sia rappresentato dai quantili (Koenker, 2005). In particolare, nel caso simmetrico la mediana minimizza la funzione di perdita, mentre quando questa è lineare e asimmetrica si preferisce la stima puntuale più probabile che si trova sul piano formato dai due rami della funzione di perdita marginale (Koenker, 2005).

La controparte campionaria della funzione vista in precedenza, si ottiene sostituendo a $F_Y(y)$ la funzione di distribuzione empirica $F_n(y) = n^{-1} \sum_{i=1}^{n} I(Y_i \leq y)$. In questo caso la funzione di perdita diventa

$$\int \rho_\tau(y - \hat{y}) dF_n(y) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - q) = V_\tau(y, q) \tag{3.6}$$

quindi per un livello di confidenza fissato $\tau \in (0, 1)$, la funzione quantile empirico è definita come

$$\hat{Q}_Y(\tau) = argmin_q V_\tau(y, q) \tag{3.7}$$

Tuttavia, la funzione di perdita $\rho_\tau(u) = u(\tau - \mathbb{1}_{(-\infty,0)}(u))$ non è differenziabile in zero, quindi gli algoritmi standard basati sulla verosimiglianza non sono in grado di minimizzare (3.6). Per ulteriori approfondimenti sui quantili e sull'ottimizzazione di cui sono soluzione si veda Koenker (2005).

Implementazione della regressione quantilica

La regressione quantilica si ottiene come estensione del problema di ottimizzazione affrontato nella sezione precedente utilizzando metodi di stima più generali per ottenere i quantili condizionati. Sia $(Y, \mathbf{X}')'$ un vettore casuale di dimensione $q + 1$ dove $X \subset \Re^q$

indica un insieme di covariate e $Y \in \Re$ la variabile risposta continua. Si assuma, inoltre, che la distribuzione congiunta $(Y, \mathbf{X}')'$ sia non specificata. Il quantile condizionato di livello τ della variabile risposta si definisce come

$$Q_\tau (Y \mid X) = \alpha_\tau + \mathbf{X}'\beta_\tau \tag{3.8}$$

, dove α_τ è la costante, $\boldsymbol{\beta}_\tau = (\beta_{1,\tau}, \beta_{2,\tau}, \ldots, \beta_{q,\tau})' \in \Re^q$ è il vettore dei q parametri ignoti, che dipendono dal livello di confidenza del quantile $\tau \in (0, 1)$, e infine $Q_\tau (Y \mid X = x) = \{\inf y \in \Re \ni F_Y (y \mid X = x) \geq \tau\}$ è l'ignota funzione quantile teorica della variabile risposta.
L'obiettivo principale della regressione quantilica è ottenere la stima dei parametri $\boldsymbol{\vartheta}_\tau = (\alpha_\tau, \boldsymbol{\beta}'_\tau)' \in \Re^{q+1}$, minimizzando la funzione quantile empirica, senza che venga fatta alcun tipo di assunzione sulla forma della distribuzione condizionata di Y.
Dato un campione $\{y_i, x_i\}, i = 1, 2, ..., n$ di osservazioni indipendenti e identicamente distribuite proveniente dalla distribuzione teorica $(Y, \mathbf{X}')'$, lo stimatore di θ_τ risolve il problema di minimizzazione della funzione di perdita quantilica empirica (quantile check function):

$$\begin{aligned} \widehat{\boldsymbol{\vartheta}}_\tau &= \arg\min_{\boldsymbol{\vartheta}_\tau} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau (y_i - \alpha_\tau - x'_i \boldsymbol{\beta}_\tau) \\ &= \arg\min_{\boldsymbol{\vartheta}_\tau} \frac{1}{n} \mathcal{V}_\tau (\boldsymbol{\vartheta}_\tau), \end{aligned} \tag{3.9}$$

dove $\rho_\tau (z) = z \left(\tau - \mathbb{1}_{(-\infty,0)} (z)\right)$ è la funzione di perdita quantilica (quantile check function).
Si noti che se $\tau = 0.5$, $\mathcal{V}_\tau(y, x, \alpha_\tau, \beta_\tau)$ è simmetrica e quindi $\hat{\beta}_{0.5}$ è lo stimatore LAD (Least Absolute Deviation).

L'algoritmo di maggiorazione-minimizzazione

Il fatto che la funzione di perdita quantilica $\rho_\tau (z) = z \left(\tau - \mathbb{1}_{(-\infty,0)} (z)\right)$ sia non differenziabile impedisce l'applicazione dei metodi di inferenza basati sulla verosimiglianza per massimizzare direttamente l'equazione (3.9) rispetto al vettore dei parametri $\boldsymbol{\vartheta}_\tau$. Di conseguenza, sono state studiate soluzioni alternative per superare il problema. Koenker e Bassett Jr (1978) propongono di minimizzare la devianza asimmetrica assoluta utilizzando

metodi di programmazione lineare. Ortega e Rheinboldt (1970) utilizzano la maggiorazione di una certa funzione per minimizzarla. L'algoritmo di maggiorazione-minimizzazione (MM), che verrà qui utilizzato, fa parte di quest'ultima tipologia di algoritmi.
Generalmente questo algoritmo divide un problema complesso di ottimizzazione in una sequenza di ottimizzazioni più semplici (Hunter e Lange, 2000).
Si suppone innanzi tutto di voler minimizzare la funzione obiettivo $\mathcal{L}(\boldsymbol{\vartheta}) : \Re^{p+1} \to \Re$ e si designa con $\boldsymbol{\vartheta}^{(k)}$ l'iterazione a cui l'algoritmo si trova adesso. In generale l'algoritmo prevede i due passi seguenti (Hunter e Lange, 2000):

(i) la scelta di una funzione maggiorante (majorizer) del tipo $\mathcal{G}\left(\boldsymbol{\vartheta} \mid \boldsymbol{\vartheta}^{(k)}\right) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ che soddisfi le seguenti condizioni:

(i.1) $\mathcal{G}\left(\boldsymbol{\vartheta}^{(k)} \mid \boldsymbol{\vartheta}^{(k)}\right) = \mathcal{L}\left(\boldsymbol{\vartheta}^{(k)}\right)$;

(i.2) $\mathcal{G}\left(\boldsymbol{\vartheta} \mid \boldsymbol{\vartheta}^{(k)}\right) \geq \mathcal{L}(\boldsymbol{\vartheta})$, per ogni $\boldsymbol{\vartheta}$;

(ii) il calcolo del minimo $\widehat{\boldsymbol{\vartheta}}^{(k+1)} = \arg\min_{\boldsymbol{\vartheta} \in \Re^{p+1}} \mathcal{G}\left(\boldsymbol{\vartheta} \mid \boldsymbol{\vartheta}^{(k)}\right)$.

Se si definisce $\boldsymbol{\vartheta}^{(k+1)}$ l'iterazione successiva alla k-esima $\boldsymbol{\vartheta}^{(k)}$, allora la minimizzazione di $\mathcal{G}\left(\boldsymbol{\vartheta} \mid \boldsymbol{\vartheta}^{(k)}\right)$ implica che $\mathcal{G}\left(\boldsymbol{\vartheta}^{(k+1)} \mid \vartheta^{(k)}\right) \leq \mathcal{G}\left(\boldsymbol{\vartheta}^{(k)} \mid \boldsymbol{\vartheta}^{(k)}\right)$. Questa disuguaglianza e le condizioni (i) e (ii), implicano che $L(\boldsymbol{\vartheta})$ sia decrescente $\mathcal{L}\left(\boldsymbol{\vartheta}^{(k+1)}\right) \leq \mathcal{L}\left(\boldsymbol{\vartheta}^{(k)}\right)$ (Hunter e Lange, 2000).
In altri termini si è mostrato che, se la funzione da minimizzare originaria è non differenziabile, si può ottenere ugualmente una funzione maggiorante differenziabile, che ha lo stesso punto di minimo della funzione originaria, a condizione che la maggiorante sia tangente a quest'ultima. In particolare, nell'ambito della Quantile Regression, si approssima la funzione $\rho_\tau(u)$ non differenziabile (3.9) utilizzando la ε-perturbation

$$\rho_\tau^\varepsilon(u) = \rho_\tau(u) - \frac{\varepsilon}{2}\log\left(\varepsilon + |u|\right), \tag{3.10}$$

proposta da Hunter e Lange (2000) per rendere più liscia la quantile check function $\rho_\tau(u)$ originale. In questa maniera si ottiene

$$\mathcal{V}_\tau^\varepsilon(\boldsymbol{\vartheta}_\tau) = \sum_{i=1}^{n} \rho_\tau^\varepsilon(\boldsymbol{\vartheta}_\tau) \tag{3.11}$$

che approssima $\mathcal{V}_\tau(\boldsymbol{\vartheta}_\tau)$, per ogni $\varepsilon > 0$. Per quello che riguarda la scelta della funzione maggiorante, dato il vettore dei residui quantilici alla k-esima iterazione $\widehat{r}_\tau^{(k)} = y - \alpha_\tau -$

$X'\widehat{\beta}_\tau^{(k)}$, Hunter e Lange (2000) propongono di maggiorare (3.10) con la seguente forma quadratica

$$\zeta_\tau^\varepsilon\left(r_i \mid \widehat{r}_{i,\tau}^{(k)}\right) = \frac{1}{4}\left[\frac{r_{i,\tau}^2}{\varepsilon + |\widehat{r}_{i,\tau}^{(k)}|} + (4\tau - 2)r_{i,\tau} + c_\tau\right], \tag{3.12}$$

dove $r_i = y_i - \alpha_\tau - x_i'\beta_\tau^{(k)}$, $\widehat{r}_{i,\tau}^{(k)}$ il valore i-esimo di $\widehat{r}_\tau^{(k)}$ e c_τ costante tale che $\sum_{i=1}^n \zeta_\tau^\varepsilon\left(\widehat{r}_{i,\tau}^{(k)} \mid \widehat{r}_{i,\tau}^{(k)}\right) = \sum_{i=1}^n \rho_\tau^\varepsilon\left(\widehat{r}_{i,\tau}^{(k)}\right)$, in modo che la funzione da approssimare e la funzione maggiorante siano tangenti. In realtà, dato che la costante c_τ non dipende dai parametri da stimare, la funzione maggiorante in questo caso ha lo stesso punto di minimo della funzione approssimata anche se queste non sono tangenti.
In seguito si opera la minimizzazione della funzione maggiorante (majoriser)

$$\begin{aligned}\widehat{\boldsymbol{\vartheta}}_\tau^{(k+1)} &= \arg\min_{\boldsymbol{\vartheta}_\tau} \mathcal{G}_\tau^\varepsilon\left(\boldsymbol{\vartheta}_\tau \mid \widehat{\boldsymbol{\vartheta}}_\tau^{(k)}\right), \\ \mathcal{G}_\tau^\varepsilon\left(\boldsymbol{\vartheta}_\tau \mid \widehat{\boldsymbol{\vartheta}}_\tau^{(k)}\right) &= \frac{1}{n}\sum_{i=1}^n \zeta_\tau^\varepsilon\left(r_i \mid \widehat{r}_i^{(k)}\right),\end{aligned} \tag{3.13}$$

rispetto a $\boldsymbol{\vartheta}_\tau$.
Entrando maggiormente nei dettagli, dati $r_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, $\widehat{r}_i^{(k)} = y_i - \mathbf{x}'\widehat{\boldsymbol{\beta}}^{(k)}$ e ipotizzando di trovarci alla k iterazione, si ha che

$$\mathcal{V}_\tau(\boldsymbol{\beta}) \to \sum_{i=1}^n \zeta_\tau^\varepsilon\left(r_i \mid \widehat{r}_{i,\tau}^{(k)}\right) = \sum_{i=1}^n \frac{1}{4}\left[\frac{r_{i,\tau}^2}{\epsilon + |\widehat{r}_{i,\tau}^{(k-1)}|} + (4\tau - 2)r_{i,\tau} + c_\tau\right], \tag{3.14}$$

Per minimizzare (3.14) si devono individuare le condizioni di primo ordine di $\mathcal{V}_\tau(\boldsymbol{\beta})$ derivandola rispetto a β, ovvero risolvere la seguente equazione:

$$\frac{\partial \mathcal{V}_\tau(\beta)}{\partial \beta} = 0 \tag{3.15}$$

dove

$$\frac{\partial \mathcal{V}_\tau(\beta)}{\partial \beta} = \sum_{i=1}^n \frac{1}{4}\left[\frac{1}{\epsilon + |\widehat{r}_{i,\tau}^{(k-1)}|}\frac{\partial r_{i,\tau}^2}{\partial \beta} + (4\tau - 2)\frac{\partial r_{i,\tau}}{\partial \beta}\right] \tag{3.16}$$

con

$$\frac{\partial r_{i,\tau}}{\partial \beta} = \frac{\partial(y_i - \mathbf{x}_i'\boldsymbol{\beta})}{\partial \beta} = -x_i \tag{3.17}$$

$$\frac{\partial r_{i,\tau}^2}{\partial \beta} = -2x_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \tag{3.18}$$

Sostituendo poi (3.17) e (3.18) in (3.16) si ottiene

$$\sum_{i=1}^{n} \frac{1}{4}\left[\frac{-2x_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})}{\epsilon + |\widehat{r}_{i,\tau}^{(k-1)}|} - (4\tau - 2)x_i\right] = 0 \tag{3.19}$$

Infine si va ad isolare β

$$\sum_{i=1}^{n}\left[\frac{x_i x_i'\beta}{\epsilon + |\widehat{r}_{i,\tau}^{(k-1)}|}\right] = \sum_{i=1}^{n}\left[\frac{x_i y_i}{\epsilon + |\widehat{r}_{i,\tau}^{(k-1)}|} + (2\tau - 1)x_i\right] \tag{3.20}$$

$$\hat{\beta}^{(k)} = \sum_{i=1}^{n}\left[\frac{x_i x_i'}{\epsilon + |\widehat{r}_{i,\tau}^{(k-1)}|}\right]^{-1} \sum_{i=1}^{n}\left[\frac{x_i y_i}{\epsilon + |\widehat{r}_{i,\tau}^{(k-1)}|} + (2\tau - 1)x_i\right] \tag{3.21}$$

Riscritto in termini matriciali si ottiene

$$\hat{\beta}^{(k)} = \left[X' diag\left(\frac{1}{\epsilon + |\widehat{r}_{i,\tau}^{(k-1)}|}\right) X\right]^{-1}\left[X' diag\left(\frac{1}{\epsilon + |\widehat{r}_{i,\tau}^{(k-1)}|}\right)\mathbf{y} + X' diag(2\tau - 1)\mathbb{1}\right] \tag{3.22}$$

per ogni $\varepsilon > 0$, dove X matrice n×k che contiene le variabili esplicative, y vettore n×1 che contiene la variabile risposta, $diag\left(\frac{1}{\epsilon + |\widehat{r}_{i,\tau}^{(k-1)}|}\right)$ matrice diagonale n×n, $diag(2\tau - 1)$ matrice diagonale n×n, $\mathbb{1}$ vettore n×1.

Estensione del modello per dati spaziali

L'obiettivo primario di questo studio è fondamentalmente quello di individuare come incidono fattori legati al contesto socio-economico – come per esempio la vicinanza di un determinato servizio a una proprietà, o il livello di criminalità nella zona in cui si trova una certo appartamento – sui livelli degli affitti in Airbnb. Tuttavia, quando si fa questo tipo di analisi è anche importante tenere conto del fatto che ogni affitto è riferito a uno specifico punto nello spazio, non sarebbe quindi irragionevole ipotizzare che i prezzi di affitto di un appartamento possano dipendere anche dai prezzi di affitto delle proprietà vicine. In altri termini si può supporre che esista un effetto di spillover, ovvero di contagio, della variabile di interesse (il livello dell'affitto Airbnb) tra osservazioni spazialmente attigue.

Questo può creare dipendenza spaziale tra le osservazioni, infatti generalmente accade che osservazioni che si riferiscono ad aree vicine mostrino meno variabilità di dati relativi ad aree che risultano lontane tra loro. Un'altra possibile causa della dipendenza spaziale tra le osservazioni è la presenza di variabili omesse correlate con la risposta o con variabili esplicative già presenti all'interno del modello. In questo caso si è in presenza di correlazione spaziale causata dall'errata specificazione del modello. L'utilizzo dei modelli della statistica classica, come per esempio il modello lineare, in queste circostanze può causare a stime distorte e procedure inferenziali non valide. Spesso, per trattare questo tipo di dipendenza si ricorre a modelli di tipo econometrico, come il modello SAR (Spatial Autoregresive Model) e il modello SEM (Structural Equation Modeling) descritti nel libro Introduction Spatial Econometrics di Pace e Lesage pubblicato nel 2009. Tuttavia, in questo caso si è scelto di modellare la dipendenza spaziale presente all'interno dei dati qui utilizzati, attraverso l'inserimento nei modelli di variabili che tenessero conto della suddivisione della città di New York in sottoquartieri. In particolare si è utilizzato un caso specifico dell'estensione della quantile regression proposta da Yue e Rue (2011).

$$Q_\tau\left(y_i \mid x_i\right) = \eta_{\tau i} = x_i^T \beta_\tau + \sum_{i=1}^{q} f_{\tau i}(u_{ij}) + b_{\tau g_i} \tag{3.23}$$

dove $x_i^T \beta_\tau$ è la componente lineare, $u_i = (u_{i1}, ..., u_{iq})^T$ è un vettore di covariate continue che si ipotizza abbiano una relazione non lineare con la risposta, descritta dalle funzioni di lisciamento $f_{\tau i}$ per $j = 1, ..., q$, b_{g_i} è un effetto casuale specifico per ogni gruppo tale che $b_{g_i} = b_g$ se l'unità i-esima appartiene ad uno specifico gruppo g con $g = 1, ..., G$. Yue e Rue (2011) propone un modello, in cui oltre alla componente lineare e agli effetti casuali dati dall'appartenenza ad un certo gruppo, è presente una componente non lineare definita come un modello additivo GAM, che può essere utilizzata per modellare effetti stagionali, trend e altri tipi di dipendenze. Il modello che verrà utilizzato per l'analisi è il seguente

$$Q_{y_i}\left(\tau \mid x_i\right) = \eta_{\tau i} = x_i^T \beta_\tau + \mu_s \tag{3.24}$$

dove $\mu_s = D_s \alpha_s$, con D_s matrice dei sottoquartieri di dimensione $n \times$q, con n numero di osservazioni e q numero di sottoquartieri, tale che l'elemento $D_{is} = 1$ se la i-esima proprietà appartiene al s-esimo sottoquartiere, 0 altrimenti. Ogni riga di D sommerà a 1, dato che ogni appartamento si trova in un solo sottoquartiere.

Ci si può ricondurre alla specificazione 3.24 partendo dalla specificazione 3.23 semplicemente utilizzando un modello che non includa gli effetti non-lineari e inserendo nel modello un effetto di gruppo per ogni sottoquartiere. In particolare, nel caso in cui la i-esima proprietà appartenga al sottoquartiere s-esimo e quindi $D_{is} = 1$ si avrà $b_{g_i} = b_g = \alpha_s$, per $i = 1, ...n$, $s = 1, ..., q$ e $g = 1, ..., 265$. Per i dettagli relativi all'implementazione del modello con il software R si veda l'Appendice E.
Nella prossima sezione verrà illustrata un'applicazione del modello descritto sopra.

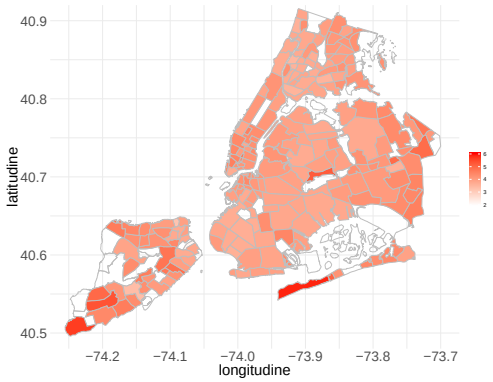
3.2 Applicazione

Specificazione del modello e interpretazione dei risultati

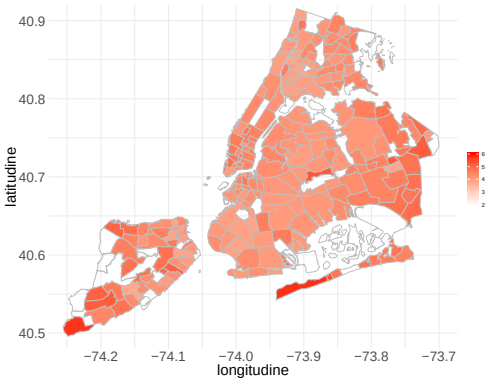
Il modello che verrà utilizzato è una regressione quantilica con componente spaziale, come descritto nella sezione precedente, la cui specificazione è la seguente

$$Q_\tau \left(y_i \mid x_i \right) = D_i^T \alpha_{s,\tau} + x_i^T \beta_\tau + \epsilon_i, \;\; i = 1, 2, ...n \tag{3.25}$$

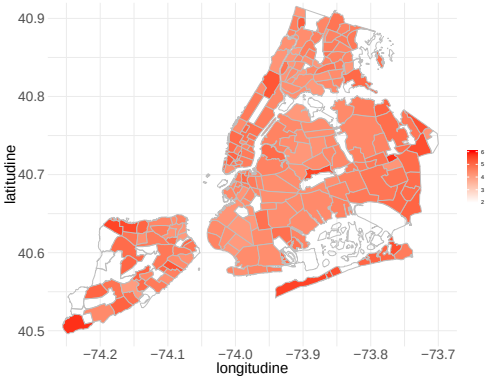
dove D_i elemento della matrice dei sottoquartieri D_s, tale che $D_i = 1$ se la i-esima proprietà appartiene al s-esimo sottoquartiere, 0 altrimenti. $\alpha_{s,\tau}$ è l'intercetta spaziale che si riferisce alla i-esima osservazione per il quantile τ, x_i^T vettore riga 1×p contenente i valori delle variabili esplicative riferiti alla i-esima osservazione, con p numero di parametri da stimare. β_τ vettore p×1 dei parametri da stimare. In particolare, si è partiti stimando il modello 3.25, denotato come modello 1, che ha come risposta la trasformata logaritmica del prezzo, con tutte le esplicative e le intercette spaziali. In seguito sono state escluse tutte le variabili che risultavano non significative ad un livello di confidenza pari a 0.05 per tutti e tre i quantili considerati e infine si è stimato il modello definito come modello 2.
Tuttavia, prima di commentare i risultati ottenuti nelle tabelle 3.1 e 3.2, è bene ricordare che, nel caso dei modelli di tipo log-lineare come quelli qui utilizzati, i coefficienti misurano una semielasticità. In particolare se una certa variabile x_i aumenta di un'unità la variabile y_i varia del $\beta * 100\%$ dove β è il coefficiente della variabile esplicativa x_i e y_i rappresenta la variabile risposta, in questo caso trasformata tramite la funzione logaritmo naturale. Innanzi tutto, se si osserva la figura 3.1 che mostra i valori assunti dalle intercette spaziali del modello 2, che rappresentano i log-prezzi medi stimati di affitto per ogni sottoquartiere, si nota che questi hanno un andamento crescente man mano che ci si muove dalla



(a) Quantile 0.15



(b) Quantile 0.50



(c) Quantile 0.85

Figura 3.1: Intercette spaziali

coda sinistra verso la coda destra della distribuzione condizionata, oltre al fatto che la variabilità dei log-prezzi medi stimati per sottoquartiere aumenta muovendosi sempre nella stessa direzione descritta in precedenza. Se si confrontano i due modelli si nota che l'unico coefficiente non significativo nell'insieme delle variabili riguardanti le caratteristiche delle proprietà è il fatto che il proprietario dell'appartamento richieda o meno che l'ospite abbia l'immagine del profilo sul sito di Airbnb. Per quello che riguarda la misura della bontà di adattamento è stato utilizzato l'R^2 aggiustato per la regressione quantilica proposto da Koenker e Machado (1999). Per il modello 1 si ha R^2 aggiustato migliore per il quantile 0.5 (R^2=0.5275), l'adattamento peggiora leggermente per i quantili della distribuzione più estremi in particolare per quelli sulla coda sinistra della distribuzione condizionata, con valori di R^2=0.4931 e di R^2=0.5083, rispettivamente per il quantile 0.15 e 0.85. Lo stesso andamento si osserva anche per il modello 2, con valori di R^2 pari a 0.4929, 0.5274, 0.5827 rispettivamente per i quantili 0.15, 0.50, 0.85. Inoltre in generale se si confronta per ogni quantile considerato il valore di R^2 di Koenker si nota che in tutti i casi il modello 1 presenta sempre un adattamento ai dati leggermente migliore rispetto al modello 2.
Per quello che riguarda gli aspetti interpretativi in generale si può notare che essere un superhost ha sempre un effetto positivo significativo sull'affitto per tutti i quantili della distribuzione condizionata, che raggiunge il massimo per gli affitti più economici (quantile 0.15). Essere un superhost porta ad un aumento dell'affitto dello 7.1% per cento rispetto a chi non lo è. In altre parole gli host che offrono un servizio migliore tendono a farsi pagare di più in tutti i casi. Lo stesso avviene nel caso in cui l'host permetta la verifica della sua identità. In altri termini essere più trasparente porta l'host a chiedere un affitto più alto rispetto al caso in cui non adotti una politica di trasparenza, questo vale per tutti i quantili condizionati della distribuzione, con un effetto massimo sui livelli di affitto bassi. Si può inoltre notare che in generale il fatto che la proprietà affittata sia una casa e non un appartamento non ha un effetto significativo per i livelli di affitto medio-bassi, mentre per livelli alti l'affitto aumenta del 2.1% rispetto al caso in cui la proprietà affittata sia un appartamento. Invece se la proprietà è un loft, l'affitto è mediamente più alto rispetto a quello di un appartamento per tutti i quantili. In altri termini per qualsiasi livello di affitto, affittare un loft costa mediamente di più che un appartamento, come si era notato anche dalle analisi esplorative svolte nel capitolo 2, questo succedeva soprattutto nei quartieri di Manhattan, Staten Island e Bronx come si vede dal figura 2.25. Per quello che riguarda

la tipologia di soggiorno mediamente costa sempre meno affittare una stanza condivisa o una stanza singola piuttosto che un intera proprietà per tutti i livelli di affitto ed è sensato che si riscontri questo. Si può anche notare che il fatto di avere una camera da letto, un bagno o una stanza in più ha sempre un effetto positivo sugli affitti, soprattutto i livelli di affitto più alti. In particolare si ha un aumento dell'affitto medio pari a 13.2% rispetto ad una proprietà con una camera da letto in meno, un aumento del 15.7% dell'affitto medio rispetto ad una proprietà con un bagno in più, mentre una stanza in più fa aumentare l'affitto medio del 9.3%. Avere un posto letto in più ha sempre un effetto negativo sugli affitti in particolare sui livelli più bassi, questo perchè possono esserci dei proprietari che hanno molti posti letto e pur di riempirli sono disposti a fare dei prezzi più bassi. Il fatto di includere un ospite in più ha un effetto significativo e positivo sugli affitti soprattutto per i livelli di affitto più alti. È interessante notare poi che all'aumentare di un dollaro del prezzo di portare un ospite non incluso aumenti anche l'affitto per tutti i quantili, un po' come succede nei soggiorni in albergo, dove se si porta un ospite non incluso nella prenotazione generalmente aumenta anche il prezzo della camera. Generalmente all'aumentare di 1 punto del voto della pulizia, della location e della soddisfazione generale, l'effetto aumenta in maniera significativa per tutti i livelli di affitto e l'effetto massimo viene raggiunto per i livelli più di prezzo più alti. In altri termini un proprietario che mantiene bene la sua proprietà e viene valutato bene dai suoi ospiti richiede un affitto più alto, probabilmente questo è un segno del fatto che gli ospiti sono disposti a pagare di più per avere un servizio migliore. Il fatto che ci sia un minimo di notti obbligatorio da prenotare ha un effetto significativo negativo sul prezzo di affitto per tutti i livelli soprattutto per i livelli di affitto più bassi, infatti per ogni notte in più che è obbligatorio prenotare l'affitto cala mediamente del 0.5% per livelli di affitto bassi. Il fatto che una proprietà sia subito disponibile per essere affittata ha un effetto negativo significativo sul prezzo di affitto per tutti i livelli, in particolare per quelli più bassi con un calo medio del 1.6% dell'affitto. È interessante poi notare come all'aumentare di un metro della distanza minima dall'attrazione turistica più vicina l'affitto cali per tutti i livelli considerati. In particolare si ha un calo mediamente dello 0.007% dell'affitto nel caso in cui le attrazioni turistiche siano palazzi storici e dello 0.004% se sono punti panoramici per tutti i livelli di affitto. Un altro aspetto notevole riguarda la distanza dagli alberghi, si può infatti notare che all'aumentare di un metro della distanza dall'albergo più vicino il livello di affitto cali mediamente con un effetto massimo

del 0.004% per livelli di affitto alti, questo è dovuto al fatto che la maggior parte degli alberghi di New York sono concentrati a Manhattan (come si vede nella figura 2.2a), che è anche la zona dove gli affitti per notte Airbnb sembrano essere mediamente più alti come si vede nella figura 2.23c, di conseguenza è logico che man mano che ci si allontana da Manhattan il livello medio degli affitti cali. Per quello che riguarda i tassi di criminalità si nota come un aumento di reati violenti come violenze sessuali e rapine abbia forte impatto negativo sui livelli dell'affitto, che cresce all'aumentare del livello di affitto. Tuttavia il reato che sembra avere il maggiore impatto in assoluto sugli affitti è l'omicidio che ha per vittime le donne, anche se il suo effetto cala man mano che aumenta il livello di affitto fino a diventare non significativo per i livelli di affitto più alti, perché non è un reato comune nei sottoquartieri che presentano questi livelli di affitto. Infine si può notare che all'aumentare dei furti l'effetto sugli affitti è positivo ed sempre maggiore all'aumentare del loro livello, poichè generalmente nei sottoquartieri con i livelli di affitto più alti si verifica un maggior numero di furti come si può vedere anche dalla figura 2.8 (Manhattan è il distretto dove avviene il maggior numero di furti).

Algorithm 1: Ottenere le stime $\boldsymbol{\beta}_j = (\beta_{1,\tau}, \beta_{2,\tau}, \ldots, \beta_{p,\tau})'$ utilizzando MM

Inizializzazione:

$$\widehat{\beta} = (X'X)^{-1}X'y$$
$$dTolX = \sqrt{\widehat{\beta}'\widehat{\beta}}$$
$$\hat{\beta}_{old} = \hat{\beta}$$
$$Dl_{old} = -\infty$$
$$Dl_{dif} = \infty$$

Calcolo dei residui:

$$\hat{r}_{old} = y - X\hat{\beta}_{old}$$

Calcolo della costante c:

$$\rho_\tau^\varepsilon(\hat{r}_{old}) = \rho_\tau(\hat{r}_{old}) - \frac{\varepsilon}{2}\log(\varepsilon + |\hat{r}_{old}|)$$
$$\zeta_\tau^\varepsilon(\hat{r}_{old}) = \frac{1}{4}\left[\frac{\hat{r}_{old}^2}{\varepsilon + |\hat{r}_{old}|} + (4\tau - 2)\hat{r}_{old}\right]$$
$$c = \sum_{i=1}^{n}(\rho_\tau^\varepsilon(\hat{r}_{old,i}) - \zeta_\tau^\varepsilon(\hat{r}_{old,i}))$$

Algoritmo MM:

while $(\| \hat{\beta}_j - \hat{\beta}_{j-1} \|^2 \geq 0.00000001 \cup |\zeta_\tau^\varepsilon(j) - \zeta_\tau^\varepsilon(j-1)| \geq \epsilon n|log(\epsilon)|) \cap (j \neq MaxIt)$ **do**

- **[1]** aggiornare il contatore j=j+1;
- **[2]** calcolare $\widehat{\beta}_j = \sum_{i=1}^n \left[\frac{x_i x_i'}{\epsilon + |\hat{r}_{old}|}\right]^{-1} \sum_{i=1}^n \left[\frac{x_i(x_i + \hat{r}_{old})}{\epsilon + |\hat{r}_{old}|} + (2\tau - 1)x_i\right]$;
- **[3]** aggiornamento residui $\hat{r} = y - X\hat{\beta}_j$;
- **[4]** calcolare la funzione di perdita $\zeta_\tau^\varepsilon(\hat{r}) = \frac{1}{4}\left[\frac{\hat{r}^2}{\varepsilon + |\hat{r}|} + (4\tau - 2)\hat{r} + c\right]$;
- **[5]** calcolare la costante c

  $\rho_\tau^\varepsilon(\hat{r}) = \rho_\tau(\hat{r}) - \frac{\varepsilon}{2}\log(\varepsilon + |\hat{r}|)$

  $\zeta_\tau^\varepsilon(\hat{r}) = \frac{1}{4}\left[\frac{\hat{r}^2}{\varepsilon + |\hat{r}|} + (4\tau - 2)\hat{r}\right]$

  $c = \sum_{i=1}^n(\rho_\tau^\varepsilon(\hat{r}_i) - \zeta_\tau^\varepsilon(\hat{r}_i))$;
- **[6]** aggiornamento parametri

  $Dl_{dif} = |\zeta_\tau^\varepsilon(\hat{r}) - Dl_{old}|$

  $Dl_{old} = Dl_{dif}$

  $\hat{r}_{old} = \hat{r}$

  $dTolX = \sqrt{(\widehat{\beta} - \hat{\beta}_{old})'(\widehat{\beta} - \hat{\beta}_{old})}$

end

	$\tau = 0.15$			$\tau = 0.50$			$\tau = 0.85$		
Variable	Coefficient	t–stat	p–Value	Coefficient	t–stat	p–Value	Coefficient	t–stat	p–Value
host_response_timewithin_a_day	0.003	0.37479	0.70782	-0.01089	-1.64903	0.09914	-0.02547	-2.97956	0.00289
host_response_timewithin_a_few_hours	0.00457	0.57486	0.56539	-0.00642	-0.9796	0.32728	-0.02211	-2.60662	0.00915
host_response_timewithin_an_hour	0.01098	1.38933	0.16474	0.00181	0.27715	0.78166	0.00178	0.21072	0.83311
host_is_superhostt	0.07069	14.202	0.00000	0.06053	14.75235	0.00000	0.05667	10.66987	0.00000
host_identity_verifiedt	0.01568	4.51857	0.00001	0.01413	4.94019	0.00000	0.01044	2.82064	0.00479
property_typeHouse	-0.01012	-1.4093	0.15875	-0.00276	-0.46542	0.64163	0.02093	2.73016	0.00633
property_typeLoft	0.0277	2.83896	0.00453	0.08805	10.94576	0.00000	0.16824	16.15953	0.00000
property_typeOther	0.02006	2.90893	0.00363	0.05315	9.34931	0.00000	0.12355	16.79217	0.00000
room_typePrivate_room	-0.52868	-127.9632	0.00000	-0.51451	-151.05157	0.00000	-0.47634	-108.05287	0.00000
room_typeShared_room	-1.0144	-107.38381	0.00000	-0.96706	-124.17242	0.00000	-0.83297	-82.63831	0.00000
accommodates	0.07798	45.47123	0.00000	0.08545	60.43443	0.00000	0.09322	50.94138	0.00000
bathrooms	0.05256	11.9194	0.00000	0.09811	26.98504	0.00000	0.15788	33.55307	0.00000
bedrooms	0.10038	31.68399	0.00000	0.11828	45.28626	0.00000	0.13287	39.30548	0.00000
beds	-0.03292	-12.73954	0.00000	-0.02148	-10.08157	0.00000	-0.01346	-4.88317	0.00000
bed_typeReal_Bed	0.03377	3.36595	0.00076	0.00149	0.18048	0.85678	0.00455	0.42468	0.67107
guests_included	0.0323	17.97678	0.00000	0.02305	15.5582	0.00000	0.01552	8.0931	0.00000
extra_people	0.00032	4.43557	0.00001	0.00042	7.10302	0.00000	0.00078	10.06514	0.00000
minimum_nights	-0.00548	-36.57917	0.00000	-0.00444	-35.93601	0.00000	-0.00226	-14.12792	0.00000
maximum_nights	-0.00000	-2.05629	0.03976	-0.00000	-1.27479	0.20239	-0.00000	-0.8372	0.40248
number_of_reviews	0.00028	4.9682	0.00000	0.00011	2.34256	0.01915	-0.00037	-6.29934	0.00000
review_scores_rating	0.00293	8.41336	0.00000	0.00341	11.86665	0.00000	0.00383	10.29684	0.00000
review_scores_accuracy	0.00861	3.15413	0.00161	0.00512	2.27359	0.02299	0.00082	0.2814	0.77841
review_scores_cleanliness	0.03587	17.52532	0.00000	0.03559	21.09146	0.00000	0.03733	17.09061	0.00000
review_scores_checkin	0.00714	2.54459	0.01094	0.00267	1.15406	0.24848	0.00284	0.94889	0.34268
review_scores_communication	-0.0028	-0.93197	0.35136	-0.004	-1.61596	0.10611	-0.00664	-2.07047	0.03841
review_scores_location	0.00518	2.12479	0.03361	0.01193	5.93062	0.00000	0.01771	6.80335	0.00000
review_scores_value	-0.04283	-15.27793	0.00000	-0.0463	-20.03362	0.00000	-0.0495	-16.55081	0.00000
instant_bookablet	-0.01639	-4.08608	0.00004	-0.0075	-2.26969	0.02323	-0.00226	-0.52807	0.59745
cancellation_policymoderate	0.00649	1.36094	0.17354	-0.01264	-3.2141	0.00131	-0.0285	-5.60152	0.00000
cancellation_policystrict	-0.02041	-4.30649	0.00002	-0.01767	-4.52271	0.00001	-0.02974	-5.88108	0.00000
cancellation_policystrict_14_with_grace_period	-0.01705	-3.37334	0.00074	-0.02552	-6.12503	0.00000	-0.04166	-7.72665	0.00000
require_guest_profile_picturet	0.0014	0.09827	0.92171	-0.00978	-0.83134	0.40578	-0.02008	-1.31912	0.18713
require_guest_phone_verificationt	-0.00117	-0.08565	0.93174	0.01358	1.21021	0.2262	0.04209	2.89814	0.00375
reviews_per_month	-0.02049	-15.49664	0.00000	-0.02508	-23.00643	0.00000	-0.03499	-24.79438	0.00000
d_fast	0.00001	0.50584	0.61297	0.00000	-0.35616	0.72172	0.00000	-0.07156	0.94295
d_hotel	-0.00003	-4.12262	0.00004	-0.00003	-4.41148	0.00001	-0.00004	-5.68172	0.00000
d_ristoranti	-0.00004	-1.79368	0.07287	-0.00003	-1.425	0.15416	-0.00003	-1.34598	0.17831
d_individual	-0.00001	-0.70882	0.47844	0.00000	-0.39662	0.69165	-0.00001	-0.57477	0.56545
d_parchi	0.00000	0.61185	0.54064	0.00001	0.97777	0.32819	0.00000	0.58373	0.5594
d_interior	-0.00007	-17.80778	0.00000	-0.00007	-22.27533	0.00000	-0.00007	-17.58846	0.00000
d_scenic	-0.00003	-6.7604	0.00000	-0.00004	-10.06183	0.00000	-0.00004	-8.08369	0.00000
d_metro	0.00001	0.88039	0.37865	0.00003	2.59207	0.00954	0.00006	3.9811	0.00007
d_luoghi_culto	0.00005	1.71547	0.08626	0.0001	4.31977	0.00002	0.00016	5.28897	0.00000
d_scuole	0.00001	0.4776	0.63293	0.00000	0.04171	0.96673	0.00003	1.08767	0.27675
droga_500	-0.20757	-0.86669	0.38611	-0.36741	-1.86078	0.06278	-0.23272	-0.91066	0.36248
omicidi_donne500	-132.28359	-3.34462	0.00082	-99.68119	-3.05699	0.00224	-22.94336	-0.54365	0.58668
furti_500	0.95575	1.63367	0.10233	2.56506	5.31811	0.00000	2.93743	4.70554	0.00000
rapine_500	-5.44207	-3.71742	0.0002	-4.09168	-3.39015	0.00007	-4.82084	-3.08618	0.00203
stupri_500	-0.50546	-3.61336	0.0003	-0.37687	-3.26779	0.00108	-0.63796	-4.27407	0.00002
rapimenti_500	27.90755	0.73156	0.46444	-1.20034	-0.03817	0.96956	-41.13248	-1.0105	0.31226
multe_500	0.95123	0.93402	0.3503	0.27026	0.32187	0.74755	0.37008	0.34055	0.73344
aggressioni_500	0.62982	1.33928	0.18048	0.30913	0.79734	0.42526	0.18712	0.3729	0.70922
reati_persona500	9.02605	2.94041	0.00328	4.47647	1.76883	0.07693	-0.89818	-0.27422	0.78392

Tabella 3.1: Regressione quantilica: modello 1

	$\tau = 0.15$			$\tau = 0.50$			$\tau = 0.85$		
Variable	Coefficient	t–stat	p–Value	Coefficient	t–stat	p–Value	Coefficient	t–stat	p–Value
host_response_timewithin_a_day	0.0039	0.50402	0.61425	-0.01134	-1.74456	0.08106	-0.02398	-2.80455	0.00504
host_response_timewithin_a_few_hours	0.00412	0.53712	0.59119	-0.00622	-0.96414	0.33498	-0.02055	-2.42226	0.01543
host_response_timewithin_an_hour	0.01038	1.36044	0.17369	0.00253	0.39452	0.6932	0.00301	0.35629	0.72163
host_is_superhostt	0.07119	14.82503	0.00000	0.0607	15.03153	0.00000	0.05569	10.48506	0.0000
host_identity_verifiedt	0.01572	4.6976	0.00000	0.01499	5.32657	0.00000	0.01039	2.8057	0.00502
property_typeHouse	-0.00901	-1.30129	0.19316	-0.00319	-0.54783	0.58381	0.02146	2.80209	0.00508
property_typeLoft	0.0287	3.05127	0.00228	0.08705	11.00378	0.00000	0.16857	16.203	0.00000
property_typeOther	0.02062	3.10055	0.00193	0.05413	9.67876	0.00000	0.12371	16.81962	0.00000
room_typePrivate_room	-0.52938	-132.87033	0.00000	-0.51433	-153.50138	0.00000	-0.47666	-108.17242	0.00000
room_typeShared_room	-1.01501	-111.41319	0.00000	-0.96708	-126.22146	0.00000	-0.83268	-82.64083	0.00000
accommodates	0.07776	46.99225	0.00000	0.08548	61.43098	0.00000	0.09311	50.87918	0.00000
bathrooms	0.05181	12.18091	0.00000	0.09839	27.50332	0.00000	0.15774	33.53008	0.00000
bedrooms	0.10078	32.97685	0.00000	0.11792	45.87706	0.00000	0.13256	39.21658	0.00000
beds	-0.03244	-13.01336	0.0000	-0.02115	-10.08605	0.00000	-0.01298	-4.70598	0.00000
bed_typeReal_Bed	0.03517	3.63325	0.00028	0.00349	0.42859	0.66822	0.00407	0.38056	0.70353
guests_included	0.03174	18.30491	0.00000	0.02284	15.66427	0.00000	0.01566	8.16678	0.00000
extra_people	0.00033	4.74946	0.00000	0.00043	7.25819	0.00000	0.00076	9.80689	0.00000
minimum_nights	-0.00544	-37.66276	0.00000	-0.0044	-36.24593	0.000000	-0.00229	-14.3247	0.00000
maximum_nights	-0.00000	-2.14025	0.03234	-0.00000	-1.30422	0.19216	-0.00000	-0.84666	0.39719
number_of_reviews	0.00027	5.09701	0.00000	0.0001	2.2882	0.02213	-0.00037	-6.26481	0.0000
review_scores_rating	0.003	8.91915	0.00000	0.00342	12.10159	0.00000	0.00386	10.37239	0.00000
review_scores_accuracy	0.00855	3.24445	0.00118	0.00502	2.26461	0.02354	0.00087	0.29979	0.76434
review_scores_cleanliness	0.03558	18.01695	0.00000	0.03559	21.42687	0.00000	0.03704	16.95521	0.00000
review_scores_checkin	0.0071	2.62296	0.00872	0.00242	1.06018	0.28907	0.00238	0.79474	0.42677
review_scores_communication	-0.00244	-0.84198	0.3998	-0.00375	-1.5387	0.12388	-0.00685	-2.138	0.03252
review_scores_location	0.00521	2.21654	0.02666	0.01188	6.01206	0.00000	0.01818	6.99524	0.00000
review_scores_value	-0.04308	-15.9257	0.00000	-0.04635	-20.37551	0.00000	-0.04965	-16.59802	0.00000
instant_bookablet	-0.01567	-4.04845	0.00005	-0.00734	-2.25621	0.02406	-0.00164	-0.38298	0.70174
cancellation_policymoderate	0.00588	1.27833	0.20114	-0.01336	-3.45342	0.00055	-0.02841	-5.58316	0.00000
cancellation_policystrict	-0.02054	-4.49175	0.00001	-0.01874	-4.87446	0.00000	-0.02927	-5.78869	0.00000
cancellation_policystrict_14_with_grace_period	-0.01766	-3.62298	0.00029	-0.02628	-6.40933	0.00000	-0.04114	-7.63092	0.00000
require_guest_phone_verificationt	0.00064	0.07879	0.9372	0.00726	1.06247	0.28802	0.02533	2.82	0.0048
reviews_per_month	-0.02014	-15.79335	0.00000	-0.02517	-23.4648	0.00000	-0.03495	-24.77632	0.00000
d_hotel	-0.00003	-4.39404	0.00001	-0.00003	-4.7055	0.00000	-0.00004	-6.05981	0.00000
d_interior	-0.00007	-21.19202	0.00000	-0.00007	-25.04216	0.00000	-0.00007	-19.50325	0.00000
d_scenic	-0.00004	-7.90171	0.00000	-0.00004	-11.24819	0.00000	-0.00004	-8.66548	0.00000
d_metro	0.00001	0.9902	0.32208	0.00002	2.21943	0.02646	0.00005	3.68354	0.00023
d_luoghi_culto	0.00004	1.59527	0.11066	0.0001	4.21801	0.00002	0.00016	5.39477	0.00000
omicidi_donne500	-124.58576	-3.45982	0.00054	-85.25567	-2.81522	0.00488	-12.52328	-0.31445	0.75318
furti_500	0.92135	1.79216	0.07311	2.50548	5.79494	0.00000	3.18138	5.59524	0.00000
rapine_500	-2.38339	-3.34363	0.00083	-3.50042	-5.83912	0.00000	-5.18264	-6.57391	0.00000
stupri_500	-0.26393	-2.10775	0.03506	-0.2579	-2.44901	0.01433	-0.57954	-4.18478	0.00003

Tabella 3.2: Regressione quantilica: modello 2

Capitolo 4

Conclusioni

Come si è illustrato in precedenza il fenomeno di Airbnb è estremamente complesso da affrontare e comprendere i fattori che vanno a influenzarne gli affitti per notte può aiutare a fare luce su questa tematica. Il lavoro svolto in questa tesi va esattamente in questa direzione. A partire da dati open disponibili si è costruito un dataset che contenesse tutte le informazioni riguardanti le proprietà affittate su Airbnb nella città di New York e i servizi forniti nel loro sottoquartiere di appartenenza. Si è scelto di operare a livello di sottoquartiere per tenere conto della realtà molto eterogenea e diversificata presente nella città di New York. I risultati ottenuti mostrano che il metodo della regressione quantilica può dimostrarsi efficace per trattare questo tipo di eterogeneità nell'analisi, essendo in grado di tenere conto della distribuzione a code pesanti dei prezzi degli affitti, ma anche dell'eterogeneità data dalla dipendenza spaziale presente all'interno dei dati. Per avere un'ulteriore conferma della bontà del metodo utilizzato si è proceduto a fare un confronto tra gli errori di previsione della regressione sul quantile $\tau = 0.5$ e alcuni modelli ritenuti potenzialmente concorrenti, selezionando le variabili utilizzate in modo automatico e dividendo il campione in maniera casuale in insieme di stima e di verifica. I risultati di questo confronto sono presentati nella tabella 4.1. A partire da questi si può osservare una miglior performance del modello prescelto rispetto agli altri in termini di MSE. Vale la pena di notare che altre tipologie di modelli come MARS e GAM non sono stati inseriti in quanto il loro MSE ha un ordine di grandezza non confrontabile con i modelli qui considerati. Si può quindi affermare che il metodo adottato in questa tesi sia efficace sia dal punto di vista interpretativo–dato che il modello fornisce risultati sensati dal punto di vista economico–sia dal punto

di vista predittivo, in quanto il metodo prescelto risulta presentare un valore di MSE più basso. Ulteriori sviluppi di questa analisi sono ovviamente legati ad una più approfondita interpretazione dei risultati conseguiti, ad una miglior calibrazione del modello considerato, eventualmente arricchendola con l'aggiunta di nuove variabili o tramite il confronto con strumenti di analisi che tengano conto della dipendenza spaziale (modelli econometrici) o dell'ordine presente nei dati (modelli gerarchici).

modello	MSE
regressione quantilica τ=0.5	0.13531
regressione lineare (stepwise)	0.13538
lasso	0.13547
regressione ridge	0.13648
albero di regressione	0.21726

Tabella 4.1: MSE ottenuti per i vari metodi di previsione utilizzati

Appendice A

Appendice codici utilizzati

A.1 Passaggio dal sistema di riferimento dello stato di New York a quello GPS

In questa appendice si trova il codice utilizzato per la conversione delle coordinate basate sul sistema di riferimento dello stato di New York in coordinate GPS.

```
####################passaggio da sistema coordinate stato di New York a lat/long
#caricamento librerie
library(rgdal)
library(sp)
##costruzione del dataset contenente XCoord e YCoord
x<-data.frame(dataset$XCoord,dataset$YCoord)
names(x)<-c("a","b")
##conversione del dataframe in SpatialPoints
coordinates(x) <- ~ a+b
#imposta o recupera gli attributi di proiezione per convertire i dati in formato SpatialData
proj4string(x) <- CRS("+init=epsg:2263")#sistema di riferimento stato di New York
#conversione in coordinate GPS
latlong = data.frame(spTransform(x, CRS("+init=epsg:4326")))
names(latlong)<-c("long","lat")
buildings_brooklyn<-cbind(buildings_brooklyn,latlong)
```

A.2 Codice per l'individuazione del sottoquartiere a partire da latitudine e longitudine

In questa appendice si trova il codice utilizzato per per inviduare il sottoquartiere dove si trova un punto designato da latitudine e longitudine.

```
#caricamento librerie da utilizzare
library(tigris)
library(ggplot2)
library(dplyr)
library(leaflet)
library(sp)
library(ggmap)
library(maptools)
library(broom)
library(httr)
library(rgdal)
###scaricamento del file .geojson contenente le informazioni sui sottoquartieri dal sito http://
    data.beta.nyc//
r <- GET('http://data.beta.nyc//dataset/0ff93d2d-90ba-457c-9f7e-39e47bf2ac5f/resource/
35dd04fb-81b3-479b-a074-a27a37888ce7/download/d085e2f8d0b54d4590b1e7
d1f35594c1pediacitiesnycneighborhoods.geojson')
#caricamento del file
nyc_neighborhoods <- readOGR(content(r,'text'), 'OGRGeoJSON', verbose = F)
nyc_neighborhoods@data
lats <- dataset$latitudine
lngs <-dataset$longitudine
##costruzione del dataset contenente latitudine e longitudine
points <- data.frame(lat=lats, lng=lngs)
points_spdf <- points
##conversione del dataframe in SpatialPoints
coordinates(points_spdf) <- ~lng + lat
#imposta o recupera gli attributi di proiezione per convertire i dati in formato  SpatialData
proj4string(points_spdf) <- proj4string(nyc_neighborhoods)
#funzione che abbina ad ogni punto il sottoquartiere di appartenenza
matches <- over(points_spdf, nyc_neighborhoods)
points <- cbind(points, matches)
points
```

A.3 Calcolo delle distanze

In questa appendice si trova il codice utilizzato per il calcolo della distanza di Haversine.

```
##funzione calcolo di tutte le distanze e della minima
function(lat1, long1, lat2, long2) {

  # definizione dimensioni
  dim1 <- length(lat1)
  dim2 <- length(lat2)

  # creazione vettori output
  dist_min <- rep(0, dim1)
  dist_mat <- matrix(0, dim1, dim2)

  D1 <- cbind(lat1, long1)
  D2 <- cbind(lat2, long2)
  ###calcolo distanze
  for (j in 1:dim1) {
    x            <- D1[j,]
    dist         <- apply(D2, 1, FUN = function(s) distHaversine(x, s))
    dist_min[j]  <- min(dist)
    dist_mat[j,] <- t(dist)
    print(j)
  }
  return (list(dist_min, dist_mat))
}
```

A.4 Calcolo del numero dei reati commessi

In questa appendice si trova il codice utilizzato per il conteggio dei crimini commessi nel raggio di 500,100 e 1500 di un certo punto in una certo intervallo di tempo.

```
#caricamento dei pacchetti per il calcolo parallelo
install.packages("foreach")
install.packages("doMC")
library(foreach)
library(geosphere)
```

```
library(doMC)
registerDoMC(16)
fun_reati_foreach <- function(date_annuncio, annuncio_lat1, annuncio_long1, crime_lat2,
    crime_long2, date_crime) {
  # inizializzazione parametri
  dim1         <- length(date_annuncio)
  ncrimini500  <- rep(0, dim1)
  ncrimini1000 <- rep(0, dim1)
  ncrimini1500 <- rep(0, dim1)

  lista<-foreach (i = 1:dim1) %dopar% {

    # Modulus operation
    if (i %% 500 == 0) {
      # Print on the screen some message
      cat(paste0("iteration: ", i, "\n"))
    }

    data_cont_fine   <- date_annuncio[i]
    data_cont_inizio <- (data_cont_fine -730)

    # calcolo degli indici dei crimini che sono stati commessi nel periodo
    date_crime_index = date_indice(data_cont_inizio, data_cont_fine, date_crime)

    # calcolo dei vettori che servono
    crime_lat2_  = crime_lat2[date_crime_index]
    crime_long2_ = crime_long2[date_crime_index]

    # calcoliamo le distanze
    distanze<-crime_dist_calculate(annuncio_lat1[i], annuncio_long1[i], crime_lat2, crime_long2)

    # determiniamo il numero di crimini che si trovano nei 3 raggi di 500, 1000 e 1500 mt
    #ncrimini500[i]<-length(distanze[distanze<=500])
    #ncrimini1000[i]<-length(distanze[distanze<=1000])
    #ncrimini1500[i]<-length(distanze[distanze<=1500])
    ncrimini500<-length(distanze[distanze<=500])
    ncrimini1000<-length(distanze[distanze<=1000])
    ncrimini1500<-length(distanze[distanze<=1500])
    ncrimini<-cbind(ncrimini500, ncrimini1000, ncrimini1500)
```

```
  }
  #ncrimini<-cbind(ncrimini500, ncrimini1000, ncrimini1500)
  ncrimini<-lista
  return(ncrimini)
}


fun_reati2 <- function(date_annuncio, annuncio_lat1, annuncio_long1, crime_lat2, crime_long2,
    date_crime) {

  # inizializzazione parametri
  dim1 <- length(date_annuncio)
  ncrimini500<-rep(0,dim1)
  ncrimini1000<-rep(0,dim1)
  ncrimini1500<-rep(0,dim1)
  for (i in 1:dim1) {

    # Modulus operation
    if (i %% 500 == 0) {
      # Print on the screen some message
      cat(paste0("iteration: ", i, "\n"))
    }

    data_cont_fine   <- date_annuncio[i]
    data_cont_inizio <- (data_cont_fine -730)

    # calcolo degli indici dei crimini che sono stati commessi nel periodo
    date_crime_index = date_indice(data_cont_inizio, data_cont_fine, date_crime)

    # calcolo deii vettori che  servono
    crime_lat2_  = crime_lat2[date_crime_index]
    crime_long2_ = crime_long2[date_crime_index]

    # calcolo delle distanze
    distanze<-crime_dist_calculate(annuncio_lat1[i], annuncio_long1[i], crime_lat2, crime_long2)
    # determiniamo il numero di crimini che si trovano nei 3 raggi di 500, 1000 e 1500 mt
    ncrimini500[i]<-length(distanze[distanze<=500])
    ncrimini1000[i]<-length(distanze[distanze<=1000])
    ncrimini1500[i]<-length(distanze[distanze<=1500])
```

```
  }
  ncrimini<-cbind(ncrimini500,ncrimini1000,ncrimini1500)
  return(ncrimini)
}

 #estrarre i dati dalla lista per metterli in un elemento di formato dataset
fun_recupera_lista <- function(lista) {

  n            <- length(lista)
  ncrimini500  <- rep(0, n)
  ncrimini1000 <- rep(0, n)
  ncrimini1500 <- rep(0, n)
  for (i in 1:n) {

    # Modulus operation
    if (i %% 500 == 0) {
      # Print on the screen some message
      cat(paste0("iteration: ", i, "\n"))
    }
    ll              <- unlist(lista[i], use.names = FALSE)
    ncrimini500[i]  <- ll[1]
    ncrimini1000[i] <- ll[2]
    ncrimini1500[i] <- ll[3]
  }
  ncrimini <- cbind(ncrimini500, ncrimini1000, ncrimini1500)
  return(ncrimini)
}
```

A.5 Quantile Regression

Questa appendice contiene il codice utilizzato per effettuare le stime tramite la Quantile Regression, implementato sulla base dell'algoritmo 1

```
quantile_check_fun=function(U, Tau){
                                    Loss = U * Tau - (U * (U < 0.0))
                                    return(Loss)
                                     }
LinQReg_MM = function(y, X, Tau, dTol_EPS, dTolX_MIN, dTolFun_MIN, dMax_IT){
```

```
#inizializzazione dei parametri dell'algoritmo MM
dL_DIF = Inf;
dItCounter = 0;
dL_OLD     = -Inf;
quantile_fun<-rep(0,dMax_IT)
 #LL=zeros(dMax_IT+1,cK);
#reg=zeros(dMax_IT,cK);
#inizializzazione dei parametri della quantile regression
beta=solve(t(X)%*%X)%*%t(X)%*%y;
dTolX= sqrt(t(beta)%*%beta);
beta_OLD =beta;
#calcolo dei residui
R_OLD =y - X %*% beta;
#calcolo costante c
Den          = dTol_EPS + abs(R_OLD);
Check_APPROX = quantile_check_fun(R_OLD, Tau) - 0.5 * dTol_EPS * log(dTol_EPS + abs(R_OLD));
ObjFun_MM    = 0.25 * ((R_OLD^2)/ Den + (4.0 * Tau - 2.0) * R_OLD);
Cost         = sum(Check_APPROX - ObjFun_MM);
####algoritmo MM
while ( (((dTolX > dTolX_MIN)&&(dL_DIF > dTolFun_MIN)) && (dItCounter < dMax_IT))==TRUE){
##contatore
dItCounter = dItCounter + 1;
 print(dItCounter);
#aggiornamento di ogni parametro
for(jt in 1:ncol(X)){
 #calcolo delle quantit\'a rilevanti
 Den = dTol_EPS + abs(R_OLD);
 vX   = X[,jt];
 vX_2 = vX^2;
 dA   = sum(vX_2/Den);
dB   = sum((R_OLD * vX + vX_2*beta[jt])/ Den);
dC   = sum(vX);
 # aggiornare i parametri di regressione
 dRegP_j_U = (dB + (2.0 * Tau - 1.0)*dC)/dA;
  beta[jt] = dRegP_j_U;
   }
 #calcolo funzione di perdita
 R = y-X%*%beta;
 dL = 0.25 * sum((R^2)/Den + (4.0 * Tau - 2.0) * R + Cost);
```

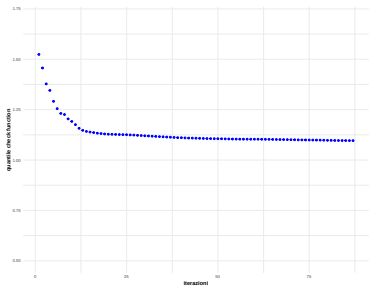
```
  #calcolo costante c
 vCheck_APPROX = quantile_check_fun(R,Tau)-0.5 *dTol_EPS *log(dTol_EPS +abs(R));
 vObjFun_MM    = 0.25 * ((R^2)/ Den + (4.0 * Tau - 2.0) *R);
 Cost= sum(vCheck_APPROX - vObjFun_MM);
#memorizzare loss
 dL_DIF    = abs(dL - dL_OLD);
dL_OLD    = dL;
 R_OLD    = R;
  dTolX= sqrt(t(beta - beta_OLD)%*%(beta - beta_OLD))
  return(beta)
        }
```

```
#funzione per il calcolo della matrice di varianza e covarianza asintotica
fQRAsyVarCovMat_1Q_IID = function (vY, mX, vRegP, dTau) {
  # ::::::::::::::::::::::::::::::::::::::::::
  # Set dimensions
  n = nrow(mX)
  # ::::::::::::::::::::::::::::::::::::::::::
  # %% calcolo dei residui.
  vRes = vY - mX %*% vRegP
  # ::::::::::::::::::::::::::::::::::::::::::
  # Compute matrices mQ and mD
  mQ = (t(mX) %*% mX) / n
  mD = fQVarCovMat(vRes, dTau)
  # ::::::::::::::::::::::::::::::::::::::::::
  # calcolo matrice di varianza e covarianza
  mQ_INV    = solve(mQ)
  mQR_VarC  = (mQ_INV * mD) / n
  return (mQR_VarC)
}
fQVarCovMat = function(vY, dTau) {
  vS_HAT = sparsity_function(vY, dTau);
  mOmega = dTau * (1.0 - dTau) * (vS_HAT %*% t(vS_HAT))
  return (mOmega)
}
sparsity_function = function (vY, dTau) {
  vX        = sort(vY)
  cH        = 0.15
  dMethKern = 4
```

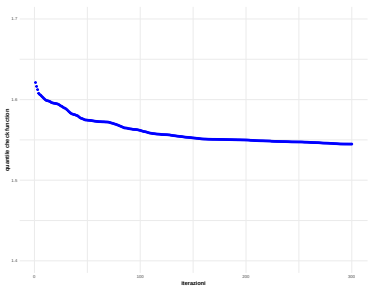
```
  cN        = length(vY)
  vK        = seq(0, 1, 1.0 / cN)
  vK_Y      = fkernDen(dTau, vK[1:cN], cH, dMethKern) - fkernDen(dTau, vK[2:(cN+1)], cH,
      dMethKern)
  dQ_HAT    = t(vK_Y) %*% vX
  return (dQ_HAT)
}
fkernDen = function (vY, vX, cH, dMethKern) {
  vU = (vY - vX) / cH;
  vF = fGetKernel(vU, dMethKern) / cH;

  return (vF)
}
fGetKernel = function(vX, dMethKern) {
  n  <- length(vX)
  vY <-rep(0, n)
  if (dMethKern == 1) {vY=(1/sqrt(2*pi))*exp(-0.5*(vX^2))}
  else if (dMethKern == 2) {vY = 0.5*(abs(vX) <= 1)}
  else if (dMethKern == 3) {vY = (1-abs(vX))*(abs(vX) <= 1)}
  else if (dMethKern == 4) {vY = (3/4)*(1-(vX^2))*(abs(vX) <= 1)}
  else if (dMethKern == 5) {vY = (35/32)*((1-(vX^2))^3)*(abs(vX) <= 1)}
  else if (dMethKern == 6) {vY = (70/81)*((1-(abs(vX)^3))^3)*(abs(vX) <= 1)}
  else if (dMethKern == 7) {vY = (pi/4)*cos(pi/2*vX)*(abs(vX) <= 1)}
  else if (dMethKern == 8) { vY = ones(1, n) }; #no weights
  return (t(vY))
}
```

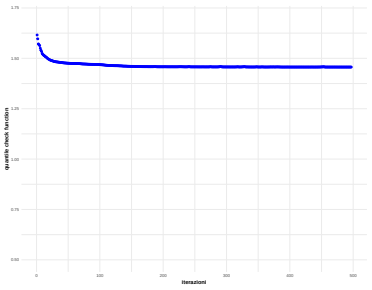
Inoltre per mostrare che la funzione opera correttamente si riportano i grafici dei valori che assume la quantile check function a ogni iterazione quando viene applicata al dataset Boston, contenuto nella libreria MASS di R. Se questa ha un andamento decrescente per ogni quantile –come avviene in questo caso– la funzione sta operando correttamente perché l'approssimazione si sta avvicinando sempre più alla funzione vera.



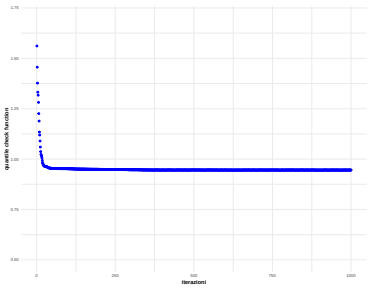
(a) Quantile 0.25



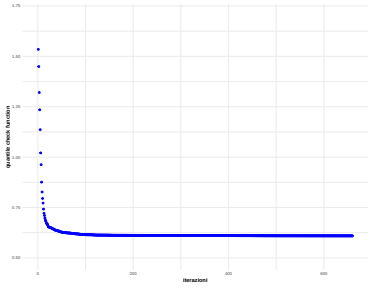
(b) Quantile 0.50



(c) Quantile 0.75



(d) Quantile 0.90



(e) Quantile 0.95

Figura A.1: Valore assunto della check quantile function stimata per ogni iterazione

Bibliografia

Azzalini, Adelchi e Bruno Scarpa (2012). *Data analysis and data mining: An introduction*. OUP USA.

Baden-Fuller, Charles e Vincent Mangematin (2013). «Business models: A challenging agenda». In: *Strategic Organization* 11.4, pp. 418–427.

Botsman, Rachel (2015). «Defining the sharing economy: what is collaborative consumption–and what isn't». In: *Fast Company* 27, p. 2015.

Botsman, Rachel e Roo Rogers (2010). «Beyond zipcar: Collaborative consumption». In: *Harvard Business Review* 88.10, p. 30.

— (2011). *What's mine is yours: how collaborative consumption is changing the way we live*. Vol. 5. Collins London.

Cervero, Robert, Aaron Golub e Brendan Nee (2007). «City CarShare: longer-term travel demand and car ownership impacts». In: *Transportation Research Record: Journal of the Transportation Research Board* 1992, pp. 70–80.

Dempster, Arthur P, Nan M Laird e Donald B Rubin (1977). «Maximum likelihood from incomplete data via the EM algorithm». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

Fang, Bin, Qiang Ye, Rob Law et al. (2016). «Effect of sharing economy on tourism industry employment». In: *Annals of Tourism Research* 57.3, pp. 264–267.

Frenken, Koen et al. (2015). «Smarter regulation for the sharing economy». In: *The Guardian* 20.

Gansky, Lisa (2010). *The mesh: Why the future of business is sharing*. Penguin.

Glind, PB Van de (2013). «The consumer potential of Collaborative Consumption: Identifying the motives of Dutch Collaborative Consumers & Measuring the consumer potential

of Collaborative Consumption within the municipality of Amsterdam». Tesi di laurea mag.

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.

Holt, Douglas B (2006). *Toward a sociology of branding*.

Hunter, David R e Kenneth Lange (2000). «Quantile regression via an MM algorithm». In: *Journal of Computational and Graphical Statistics* 9.1, pp. 60–77.

Ikkala, Tapio e Airi Lampinen (2015). «Monetizing network hospitality: Hospitality and sociability in the context of Airbnb». In: *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM, pp. 1033–1044.

infoplease (2015). *https://www.infoplease.com/us/states/idaho*.

Johnson, AnnDee (2014). «Idaho 2013 visitor profile Idaho». In: *IIdaho Conference on Recreation and Tourism*.

Koenker, Roger (2005). «Frontmatter». In: *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, pp. i–vi.

Koenker, Roger e Gilbert Bassett Jr (1978). «Regression quantiles». In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Koenker, Roger e Jose AF Machado (1999). «Goodness of fit and related inference processes for quantile regression». In: *Journal of the american statistical association* 94.448, pp. 1296–1310.

Lehr, Dean D (2015). «An analysis of the changing competitive landscape in the hotel industry regarding Airbnb». In:

Liang, Lena Jingen, HS Chris Choi e Marion Joppe (2018). «Understanding repurchase intention of Airbnb consumers: perceived authenticity, electronic word-of-mouth, and price sensitivity». In: *Journal of Travel & Tourism Marketing* 35.1, pp. 73–89.

Malhotra, Arvind e Marshall Van Alstyne (2014). «The dark side of the sharing economy... and how to lighten it». In: *Communications of the ACM* 57.11, pp. 24–27.

Martin, Elliot, Susan A Shaheen e Jeffrey Lidicker (2010). «Impact of carsharing on household vehicle holdings: Results from North American shared-use vehicle survey». In: *Transportation Research Record* 2143.1, pp. 150–158.

Mikhalkina, Tatiana e Laure Cabantous (2015). «Business model innovation: How iconic business models emerge». In: *Business models and modelling*. Emerald Group Publishing Limited, pp. 59–95.

Mosteller, Frederick e John Wilder Tukey (1977). «Data analysis and regression: a second course in statistics.» In: *Addison-Wesley Series in Behavioral Science: Quantitative Methods*.

Neeser, Dâvid, Martin Peitz e Jan Stuhler (2015). «Does Airbnb hurt hotel business: Evidence from the Nordic countries». In: *Universidad Carlos III de Madrid*, pp. 1–26.

Olma, Sebastian (2014). «Never mind the sharing economy: here's platform capitalism». In: *Institute of network cultures blog* 16.

Ortega, James M e Werner C Rheinboldt (1970). *Iterative solution of nonlinear equations in several variables*. Vol. 30. Siam.

Oskam, Jeroen e Albert Boswijk (2016). «Airbnb: the future of networked hospitality businesses». In: *Journal of Tourism Futures* 2.1, pp. 22–42.

Rifkin, Jeremy (2014). *The zero marginal cost society: The internet of things, the collaborative commons, and the eclipse of capitalism*. St. Martin's Press.

Rochet, Jean-Charles e Jean Tirole (2004). «Two-sided markets: an overview». In: *Toulouse, France: IDEI, mimeo, March*.

Rothkopf, E (2014). «CCTP-725: remix and dialogic culture». In: *Retrieved September* 20, p. 2017.

Sabatier, Valérie, Vincent Mangematin e Tristan Rousselle (2010). «From recipe to dinner: business model portfolios in the European biopharmaceutical industry». In: *Long Range Planning* 43.2-3, pp. 431–447.

Stern, Joseph (2010). «AirBnb benefits from social proof theory». In: *Retrieved August* 24, p. 2015.

Stors, Natalie e Andreas Kagermeier (2015). «Motives for Using Airbnb in Metropolitan Tourism—Why do People Sleep in the Bed of a Stranger?» In: *Regions Magazine* 299.1, pp. 17–19.

Tuttle, B (2015). «Marriott's CEO just made a pretty good sales pitch for... Airbnb?» In: *Money. com* 9.

Yannopoulou, Natalia, Mona Moufahim e Xuemei Bian (2013). «User-generated brands and social media: Couchsurfing and AirBnb». In: *Contemporary Management Research* 9.1.

Yue, Yu Ryan e Håvard Rue (2011). «Bayesian inference for additive mixed quantile regression models». In: *Computational Statistics & Data Analysis* 55.1, pp. 84–96.

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.

Zott, Christoph e Raphael Amit (2010). «Business model design: an activity system perspective». In: *Long range planning* 43.2-3, pp. 216–226.

Ringraziamenti

Trovarsi alle battute conclusive di un percorso di studi, quello universitario, lascia indubbiamente spazio a qualche momento di riflessione su quanto l'università e tutto ciò che le gravita attorno abbia influenzato ed arricchito la mia crescita personale negli ultimi cinque anni. Ci sono alcune persone che desidero ringraziare e nominare, per aver contribuito in un modo e nell'altro ad aiutarmi nell'arrivare al termine di questo percorso.

Il ringraziamento più grande e doveroso va ai miei genitori, senza i quali il percorso universitario sarebbe probabilmente stato molto più difficile: non solo per avermi sempre spinto a dare il meglio di me stessa, ma per avermi. Un ringraziamento che voglio naturalmente estendere ai miei parenti che mi hanno sempre supportato.

Un secondo ringraziamento va alla prof.ssa Guidolin e al prof. Bernardi, per l'infinita disponibilità nell'aiutarmi a svolgere questo lavoro e per il tempo dedicatomi in questi mesi, grazie davvero senza di voi non ce l'avrei fatta a fare un lavoro così soddisfacente.

Un ringraziamento speciale a Sara per non avermi mai fatto mancare la sua amicizia e il suo supporto. Un menzione speciale per Ester grazie di esserci sempre stata in ogni momento, l'ho apprezzato molto. Altra menzione speciale per Laura grazie per il tuo affetto e la tua simpatia. Un ringraziamento ad Elena per il tuo affetto che non è mai mancato Un ringraziamento a Laura e Francesca per la vostra amicizia che ormai dura da una vita.

Un ringraziamento a tutti gli amici padovani e a tutti i compagni di corso : nomi ed elenchi sarebbero troppo lunghi e rischierei ingiustamente di dimenticare qualcuno.

Un ultimo ringraziamento, più formale e diretto al lavoro svolto in questa tesi, a Paolo Scopelliti e Vincenzo Agosto per avermi permesso di sfruttare le risorse computazionali messe a disposizione dell'ateneo sulle quali eseguire i calcoli computazionalmente più onerosi svolti in questo lavoro.