Università degli studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche

Tesi di laurea

Open data al servizio del bene pubblico: analisi di un caso pratico a New York

Relatore Dott.ssa Mariangela Guidolin
Dipartimento di Scienze Statistiche

Laureanda Clelia De Michieli
Matricola N 1202697

Anno Accademico 2019/2020

Indice

Elenco dei codici

Elenco delle figure

Elenco delle tabelle

Introduzione

Negli ultimi decenni si è assistito ad un rapido avanzamento tecnologico che ha permesso di raccogliere e gestire online grandi moli di dati. È in tale contesto che hanno iniziato a diffondersi gli Open Data. Sebbene non esista una definizione unica ed universalmente accettata del termine, è invece chiaro quali siano i requisiti necessari affinché tali dati possano essere effettivamente considerati "open". Le caratteristiche imprescindibili che un dato aperto deve possedere possono essere riassunte nelle seguenti proprietà: *completezza, accessibilità e gratuità, integrabilità, aggiornamento continuo, machine-readable*.
Tuttavia, nonostante il crescente interesse verso questa tematica e l'apparente abbondanza di dati aperti reperibili online, solo una piccola percentuale di questi dati può realmente essere considerata "open". Infatti spesso vengono dichiarati Open Data dati che rispettano solo in parte le caratteristiche sopraelencate che dovrebbero renderli tali. Ciò ne limita fortemente l'utilizzo e fa sì che il potenziale dei dati aperti sia ad oggi largamente inespresso.
Nonostante le varie limitazioni, nella sola Europa il valore degli Open Data era di 184 miliardi di euro nel 2019 e si stima che possa crescere fino a 334,21 miliardi di euro nel 2025. Ma i benefici non si riscontrano solo in ambito economico. L'utilizzo di Open Data nel settore pubblico può ad esempio facilitare un maggior coinvolgimento dei cittadini nella vita pubblica, favorendo l'insorgere di un'amministrazione più controllata ed efficiente.

In questo elaborato si vuole far emergere come l'utilizzo di Open Data possa portare a risultati significativi nello studiare tematiche relative al benessere pubblico. A tale scopo, nel modellare il fenomeno oggetto di studio, saranno impiegate tecniche non parametriche di analisi dei dati. Una delle

caratteristiche proprie dell'approccio non parametrico è infatti quella di svincolarsi dalle assunzioni circa il fenomeno generatore dei dati, lasciando che siano i dati stessi a far emergere eventuali relazioni.
Tra le tematiche attuali e di interesse pubblico si è quindi scelto di concentrarsi sul problema della sicurezza nelle grandi città, ritenendo di interesse valutare quali fattori possano incidere sulle dinamiche relative alla criminalità. Focalizzandosi sul caso concreto delle aggressioni nella città di New York, l'elaborato si pone quindi l'obiettivo di mostrare come l'utilizzo di soli Open Data, accompagnati da tecniche non parametriche di analisi dei dati, sia fondamentale nella comprensione del fenomeno. I risultati delle analisi possono essere ritenuti di interesse anche dal punto di vista pubblico: possono infatti essere d'aiuto per le amministrazioni nel prendere iniziative volte a migliorare la qualità della vita dei cittadini.

Per raggiungere tale obiettivo, l'elaborato è stato così strutturato:

- Il primo capitolo introduce il concetto di Open Data. Vengono proposte alcune definizioni del termine e si descrivono le caratteristiche principali che rendono i dati "open". Si procede poi con illustrarne lo sviluppo temporale e i limiti attuali di utilizzo, ma anche il potenziale in termini di benefici economici e di benessere sociale. Infine si riportano alcuni casi pratici di utilizzo degli Open Data e si introduce il caso di studio affrontato in questo elaborato.

- Il secondo capitolo si compone di due parti. Nella prima si fa riferimento alla letteratura che si occupa di studiare fenomeni di criminalità, descrivendo quali fattori incidono maggiormente sulla propensione a commettere reati. Viene inoltre presentato in maggior dettaglio il caso di studio considerato, soffermandosi sulle motivazioni che hanno portato alla scelta della città di New York. La seconda parte del capitolo è invece dedicata alla costruzione del dataset e alle analisi esplorative.

- Il terzo capitolo è dedicato ad analizzare il fenomeno tramite modellazione statistica. Dapprima si illustra l'approccio utilizzato per studiare il caso di studio e si descrivono brevemente i modelli selezionati, in se-

guito si presentano e discutono i risultati ottenuti dall'adattamento di questi modelli.

- Infine il quarto capitolo è dedicato ad una serie di considerazioni conclusive. Si ripercorrono le scelte compiute durante l'elaborato, evidenziando limiti e possibili sviluppi futuri delle analisi svolte.

Capitolo 1

Open Data

Il primo capitolo di questo elaborato è volto alla trattazione degli Open Data. Nei paragrafi 1.1 e 1.2 vengono illustrate alcune definizioni e caratteristiche dei dati aperti, ripercorrendone sviluppo e diffusione fino ai limiti d'utilizzo che derivano da una ancora non piena consapevolezza delle potenzialità; nei paragrafi 1.3 e 1.4 viene dapprima descritto il potenziale degli Open Data, in termini di benefici sia diretti che indiretti, e in seguito si presentano alcuni casi pratici relativi al loro utilizzo. Infine nel paragrafo 1.5 viene introdotto il caso di studio affrontato, chiarendo i motivi di tale scelta.

1.1 Definizioni e caratteristiche

Sebbene il termine Open Data si sia sempre più diffuso negli anni, attualmente non ne esiste una definizione unica e universalmente accettata. Nella letteratura corrente è infatti possibile trovare diverse definizioni ugualmente valide.

Una definizione a cui si fa comunemente riferimento è quella contenuta nella Open Definition[1], secondo cui:

[1]Documento redatto dalla fondazione non profit Open Knowledge Foundation, fondata il 24 maggio 2004 a Cambridge con lo scopo di promuovere l'apertura dei contenuti e i dati aperti.

I dati aperti sono dati che possono essere liberamente utilizzati, riutilizzati e ridistribuiti da chiunque, soggetti eventualmente alla necessità di citarne la fonte e di condividerli con lo stesso tipo di licenza con cui sono stati originariamente rilasciati.

L'Agenzia per l'Italia digitale (AgID)[2] definisce invece i dati aperti nel seguente modo:

Gli open data sono dati pubblici che devono essere pubblicati in maniera che sia facile il loro riutilizzo. A tal fine sono fondamentali aspetti quali: licenze, standardizzazione, qualità, accessibilità anche attraverso applicazioni automatizzate.

Infine l'ultima definizione che si riporta è quella fornita dall'International Open Data Charter[3], secondo cui:

Gli open data sono dati digitali che vengono messi a disposizione liberamente e che presentano caratteristiche tecniche e legali necessarie per essere liberamente utilizzati, riutilizzati e ridistribuiti da chiunque, sempre e ovunque.

Ciò che accomuna tutte le definizioni è l'idea che i dati aperti appartengano alla comunità e possano essere consultati e riutilizzati liberamente. Il riutilizzo dei dati può infatti permettere di creare nuove risorse, applicazioni e servizi di pubblica utilità, generando quindi valore economico e benessere sociale. Nonostante il grande potenziale che gli Open Data possono generare, la quasi totalità dei dati accessibili liberamente proviene da enti pubblici. Spesso gli enti privati mostrano infatti renitenza di fronte alla possibilità di diffondere il proprio patrimonio informativo. Molti governi invece, in un'ottica di

[2] https://avanzamentodigitale.italia.it/it/progetto/open-data.

[3] https://opendatacharter.net/principles/. La Carta internazionale dei dati aperti è una convenzione internazionale che contiene principi e buone pratiche concernenti il rilascio di dati aperti da parte di enti governativi. La Carta è stata sottoscritta inizialmente da 17 governi nazionali (tra cui quello italiano) e locali durante l'incontro globale sull'Open Government Partnership a Città del Messico nell'ottobre 2015.

Open Government[4], hanno scelto di rendere pubblici tutti i dati della pubblica amministrazione. Ciò permette ai cittadini di avere accesso alle informazioni circa il funzionamento e l'operato delle pubbliche amministrazioni, garantendone la totale trasparenza. Inoltre rendere l'amministrazione trasparente può portare anche ad una maggiore partecipazione da parte del cittadino, favorendo l'insorgere di un'amministrazione più controllata ed efficiente.

Indipendentemente dalla natura pubblica o privata dell'ente che rende disponibili i dati, le caratteristiche che accomunano gli Open Data possono essere riassunte nei successivi cinque punti:

- *Disponibili e completi.* I dati devono essere disponibili nel loro complesso (comprendendo tutte le componenti tra cui i metadati), per un prezzo non superiore ad un ragionevole costo di riproduzione.

- *Accessibili senza restrizioni.* I dati devono essere disponibili in formati aperti e pubblici e senza il ricorso a piattaforme proprietarie. Devono essere inoltre resi disponibili senza alcuna sottoscrizione di contratto, pagamento, registrazione o richiesta.

- *Riutilizzabili e integrabili.* I dati non devono devono essere caratterizzati da licenze che ne limitino l'uso, la diffusione o la redistribuzione. È inoltre necessario che siano presentati in maniera sufficientemente granulare, così che possano essere integrati e aggregati con altre basi dati (interoperabilità).

- *Aggiornati.* È necessario che i dati siano aggiornati periodicamente per garantire continuità e possibilità di analisi.

- *Leggibili da computer.* Per garantire agli utenti la piena libertà di accesso e soprattutto di utilizzo e integrazione dei contenuti digitali, è necessario che i dati siano machine-readable, ovvero processabili in automatico dal computer.

[4]La dottrina dell'Open Government si basa sul principio per cui tutte le attività dei Governi e delle Amministrazioni dello Stato devono essere aperte e disponibili.

Un concetto fondamentale sui cui soffermarsi è quello della interoperabilità. Infatti affinché i dati possano effettivamente generare valore, gli utenti devono essere messi in condizione di riutilizzarli e integrarli. Spesso però si riscontrano gravi problemi di interoperabilità che ostacolano la combinazione di dati provenienti da fonti diverse. Ciò può essere dovuto sia ai formati con cui vengono resi disponibili i dati sia alla qualità dei dati stessi.
Per classificare il formato e la qualità dei dati è pratica comune fare riferimento alla scala a cinque stelle di Tim Berners-Lee, l'inventore del World Wide Web. Secondo questo riferimento, adottato anche dell'Agenzia per l'Italia Digitale per la pubblicazione degli Open Data italiani, i dati possono essere classificati in cinque categorie, da una stella (dati non strutturati) a cinque stelle (Linked Open Data, LOD). Quest'ultimo formato rappresenta la tipologia di dati aperti che consente il massimo livello di interoperabilità tra dataset, rendendo possibile effettuare correlazioni tra più dataset indipendenti l'uno dall'altro.

★ I dati non sono strutturati. Sono disponibili tramite una licenza aperta, ma sono inclusi in documenti leggibili e interpretabili solo grazie a un significativo intervento umano (ad esempio in formato PDF).

★★ I dati sono disponibili in forma strutturata e con licenza aperta. Tuttavia, i formati sono proprietari (ad esempio Excel) e un intervento umano è fortemente necessario per un'elaborazione dei dati.

★★★ I dati sono strutturati come nel livello precedente, ma in un formato non proprietario (e.g., CSV, JSON, geoJSON). Sono inoltre leggibili da programmi, ma l'intervento umano è necessario per una qualche elaborazione degli stessi.

★★★★ Oltre a rispettare tutti i criteri precedenti, i dati fanno uso di standard aperti (ad esempio RDF e SPARQL) e sono dotati di un identificativo unico di risorsa (URI) che li rende indirizzabili sulla rete e quindi utilizzabili direttamente online, consentendo a determinati programmi di elaborarli senza quasi ulteriori interventi umani.

★★★★★ I dati rispettano tutti gli altri criteri e inoltre contengono collegamenti ad altri dati al fine di fornire un contesto alle proprie informazioni. In questo caso si può effettivamente parlare di Linked Open Data.

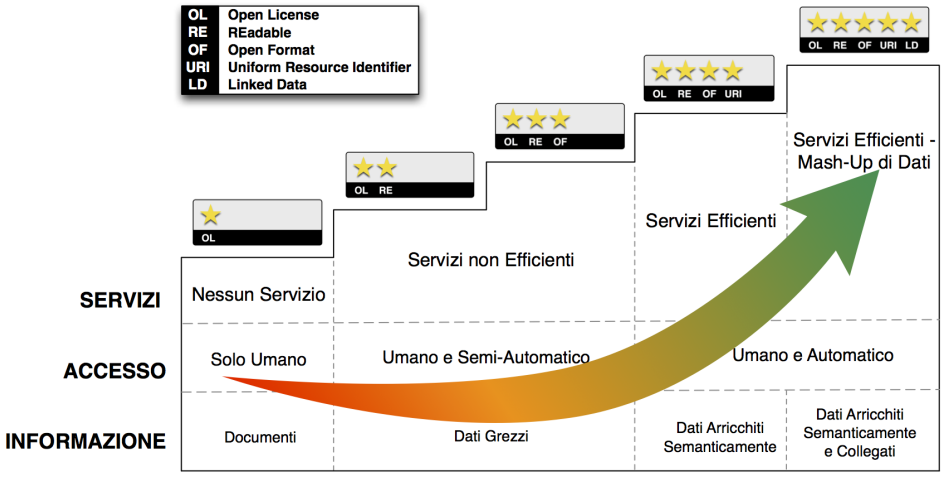


Figura 1.1: Modello a cinque stelle per i dati aperti, https://docs.italia.it/italia/daf/lg-patrimonio-pubblico/it/bozza/modellodati.html

1.2 Diffusione degli Open Data e limiti di utilizzo

Sebbene la Open Definition risalga al 2005, la diffusione di portali da cui avere accesso ad Open Data non è stata rapidissima. La svolta decisiva è avvenuta nel 2009, quando l'allora presidente degli Stati Uniti Barack Obama al suo primo giorno di presidenza ha siglato il "Memorandum on Transparency and Open Government" [5], ponendo le basi dell'Open Government. Nel documento vengono elencati i tre principi chiave che devono essere alla base di un governo aperto: trasparenza, partecipazione e collaborazione. Al Memorandum viene dato un seguito pratico nel maggio dello stesso anno con la creazione del portale Data.gov. Il portale nasce con l'intenzione di raccogliere i dati governativi statunitensi per metterli a disposizione di tutti i cittadini,

[5] Per maggiori informazioni si veda https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government.

permettendo loro di produrre beni e servizi usando come materia prima i dati aperti. Seguendo l'esempio americano, nello stesso anno sono nati anche i portali governativi di Regno Unito e Nuova Zelanda, seguiti a breve distanza da quelli di Norvegia, Australia, Marocco, Kenya, Cile, Olanda, Spagna, Slovenia, Italia, Belgio, Estonia, Francia e Portogallo. Alla fine del 2015 ventisette Paesi europei possiedono un portale governativo da cui è possibile avere accesso a dati aperti. Nel novembre dello stesso anno viene inoltre istituito l'European Data Portal, allo scopo di raccogliere in un unico portale le informazioni del settore pubblico disponibili sui portali dei vari paesi europei. Sul sito sono inoltre presenti diverse relazioni volte a far conoscere il potenziale non solo economico degli Open Data e lo stato di maturità dei Paesi europei circa l'apertura e disponibilità dei propri dati.

Nonostante l'apparente moltitudine di portali di Open Data ormai diffusi in quasi tutto il mondo, gli Open Data costituiscono un'eccezione e non una regola. È quanto emerge dall'ultimo studio condotto dall'Open Data Barometer[6] che considera i dati presenti sui portali governativi di 115 Paesi. Dall'analisi[7] risulta che solo il 7% dei dataset considerati può realmente essere definito aperto secondo le caratteristiche definite nella sezione 1.1. Ciò è dovuto principalmente alla qualità dei dati messi a disposizione sui portali: molti dati sono incompleti, frammentati, non aggiornati o aggiornati manualmente e sprovvisti dei metadati necessari alla loro corretta interpretazione. Inoltre tematiche di interesse per il cittadino vengono ignorate, andando contro il principio di trasparenza che dovrebbe caratterizzare un portale governativo. Ad esempio i dati sulla spesa pubblica sono disponibili solo nel 3% dei Paesi considerati, mentre i dati relativi agli appalti pubblici, alla proprietà delle imprese e alla proprietà terriera sono tra i meno aperti e sono spesso di scarsa qualità. Affinché si possa trarre valore dall'utilizzo degli Open Data è invece necessario che questi siano di interesse per il cittadino. Molti portali vantano infatti tanti dati ma di questi pochi sono relativi a tematiche di interesse

[6] Si tratta di un programma in seno alla World Wide Wef Foundation che mira ad individuare la reale diffusione e l'impatto delle iniziative open data in tutto il mondo. https://opendatabarometer.org/barometer/.

[7] Per ulteriori informazioni sullo studio si veda https://opendatabarometer.org/doc/4thEdition/ODB-4thEdition-GlobalReport.pdf.

collettivo. Una grande quantità di dati non è sempre un bene: quando i dati diventano addirittura troppi o non sono organizzati opportunamente rischiano di creare confusione nell'utilizzatore sfociando in una minore comprensione delle azioni della pubblica amministrazione. Tuttavia, anche in presenza di Open Data di qualità e relativi ad ambiti di interesse, sussistono dei limiti di utilizzo dovuti al basso livello di conoscenza della popolazione verso questa tematica. Purtroppo gran parte dei cittadini ignora il significato del termine Open Data, né conosce le potenzialità che tali dati possono rappresentare, oppure non ha gli strumenti o le conoscenze tecniche per poterli analizzare estraendone valore.
Infine sulla pagina dell'Open Data Barometer è possibile trovare una classifica dei Paesi coinvolti nell'analisi. Ad ogni Paese viene attribuito un punteggio tenendo conto dello stato di maturità delle iniziative basate sull'utilizzo di Open Data, della loro attuazione e dell'impatto economico, politico e sociale da queste generato. I primi dieci Paesi per punteggio sono Regno Unito, Canada, Francia, Stati Uniti, Corea, Australia, Nuova Zelanda, Giappone, Paesi Bassi e Norvegia. L'Italia si colloca al ventesimo posto ma con un punteggio di soli 56 punti su un totale di 100.
Dalla cartina in figura 1.2 si può inoltre osservare come le criticità maggiori si riscontrino in Africa e Asia. Questo rischia di portare ad un ulteriore divario tra Paesi ricchi e Paesi in via di sviluppo. Gli Open Data per loro natura dovrebbero invece essere una risorsa accessibile a tutti.

1.3 Valore degli Open Data

Gli Open Data non rappresentano solo un importante elemento di trasparenza delle amministrazioni verso i cittadini, ma anche un'opportunità di crescita economica e di occupazione. Dal rapporto 2020 sul Valore economico dei dati aperti[8] emerge che il valore di mercato degli Open Data in Europa è in forte crescita. Le dimensioni del mercato europeo[9] dei dati aperti sono sti-

[8]European Data Portal, Open Data Impact https://www.europeandataportal.eu/sites/default/files/the-economic-impact-of-open-data.pdf.

[9]Nello studio si considerano i 27 Paesi membri dell'UE più i Paesi appartenenti all'associazione europea di libero scambio (Norvegia, Islanda, Liechtenstein e Svizzera).

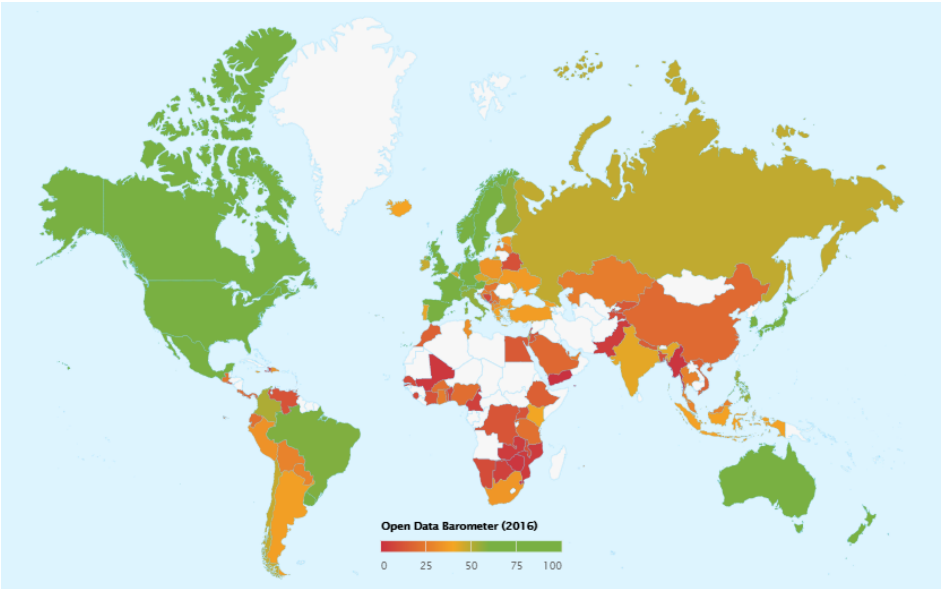


Figura 1.2: Open Data Barometer, risultati 2016, https://opendatabarometer.org/4thedition/?_year=2016&indicator=ODB

mate in 184 miliardi di euro nel 2019 e secondo le previsioni raggiungeranno un valore compreso tra i 199,51 e 334,21[10] miliardi di euro nel 2025. L'imponenza del fenomeno si riflette anche sul mercato occupazionale. In questo caso le stime parlano di 1,12 milioni di posizioni lavorative nel 2019 e 1,12-1,97[11] milioni previste per il 2025. Infine non meno importanti sono i benefici previsti in termini di risparmi ed efficienza. In questi casi le stime parlano ad esempio di 1,1 miliardi di euro risparmiati nel settore pubblico, di 27 milioni di ore guadagnate nell'attesa dei trasporti pubblici e di una riduzione di 5,8 Mtoe[12] nei consumi energetici domestici.

Il valore generato dagli Open Data può essere inoltre distinto in valo-

[10]Le previsioni sono calcolate utilizzando il tasso base di crescita del prodotto interno lordo europeo (1,0-1,4%) e quello ottimistico (10,4%)

[11]Le previsioni sono calcolate utilizzando il tasso base dello 0,5% stimato dai tassi forniti dalla Banca Centrale europea (BCE), dall'European Economic Forecast e dal Centro europeo per lo sviluppo della formazione professionale (Cedefop) e quello ottimistico (10,4%)

[12]Tonnellata equivalente di petrolio, è un'unità di misura dell'energia.

re diretto e indiretto. Nel primo caso si considerano i benefici propriamente monetizzati sotto forma di ricavi e di valore aggiunto lordo dell'UE. Il valore indiretto è invece strettamente collegato al potenziale degli Open Data: creazione di nuovi posti di lavoro, beni e servizi, risparmio di tempo per gli utenti che utilizzano applicazioni basate su dati aperti, maggiore efficienza nei servizi pubblici e crescita dei mercati correlati. Un altro studio[13] finanziato dall'European Data Portal rivela quali sono i settori che possono trarre maggiori benefici dal riutilizzo dei dati aperti. A fronte di un valore diretto del mercato stimato di 75,7 miliardi di euro, la Pubblica Amministrazione è il settore che maggiormente beneficia dall'adozione del modello "open". Il valore stimato per la PA è infatti di 22,1 miliardi di euro, più del doppio del giro d'affari attribuito al comparto industriale, (10 miliardi). Seguono poi i settori del commercio e dei trasporti (9,9 miliardi), dell'immobiliare (9 miliardi) e quello dei servizi professionali (8,3 miliardi). Per l'agricoltura, l'arte e l'intrattenimento, i benefici stimati sono minori, ma ciò non significa che gli Open Data non abbiano potenziale in questi settori, semplicemente sarà necessario più tempo per raggiungere la piena maturità.

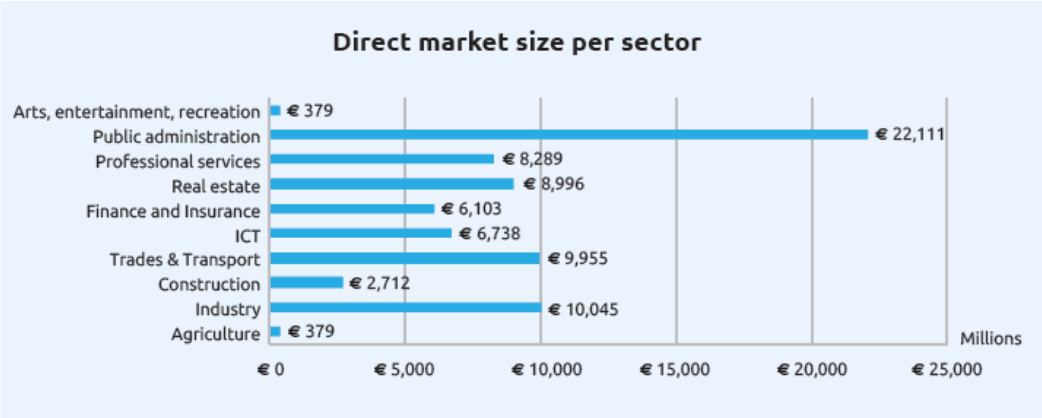


Figura 1.3: Valore del mercato diretto stimato per il 2020.
Fonte: Creating Value through Open Data, https://www.europeandataportal.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf

[13] Creating Value through Open Data, https://www.europeandataportal.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf.

1.4 Possibili applicazioni

Come si è visto nella sezione precedente, gli Open Data rappresentano un enorme potenziale in svariati settori, sia pubblici che privati. Tuttavia l'innovazione non è prevedibile e può dare origine a sviluppi inaspettati in settori ad oggi non immaginabili. Affinché ciò possa realizzarsi, è però necessario che più dati possibili siano resi aperti secondo gli standard descritti nella sezione 1.1.

Di seguito si illustrano alcuni casi pratici in cui la generazione e/o l'utilizzo di Open Data comporta benefici sia in termini economici, sia in termini di pubblica utilità.

OpenStreetMap

OpenStreetMap (OSM) è un progetto collaborativo, nato in seno alla OpenStreetMap Foundation, finalizzato a creare mappe del mondo a contenuto libero. La volontà di fornire dati geografici con una licenza libera è nata dalla constatazione che la maggior parte delle mappe online che si credono liberamente utilizzabili hanno invece restrizioni legali. Ciò costituisce un forte vincolo allo sviluppo, in quanto impedisce di riutilizzare i dati. I dati presenti in OSM sono invece utilizzabili liberamente per qualsiasi scopo, anche commerciale, con il solo vincolo di citare la fonte e usare la stessa licenza per eventuali lavori derivati. Da non sottovalutare è anche il fatto che dal sito è possibile scaricare gratuitamente i dati grezzi e non solo le mappe, dando vita ad una infinità di possibili utilizzi, dal routing alla mappa per videogiochi passando per svariate tipologie di analisi.
Un altro aspetto cruciale è il fatto che tutti possono contribuire arricchendo o correggendo i dati. Ciò permette di avere a disposizione dati continuamente aggiornati e controllati. Inoltre ai contributori viene data la possibilità di mappare qualsiasi tipo di oggetto. Oltre alle caratteristiche presenti su qualsiasi mappa, è infatti possibile visualizzare panchine, fontanelle dell'acqua o anche rifugi alpini. Ciò permette di creare prodotti di interesse per tantissimi settori. OpenStreetMap fornisce infatti dati geografici per migliaia di siti web, applicazioni mobili e dispositivi hardware. Un esempio interessante è costituito dal riutilizzo delle mappe su dispositivi GPS cartografici. Gli amanti del

trekking, del ciclismo o del geocaching possono così avere a disposizione la mappatura di sentieri di montagna, piste e percorsi ciclabili e altri elementi di interesse.
Oltre che per usi amatoriali, le mappe di OSM vengono impiegate anche per usi professionali. Ad esempio l'Agenzia delle Entrate utilizza le mappe di OpenStreetMap per permettere agli utenti di consultare le quotazioni dell'Osservatorio del Mercato Immobiliare, mentre la nota applicazione Moovit le utilizza per mostrare il percorso da seguire per raggiungere la destinazione scelta, le fermate dei trasporti pubblici e qualunque altra ricerca su mappa.
Infine il progetto OpenStreetMap si è rivelato fondamentale in ambito umanitario in seguito al terremoto di Haiti del 2010. In quella occasione gli utenti di OSM hanno provveduto alla digitalizzazione delle ortofoto[14], messe a disposizione da Google, potendo così segnalare in tempo reale la presenza di campi di soccorso, di ponti e strade distrutti e di altri elementi utili per i soccorritori. L'esperienza ha inoltre mostrato al mondo come i dati creati dal basso possono essere essenziali in certe situazioni. Tutti i dati provenienti da enti privati non erano infatti aggiornati e non potevano quindi essere utilizzati per favorire i soccorsi.

Emergenza Covid-19

L'emergenza Covid-19 sta permettendo di comprendere sempre più il grande valore degli Open Data. In tutto il mondo i dati sul contagio sono stati prontamente messi a disposizione in modo aperto dalle Pubbliche Amministrazioni, consentendo di utilizzarli non solo per capire la portata del fenomeno, ma anche per fare previsioni sull'andamento della pandemia.
Per quanto riguarda l'Italia, i numeri del contagio vengono resi disponibili dalla Protezione Civile su un repository pubblico in GitHub e aggiornati quotidianamente alle 18:30. A partire da questi dati è stato inoltre creato un cruscotto geografico interattivo che permette a tutti i cittadini di essere informati sullo stato dell'epidemia in Italia. È infatti possibile indagare il numero totale di casi, di guariti e di deceduti, ma anche visualizzare gli incrementi

[14] È una fotografia aerea georeferenziata in modo tale che la scala di rappresentazione della fotografia sia uniforme, potendo così considerare la fotografia equivalente ad una carta geografica.

giornalieri e gli spaccati relativi alle varie regioni. Tramite grafici e mappe è inoltre possibile osservare l'evoluzione del fenomeno sia nel tempo che nello spazio. Non meno importante è il fatto che i dati siano rilasciati con licenza "CC-BY-4.0" che ne permette la loro completa ridistribuzione e modifica. Grazie a questo, una piattaforma open source di business analytics, Knowage, ha potuto utilizzare i dati per creare cruscotti che mostrano previsioni a 20 giorni sugli sviluppi dell'epidemia, permettendo anche di fare confronti tra le varie regioni d'Italia.
Un'altra iniziativa interessante è quella avviata a marzo 2020 dal Complexity Science Hub (CHS) di Vienna. In questo caso si è provveduto a raccogliere a livello mondiale tutte le misure governative che sono state messe in campo per contrastare la pandemia. I primi risultati sono stati pubblicati il 27 agosto in un articolo sulla rivista Nature Scientific Data[15] e possono essere navigati liberamente sulla piattaforma dedicata del CHS. Tutti i dati sono stati raccolti da fonti pubbliche: fonti governative ufficiali, articoli scientifici, comunicati stampa, comunicazioni governative e social media. Come affermato dalla leader del progetto e prima autrice dell'articolo, Amelie Desvars-Larrive, lo studio ha permesso di quantificare l'impatto delle politiche di controllo individuali messe in campo dai vari governi nel contrastare la diffusione del coronavirus. Risultati del genere possono aiutare a rispondere in modo migliore all'emergenza, seguendo gli esempi virtuosi che meglio hanno funzionato.

Amministrazione trasparente

Nell'ambito della trasparenza, un esempio significativo è rappresentato dal portale Soldi Pubblici realizzato nel 2014 dall'AgID. Il sito è nato per "promuovere e migliorare l'accesso e la comprensione dei cittadini sui dati della spesa della Pubblica Amministrazione, in un'ottica di maggiore trasparenza e partecipazione alla cosa pubblica", come si legge dal notiziario presente sul sito dell'AgID, e rende disponibili i dati di Ministeri, Regioni, Aziende Sanitarie Regionali, Province e Comuni. Tali dati sono tratti dal sistema SIOPE [16]

[15] Per maggiori informazioni sull'articolo si veda https://www.nature.com/articles/s41597-020-00609-9.

[16] Si tratta del Sistema informativo delle operazioni degli enti pubblici che aggrega i pagamenti giornalieri delle diverse PA attraverso una serie di circa 250 codifiche gestionali.

in collaborazione con la Banca d'Italia e la Ragioneria Generale dello Stato e vengono aggiornati settimanalmente e aggregati sul mese precedente a quello in corso.
Consultare il sito è semplice e alla portata di tutti. Dalla home page del portale è infatti possibile selezionare un ente di interesse (Comune, Provincia o Regione ad esempio) e una voce di spesa per poi visualizzare grafici che riportano l'andamento della spesa nel tempo confrontandolo con quello dell'anno passato e con la media nazionale relativa a quella voce di costo. Inoltre è possibile scaricare il risultato della ricerca in formato csv. Nelle FAQ è infatti espressamente riportato che tutti i dati presenti sul portale possono essere scaricati e riutilizzati.

Si sono riportati solo tre esempi di casi in cui gli Open Data sono generati e/o utilizzati per produrre valore, ma le applicazioni possibili sono quasi infinite e spesso anche fantasiose. In Danimarca esiste ad esempio un sito che migliora la vita delle persone incontinenti, FindToilet. Tramite una mappa è infatti possibile visualizzare la posizione di tutti i bagni pubblici nell'arco del raggio chilometrico selezionato dall'utente. Invece in Scozia tramite la piattaforma Air Quality in Scotland è possibile salvaguardare la propria salute conoscendo la qualità dell'aria. Registrandosi al sito si possono anche ricevere avvisi giornalieri sui valori degli inquinanti e sui possibili sviluppi nei giorni seguenti. Nel panorama italiano si evidenzia BicinCittà, il servizio di bikesharing che utilizzando le mappe libere di OpenStreetMap permette di trovare le biciclette disponibili in oltre 100 città. Infine per gli amanti dello sci esiste la piattaforma MySnowMaps che tramite i dati delle stazioni meteorologiche, delle previsioni meteorologiche e di altre misure meteorologiche aperte ufficiali permette agli utenti di verificare quanta neve è presente sulle Alpi.

1.5 Obiettivo dell'elaborato

Dalle sezioni precedenti è emerso come gli Open Data rappresentino una preziosa opportunità di cui non si ha però ancora piena consapevolezza. In

questa tesi ci si vuole focalizzare sul valore che i dati aperti possono generare in ambito pubblico, se analizzati tramite modelli statistici appropriati.

L'analisi dei dati, ed in particolare il data mining, sono stati oggetto di forte sviluppo negli ultimi decenni. Il rapido avanzamento tecnologico ha infatti reso possibile l'analisi di grandi moli di dati e l'implementazione di modelli che prima erano solo stati teorizzati.
Il settore che ha tratto maggior beneficio dall'impiego di queste tecniche è quello aziendale (Shmueli et al. 2017). Tra le molteplici applicazioni possibili vi sono infatti la *sentiment analysis* volta ad estrarre conoscenza dai testi, l'*analisi del churn* che mira ad individuare i clienti a rischio di abbandono, la *market basket analysis* che finalizza le strategie di marketing grazie all'analisi delle abitudini di acquisto dei clienti e la *segmentazione della clientela* che permette di offrire agli acquirenti offerte mirate sulla base del loro profilo. Tutte le applicazioni elencate hanno come fine ultimo quello di conoscere al meglio la clientela, così da poter instaurare con essa rapporti quanto più personalizzati, duraturi e di conseguenza proficui.
Sebbene le tecniche di data mining siano più comunemente impiegate per analizzare dataset di grandi dimensioni in contesti di business, le applicazioni sono però possibili anche in altri settori[17] e in presenza di insiemi di dati più ristretti. Inoltre, il successo di tali tecniche è in gran parte dovuto al fatto che non necessitano di assunzioni circa il fenomeno generatore dei dati. Ciò permette di svincolarsi dalla rigida struttura dei modelli parametrici, lasciando che siano i dati stessi a far emergere eventuali patterns (Azzalini e Scarpa 2012). Al contrario, i modelli parametrici, proprio a causa della loro rigida struttura, spesso non sono in grado di cogliere al meglio la complessità dei dati. Ad esempio il modello lineare è soggetto a forti assunzioni sulla distribuzione della risposta. Quando tali ipotesi non sono soddisfatte, le stime dei coefficienti del modello restano non distorte, ma non godono più delle usuali proprietà in termini di efficienza (Pace e Salvan 2001). Ciò incide sull'interpretazione dei risultati e di conseguenza sulla comprensione del

[17] Ad esempio in ambito medico possono essere utilizzate per la scelta dei protocolli di cura, in ambito astronomico per la classificazione e individuazione di numerosi corpi celesti, in ambito meteorologico per l'analisi dei dati trasmessi dai satelliti.

fenomeno.

L'obiettivo dell'elaborato è quindi quello di utilizzare soli Open Data, accompagnati da tecniche non parametriche di analisi dei dati, per cercare di comprendere meglio fenomeni attuali di interesse pubblico. La comprensione di tali fenomeni può infatti indurre governi o enti locali a prendere iniziative volte a migliorare la qualità della vita dei cittadini. Una tematica rilevante in tal senso è quella della sicurezza nelle grandi città ed in particolare si è scelto di concentrarsi sullo studio dei fattori che influenzano le aggressioni nella metropoli di New York.
La scelta della città è stata dettata dalla necessità di avere a disposizione dati adatti a svolgere analisi sufficientemente complesse ed approfondite. Nel panorama mondiale, gli Stati Uniti rappresentano il Paese che più ha trovato negli Open Data una risorsa da valorizzare. Infatti varie città statunitensi vantano numerosi portali di Open Data da cui è possibile reperire gratuitamente informazioni aggiornate, puntuali e integrabili tra loro.
Purtroppo, nonostante in Italia si stia sviluppando una crescente attenzione verso la tematica degli Open Data, si è ancora lontani dall'avere dati che permettano di svolgere analisi statistiche dettagliate. Un esempio è rappresentato dal dataset "Reati denunciati all'autorità giudiziaria dalla forze di polizia (2004 – 2018)" presente sul portale di Open Data del Comune di Milano[18]. Tale dataset contiene solo il numero complessivo di reati, divisi per macro tipologie, avvenuti sull'intero territorio comunale nei vari anni. Un livello di aggregazione così alto non permette di avere informazioni relative né ai reati stessi, posizione e orario, né inerenti a vittime o colpevoli, rendendo così impossibile uno studio approfondito del fenomeno.

[18] https://dati.comune.milano.it.

Capitolo 2

Aggressioni a New York: costruzione del dataset

Il secondo capitolo è organizzato in due parti. Nel paragrafo 2.1 si fa riferimento alla letteratura che si occupa di studiare fenomeni di criminalità, descrivendo quali fattori incidono maggiormente sulla propensione a commettere reati e nel paragrafo 2.2 viene illustrato il caso di studio, soffermandosi sulle caratteristiche di New York che l'hanno resa adatta per l'analisi del fenomeno. La seconda parte del capitolo è invece dedicata alla costruzione e presentazione del dataset. Nel paragrafo 2.3 vengono dapprima descritti i dati a disposizione, concentrandosi sulle procedure di pulizia e integrazione che hanno portato alla costruzione del dataset finale, in seguito vengono presentate alcune analisi descrittive volte a comprendere meglio il fenomeno in esame.

2.1 Premessa

La criminalità rappresenta un fenomeno di interesse pubblico dinamico e complesso. I fattori che inducono gli individui a commettere crimini sono infatti molteplici e di diversa natura e la letteratura a tale riguardo è amplissima. Sebbene non sia obiettivo dell'elaborato fornire una trattazione esaustiva di questa tematica, si è ugualmente ritenuto utile richiamare alcuni

studi presenti in letteratura, privilegiando quelli più attinenti al caso di studio in esame.
La prima teoria a legare motivazioni economiche ad atti criminali è stata formulata in Becker (1968), e si basa sul dipolo costo-beneficio. Secondo la teoria economica del comportamento criminale, un individuo sceglie di commettere un crimine se il beneficio atteso derivante da tale azione supera il beneficio atteso che otterrebbe da un'attività alternativa legale che comporti lo stesso dispiego di risorse. Citando Becker: «alcune persone diventano "criminali" non perché le loro motivazioni di fondo differiscono da quelle delle altre persone, ma perché i loro costi e benefici sono diversi». A questa teoria pionieristica si sono ispirati diversi ricercatori che hanno quindi studiato la relazione tra criminalità e fattori vari, tra cui disuguaglianze di reddito, disoccupazione, sesso ed etnie, immigrazione, livello di istruzione ed altri indicatori socioeconomici. In particolare numerosi studi affermano che esiste una forte associazione tra benessere economico e criminalità (Coccia 2018; Fleisher 1966; Kelly 2000; Lin 2008; Sachsida et al. 2010; Savage, Ellis e Wozniak 2019). Ad esempio Kelly (2000) ha mostrato che le disuguaglianze economiche hanno un significativo effetto positivo sull'incidenza dei crimini violenti. Fleisher (1966) ha invece studiato l'effetto del reddito e della disoccupazione sulla delinquenza, evidenziando come un incremento dell'1% nel reddito faccia diminuire del 2.5% il tasso di delinquenza delle aree più soggette a criminalità nella città di Chicago. Mentre da un'analisi condotta da Lin (2008) sui dati statunitensi forniti dall'Uniform Crime Reporting, è emerso che un incremento percentuale dell'1% nel tasso di disoccupazione si traduce in un aumento dell'1.8% dei crimini di proprietà.

2.2 Inquadramento del fenomeno di studio

In questo elaborato si è scelto di focalizzarsi in particolare sul fenomeno delle aggressioni, in quanto negli Stati Uniti costituiscono il primo reato per diffusione tra quelli violenti. I motivi per cui si è invece scelto di studiare il fenomeno focalizzandosi sulla città di New York sono principalmente due. Da una lato la grande disponibilità di dati presenti sul portale principale di Open

Data della città, da cui è possibile reperire più di 2900 dataset[1]; dall'altro la complessità della città stessa. La città di New York è infatti suddivisa amministrativamente, ma anche socio-economicamente, in cinque distretti[2] (figura 2.1) che presentano caratteristiche diverse. Ciò permette quindi di studiare il problema da più prospettive, il che potenzialmente potrebbe far emergere conclusioni differenti ed interessanti.

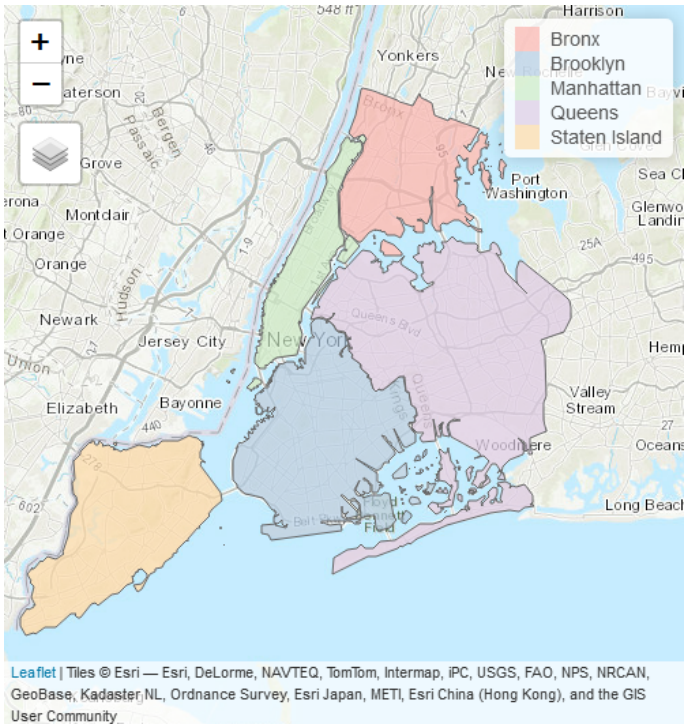


Figura 2.1: Cartina di New York con i cinque distretti evidenziati

Nello specifico si è scelto di considerare come unità statistica il numero di aggressioni avvenute nel periodo 2014-2019 all'interno delle circoscrizioni elettorali (census tracts) residenziali di New York. Si è scelto di considerare i census tracts in quanto rappresentano la più piccola entità territoriale per la

[1] In data 06/09/2020 sul portale `https://data.cityofnewyork.us` sono presenti 2918 dataset divisi in 11 macrocategorie: affari, amministrazione della città, educazione, ambiente, salute, alloggi, NYC BigApps, sicurezza pubblica, svago, servizi sociali, trasporti.

[2] Manhattan, Bronx, Queens, Brooklyn e Staten Island

quale è possibile reperire dati relativi alla popolazione. Tali informazioni sono raccolte ogni cinque anni grazie al programma American Community Survey (ACS)[3] che copre una buona varietà di argomenti di interesse generale sugli Stati Uniti: non solo dati demografici di base come età, sesso ed etnia, ma anche questioni economiche, sociali e abitative. Il sito Census Reporter, in linea con il principio di trasparenza nei confronti dei cittadini, mette a disposizione queste informazioni in formato "open". È infatti possibile consultare e scaricare qualunque tabella presente sul sito.

2.3 Costruzione del dataset

Per studiare il numero di aggressioni a New York tenendo in considerazione sia fattori relativi ai residenti, sia fattori relativi a servizi e specificità dei census tracts, è stato necessario aggregare informazioni contenute in 16 diversi dataset di formati differenti e provenienti da più portali. Ciò che accomuna tutti i dati utilizzati è il fatto di poter essere definiti "open" secondo i principi fondamentali di accessibilità, completezza, gratuità, aggiornamento periodico.

La tabella 2.1 riassume i dati utilizzati specificando il portale da cui sono stati scaricati.

Tabella 2.1: Open Data utilizzati

Portale	**Open Data utilizzati**
`https://data.cityofnewyork.us`	Dataset relativi al numero di aggressioni, alla posizione di fermate di metro e bus, alla codifica dei quartieri e dei census tracts.
`https://censusreporter.org`	Dataset contenenti informazioni demografiche, economiche e sociali.
`https://data.ny.gov`	Dataset sulla posizione degli ATM.

[3] L'American Community Survey (ACS) è un programma di sondaggi demografici condotto dallo U.S. Census Bureau e copre una buona varietà di argomenti di interesse generale sugli Stati Uniti. Le ultime informazioni disponibili a livello di census tracts sono ricavate con i dati relativi agli anni 2014-2018.

2.3.1 Pulizia e integrazione degli Open Data

Aggressioni

Il dataset principale contiene più di 7 milioni di crimini verificatisi a New York tra il 2006 e il 2019 e segnalati al New York City Police Department (NYPD). Ciascun reato è classificato secondo il diritto penale dello Stato di New York tramite due codici numerici che permettono di risalire al tipo di offesa e alla sua gravità. In questo modo è stato possibile estrarre le informazioni di interesse sulle sole aggressioni oggetto di questo studio (valori 106 e 344 della variabile `ky_cd`).

Prima di analizzare i dati si è proceduto con alcune operazioni di pulizia. In particolare si sono eliminate tutte le osservazioni che risultavano duplicate e quelle che non contenevano informazioni sul luogo del reato a causa di segnalazioni errate degli indirizzi[4]. Si è inoltre provveduto a ricodificare alcuni valori relativi all'età di vittime ed aggressori, in quanto risultavano casi in cui l'età riportata era negativa o superiore ai 900 anni. In questi casi le osservazioni sono state accorpate nella categoria "sconosciuto". Lo stesso è stato fatto per le variabili relative ad etnia e sesso quando i valori non erano disponibili.

Al fine di rendere possibile l'integrazione con le altre fonti dati si è scelto di considerare solo le aggressioni riportate nel periodo 2014-2019. Al termine della pulizia risultano così 431192 righe e 19 variabili di interesse. In tabella 2.2 si riporta l'elenco delle variabili con la relativa descrizione.

Variabili sociodemografiche ed economiche

Le motivazioni che spingono un uomo ad aggredirne un altro sono indubbiamente molteplici, complesse e spesso non comprensibili. Come visto nel paragrafo 2.1 vari studi mostrano però l'esistenza di relazioni tra fattori sociodemografici ed economici e l'attitudine a commettere reati violenti.

[4]NYPD Complaints Incident Level Data Footnotes, https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i.

Tabella 2.2: Descrizione variabili relative alle aggressioni

Nome variabile	Tipo variabile	Descrizione
cmplnt_num	Numerica	Identificativo del crimine
time	Data	Ora del reato
rpt_dt	Data	Data del reato
prem_typ_desc	Categoriale	Luogo del reato
ky_cd	Numerica	Codice del reato
pd_cd	Numerica	Codice del reato con dettaglio maggiore
ofns_desc	Testuale	Descrizione del reato
pd_desc	Testuale	Descrizione del reato con dettaglio maggiore
vic_sex	Dicotomica	Sesso della vittima
vic_age_group	Dicotomica	Categoria di età della vittima
vic_race2	Categoriale	Etnia della vittima
vic_race3	Dicotomica	Origini della vittima
susp_sex	Dicotomica	Sesso dell'aggressore
susp_age_group	Categoriale	Categoria di età dell'aggressore
susp_race2	Categoriale	Etnia dell'aggressore
susp_race3	Dicotomica	Origini dell'aggressore
boro	Categoriale	Quartiere in cui è avvenuta l'aggressione
latitude	Numerica	Latitudine del luogo dell'aggressione
longitude	Numerica	Longitudine del luogo dell'aggressione

Per questo motivo, nella scelta degli Open Data da integrare, si è scelto di considerare fattori demografici, sociali ed economici.

La tabella 2.3 riporta le fonti selezionate dal sito Census Reporter da cui sono state estratte le informazioni di interesse. I dati sono stati scaricati in formato shapefile[5] selezionando il più basso livello di aggregazione disponibile. Ogni riga di ciascuna tabella corrisponde quindi ad un census tract, identificato univocamente attraverso un codice alfanumerico di 18 caratteri. Purtroppo questo identificativo non è lo stesso di quello utilizzato dal Department of City Planning[6] per identificare i census tracts, quindi per integrare i dati prove-

[5]Il formato shapefile permette di descrivere spazialmente punti, linee e poligoni conservando la struttura spaziale dei dati.

[6]Dipartimento di New York responsabile della pianificazione fisica e socioeconomica

nienti dalle due fonti si è reso necessario trovare una chiave (maggiori dettagli in proposito sono contenuti nell'appendice Codice A.1).

Tabella 2.3: Tabelle presenti sul sito Census Reporter contenenti informazioni sulla popolazione

Identificativo tabella	Informazioni contenute
B01001	Informazioni sul sesso
B01002	Informazioni sull'età
B01003	Informazioni sul totale della popolazione
B02001	Informazioni sulle etnie
B03003	Informazioni sulle origini della popolazione
B17001	Informazioni sulla popolazione sotto la soglia di povertà
B19013	Informazioni sul reddito medio
B23025	Informazioni sulla disoccupazione
B25008	Informazioni sulla proprietà della casa in cui si vive
B08301	Informazioni su come è raggiunto il luogo di lavoro

A partire dalle informazioni a disposizione nelle varie tabelle, si è poi proceduto con la costruzione di alcuni indicatori relativi ai residenti dei census tracts. Si sono calcolate le percentuali di popolazione maschile, ispanica, bianca e afroamericana dividendo le rispettive numerosità per il totale della popolazione residente. Il tasso di disoccupazione è stato calcolato dividendo il numero dei disoccupati per il numero di persone in età lavorativa, mentre la variabile che tiene conto della percentuale di minorenni e pensionati è stata ottenuta dividendo il numero di persone in età non lavorativa per il numero di residenti. Come indicatori di benessere economico si sono inoltre calcolate la percentuale di residenti che vivono sotto la soglia di povertà e la percentuale di residenti che vivono in case non di proprietà. Infine si sono ottenuti alcuni indicatori relativi alla modalità di raggiungimento della sede lavorativa: per ogni census tract si sono calcolate le percentuali di residenti che si recano a lavoro a piedi o con i mezzi pubblici o che lavorano da casa.
Le variabili così costruite sono state riportate in tabella 2.4.

della città.

Tabella 2.4: Variabili ricavate a partire dalle informazioni contenute in tabella 2.3

Nome variabile	Tipo variabile	Descrizione
`maschi.perc`	Numerica	Percentuale di popolazione maschile
`perc.isp`	Numerica	Percentuale di ispanici
`white.perc`	Numerica	Percentuale di bianchi
`black.perc`	Numerica	Percentuale di afroamericani
`renter_occupied.perc`	Numerica	Percentuale di residenti in affitto
`tasso_disoccupazione`	Numerica	Tasso di disoccupazione
`not_in_labor_force.perc`	Numerica	Percentuale di popolazione non in età lavorativa (minorenni e pensionati)
`reddito_poverta.perc`	Numerica	Percentuale di popolazione che vive sotto la soglia di povertà
`mezzi_pubblici.perc`	Numerica	Percentuale di popolazione che si reca a lavoro con i mezzi pubblici
`piedi.perc`	Numerica	Percentuale di popolazione che si reca a lavoro a piedi
`lavoro_da_casa.perc`	Numerica	Percentuale di popolazione che lavora da casa

Infine, utilizzando l'identificativo dei census tracts, è stato possibile unire le varie tabelle, ottenendo un unico dataset di 2101 righe. Nella tabella 2.5 si sono riportate le variabili che insieme a quelle della tabella 2.4 costituiscono il dataset ottenuto combinando le informazioni provenienti dalle diverse fonti.

Tabella 2.5: Variabili aggiuntive

Nome variabile	Tipo variabile	Descrizione
`geoid`	Testuale	Identificativo della circoscrizione
`name`	Testuale	Nome della circoscrizione
`ntaname`	Testuale	Descrizione del sottoquartiere
`boro`	Testuale	Quartiere in cui si trova la circoscrizione
`popolazione`	Numerica	Numero di abitanti
`median_age`	Numerica	Età media della popolazione
`median_income`	Numerica	Reddito medio

In figura 2.2 si sono riportati i census tracts di New York, distinguendo le aree residenziali prese in considerazione da quelle non residenziali.

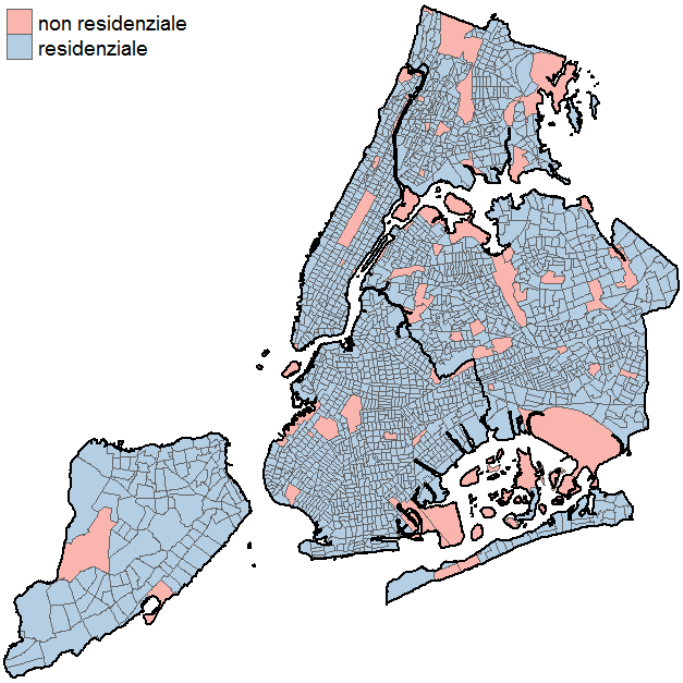


Figura 2.2: Cartina di New York con divisione tra le 2101 aree residenziali e le 64 non residenziali

Fermate dei mezzi pubblici e ATM

Per tenere in considerazione anche le dinamiche cittadine di una metropoli come New York, si è deciso di porre l'attenzione anche su aspetti legati ai servizi presenti sul territorio. In aggiunta alle variabili precedenti si sono quindi considerate le fermate dei mezzi di trasporto e la posizione degli sportelli bancari.

I dati relativi alle fermate di metro e bus sono forniti e mantenuti rispettivamente dalla Metropolitan Transportation Authority (MTA) e dal NYC Department of Transportation e possono essere scaricati in formato shapefile. Sfruttando la spazialità dei dati è stato possibile raggruppare il numero complessivo di fermate in ciascun census tract, ottenendo la variabile `tot_fermate` (si veda appendice Codice A.2).

I dati relativi alla posizione degli ATM sono invece stati reperiti sul sito di Open Data dello Stato di New York e sono forniti dal Department of Financial Services. Per poter selezionare solo gli ATM relativi alla città di New York, è stato necessario estrarre le coordinate geografiche dalla variabile contenente la posizione sotto forma di indirizzo. Una volta ricavate latitudine e longitudine è stato possibile convertire il dataset in formato spaziale e raggruppare gli ATM nei rispettivi census tracts di appartenenza, ottendendo la variabile `n_ATM` (si veda appendice Codice A.3).

Caratteristiche spaziali dei census tracts

Infine, sfruttando la struttura spaziale del dataset costruito, è stato possibile ricavare alcune informazioni geografiche relative ai vari distretti. In particolare sono state estratte le coordinate geografiche del centroide di ciascun census tract, espresse come latitudine e longitudine nel sistema di riferimento WGS 1984 (EPSG 4326), e si è calcolata la superficie in m^2 (si veda appendice Codice A.4).

Dataset finale

Una volta ottenute tutte le variabili esplicative ritenute di interesse, si è proceduto con la creazione del dataset finale. Per ciascuna circoscrizione elettorale si sono raggruppate le aggressioni verificatesi in ciascun anno, ottenendo così un dataset di 12606 righe e 25 colonne.

2.3.2 Analisi esplorative

Per poter comprendere quali fattori influiscono sul numero di aggressioni non si può prescindere da un'analisi descrittiva che tenga in considerazione vari aspetti relativi al fenomeno. Per questo motivo, nel condurre questo tipo di analisi, si è scelto di utilizzare sia il dataset sulle aggressioni prima del raggruppamento, sia quello aggregato a livello di census tracts. Ciò permette di cogliere aspetti che si perderebbero utilizzando i soli dati aggregati, come ad esempio alcune caratteristiche relative a vittime ed aggressori o su luogo e orario delle aggressioni.

In primo luogo si è scelto di presentare una panoramica generale del fenomeno. Dal grafico 2.3 si può osservare come le aggressioni rappresentino il secondo crimine più frequente (circa il 30%) a New York, tra quelli più gravi[7]. La tabella 2.6 mostra invece i primi dieci luoghi teatro del reato. Ad un pri-

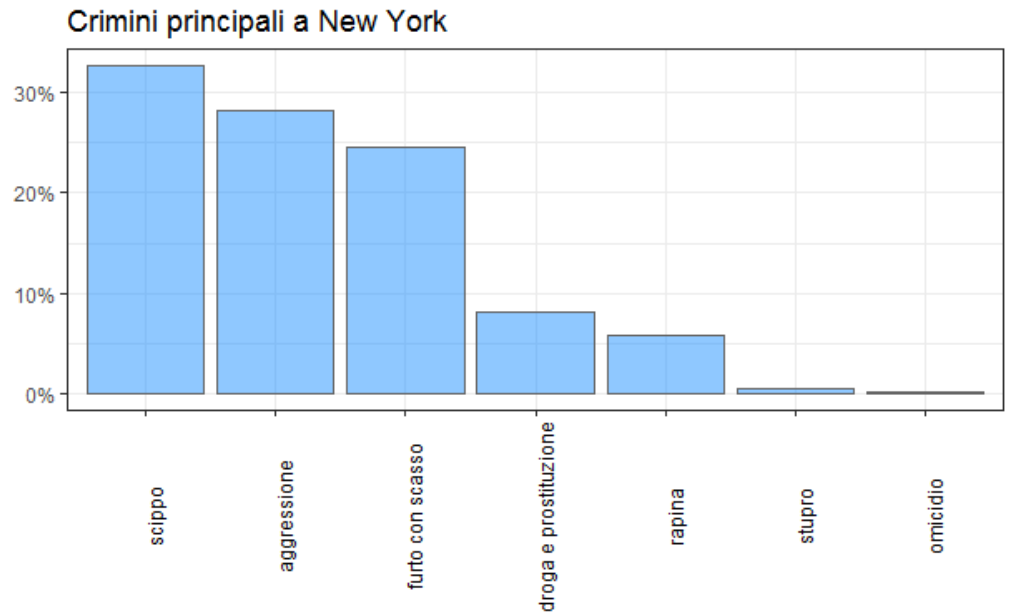


Figura 2.3: Crimini principali avvenuti a New York tra il 2014 e il 2019

mo sguardo si nota subito una grande disparità. È infatti interessante notare come le prime due voci coprano da sole quasi il 70% del totale. In particolare oltre la metà delle aggressioni avviene tra le mura domestiche e quasi il 30% in strada. Nonostante le altre voci coprano basse percentuali, è di interesse la varietà di luoghi in cui avvengono le aggressioni.
Oltre al luogo delle aggressioni si è ritenuto interessante indagare anche l'orario. Dalla figura 2.4 emergono infatti differenze notevoli tra le fasce orarie. In particolare si può osservare come la maggior parte delle aggressioni si verifichi tra il pomeriggio e la sera.
Infine per osservare l'evolversi del fenomeno nel corso degli anni, si è scelto

[7]La classificazione dei crimini statunitensi è contenuta nell'Uniform Crime Reporting Handbook redatto dall' U.S. Department of Justice.

Tabella 2.6: Primi 10 luoghi dove sono avvenute le aggressioni

Luogo	Numero di aggressioni	Percentuale
Casa	221613	51.40 %
Strada	116721	27.07 %
Bar/Night club	9507	2.21 %
Negozi alimentari	8556	1.98 %
Metropolitana	8545	1.98 %
Parchi	5615	1.30 %
Ristoranti	5147	1.19 %
Scuole pubbliche	4945	1.15 %
Ospedali	3018	0.70 %
Fast food	2853	0.66 %
Altro	44672	10.36 %
Totale	431192	100 %

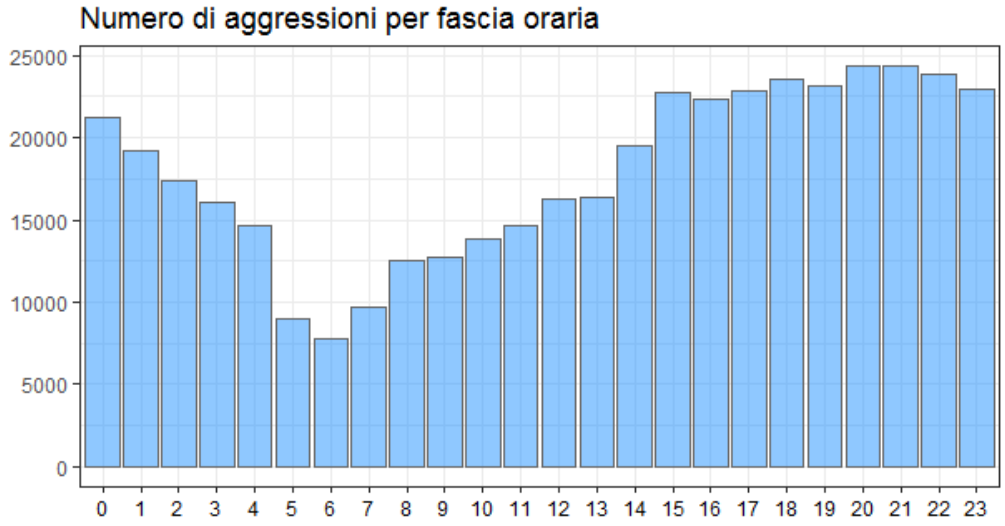


Figura 2.4: Grafico a barre relativo al numero di aggressioni diviso per fasce orarie

di calcolare i tassi di criminalità relativi alle aggressioni dal 2014 al 2019. I valori in tabella 2.7 sono stati ottenuti dividendo il numero di aggressioni di ciascun anno per il totale della popolazione, moltiplicando per 100000 abitanti e arrotondando all'intero più vicino. Si è inoltre calcolata la variazione percentuale tra il 2014 e il 2019. Da una prima lettura dei valori sembrerebbe che il fenomeno sia stazionario. Non si osservano infatti differenze significative tra un anno e l'altro.

Tabella 2.7: Tassi di criminalità per 100000 abitanti per gli anni 2014-2019

	popolazione	2014	2015	2016	2017	2018	2019	variazione %
New York	8429669	848	833	841	822	845	845	-0.28%

Dopo aver descritto il fenomeno da un punto di vista generale, ci si è soffermati sulle caratteristiche di vittime ed aggressori, al fine di coglierne eventuali differenze. La figura 2.5 mostra come la maggior parte di vittime ed aggressori abbia un'età compresa tra i 25 e i 44 anni. Bisogna però considerare l'incertezza dovuta ai valori mancanti, soprattutto per quanto riguarda gli aggressori. Ciò è dovuto anche al fatto che non sempre è immediato riconoscere le caratteristiche fisiche del proprio aggressore. Lo stesso può dirsi in riferimento alla variabile sesso. In questo caso si notano però evidenti differenze, al netto dei valori mancanti, tra vittime e aggressori. Infatti, se nel caso delle vittime la situazione pare bilanciata, lo stesso non si può dire nel caso degli aggressori, dove sembrerebbero essere i maschi i principali autori dei reati. Si sono poi indagate le caratteristiche in termini di origini ed etnie, riportando i risultati in figura 2.6. Al netto dei valori mancanti, si osserva che nel caso delle vittime afroamericani e bianchi condividono percentuali simili, mentre nel caso degli aggressori gli afroamericani mostrano una percentuale maggiore. A differenza di quanto osservato sull'etnia di vittime ed aggressori, non si riscontrano invece differenze significative per quanto riguarda le origini.

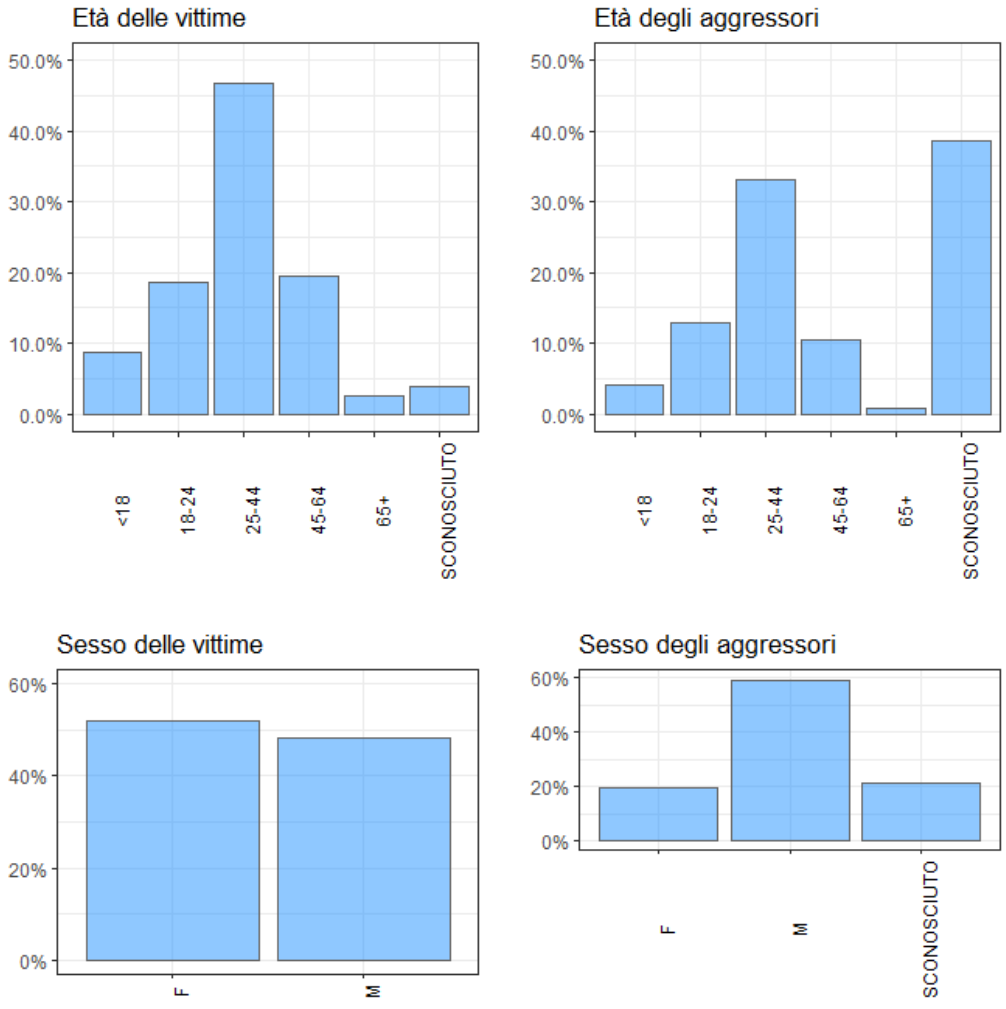


Figura 2.5: Grafici a barre relativi a sesso ed età di vittime ed aggressori

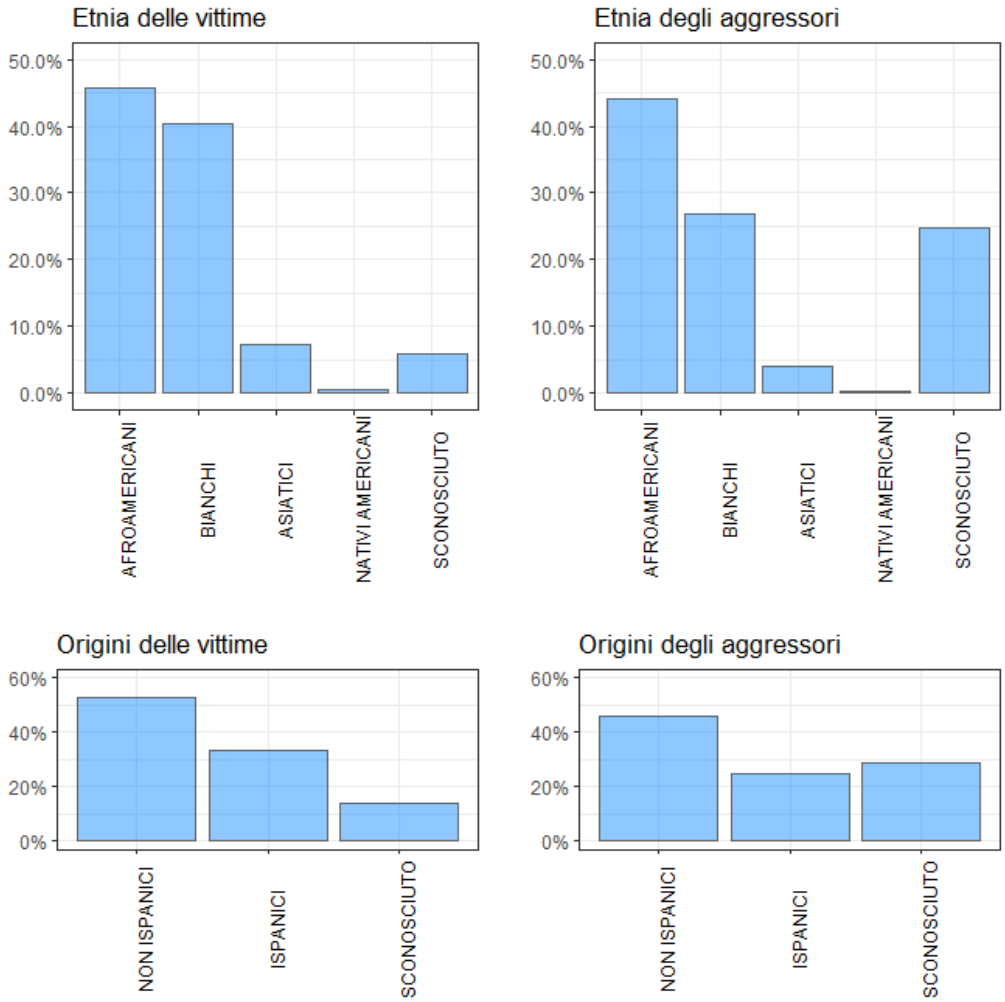
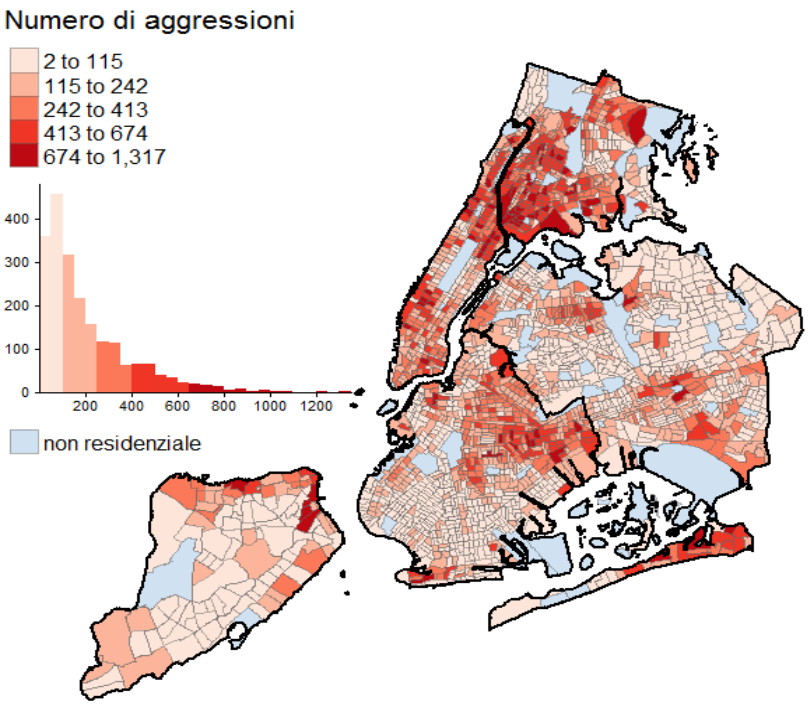


Figura 2.6: Grafici a barre relativi ad etnie e origini di vittime ed aggressori

Concluse le analisi descrittive relative al fenomeno in generale, si è utilizzato il dataset aggregato per esplorare possibili relazioni tra variabili esplicative e numero di aggressioni per census tract. Si è inoltre cercato di inquadrare il fenomeno nei cinque quartieri della città per fare emergere le specificità di questi ultimi.

In primo luogo si è studiata la distribuzione spaziale del numero di aggressioni per il periodo 2014-2019. In figura 2.7 si sono riportate una mappa delle aggressioni, con relativo istogramma, sull'intero territorio di New York e un grafico a violino che mostra la distribuzione del numero di aggressioni nei cinque quartieri. Dalle due rappresentazioni è immediato notare come il Bronx e Manhattan siano i quartieri con il più alto numero di aggressioni, seguiti da Brooklyn, Queens e Staten Island. L'istogramma mostra inoltre una distribuzione fortemente asimmetrica. Possiamo infatti immaginare che ci siano poche zone altamente pericolose e molte zone di pericolosità basso-media. Dalla cartina si può inoltre osservare che le zone fortemente pericolose tendono ad essere raggruppate, facendo intendere una forte correlazione spaziale tra le osservazioni.
Successivamente si sono calcolati i tassi di criminalità relativi alle aggressioni dal 2014 al 2019 e la variazione percentuale tra il primo e l'ultimo anno, riportando i valori in tabella 2.8. È immediato osservare come il Bronx presenti i tassi più elevati, discostandosi di molto anche dalla media della città (riportata nella tabella 2.7), mentre Queens e Staten Island si collocano al di sotto della media. La divisione delle aggressioni tra i vari quartieri permette anche di notare differenze significative tra gli anni che non erano emerse dall'analisi della città nel complesso. Infatti, se per New York emergeva un quadro pressoché stazionario, lo stesso non si può dire per i quartieri di Manhattan e Brooklyn. Nello specifico si osserva un andamento dei tassi monotono crescente e decrescente rispettivamente, con una variazione percentuale tra il 2014 e il 2019 in valore assoluto di oltre il 14 %. Per quanto riguarda gli altri quartieri l'andamento risulta meno definito: sembrerebbe globalmente decrescente per Staten Island, crescente per il Bronx e stazionario per il Queens.



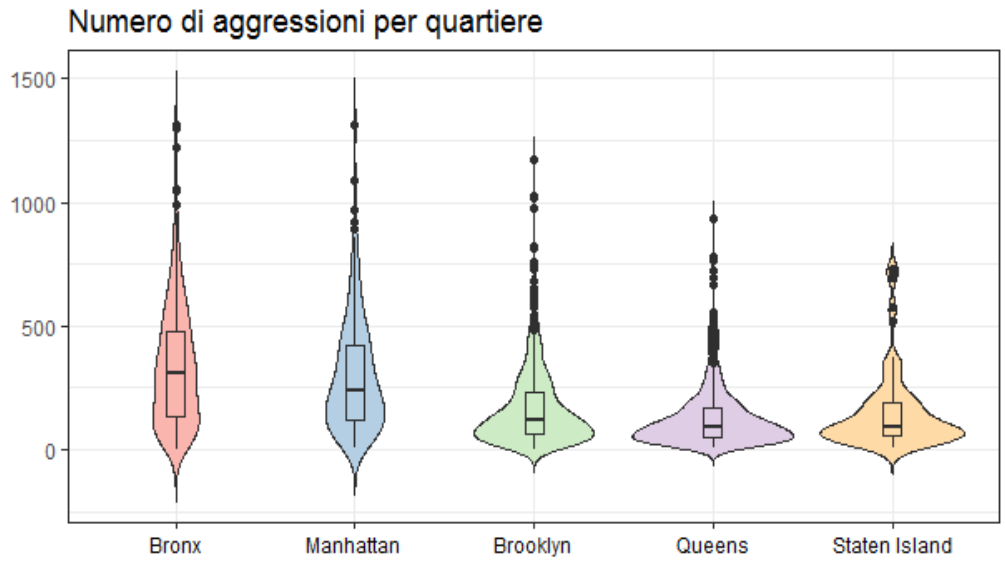


Figura 2.7: Numero di aggressioni a New York e divisione nei 5 quartieri

Tabella 2.8: Tassi di criminalità per 100000 abitanti per gli anni 2014-2019 nei cinque quartieri

	popolazione	2014	2015	2016	2017	2018	2019	variazione %
Bronx	1428458	1264	1258	1297	1277	1354	1341	5.80%
Manhattan	1630533	783	812	840	840	873	899	12.90%
Brooklyn	2598764	885	846	839	792	798	772	-14.63%
Queens	2297825	642	619	613	613	615	647	0.73%
Staten Island	474089	612	595	587	570	585	532	-14.94%

Infine, per completare le analisi, si sono rappresentate le distribuzioni delle variabili esplicative sotto forma di mappe, corredate dai rispettivi istogrammi. Si è scelto di rappresentare questi dati tramite mappe per poter offrire una visualizzazione immediata. Osservando le figure da 2.8 a 2.15 è infatti possibile individuare differenze significative sia tra i cinque quartieri, sia al loro interno a livello di census tracts.
Le prime due mappe (figura 2.8) forniscono una prima panoramica dei residenti di New York. In primo luogo si può osservare una discreta discrepanza nel numero di abitanti nei vari census tracts. Infatti, sebbene il Census Bureau definisca i census tracts come divisioni territoriali di circa 4000 abitanti, dalla mappa risulta che la popolazione residente nelle diverse circoscrizioni varia da 36 a 28272 abitanti. Un'analisi più approfondita rivela che il census tract con più abitanti fa parte del complesso residenziale "Co-op City", nel Bronx. Questa zona viene definita una "città nelle città" in quanto sono presenti 15.372 unità residenziali, suddivise in 35 grattacieli e 7 gruppi di case a schiera, 8 grandi parcheggi pubblici, 3 centri commerciali, un parco didattico di 25 acri, che comprende una scuola superiore, due scuole medie e tre scuole elementari, ed anche una centrale elettrica. All'estremo opposto si colloca il census tract del Queens con soli 36 abitanti, ma ciò è dovuto al fatto che la maggior parte dell'area è occupata dal cimitero di Mount Olivet.
Per quanto riguarda l'età media dei residenti la distribuzione pare invece simmetrica e centrata su un valore medio di 35 anni. La mappa evidenzia come le zone con la popolazione mediamente più anziana siano concentrate a Staten Island e nella parte ad ovest del Queens. Viceversa la situazione appare bilanciata per quanto riguarda il sesso dei residenti (prima mappa della figura

2.9).
Le differenze di maggior impatto visivo si riscontrano invece a livello di etnie ed origini dei residenti. Le relative mappe (figure 2.9 e 2.10) mostrano infatti evidenti cluster relativi a queste variabili. In particolare si nota come gli afroamericani si concentrino nella zona a nord del Bronx e nelle zone di sud-est di Brooklyn e Queens. La popolazione bianca è invece fortemente concentrata a Staten Island, a sud di Manhattan e a ovest di Brooklyn e Queens. La popolazione ispanica tende a concentrarsi in particolare nel Bronx e nella zona di Manhattan di confine, mentre negli altri quartieri sono presenti piccoli cluster isolati.
È anche interessante soffermarsi sulle mappe che indicano il benessere economico dei residenti (figura 2.11 e prima mappa della figura 2.12). Dalla distribuzione del reddito si nota come alcune aree della città possano essere definite molto più ricche di altre. In particolare, come prevedibile, spicca la zona a sud di Manhattan. Mediamente anche Queens e Staten Island mostrano alti valori del reddito, mentre a Brooklyn e nel Bronx sono pochi i census tracts con reddito elevato. Se si osservano invece le mappe relative alla percentuale di residenti che vivono in povertà e al tasso di disoccupazione si ha una visione del tutto speculare e coerente. Le percentuali maggiori si osservano infatti a Brooklyn e nel Bronx.
Per quanto riguarda invece la percentuale di residenti in pensione o in età non lavorativa, dalla seconda mappa della figura 2.12 si nota come i valori maggiori si concentrino a Staten Island. Nel resto di New York la distribuzione pare abbastanza omogenea.
Ulteriori considerazioni sulle dinamiche cittadine emergono dalla mappa relativa alla proprietà della casa in cui si vive. La prima mappa della figura 2.13 mostra una distribuzione fortemente diversificata tra i cinque quartieri. In particolare si nota come la quasi totalità della popolazione a Staten Island viva in case di proprietà. Viceversa la situazione pare ribaltata nei quartieri di Manhattan, Bronx e Brooklyn dove si predilige l'affitto. Per quanto riguarda il Queens invece la situazione sembra la più bilanciata: si alternano in modo pressoché uniforme census tracts dove viene prediletto l'affitto a census tracts in cui prevale la proprietà della casa.
Uno spunto interessante è anche dato dal confronto tra la mappa relativa ai

lavoratori che utilizzano mezzi pubblici e quella relativa all'ubicazione delle fermate di metropolitana e autobus (figura 2.14 e prima mappa della figura 2.15) da cui emerge un'evidente correlazione. I quartieri con una rete più capillare di trasporti sono anche quelli che presentano le più alte percentuali di utilizzo dei mezzi da parte dei lavoratori. Coerentemente con quanto detto si può osservare come Staten Island presenti le percentuali più basse di utilizzo dei mezzi; infatti questo quartiere possiede una sola linea metropolitana. Lo stesso discorso vale per la zona est del Queens in cui si può osservare una rada distribuzione dei trasporti pubblici e un conseguente scarso utilizzo. Viceversa Manhattan presenta la più densa rete di trasporti e le più alte percentuali di utilizzo dei mezzi pubblici per recarsi a lavoro.
Ulteriori osservazioni possono essere fatte anche a partire dalle mappe relative alla percentuale di residenti che lavora da casa e a quella che si reca in ufficio a piedi (seconde mappe delle figure 2.13 e 2.14). In entrambi i casi le distribuzioni sono fortemente asimmetriche e si assestano complessivamente su bassi valori. Per quanto riguarda il lavoro da casa si può notare che le zone con le percentuali più alte si trovano a Manhattan e a nord di Brooklyn. Mentre in relazione all'andare a lavoro a piedi si osserva una diffusione consistente a sud di Manhattan ed un cluster a Brooklyn.
Infine l'ultima mappa mostra la collocazione degli ATM (seconda mappa della figura 2.15). Come prevedibile le zone meglio servite si trovano a Manhattan.

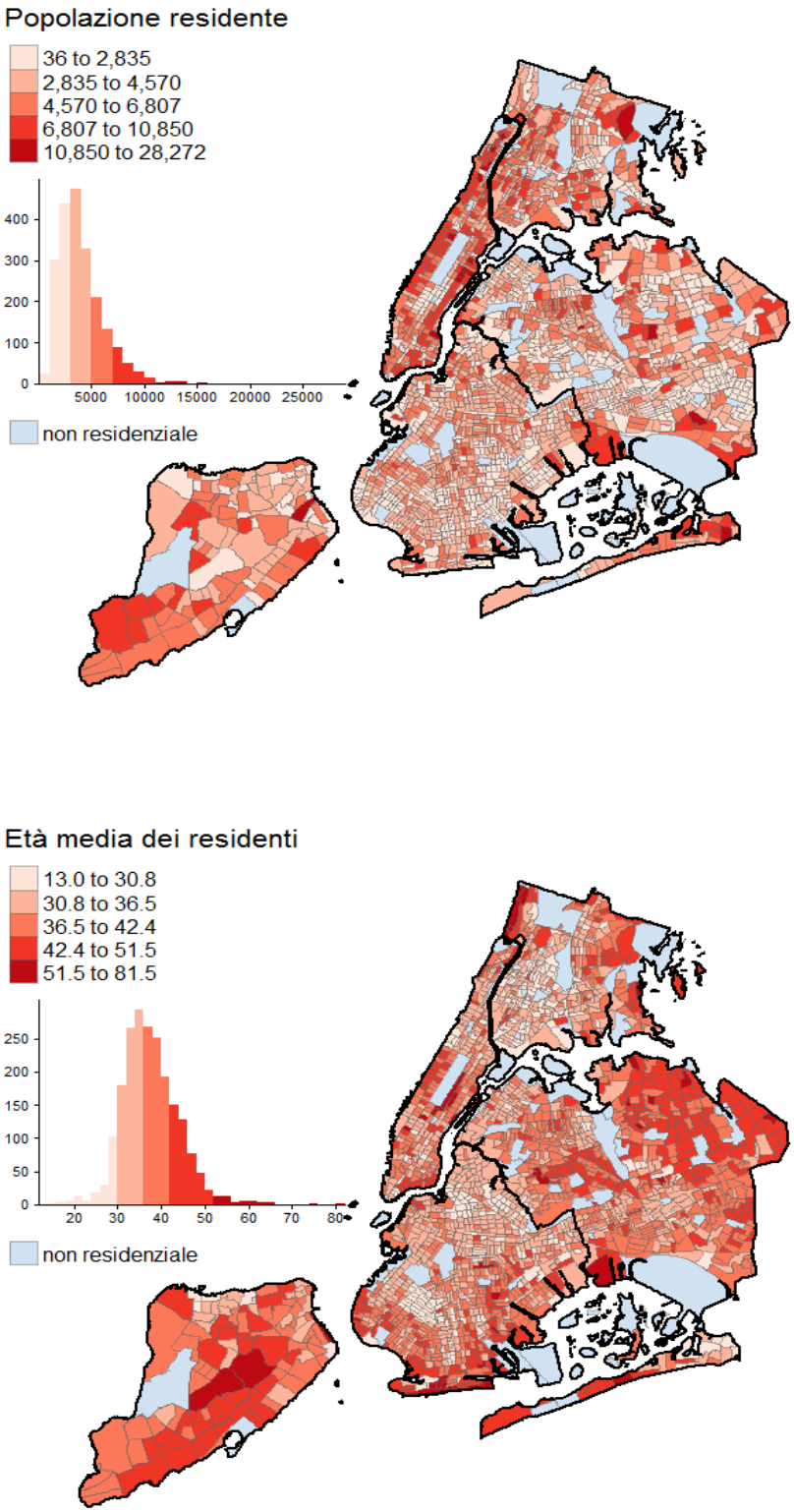


Figura 2.8: Mappe relative al numero di residenti e all'età media

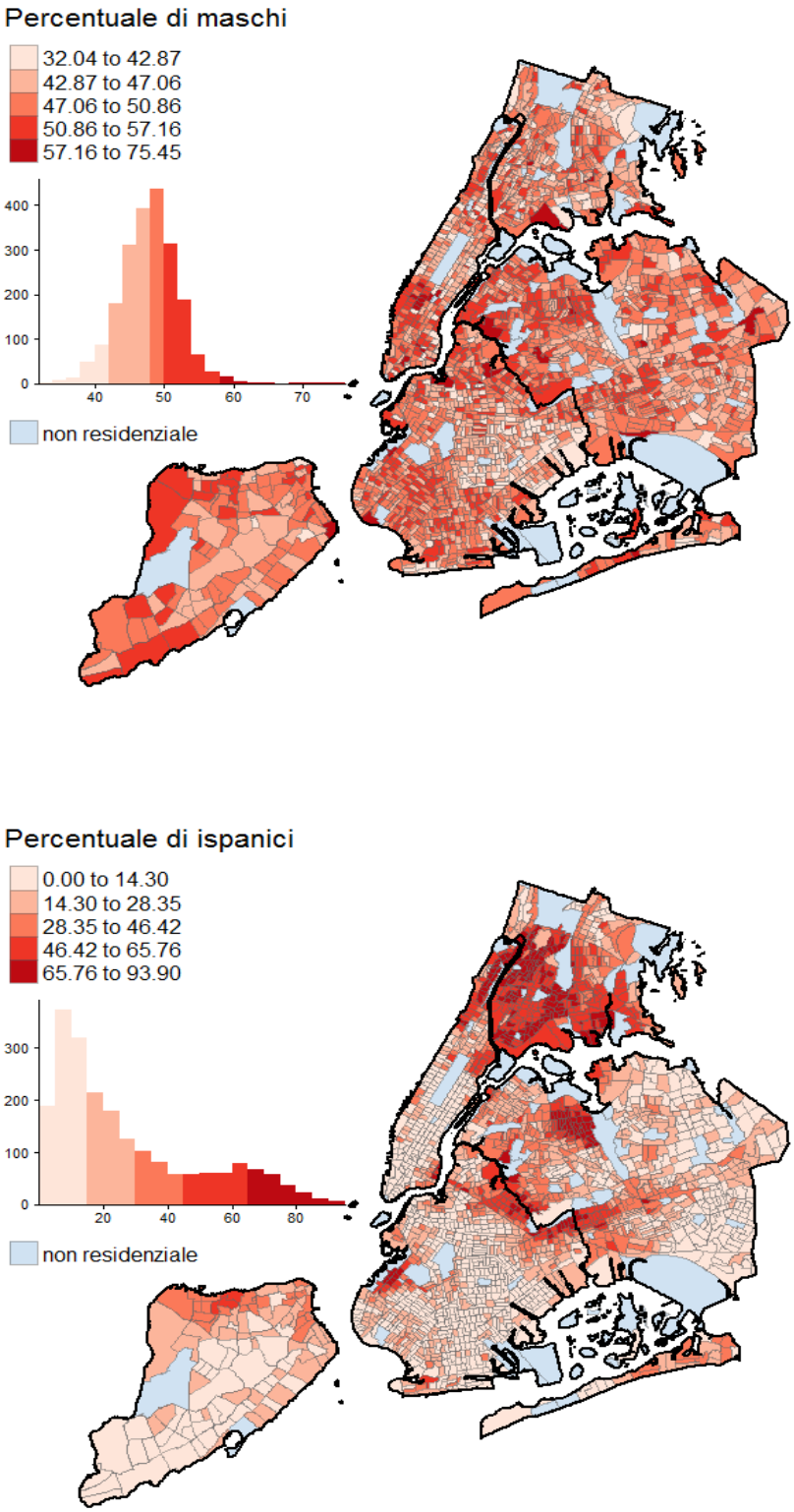


Figura 2.9: Mappe relative al sesso e alle origini dei residenti

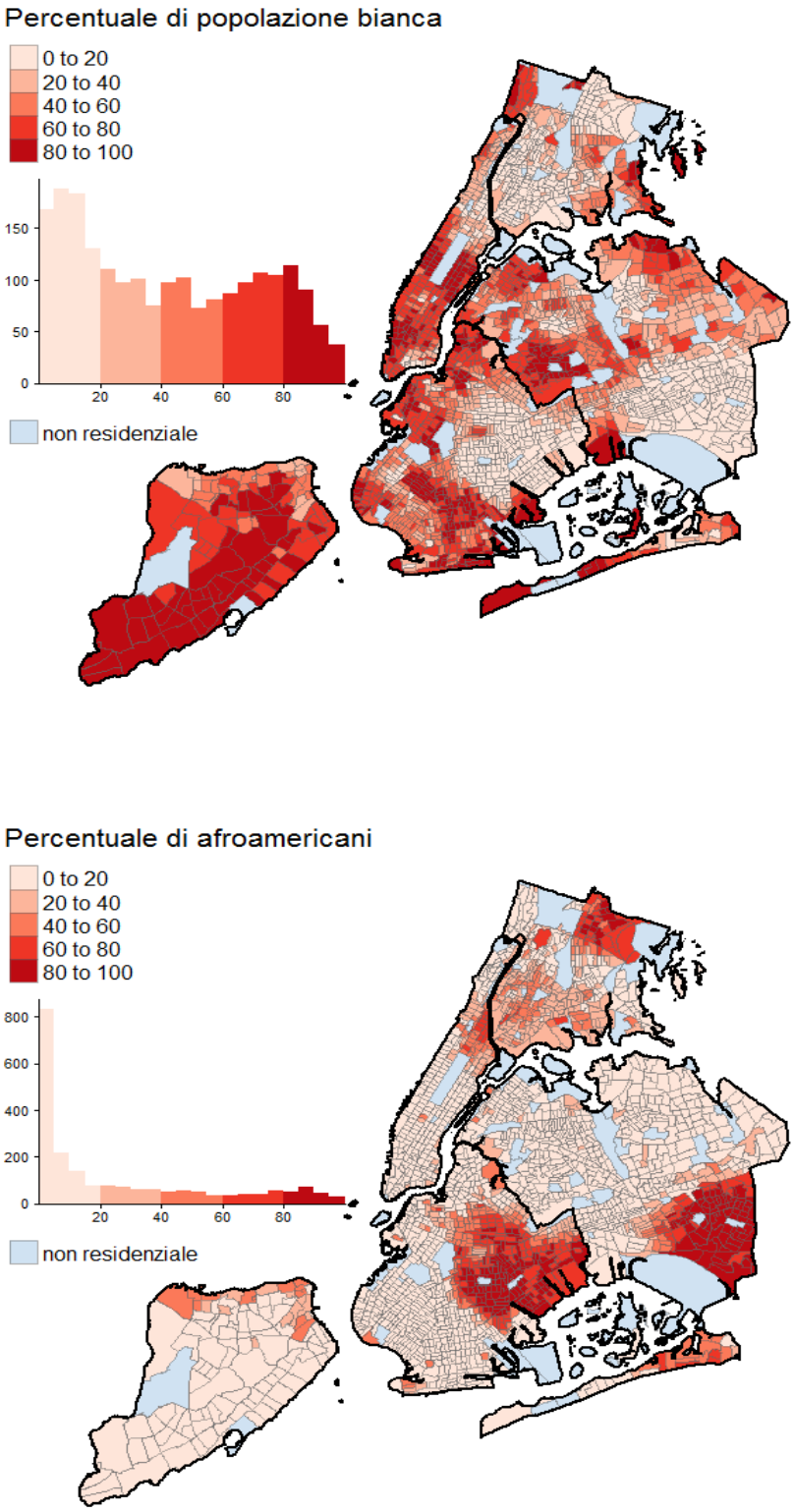


Figura 2.10: Mappe relative all'etnia dei residenti

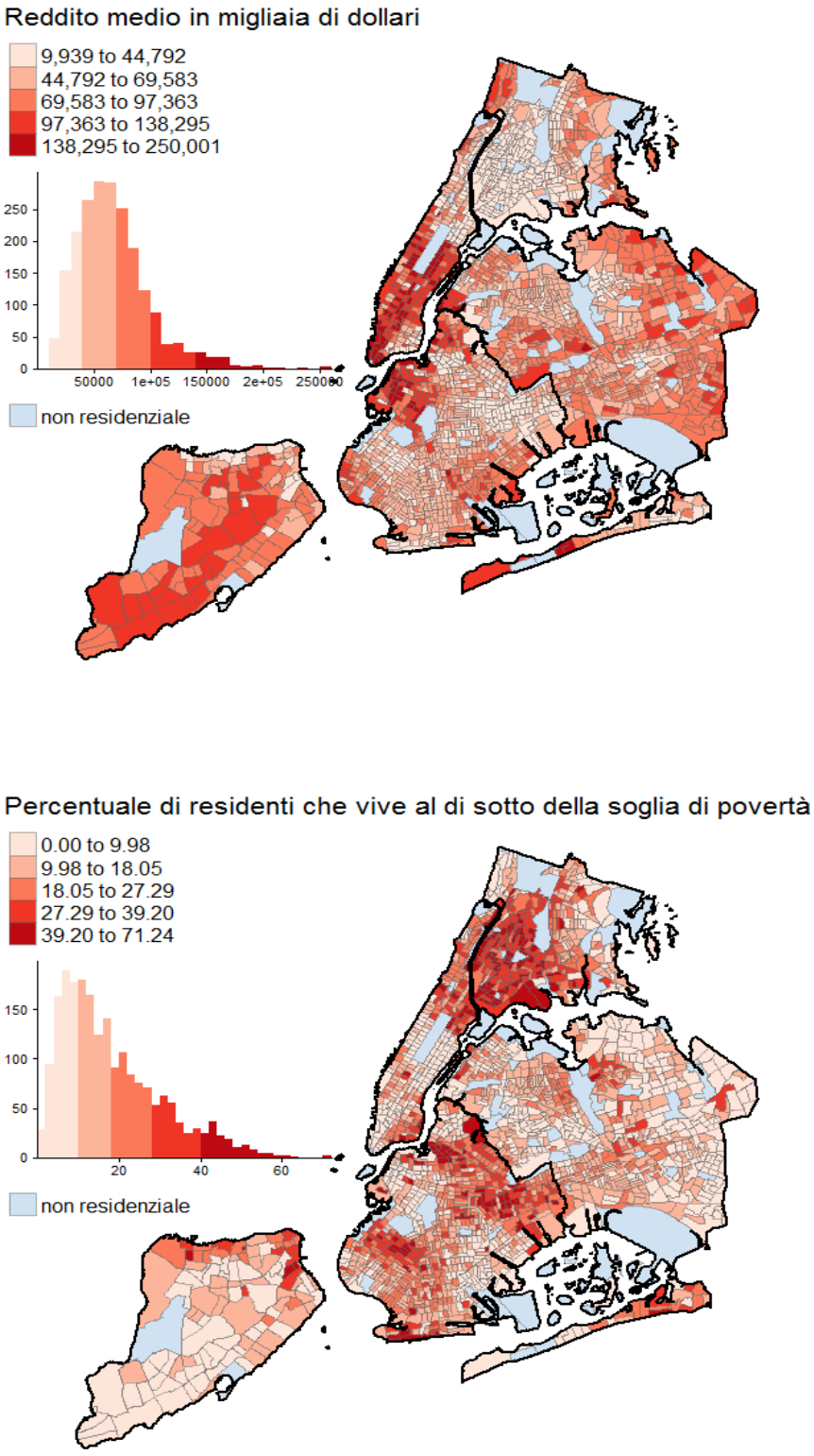


Figura 2.11: Mappe relative al reddito dei residenti

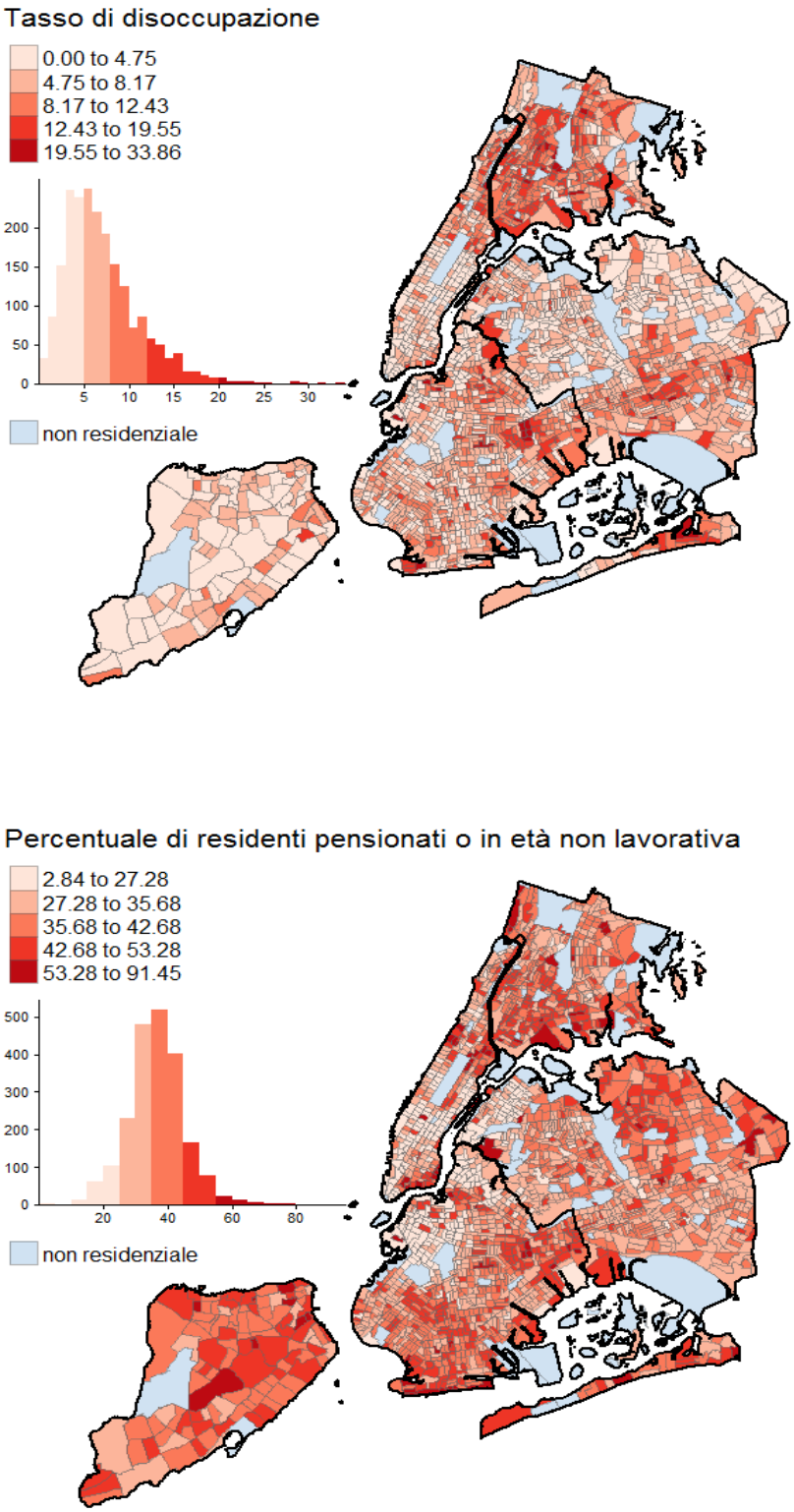


Figura 2.12: Mappe relative allo status lavorativo dei residenti

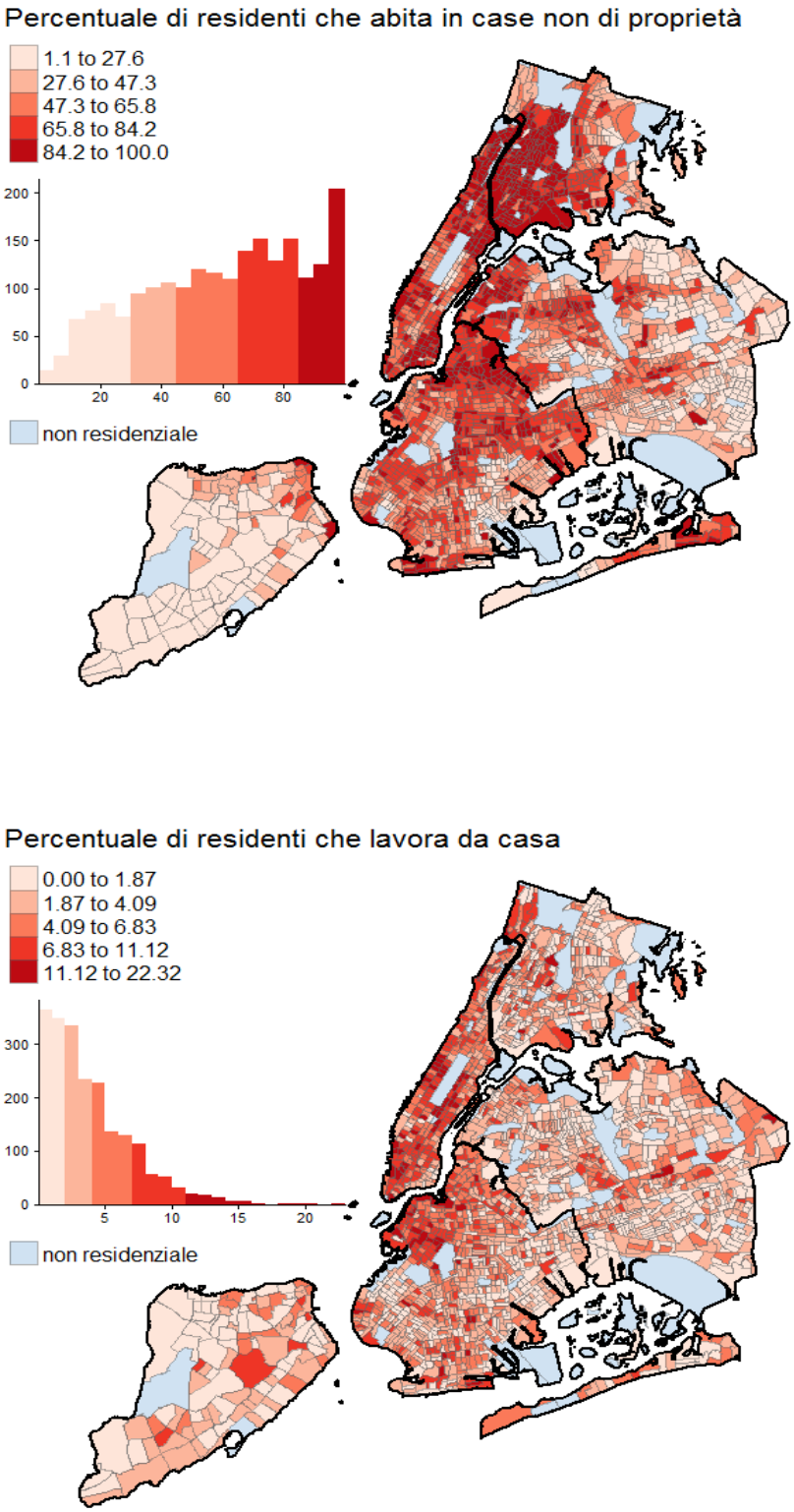


Figura 2.13: Mappe relative alla proprietà delle case dei residenti e alla percentuale di residenti che lavora da casa

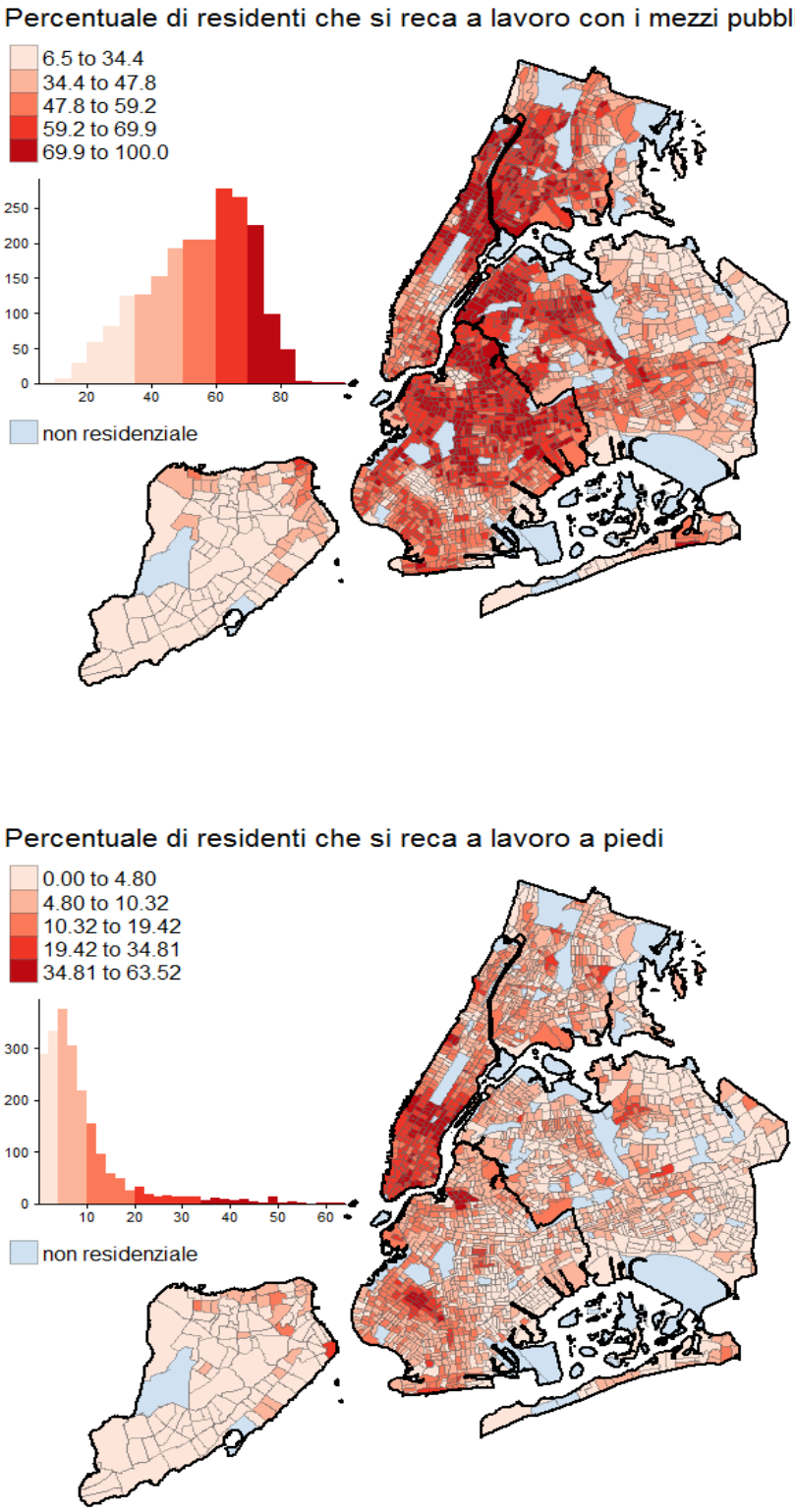
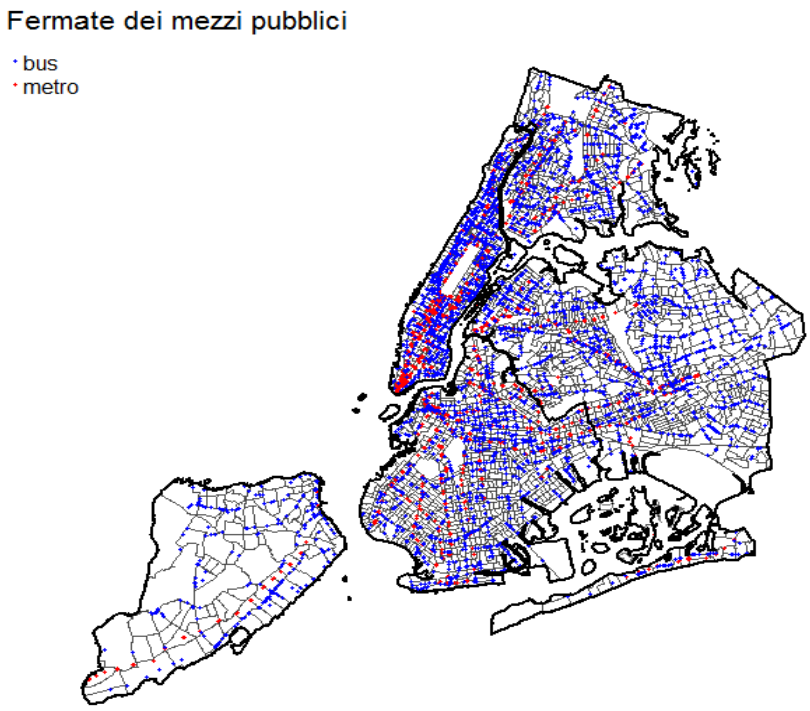


Figura 2.14: Mappe relative al modo in cui i residenti si recano a lavoro



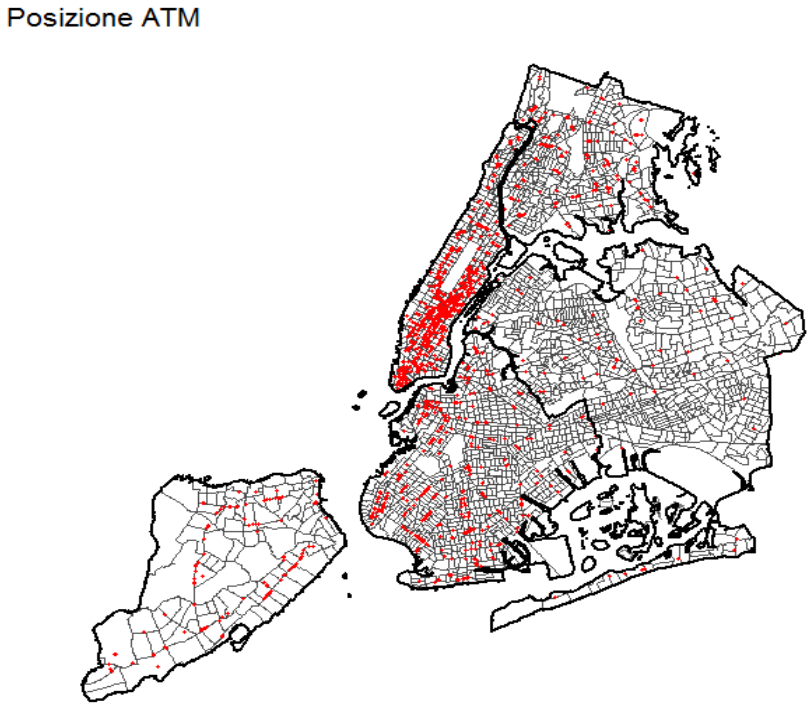


Figura 2.15: Mappe relative alla posizione delle fermate dei mezzi pubblici e degli ATM

Capitolo 3

Aggressioni a New York: modellazione

Il terzo capitolo di questo elaborato è dedicato a modellare il fenomeno delle aggressioni nella città di New York. Nei paragrafi 3.1 e 3.2 viene illustrato l'approccio utilizzato per studiare il fenomeno in esame e si descrivono le caratteristiche principali dei modelli selezionati. Il paragrafo 3.3 è invece dedicato a presentare l'analisi effettuata, soffermandosi su risultati ed interpretazione per ciascun distretto. Infine il paragrafo 3.4 riporta alcune considerazioni conclusive.

3.1 Premessa

Dalle analisi descrittive nella sezione 2.3.2 è emersa una grande eterogeneità all'interno di New York, sia per quanto riguarda il fenomeno delle aggressioni, sia per quanto riguarda le variabili che caratterizzano la popolazione e i census tracts. Per tenere conto di questa eterogeneità, si è quindi deciso di considerare ogni quartiere di New York separatamente e conseguentemente di stimare i modelli per ciascun distretto. Ciò permette non solo di interpretare meglio i risultati tenendo conto delle differenti realtà, ma anche di mitigare la correlazione spaziale presente tra le osservazioni. Al fine di catturare la spazialità del fenomeno all'interno dei singoli distretti, si è anche

scelto di inserire tra le esplicative l'interazione tra latitudine e longitudine relative al centroide di ciascun census tract. Infine come variabile risposta si è deciso di considerare una trasformazione monotona del numero di aggressioni. Infatti dalle analisi esplorative è emerso come la distribuzione del numero di aggressioni fosse fortemente asimmetrica. Si è quindi scelto di modellare la radice quadrata del numero di aggressioni, sia per rendere simmetrica e bilanciata la distribuzione della risposta, sia per garantire la positività delle stime.

La tabella 3.1 riporta il numero di osservazioni per ciascun quartiere.

Tabella 3.1: Numero di osservazioni per ciascun quartiere

	Osservazioni
Bronx	1956
Manhattan	1668
Brooklyn	4488
Queens	3852
Staten Island	642

Molti degli studi presenti in letteratura, volti ad analizzare i fattori che incidono sui fenomeni di criminalità, si avvalgono del modello lineare e delle sue naturali estensioni soprattutto in ambito econonometrico (Kelly 2000; Sachsida et al. 2010). In questo elaborato si è invece scelto di utilizzare tecniche non parametriche di analisi dei dati. In linea con quanto espresso nella sezione 1.5, tali tecniche consentono infatti di cogliere la complessità dei dati senza vincolarsi ad una scelta a priori circa il legame tra la risposta e i dati stessi e possono quindi far emergere relazioni inaspettate. Le tecniche non parametriche di analisi dei dati sono inoltre poco utilizzate nello studiare quali fattori incidano maggiormente sulla propensione a commettere reati, ma trovano largo impiego nell'analisi di altre problematiche legate alla criminalità. Ad esempio le tecniche di *natural language processing* possono essere utilizzate per estrarre informazioni di valore (come informazioni relative a persone, veicoli, indirizzi) da dati non strutturati come i verbali di polizia. Le tecniche di *clustering* possono invece essere sfruttate per aggregare sospetti o gruppi

criminali che compiono i reati in modi simili, permettendo di identificare più facilmente criminali seriali o membri appartenenti allo stesso gruppo. Anche la *social network analysis* può essere sfruttata per favorire le indagini. Gli investigatori possono infatti utilizzare tale tecnica per costruire una rete che illustra i ruoli dei criminali, la centralità di alcuni soggetti e le connessioni esistenti tra individui appartenenti alla rete. (Chauhan e Sehgal (2017), Chen et al. (2004), Hassani et al. (2016) e Nath (2006)).
Ai fini di questo elaborato si è scelto di utilizzare alcune tecniche non parametriche di analisi dei dati per studiare invece quali fattori incidano sul numero di aggressioni. Nello scegliere quali modelli fossero più adatti per il caso di studio si è cercato di tenere in considerazione due aspetti: capacità predittiva e interpretabilità. Dal momento che l'obiettivo dell'elaborato è quello di ricavare valore informativo dagli Open Data, si è scelto di prediligere l'aspetto interpretativo. Per questo motivo i primi modelli considerati saranno i modelli additivi, che tramite grafici marginali permettono di valutare la relazione tra ciascuna esplicativa e la risposta. Si è poi scelto di stimare anche foreste casuali e Gradient Boosting[1], in quanto questi modelli permettono di valutare l'importanza delle variabili ed inoltre spesso raggiungono ottimi risultati in termini di capacità predittiva.

Infine, per confrontare i modelli in termini di errore di previsione, si è scelto di calcolare l'errore quadratico medio. Al fine di ottenere errori consistenti e non troppo ottimistici si è scelto di utilizzare il metodo della convalida incrociata. Ciò permette di non dividere i dataset in due porzioni, una dedicata alla stima e una alla verifica, ma di sfruttare tutta l'informazione disponibile per adattare i modelli.

[1]Si è scelto di utilizzare questo metodo invece del più recente ed efficiente Extreme Gradient Boosting (XGBoost) a causa dell'elevato numero di parametri che sarebbe necessario ottimizzare. Dovendo stimare i modelli per ciascuno dei cinque quartieri il costo computazionale richiesto dall'XGBoost sarebbe risultato elevato.

3.2 Modelli selezionati

In questa sezione vengono descritte le caratteristiche principali dei modelli scelti per l'analisi del caso di studio. Trattandosi di un problema di regressione ci si focalizzerà quindi sugli aspetti legati a tale ambito.

Modelli additivi con splines di lisciamento

Il modello additivo costituisce un'estensione del modello lineare in quanto permette ai regressori di assumere svariate forme funzionali, mantenendo però la struttura additiva. Ciò permette al modello di essere sufficientemente flessibile ed al tempo stesso facilmente interpretabile.
Supponendo di avere a disposizione $x_1, x_2, \ldots, x_p$ variabili esplicative ed una risposta y, il modello assume la forma

$$\mathbb{E}\left(y \,|\, x_1,\, x_2, \ldots, x_p\right) = \alpha + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p) \tag{3.1}$$

dove $f_1, f_2, \ldots, f_p$ sono funzioni univariate generalmente non parametriche e α rappresenta la costante del modello. La scelta dei lisciatori f_j non è cruciale e, nonostante solitamente si scelga lo stesso metodo per ciascuna esplicativa, non ci sono vincoli a riguardo. Affinché il modello sia identificabile è però necessario porre un vincolo sulla struttura delle f_j.
In genere si richiede che le funzioni siano centrate intorno allo zero:

$$\sum_{i=1}^{n} f_j(x_{ij}) = 0, \qquad (j = 1, \ldots, p) \tag{3.2}$$

dove n è il numero di osservazioni e x_{ij} è la i-esima osservazione della j-esima variabile (Azzalini e Scarpa 2012).

Il modello additivo rientra nella categoria dei metodi non parametrici ed in principio risente quindi della *maledizione della dimensionalità*. Per limitare tale problema è quindi necessario definire una struttura per i lisciatori. In questo elaborato si è scelto di seguire l'approccio di Wood (2017) secondo cui ciascun regressore può essere rappresentato tramite splines di lisciamento

(*thin plate splines*). Il modello (3.1) può inoltre essere esteso al caso multivariato aggiungendo termini del tipo $f(x_i, x_j)$, modellati tramite splines multivariate. Ciò permette di considerare l'effetto di interazione tra le variabili di interesse.

L'utilizzo di splines di lisciamento consente inoltre di considerare il problema di stima da un altro punto di vista. Un problema comune in statistica è infatti quello di individuare regressori che siano sufficientemente lisci. Nel caso univariato tale necessità si traduce nell'individuare una funzione che minimizzi la somma dei quadrati dei residui e sia al tempo stesso continua fino alla derivata seconda e non eccessivamente sinuosa.
Considerando un solo regressore la funzione di perdita da minimizzare è data da

$$D(f, \lambda) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int f''(t)^2 \, dt \tag{3.3}$$

dove λ assume valori positivi e rappresenta il parametro di regolazione che determina l'ammontare di penalizzazione.
Si dimostra (Wood 2017) che la soluzione del problema di minimo vincolato è data appunto da un caso speciale delle *thin plate splines*: le splines cubiche naturali.
Nel caso bivariato invece la funzione di perdita da minimizzare è più complessa e la derivata seconda viene sostituita dal Laplaciano.
Il termine di penalizzazione è quindi dato da

$$\lambda \iint_{\mathbb{R}^2} \Big\{ \Big(\frac{\partial^2 f(x)}{\partial x_1^2} \Big)^2 + 2 \Big(\frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \Big)^2 + \Big(\frac{\partial^2 f(x)}{\partial x_2^2} \Big)^2 \Big\} \, dx_1 \, dx_2 \tag{3.4}$$

che anche in questo caso ha come soluzione una *thin plate spline*.

Come precedentemente accennato, il parametro λ determina l'ammontare di penalizzazione. Quando $\lambda \to 0$ l'effetto di penalizzazione va ad annullarsi e di conseguenza la curva ottenuta tenderà ad interpolare i dati, mentre quando $\lambda \to \infty$ la penalizzazione è massima e la curva risultante sarà quindi la retta stimata con i minimi quadrati. Da ciò si intuisce l'importanza di una corretta scelta del parametro di regolazione λ. Un metodo largamente utilizzato è quello della convalida incrociata generalizzata. Tuttavia si preferisce indivi-

duare il valore ottimo di un altro parametro in stretta relazione con λ: i gradi di libertà equivalenti (*edf*). Tra i due parametri sussiste infatti una relazione monotona decrescente tale per cui la conoscenza di uno dei due implica in modo univoco la conoscenza dell'altro.
Si sceglie quindi il parametro *edf* che minimizza l'indice

$$GCV(edf) = \frac{\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}{(1 - edf/n)^2} \tag{3.5}$$

Quanto esposto in precedenza si estende in modo naturale al caso con più regressori. Inoltre se il vincolo di identificabilità (3.2) è soddisfatto e la matrice delle esplicative ha rango pieno, il problema di minimizzazione è rappresentato da una funzione di perdita strettamente convessa e dunque la soluzione è unica. Tale soluzione può essere ottenuta tramite un algoritmo iterativo chiamato *backfitting* (Azzalini e Scarpa 2012).

Foreste casuali

Le foreste casuali fanno parte dei cosiddetti metodi di *ensemble*. Tali metodi si basano sul combinare più algoritmi di apprendimento per ottenere previsioni più accurate rispetto all'utilizzo di un solo modello. Nella formulazione originaria di Breiman (2001) le foreste casuali sono costruite combinando insieme molti alberi decisionali che vengono fatti crescere fino quasi alle foglie senza essere potati. Ciascun albero avrà quindi una bassa distorsione ed un'alta varianza, mentre la loro combinazione non dà luogo sovradattamento e presenta bassa varianza (Azzalini e Scarpa 2012).
Ciò è garantito dal teorema centrale del limite[2], dal momento che le previsioni si ottengono come medie dei risultati ottenuti da ciascun albero. Ma, affinché sia possibile sfruttare i risultati di tale teorema, è necessario che gli alberi siano numerosi ed inoltre indipendenti ed identicamente distribuiti.
Per ottenere alberi *i.i.d.* ci si avvale della stessa procedura utilizzata per i metodi di *bagging*. Tramite una procedura di campionamento *bootstrap* vengono

[2]Date n variabili aleatorie X_j *i.i.d.* di media $\mathbb{E}[X_j] = \mu$ e varianza finita $\mathbb{V}[X_j] = \sigma^2$ si ha che la media $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_j$ tende in legge ad una normale di media μ e varianza σ^2/n.

quindi creati B nuovi campioni di dati, selezionando casualmente con reinserimento n unità dal campione di partenza. Ogni nuovo campione avrà quindi la stessa numerosità del campione iniziale e costituirà uno dei B insiemi di stima su cui adattare gli alberi che andranno a comporre la foresta. Le osservazioni che non rientrano nei nuovi campioni, invece, saranno raccolte nei rispettivi insiemi *out-of-bag*. Inoltre ciascun albero sarà stimato utilizzando un diverso insieme di esplicative di numerosità $q < p$. Le q esplicative vengono infatti selezionate in modo randomico tra le p variabili disponibili. Ciò assicura che tutte le variabili abbiano la possibilità di essere selezionate.
L'indipendenza degli alberi permette inoltre di rendere il processo di stima parallelizzabile, riducendo il tempo computazionale.
I parametri da scegliere per adattare la foresta sono quindi il numero di alberi B e il numero di variabili q utilizzate per la costruzione di ciascun albero. Per quanto riguarda il primo parametro, Breiman (2001) stesso ha dimostrato che al crescere di B il modello non è soggetto a sovradattamento. Si osserva infatti che all'aumentare del numero di alberi l'errore di previsione converge ad una soglia inferiore, quindi scegliendo B sufficientemente grande si è certi che l'errore di previsione risultante non sarà tanto lontano dal suo limite inferiore (Azzalini e Scarpa 2012).
Per quanto riguarda il numero di esplicative q utilizzate per ciascun albero, Breiman (2001) consiglia di utilizzare un terzo delle variabili globalmente disponibili. Tuttavia, varie applicazioni hanno dimostrato che non sempre questa numerosità si rivela ottima. Di conseguenza è meglio trattare q come un parametro di regolazione da scegliere in modo adattivo, ad esempio tramite convalida incrociata o stima e verifica. Un ulteriore modo per determinare q consiste nello sfruttare l'insieme *out-of-bag*. Infatti tale insieme non viene utilizzato per stimare il modello e può quindi essere sfruttato per calcolare una stima affidabile dell'errore di previsione. Si sceglierà allora il parametro q che minimizza l'errore calcolato sull'insieme *out-of-bag*.
L'insieme *out-of-bag* non è però solo utile ai fini della stima dell'errore di previsione, permette infatti di determinare anche una misura dell'importanza delle variabili. La procedura consiste di due passi. Prima si calcola l'errore complessivo nell'insieme *out-of-bag*, poi per ciascuna variabile in tale insieme si effettua una permutazione dei valori assunti, mantenendo costanti i valori

di tutte le altre, e si calcola l'errore di previsione che ne consegue. A questo punto per ogni predittore si calcola la differenza tra l'errore generale del modello e quello ottenuto in seguito alle permutazioni. Una variabile sarà tanto più importante quanto più è grande la differenza tra i due errori. La procedura viene ripetuta per tutti gli alberi della foresta dando luogo a B differenze di errori per ciascuna variabile. L'indice di importanza finale consiste nella media di questi errori divisa per la deviazione standard. Infine, per avere confronto più immediato tra tutte le variabili, si possono riscalare i valori ottenuti, in modo da attribuire un punteggio di 100 all'esplicativa più rilevante per il modello.

Gradient Boosting

I metodi di *boosting* sono stati ideati inizialmente per risolvere problemi di classificazione. Tuttavia dato l'enorme successo ottenuto, sono presto stati estesi per essere impiegati in problemi di regressione. Nella loro accezione originaria (AdaBoost) l'obiettivo è quello di combinare classificatori deboli, dotati di bassa varianza ma alta distorsione, per ottenere risultati competitivi in termini di errore (Hastie, Tibshirani e Friedman 2009). Generalmente si fa uso di alberi decisionali per la costruzione di tali modelli, in quanto possiedono le caratteristiche sopra citate e sono di facile implementazione.
Il Gradient Boosting è quindi un metodo di *ensemble* come la foresta casuale e, al pari di questa, sfrutta gli alberi decisionali per la costruzione del modello. Tuttavia esistono sostanziali differenze tra i due metodi. Gli alberi del Gradient Boosting devono infatti essere modelli di previsione deboli a bassa varianza e quindi vengono fatti crescere solo fino ad una profondità prestabilita. Inoltre tali alberi non sono indipendenti, in quanto vengono costruiti in modo sequenziale. L'algoritmo di stima del modello si basa infatti su un processo iterativo che ad ogni passo aggiorna le previsioni generate al passo precedente. Nel dettaglio il processo prevede che si minimizzi una determinata funzione di perdita, il che equivale a minimizzare la funzione muovendosi nella direzione opposta al gradiente. Si dimostra che utilizzando la funzione di perdita quadratica il gradiente coincide con i residui cambiati di segno (Hastie, Tibshirani e Friedman 2009).

Il primo passo dell'algoritmo prevede di inizializzare il modello ad un valore costante (in genere la media delle osservazioni). In seguito si calcolano i gradienti negativi relativi alla funzione di perdita selezionata e si adatta il primo albero utilizzando questi valori come nuova risposta. L'algoritmo procede quindi in modo iterativo, alternando il calcolo dei gradienti negativi alla stima del successivo albero basato su tali valori. Ad ogni passo le previsioni vengono aggiornate sommando al valore del passo precedente la risposta generata dall'albero del passo attuale. Il processo si arresta quando viene raggiunto il numero di iterazioni prestabilito M.
Tuttavia questo procedimento potrebbe non convergere all'ottimo globale. Infatti gli alberi sono costruiti attraverso un algoritmo *greedy* che ad ogni split individua la combinazione ottima per quella determinata configurazione. Inoltre, a differenza di quanto accade per le foreste casuali, un elevato numero di iterazioni può portare al sovradattamento. Per ovviare a tali problemi è possibile introdurre un ulteriore parametro di regolazione all'interno del processo di stima (Hastie, Tibshirani e Friedman 2009). Tale parametro ν è detto "tasso di apprendimento" e ha l'obiettivo di ridurre la quantità con cui viene aggiornata la previsione in ciascun passo dell'algoritmo. Solo una frazione del risultato prodotto dall'albero al passo $j + 1$-esimo sarà quindi aggiunto al valore risultante del passo j-esimo. Il parametro di shrinkage può assumere valori compresi tra 0 e 1. Più piccolo sarà il valore del parametro e più i risultati provenienti da ciascun albero saranno penalizzati. Questo significa che la convergenza verso la soluzione di ottimo sarà più lenta e richiederà quindi un numero maggiore di iterazioni. Considerando che la procedura di stima non è parallelizzabile, a causa della costruzione sequenziale degli alberi, ciò può portare a costi computazionali non indifferenti.
Infine un aspetto interessante del Gradient Boosting è la possibilità di stilare una classifica dell'importanza delle variabili. Per ogni variabile si considera di quanto è migliorato l'errore di previsione ad ogni split in cui è stata coinvolta tale variabile. L'indice è quindi calcolato come media dei valori ottenuti considerando tutti gli alberi del processo in cui è coinvolta l'esplicativa. Come nel caso delle foreste casuali, è poi possibile riscalare i valori ottenuti, attribuendo un punteggio di 100 alla variabile più rilevante per il modello.

3.3 Applicazione dei modelli

Di seguito si riportano i risultati ottenuti per ciascun modello per ciascun quartiere. Prima di presentare i risultati, si è ritenuto opportuno descrivere brevemente le caratteristiche principali di ciascun distretto. Infatti conoscere il contesto può aiutare a comprendere meglio i risultati e a trarre conclusioni appropriate.
Inoltre, per far emergere il contributo che gli Open Data possono dare in termini di comprensione del fenomeno e di capacità predittiva, si è scelto di adattare prima i modelli con le sole caratteristiche territoriali (latitudine e longitudine del centroide, superficie del census tract, numero di residenti e anno) e poi quelli con tutte le altre variabili "open". Si sono quindi calcolati gli errori di previsione.

Nell'adattare ciascun modello sono state fatte alcune scelte, sia in relazione alle variabili, sia nella determinazione degli iperparametri.
Per quanto riguarda il modello additivo, si è scelto di non applicare i lisciatori alle variabili territoriali `area` e `popolazione`. La scelta è motivata dal fatto che si è ritenuto di maggior interesse valutare l'impatto dovuto alle caratteristiche specifiche dei residenti. Quindi, per evitare un possibile mascheramento degli effetti, si è deciso di inserire il contributo delle variabili sopra citate solo in forma parametrica. Per tenere in considerazione l'effetto congiunto di latitudine e longitudine si è invece fatto ricorso ad una *thin plate spline* bivariata.
In riferimento alle foreste casuali si è scelto di inserire l'interazione tra longitudine e latitudine attraverso nuova variabile che tiene conto del loro prodotto algebrico. Per tali modelli si sono anche compiute scelte differenti nella determinazione dei parametri B e q a seconda della numerosità di osservazioni dei vari quartieri. Infatti dalla tabella 2.1 si può notare come le numerosità siano eterogenee in relazione ai vari distretti. In particolare per Manhattan, Bronx e Staten Island si è scelto di adattare foreste casuali composte da 2000 alberi, determinando il parametro q tramite convalida incrociata a cinque. Per quanto riguarda Brooklyn e Queens, che presentano un maggior numero di osservazioni, si sono invece utilizzati 800 alberi e si è sfruttato l'insieme *out-of-bag* nella determinazione del parametro q ottimale.

Infine per il Gradient Boosting, per motivi computazionali, si è scelto di fissare arbitrariamente il parametro di shrinkage a 0.01 e di determinare quindi tramite convalida incrociata i valori ottimi per la profondità degli alberi e per il numero di iterazioni dell'algoritmo. Anche in questo caso, come per le foreste casuali, l'interazione tra latitudine e longitudine è stata inserita come prodotto tra le due variabili.

Nel presentare i risultati del modello additivo, per non appesantire troppo la trattazione, si è scelto di mostrare solo alcuni grafici delle variabili che si sono ritenute di maggior interesse nello spiegare il fenomeno. I rimanenti grafici sono riportati in appendice B.

3.3.1 Manhattan

Manhattan è il quartiere meno esteso, ma più densamente popolato di tutta New York. L'isola è circondata da tre corsi d'acqua che la dividono dal resto della città: a ovest il fiume Hudson, a est e a sud l'East River e a nord l'Harlem River; ma presenta numerosi collegamenti con gli altri distretti.
Manhattan è spesso identificata come il cuore della città. In questo quartiere si concentrano infatti quasi tutti i luoghi più importanti e famosi di New York. Per quanto riguarda le attrazioni turistiche è possibile spaziare tra la Statua della Libertà, Central Park, Times Square, il Moma, i teatri di Broadway, e tante altre attrazioni. Ma il quartiere non è famoso solo dal punto di vista culturale. La presenza della Borsa di Wall Street fa infatti sì che possa essere considerato il centro economico degli Stati Uniti.
Infine Manhattan è senz'altro famosa per il suo skyline putellato di grattacieli, tra cui spiccano l'Empire State Building, il Chrysler Building e il palazzo dell'ONU.

La tabella 3.2 mostra i risultati dei modelli stimati per Manhattan. In primo luogo si può notare che foresta casuale e Gradient Boosting hanno dato luogo ad errori di previsioni molto simili tra loro, sia per quanto riguarda il modello base sia per quanto riguarda il modello con tutte le variabili. Gli errori del modello additivo risultano invece più che doppi rispetto a quelli appena descritti. Inoltre, sempre per il modello additivo, l'errore prodotto dal modello

con tutte le covariate è maggiore anche di quelli di foresta casuale e Gradient Boosting senza le variabili "open". In ogni caso, per tutti i modelli, i risultati migliori si ottengono con l'aggiunta delle variabili esplicative, confermando il valore aggiunto degli Open Data.

Tabella 3.2: Errori di previsione per Manhattan

	MSE modello base	MSE covariate	Rapporto
Modello additivo	3.318	1.726	1.922
Foresta casuale	1.209	0.917	1.319
Gradient Boosting	1.442	0.911	1.582

Le tabelle 3.3 e 3.4 riportano le stime degli effetti rispettivamente parametrici e non parametrici del modello additivo. Dalla prima tabella si nota come tutti gli effetti parametrici siano globalmente significativi. In particolare emerge quanto ipotizzato nelle analisi esplorative sull'andamento temporale del fenomeno: i coefficienti relativi agli anni sono positivi e monotoni crescenti, ad indicare un progressivo intensificarsi del fenomeno. Osservando la seconda tabella si ha invece una visione dell'effetto delle variabili "open". È interessante notare come l'effetto di tutte queste variabili sia significativo. Inoltre le stime dei gradi di libertà equivalenti mostrano che tutte le variabili, ad eccezione di quelle relative alla percentuale di residenti che utilizzano i mezzi pubblici per raggiungere la sede lavorativa e al numero di ATM, hanno un comportamento non lineare.

I grafici in figura 3.1 confermano quanto detto. Si può infatti osservare la non linearità delle variabili relative alla percentuale di residenti in età non lavorativa e alla percentuale di residenti che vive sotto la soglia di povertà, contrapposta alla linearità delle variabili relative alla percentuale di residenti che utilizzano i mezzi pubblici per raggiungere la sede lavorativa e al numero di ATM. Per quanto riguarda le prime due variabili l'andamento globale è opposto. Infatti all'aumentare della percentuale di residenti che non lavora si osserva un calo progressivo nella radice quadrata del numero di aggressioni, mentre all'aumentare della percentuale di residenti che vive sotto la soglia di

Tabella 3.3: Stima degli effetti parametrici per Manhattan

	Estimate	Std. Error	t value	Pr(>|t|)	
intercetta	4.460e+00	1.510e-01	29.541	< 2e-16	***
popolazione	2.240e-04	1.954e-05	11.461	< 2e-16	***
area	2.450e-06	5.718e-07	4.285	1.94e-05	***
anno 2015	1.297e-01	1.049e-01	1.236	0.21648	
anno 2016	2.561e-01	1.049e-01	2.441	0.01475	*
anno 2017	3.080e-01	1.049e-01	2.936	0.00338	**
anno 2018	4.191e-01	1.049e-01	3.994	6.80e-05	***
anno 2019	5.159e-01	1.049e-01	4.917	9.75e-07	***

povertà si ha un aumento nella risposta. Nel primo caso il fenomeno è coerente con quanto osservato nelle analisi esplorative (sezione 2.3.2, figura 2.5). Infatti si era osservato che in media vittime ed aggressori hanno un'età media compresa tra i 25 e i 44 anni, quindi ben lontano dalla minore età e dall'età pensionabile. Inoltre le categorie "<18" e "85+" erano risultate quelle con il minor numero di aggressioni subite o inflitte. Per quanto riguarda il secondo grafico invece, si può ipotizzare che le aggressioni siano la conseguenza di una condizione economica sfavorevole che sfocia nella violenza, così come rilevato dagli studi riportati nel paragrafo 2.1. I rimanenti due grafici mostrano entrambi un andamento crescente nella risposta al crescere della rispettiva esplicativa. Nel primo caso si può immaginare che utilizzare frequentemente i mezzi per recarsi a lavoro possa aumentare il rischio di aggressione in quanto spesso le fermate di metro ed autobus si trovano in luoghi isolati e maggiormente soggetti a criminalità. Per quanto riguarda il numero di ATM è invece lecito immaginare che le aggressioni avvengano a seguito di un prelievo di denaro, al fine di appropriarsene.

Infine i grafici in figura 3.2 riportano l'importanza delle variabili per foresta casuale e Gradient Boosting. Il primo modello è stato ottenuto facendo crescere una foresta di 2000 alberi in cui ogni albero è costruito utilizzando 7 variabili, mentre per il Gradient Boosting la scelta dei parametri ottimali con shrinkage 0.01 ha portato ad utilizzare 1600 iterazioni e a consentire in-

Tabella 3.4: Stima degli effetti non parametrici per Manhattan

	edf	F	p-value	
s (latitudine, longitudine)	26.751	14.214	< 2e-16	***
s (% maschi)	8.824	11.828	< 2e-16	***
s (età)	8.836	8.351	< 2e-16	***
s (% ispanici)	7.089	3.021	0.00221	**
s (% popolazione bianca)	8.856	7.099	< 2e-16	***
s (% afroamericani)	8.729	7.207	< 2e-16	***
s (% residenti in affitto)	6.772	2.881	0.00586	**
s (tasso di disoccupazione)	6.916	1.923	0.04932	*
s (% residenti in età non lavorativa)	8.828	20.053	< 2e-16	***
s (reddito)	8.809	10.677	< 2e-16	***
s (% residenti in povertà)	8.737	18.208	< 2e-16	***
s (% che si reca a lavoro con mezzi pubblici)	1.000	85.147	< 2e-16	***
s (% che si reca a lavoro a piedi)	8.240	9.787	< 2e-16	***
s (% lavoratori da casa)	6.847	8.751	< 2e-16	***
s (n° di fermate dei mezzi pubblici)	9.000	16.980	< 2e-16	***
s (n° di ATM)	1.059	34.264	< 2e-16	***

terazioni fino al nono grado.
Da un confronto diretto tra i due grafici si può notare come i due modelli siano concordi nell'attribuire maggiore peso alle stesse variabili. Si nota infatti che le prime tre variabili sono esattamente le stesse. La variabile di maggior importanza risulta essere la percentuale di residenti con reddito sotto la soglia di povertà. Tra le variabili "open" giocano un ruolo fondamentale anche la percentuale di popolazione bianca e il reddito percepito.

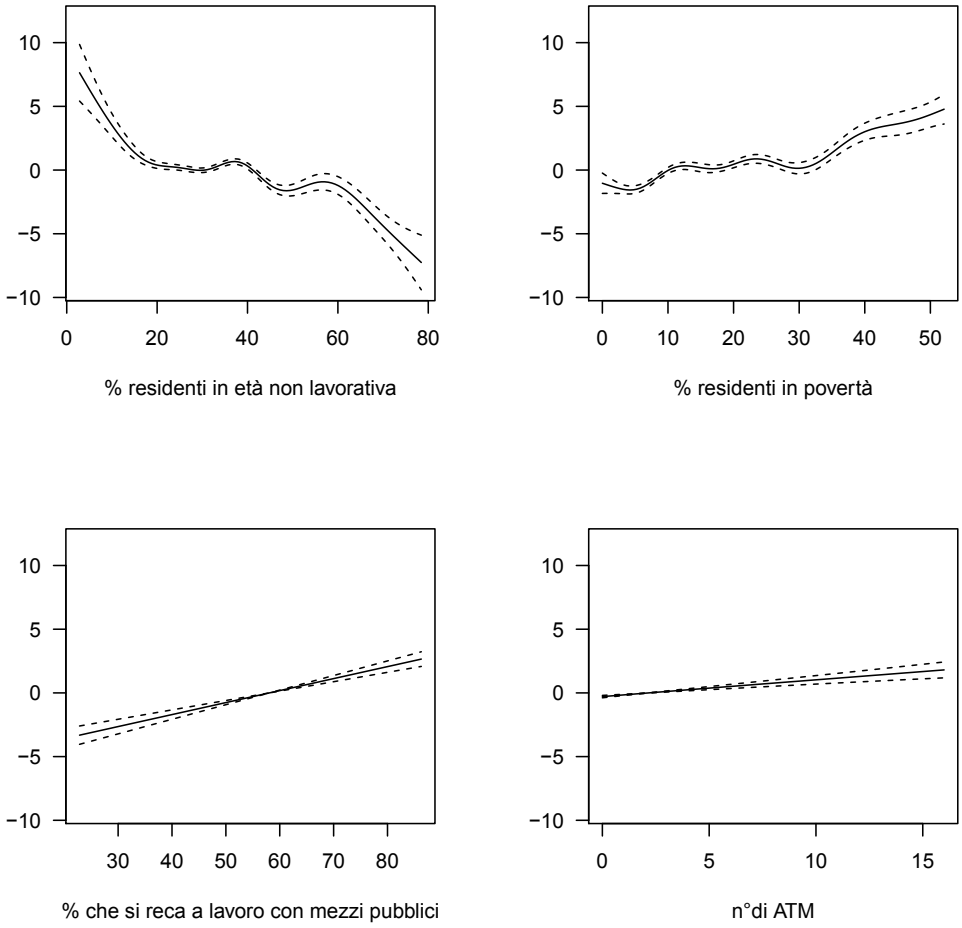


Figura 3.1: Grafici selezionati per Manhattan prodotti dal modello additivo

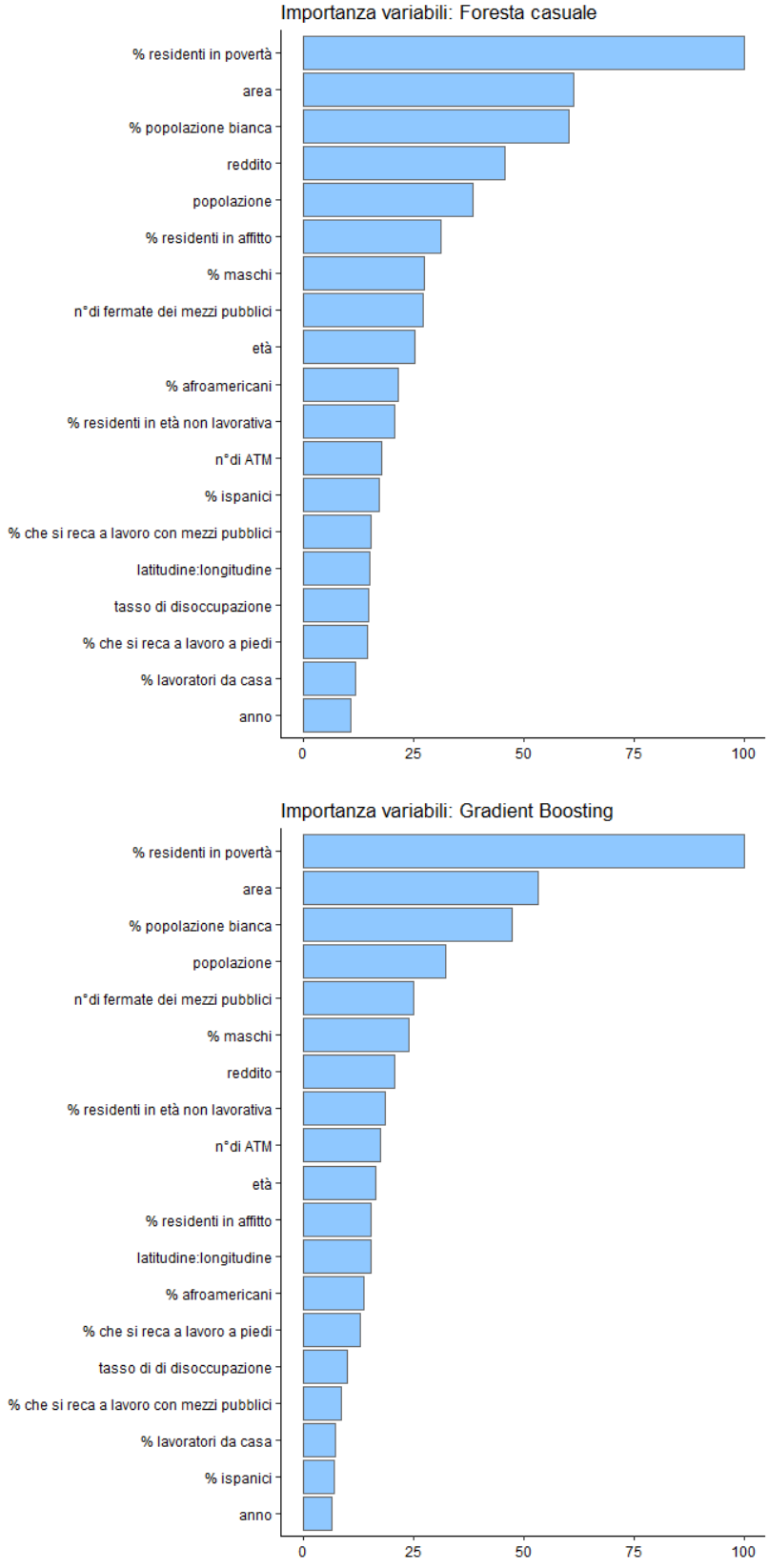


Figura 3.2: Importanza delle variabili per Manhattan

3.3.2 Brooklyn

Con una popolazione di circa 2,6 milioni di abitanti, Brooklyn è il quartiere più popoloso di New York. Si trova sull'estremità occidentale dell'isola di Long Island e ha come unico confine terrestre quello di nord-est con il Queens. I collegamenti con Manhattan e con Staten Island sono invece garantiti grazie a numerosi ponti, tra cui il celebre ponte di Brooklyn.
Il quartiere proviene da un forte passato industriale, ma oggi è un distretto vivace e alla moda. È anche noto per le ricche aree residenziali in cui vivono personaggi famosi del calibro di Woody Allen.
Dal punto di vista turistico Brooklyn è famoso soprattutto per le vedute che offre sullo skyline di Manhattan. Dal Brooklyn Bridge Park è infatti possibile godere della migliore vista sui grattacieli del quartiere degli affari. Anche il ponte di Brooklyn costituisce una grande attrazione e può essere attraversato anche a piedi e in bicicletta. Infine il sottoquartiere di Coney Island dispone di numerose spiagge e parchi divertimento.

A differenza di quanto osservato per Manhattan, per Brooklyn il modello additivo risulta competitivo come errore di previsione. Dalla tabella 3.5 si nota infatti che tutti i modelli forniscono errori confrontabili. Inoltre è di interesse il forte divario tra gli errori dei modelli senza variabili "open" e quelli dei modelli che le sfruttano. In questo caso si ha una riduzione quasi doppia per modello additivo e foresta casuale e addirittura più che doppia per Gradient Boosting.

Tabella 3.5: Errori di previsione per Brooklyn

	MSE modello base	MSE covariate	Rapporto
Modello additivo	2.288	1.188	1.926
Foresta casuale	2.283	1.195	1.911
Gradient Boosting	2.533	1.091	2.322

Le tabelle 3.6 e 3.7 riportano nel dettaglio le stime degli effetti del modello additivo per Brooklyn. Anche in questo caso si nota che tutti gli effetti parametrici e non parametrici sono significativi. A differenza di Manhattan

però, i coefficienti relativi alla variabile anno sono invertiti di segno e presentano un andamento decrescente nel tempo. Ciò è in linea con quanto si era osservato durante le analisi esplorative sull'evoluzione temporale del fenomeno e può essere contestualizzato considerando la natura del quartiere. Infatti, come accennato nella sezione descrittiva di Brooklyn, tale quartiere negli ultimi anni ha mutato profondamente la sua natura, trasformandosi da distretto industriale a realtà vivace e alla moda.

Tabella 3.6: Stima degli effetti parametrici per Brooklyn

	Estimate	Std. Error	t value	Pr(>|t|)	
intercetta	3.338e+00	7.163e-02	46.594	< 2e-16	***
popolazione	4.135e-04	1.656e-05	24.974	< 2e-16	***
area	9.500e-07	1.136e-07	8.361	< 2e-16	***
anno 2015	-7.078e-02	5.508e-02	-1.285	0.1988	
anno 2016	-1.166e-01	5.508e-02	-2.116	0.0344	*
anno 2017	-2.570e-01	5.508e-02	-4.665	3.17e-06	***
anno 2018	-2.565e-01	5.508e-02	-4.656	3.32e-06	***
anno 2019	-3.091e-01	5.508e-02	-5.612	2.12e-08	***

Per quanto riguarda i grafici in figura 3.3 si può osservare che l'etnia dei residenti non sembra giocare un ruolo chiave nella descrizione del fenomeno. I relativi grafici si mostrano abbastanza piatti, senza particolari cali o incrementi. Il discorso è invece diverso per quanto riguarda l'età. In questo caso si nota che l'effetto è pressoché costante fino ad un'età media di 55 anni per poi crollare rapidamente. Anche in questo caso possiamo immaginare che un census tract dove vivono in prevalenza anziani sia meno soggetto ad aggressioni, per quanto già detto in riferimento alle analisi descrittive. L'ultimo grafico della figura mostra invece l'andamento della risposta in funzione del tasso di disoccupazione. In questo caso si nota un andamento leggermente crescente, che lascia intuire che la mancanza di un lavoro, e quindi di un'entrata economica, può sfociare nella violenza.

I risultati di foresta casuale e Gradient Boosting in termini di importanza delle variabili sono stati riportati in figura 3.4. In questo caso la foresta casuale

Tabella 3.7: Stima degli effetti non parametrici per Brooklyn

	edf	F	p-value	
s (latitudine, longitudine)	27.040	16.046	< 2e-16	***
s (% maschi)	8.457	3.092	0.00153	**
s (età)	8.347	10.052	< 2e-16	***
s (% ispanici)	7.240	24.018	< 2e-16	***
s (% popolazione bianca)	8.706	7.330	< 2e-16	***
s (% afroamericani)	8.916	16.055	< 2e-16	***
s (% residenti in affitto)	8.037	8.203	< 2e-16	***
s (tasso di disoccupazione)	7.238	9.716	< 2e-16	***
s (% residenti in età non lavorativa)	8.711	10.594	< 2e-16	***
s (reddito)	8.012	7.913	< 2e-16	***
s (% residenti in povertà)	8.850	17.350	< 2e-16	***
s (% che si reca a lavoro con mezzi pubblici)	8.336	14.084	< 2e-16	***
s (% che si reca a lavoro a piedi)	8.502	6.255	< 2e-16	***
s (% lavoratori da casa)	8.135	8.309	< 2e-16	***
s (n° di fermate dei mezzi pubblici)	8.014	63.084	< 2e-16	***
s (n° di ATM)	4.832	21.678	< 2e-16	***

è stata adattata con 800 alberi, ciascuno costruito con 17 variabili; mentre la scelta dei parametri del Gradient Boosting con shrinkage 0.01 ha portato a 1929 iterazioni ed una profondità di 8 per gli alberi.

Anche in questo caso i modelli sono concordi nella scelta delle variabili di maggiore impatto. Per Brooklyn si ha che l'etnia dei residenti, la percentuale di residenti che vive in affitto e il reddito percepito sono fattori importanti nello spiegare il fenomeno.

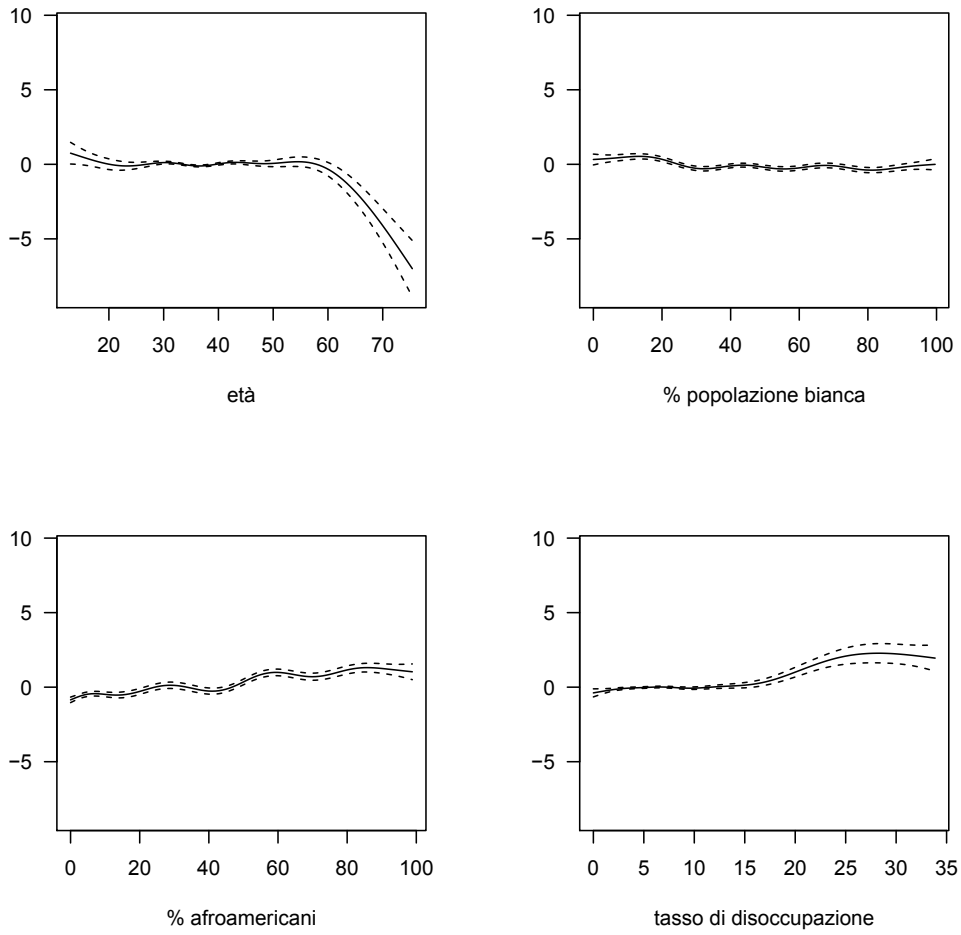


Figura 3.3: Grafici selezionati per Brooklyn prodotti dal modello additivo

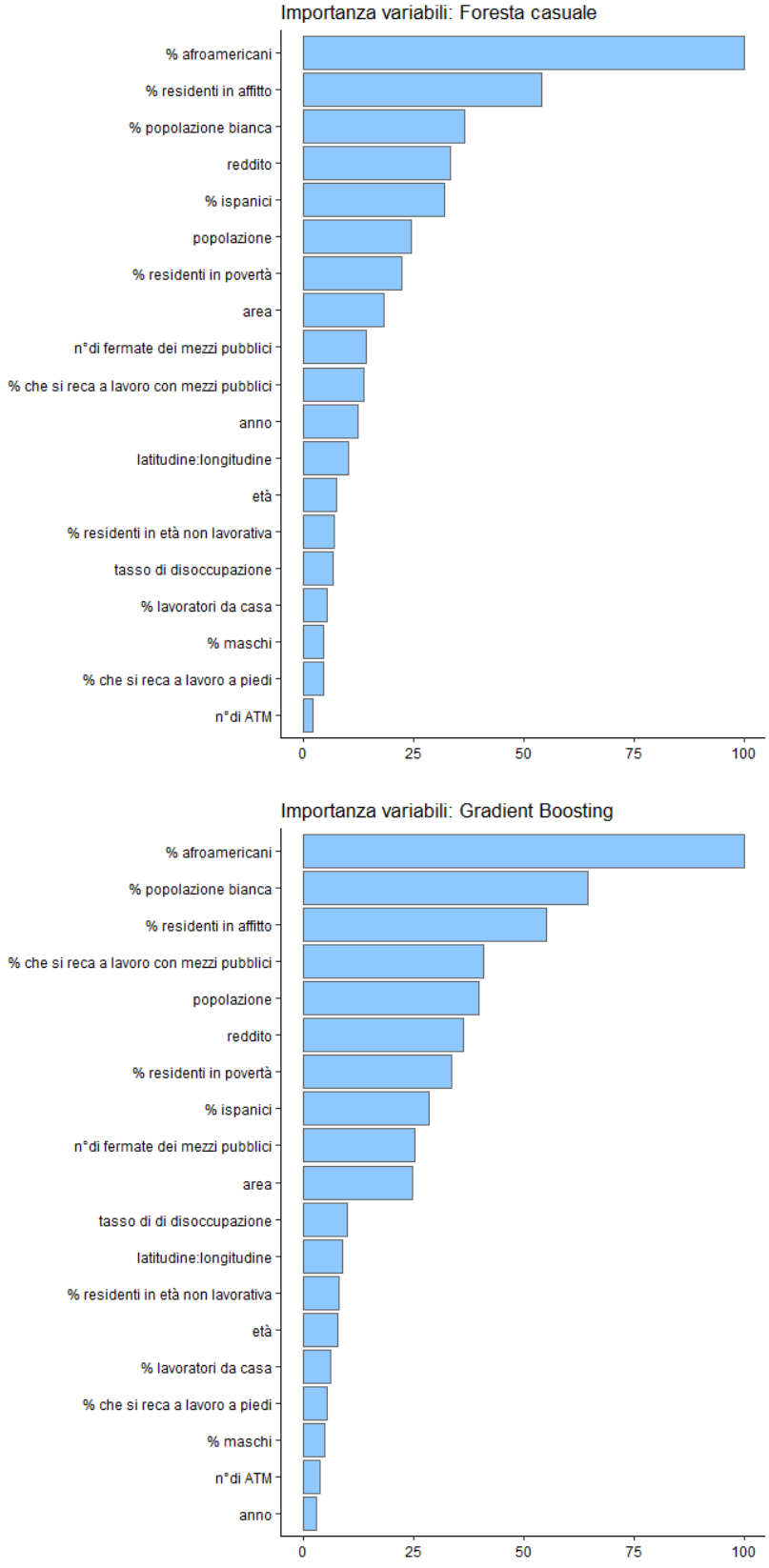


Figura 3.4: Importanza delle variabili per Brooklyn

3.3.3 Bronx

Il Bronx è il quartiere più a nord di New York ed è conosciuto in tutto il mondo per essere una zona difficile e poco sicura. Questa spiacevole fama è dovuta agli avvenimenti degli anni '70 che hanno visto il quartiere soggetto ad incendi, distruzione di interi blocchi di edifici e conseguente decadimento urbano. Povertà, criminalità ed elevati tassi di disoccupazione hanno fatto sì che si radicassero stereotipi negativi che ancora oggi sono difficili da cancellare. Infatti nonostante si sia assistendo ad una progressiva rinascita del quartiere, i turisti tendono ancora ad evitare la zona, etichettandola come pericolosa.
Il Bronx non è però solo famoso nel mondo per aspetti negativi. Qui vi hanno infatti sede i New York Yankees, una delle squadre di baseball americane più importanti e famose. Oltre agli aspetti sportivi, il quartiere è noto anche per la sua storia musicale. È stato infatti la culla di hip-hop, rap e break dance, che si sono poi rapidamente diffusi in tutto il resto del mondo.
Infine una peculiarità del quartiere è il suo plurilinguismo: per le strade del Bronx si parlano ben 75 lingue differenti, sebbene inglese e spagnolo siano quelle predominanti.

Come per Manhattan anche per il Bronx il modello additivo non riesce a competere con foresta casuale e Gradient Boosting in termine di errore di previsione. Inoltre anche in questo caso si ha che i modelli senza variabili "open" di foresta casuale e Gradient Boosting risultano migliori in termini di errore del modello additivo che considera tutte le esplicative. Dalla tabella 3.8 si nota anche che il miglior risultato in termini di errore è raggiunto dal Gradient Boosting. Anche in questo caso però, seppur in modo minore, tutti gli errori ottenuti con i modelli che considerano tutte le variabili sono inferiori a quelli che non ne fanno uso. La differenza maggiore si nota nel caso del mdello additivo, dove il rapporto tra gli errori è quasi il doppio.

A differenza di quanto visto finora per gli altri quartieri, la tabella 3.9 riporta la non significatività di molti coefficienti legati alla variabile anno. Dal test anova emerge però una significatività globale per tale variabile (p-value 0.00288), tuttavia non è possibile individuare un pattern temporale preciso. Gli effetti delle variabili in tabella 3.10 risultano invece tutti significativi.

Tabella 3.8: Errori di previsione per il Bronx

	MSE modello base	MSE covariate	Rapporto
Modello additivo	2.845	1.439	1.976
Foresta casuale	1.097	0.909	1.206
Gradient Boosting	1.167	0.858	1.360

Tabella 3.9: Stima degli effetti parametrici per il Bronx

	Estimate	Std. Error	t value	Pr(>|t|)	
intercetta	4.003e+00	1.138e-01	35.187	< 2e-16	***
popolazione	5.630e-04	2.120e-05	26.560	< 2e-16	***
area	1.613e-06	2.026e-07	7.959	3.03e-15	***
anno 2015	-3.317e-02	8.913e-02	-0.372	0.70984	
anno 2016	1.067e-01	8.913e-02	1.197	0.23142	
anno 2017	1.559e-02	8.913e-02	0.175	0.86114	
anno 2018	2.625e-01	8.913e-02	2.945	0.00327	**
anno 2019	2.049e-01	8.913e-02	2.298	0.02166	*

Anche in questo caso si sono riportati alcuni grafici relativi all'andamento delle variabili nel modello additivo. In figura 3.5 si può osservare l'andamento della risposta in funzione dell'età dei residenti, della percentuale di popolazione bianca e di afroamericani e del reddito.
Il grafico forse di maggior interesse è quello relativo all'età. Infatti, a differenza di quanto visto finora, per il Bronx si osserva un aumento nella risposta all'aumentare dell'età media. Il Bronx è infatti per molti aspetti diverso dal resto di New York e rappresenta per certi versi una realtà a sé stante con un passato difficile. Il fatto di aver analizzato il fenomeno delle aggressioni dividendo le analisi per ciascun quartiere, ha permesso di far emergere questo aspetto che altrimenti non sarebbe stato possibile cogliere.
I grafici relativi all'etnia mettono in luce un altro aspetto interessante. In entrambi i casi si osserva un incremento nella risposta al crescere delle percentuali, il che mette in evidenza un rischio maggiore di aggressioni nelle zone mono-etniche.

Tabella 3.10: Stima degli effetti non parametrici per il Bronx

	edf	F	p-value	
s (latitudine, longitudine)	27.646	15.815	$<$ 2e-16	***
s (% maschi)	8.906	23.324	$<$ 2e-16	***
s (età)	8.267	5.801	1.24e-07	***
s (% ispanici)	8.673	11.011	$<$ 2e-16	***
s (% popolazione bianca)	3.920	2.823	0.014079	*
s (% afroamericani)	8.412	10.815	$<$ 2e-16	***
s (% residenti in affitto)	8.345	4.169	2.68e-05	***
s (tasso di disoccupazione)	7.228	4.622	1.19e-05	***
s (% residenti in età non lavorativa)	8.908	9.246	$<$ 2e-16	***
s (reddito)	8.485	8.742	$<$ 2e-16	***
s (% residenti in povertà)	6.530	3.601	0.000514	***
s (% che si reca a lavoro con mezzi pubblici)	8.119	9.946	$<$ 2e-16	***
s (% che si reca a lavoro a piedi)	8.769	6.374	$<$ 2e-16	***
s (% lavoratori da casa)	8.302	10.330	$<$ 2e-16	***
s (n° di fermate dei mezzi pubblici)	8.055	38.295	$<$ 2e-16	***
s (n° di ATM)	3.488	16.285	$<$ 2e-16	***

Infine, in linea con quanto già osservato per gli altri quartieri in termini degli effetti delle variabili relative alla situazione economica dei residenti, per il Bronx si ha che all'aumentare del reddito il valore della risposta cala. Ciò può essere contestualizzato anche considerando le caratteristiche del quartiere. Infatti gli interventi di riqualifica degli ultimi anni hanno portato alla nascita di aree residenziali mediamente più ricche, più sorvegliate e di conseguenza con tassi di criminalità minori.

Per il Bronx la foresta casuale è stata stimata con 2000 alberi e 7 variabili per ciascun albero. L'ottimizzazione dei parametri svolta per il Gradient Boosting con shrinkage 0.01 ha invece portato a 1940 iterazioni con possibilità di interazione fino ad 8 livelli. Anche in questo caso si sono riportati i grafici relativi all'importanza delle variabili. Dalla figura 3.6 si può osservare una parziale sintonia tra i due metodi. Per la prima volta si nota infatti che il Gradient Boosting tende a dare importanza soltanto a due variabili, mentre la foresta casuale distribuisce i pesi in modo più graduale. In entrambi i casi

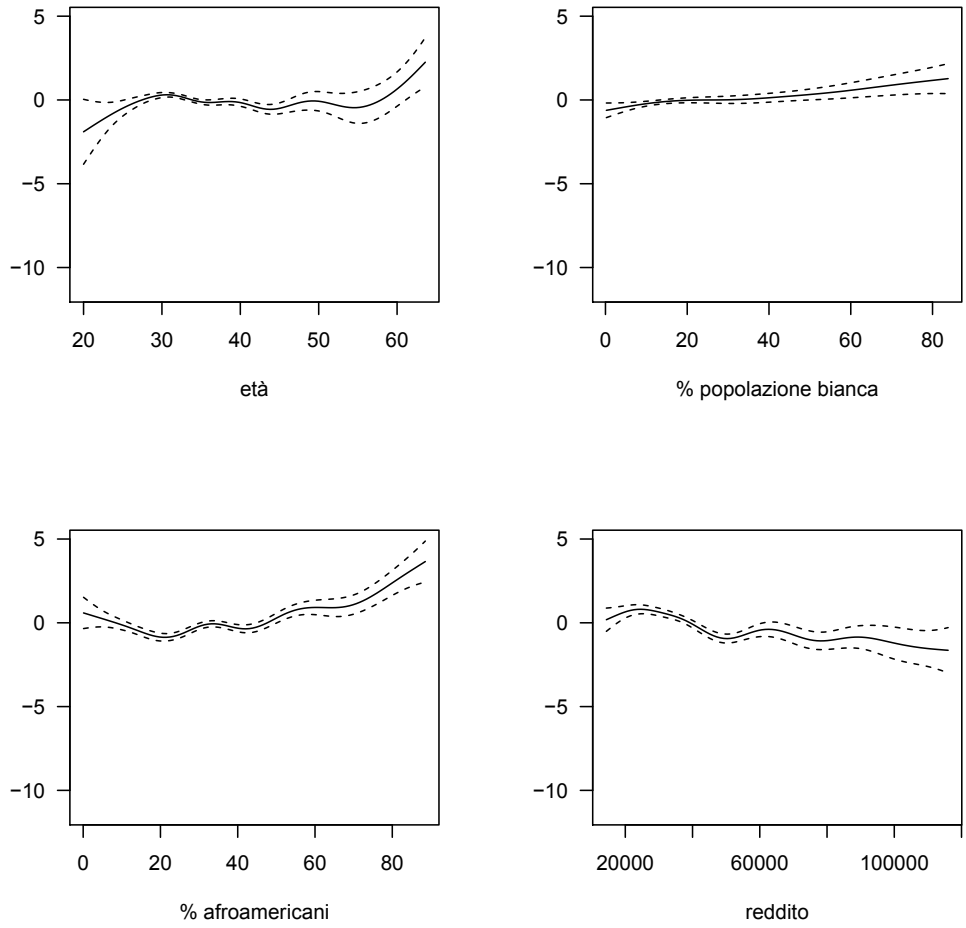


Figura 3.5: Grafici selezionati per il Bronx prodotti dal modello additivo

al primo posto si trova la variabile relativa al reddito.

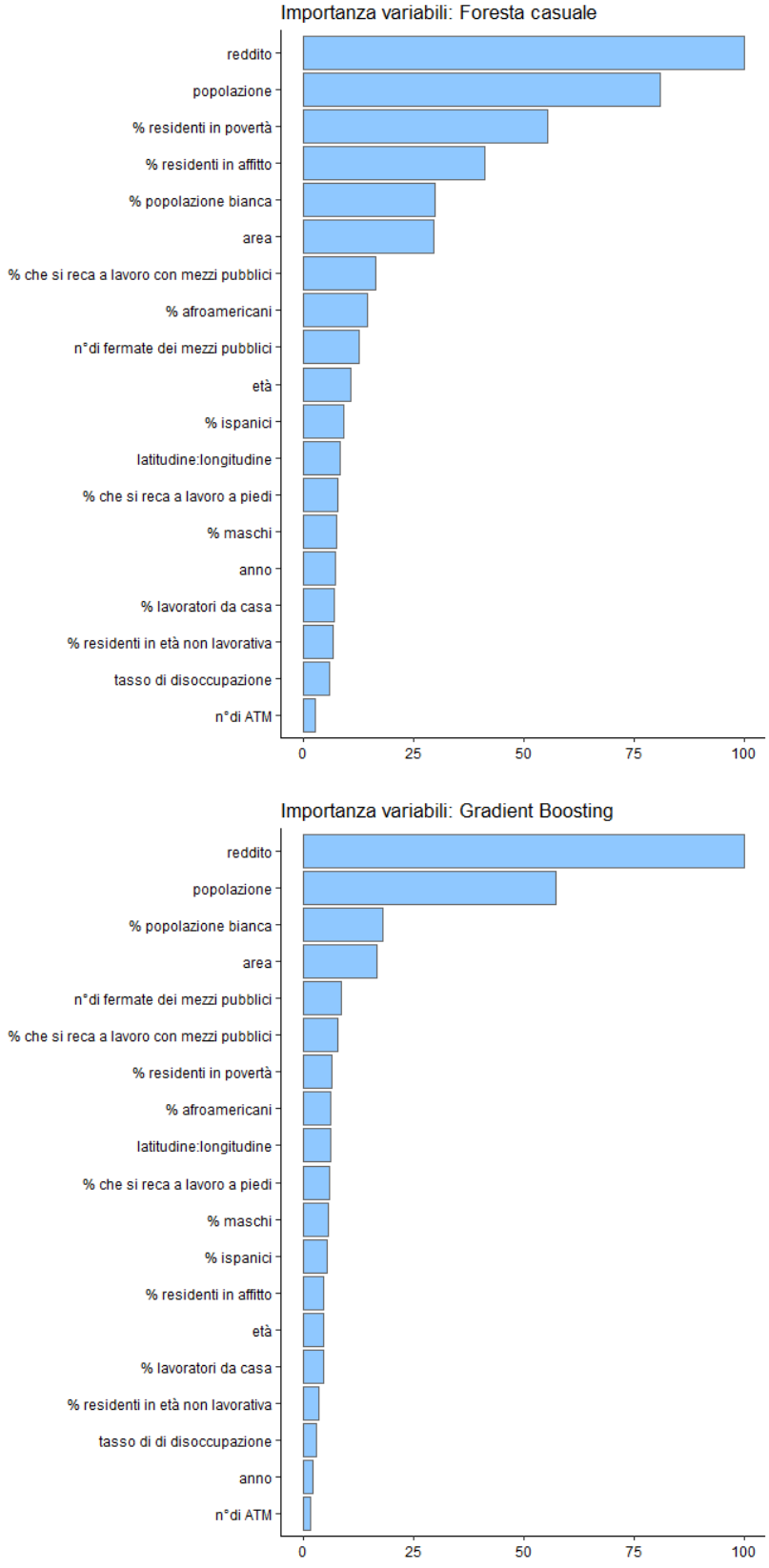


Figura 3.6: Importanza delle variabili per il Bronx

3.3.4 Queens

Il Queens è il quartiere più esteso di New York e si trova nel nord-ovest dell'isola di Long Island. Stando ai dati dell'American Community Survey è il distretto più diversificato della città dal punto di vista culturale. La marcata presenza di culture differenti fa sì che per le strade del quartiere si possano assaggiare specialità culinarie provenienti da ogni parte del mondo, così come visitare negozi con prodotti tradizionali.
Il Queens è un quartiere tranquillo a prevalenza residenziale. Non sono presenti molti grattacieli, ma si trovano prevalentemente piccole case in legno abitate da singole famiglie. Spesso viene quindi identificato come quartiere periferico di New York.
Infine il quartiere è ben noto per la musica e gli eventi sportivi. Per quanto riguarda la musica è, insieme al Bronx, uno dei centri nevralgici del rap e dell'hip hop ed inoltre è stato la culla del jazz. Dal punto di vista sportivo il quartiere è famoso in quanto vi si svolge l'U.S. Open Tennis, il quarto e ultimo dei tornei di tennis del Grande Slam, ed è la sede della squadra di baseball degli New York Mets.

Anche per il Queens vale quanto già osservato per Manhattan e Bronx circa gli errori di previsione. Infatti dalla tabella 3.11 si nota il divario tra il modello additivo e gli altri due modelli. Anche in questo caso i modelli che utilizzano tutte le variabili presentano errori inferiori a quelli che non ne fanno uso. In questo caso però il rapporto é simile per tutti i modelli.

Tabella 3.11: Errori di previsione per il Queens

	MSE modello base	MSE covariate	Rapporto
Modello additivo	1.749	1.177	1.487
Foresta casuale	1.192	0.899	1.326
Gradient Boosting	1.273	0.858	1.483

Per quanto riguarda le stime degli effetti parametrici del modello additivo, dalla tabella 3.12 si nota che tutte le variabili sono globalmente significative. Come per il Bronx non è però possibile individuare un pattern temporale

ben definito, in quanto i coefficienti relativi alla variabile anno sono simili in valore assoluto, delineando una certa stazionarietà del fenomeno. La tabella 3.13 riporta invece gli effetti non parametrici. Per la prima volta si osserva la non significatività di una variabile, in particolare quella relativa al tasso di disoccupazione.

Tabella 3.12: Stima degli effetti parametrici per il Queens

	Estimate	Std. Error	t value	Pr(>|t|)	
intercetta	2.981e+00	6.289e-02	47.403	< 2e-16	***
popolazione	3.420e-04	1.326e-05	25.794	< 2e-16	***
area	3.856e-07	8.711e-08	4.426	9.85e-06	***
anno 2015	-1.131e-01	5.953e-02	-1.900	0.0576	.
anno 2016	-1.226e-01	5.953e-02	-2.059	0.0395	*
anno 2017	-1.052e-01	5.953e-02	-1.768	0.0772	.
anno 2018	-1.159e-01	5.953e-02	-1.947	0.0516	.
anno 2019	-1.268e-03	5.953e-02	-0.021	0.9830	

La figura 3.7 riporta alcuni grafici prodotti dal modello additivo che si sono ritenuti di interesse. In particolare è riportato l'andamento della risposta in funzione della percentuale di ispanici, di afroamericani, del reddito e della percentuale di residenti sotto la soglia di povertà.
Dai primi due grafici si può osservare un consistente incremento nella risposta all'aumentare della percentuale di ispanici e un lieve calo all'aumentare della percentuale di afroamericani. I rimanenti grafici mostrano invece due aspetti complementari. Si nota infatti come il rischio di aggressioni cali all'aumentare del reddito, e all'opposto cresca all'aumentare della percentuale di residenti che vivono sotto la soglia di povertà. Ciò è concorde con quanto evidenziato in Fleisher (1966) e in Kelly (2000).

Infine la figura 3.8 riporta l'importanza delle variabili di foresta casuale e Gradient Boosting. In questo caso la foresta casuale consta di 800 alberi costruiti con 7 variabili, mentre il Gradient Boosting, fissato lo shrinkage di 0.01, è stato adattato con 1510 iterazioni e posibilità di far crescere gli alberi

Tabella 3.13: Stima degli effetti non parametrici per il Queens

	edf	F	p-value	
s (latitudine, longitudine)	27.277	20.842	< 2e-16	***
s (% maschi)	6.765	2.119	0.0265	*
s (età)	7.188	7.643	< 2e-16	***
s (% ispanici)	8.053	17.292	< 2e-16	***
s (% popolazione bianca)	8.711	7.427	< 2e-16	***
s (% afroamericani)	3.501	6.884	1.16e-05	***
s (% residenti in affitto)	8.723	13.991	< 2e-16	***
s (tasso di disoccupazione)	5.318	1.396	0.2219	
s (% residenti in età non lavorativa)	7.689	10.687	< 2e-16	***
s (reddito)	3.974	5.948	1.68e-05	***
s (% residenti in povertà)	8.061	16.446	< 2e-16	***
s (% che si reca a lavoro con mezzi pubblici)	8.027	7.261	< 2e-16	***
s (% che si reca a lavoro a piedi)	7.419	9.977	< 2e-16	***
s (% lavoratori da casa)	6.343	7.261	< 2e-16	***
s (n° di fermate dei mezzi pubblici)	7.929	28.572	< 2e-16	***
s (n° di ATM)	2.355	18.775	< 2e-16	***

fino ad una profondità di 7.
Come nei casi precedenti i modelli si mostrano concordi nell'attribuzione dell'importanza alle variabili principali. Oltre al totale della popolazione, spiccano il reddito, la percentuale di afroamericani e l'età.

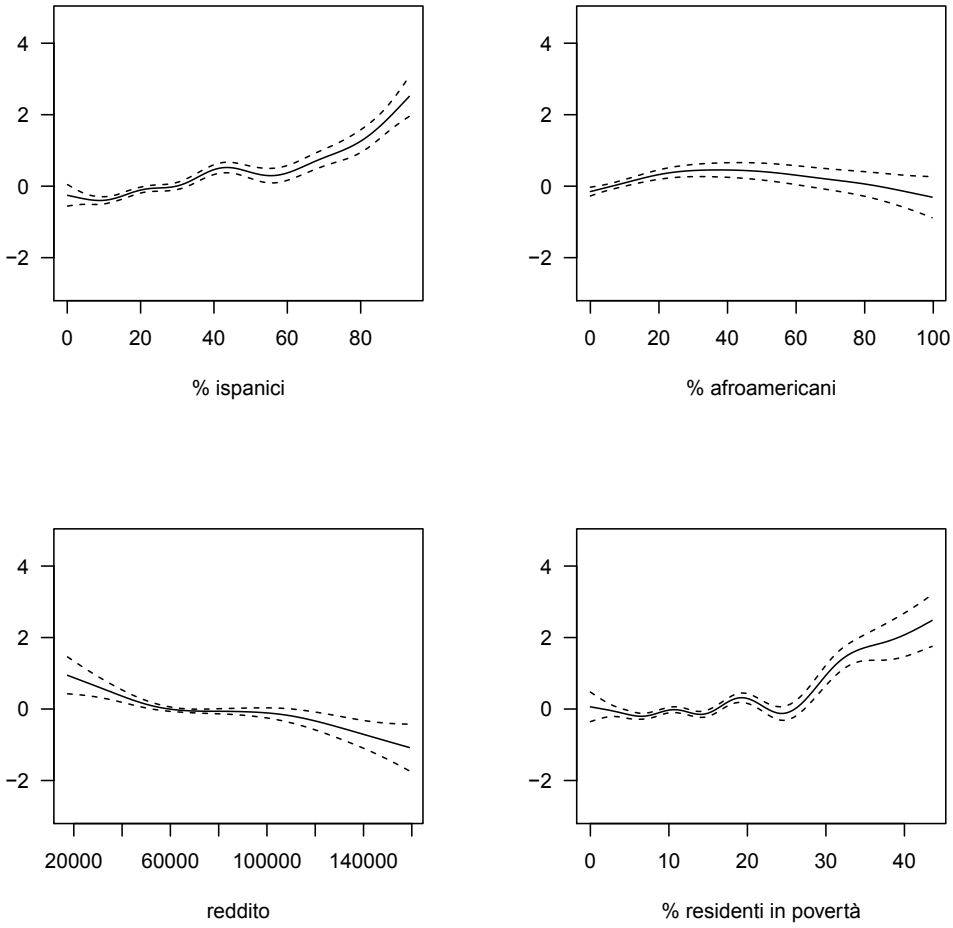


Figura 3.7: Grafici selezionati per il Queens prodotti dal modello additivo

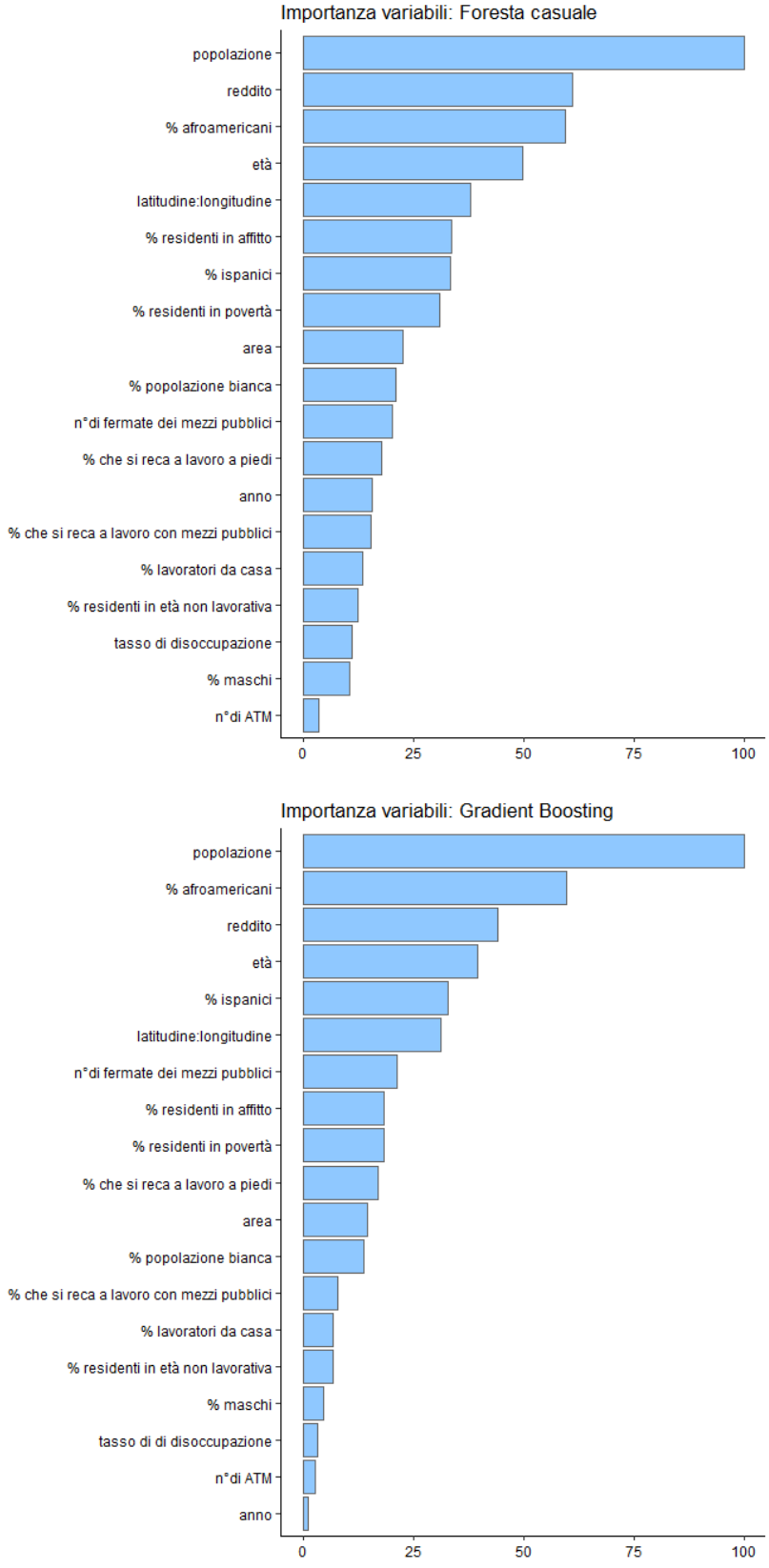


Figura 3.8: Importanza delle variabili per il Queens

3.3.5 Staten Island

Staten Island è forse il distretto meno noto di New York. È situato sull'omonima isola posta a sud-ovest di New York ed è collegato fisicamente con il resto della città solo tramite il ponte di Verrazano, che permette di raggiungere Brooklyn via terra. I collegamenti con Manhattan sono invece garantiti grazie al traghetto Staten Island Ferry. Nonostante la notevole estensione, è il quartiere meno popoloso, meno densamente abitato e con il minor numero di census tracts di tutta New York. Di contro è il quartiere con più aree verdi e vanta numerose spiagge.
Staten Island è inoltre il quartiere con la più alta percentuale di popolazione bianca, di cui più di un terzo di origine italiana. Qui è infatti possibile trovare la casa dove abitò Antonio Meucci, che oggi è sede di un museo a lui dedicato.

Come per gli altri quartieri, si sono riportati gli errori di previsione ottenuti dalla stima dei modelli. Dalla tabella 3.14 si nota come in questo caso tutti i modelli abbiano dato origine ad errori confrontabili. Per quanto riguarda le differenze tra il modello base e quello con tutte le covariate, si osserva che per Staten Island è il modello additivo a mostrare il miglioramento maggiore grazie all'uso delle variabili "open", con una riduzione dell'errore di quasi 4 volte.

Tabella 3.14: Errori di previsione per Staten Island

	MSE modello base	MSE covariate	Rapporto
Modello additivo	1.830	0.483	3.793
Foresta casuale	0.509	0.477	1.068
Gradient Boosting	0.545	0.485	1.123

Per Staten Island i risultati degli effetti pametrici e non parametrici del modello additivo sono riportati rispettivamente nelle tabelle 3.15 e 3.16. Dalla prima tabella si nota che la maggior parte dei coefficienti non sono significativi. In particolare la temporalità non sembra avere particolare influenza sul fenomeno. Lo stesso era già stato osservato durante le analisi esplorative. Per quanto riguarda gli effetti non parametrici riportati nella seconda tabella, si

nota che la variabile relativa alla percentuale di residenti con reddito sotto la soglia di povertà è non significativa. Come per Manhattan, si osserva anche che alcune variabili presentano gradi di libertà equivalenti prossimi ad uno. Le variabili relative alla percentuale di residenti che si reca a lavoro con i mezzi pubblici, alla percentuale di residenti che invece lavora da casa e al numero di ATM mostreranno quindi un andamento lineare nei grafici marginali.

Tabella 3.15: Stima degli effetti parametrici per Staten Island

	Estimate	Std. Error	t value	Pr(>|t|)	
intercetta	2.865e+00	1.721e-01	16.647	< 2e-16	***
popolazione	3.776e-04	3.210e-05	11.763	< 2e-16	***
area	9.544e-08	6.772e-08	1.409	0.159312	
anno 2015	-5.828e-02	8.623e-02	-0.676	0.499454	
anno 2016	-1.028e-01	8.623e-02	-1.192	0.233756	
anno 2017	-1.655e-01	8.623e-02	-1.919	0.055451	.
anno 2018	-1.215e-01	8.623e-02	-1.409	0.159270	
anno 2019	-3.092e-01	8.623e-02	-3.585	0.000366	***

I grafici in figura 3.9 mostrano tutti un andamento crescente della risposta in funzione delle quattro variabili scelte. In particolare si può osservare la risposta in funzione della percentuale di ispanici, del tasso di disoccupazione, della percentuale di utilizzatori dei trasporti pubblici per recarsi a lavoro e del numero complessivo di fermate di metro e autobus. Tutti gli andamenti sono concordi con quanto già osservato per gli altri quartieri in riferimento a queste variabili.

Per quanto riguarda foresta casuale e Gradient Boosting in questo caso i modelli sono stati adattati rispettivamente con 2000 alberi formati da 7 variabili, e con 850 iterazioni consentendo le interazioni fino al settimo livello con shrinkage 0.01. La figura 3.10 riporta l'importanza delle variabili per i due modelli. Anche per Staten Island vale quanto osservato per tutti gli altri quartieri circa la concordanza dei modelli nell'attribuire importanza alle

Tabella 3.16: Stima degli effetti non parametrici per Staten Island

	edf	F	p-value	
s (latitudine, longitudine)	21.6562	4.970	< 2e-16	***
s (% maschi)	4.4551	3.756	0.00211	**
s (età)	4.1664	8.126	3.83e-06	***
s (% ispanici)	5.8717	17.771	< 2e-16	***
s (% popolazione bianca)	5.9180	5.944	5.52e-06	***
s (% afroamericani)	7.7998	20.107	< 2e-16	***
s (% residenti in affitto)	6.0886	7.743	< 2e-16	***
s (tasso di disoccupazione)	5.9069	3.537	0.00104	**
s (% residenti in età non lavorativa)	3.9578	6.660	6.18e-05	***
s (reddito)	8.0392	5.180	7.91e-07	***
s (% residenti in povertà)	3.0830	1.168	0.27301	
s (% che si reca a lavoro con mezzi pubblici)	0.9954	24.054	1.80e-06	***
s (% che si reca a lavoro a piedi)	3.0663	3.513	0.01303	*
s (% lavoratori da casa)	0.9996	4.187	0.04124	*
s (n° di fermate dei mezzi pubblici)	1.6574	29.120	< 2e-16	***
s (n° di ATM)	0.9994	25.634	8.34e-07	***

stesse variabili. In questo caso sono particolarmente rilevanti la percentuale di ispanici e la percentuale di abitanti che vive in uno stato di povertà.

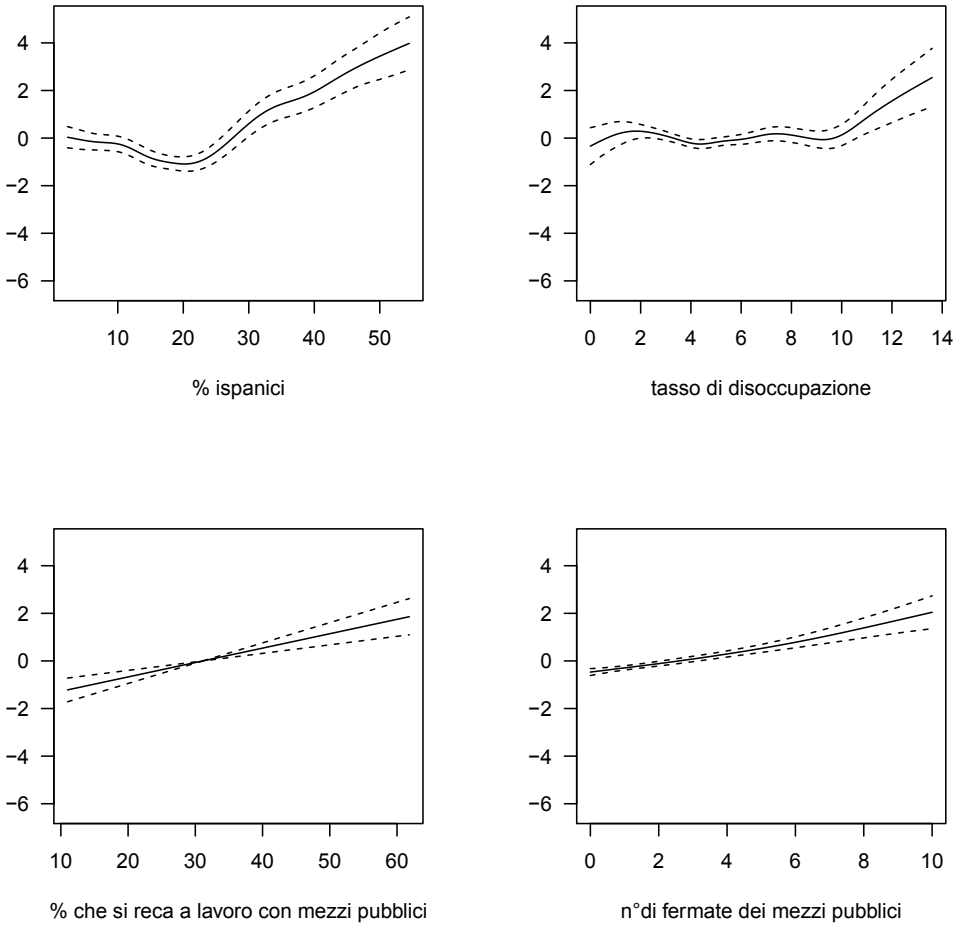


Figura 3.9: Grafici selezionati per Staten Island prodotti dal modello additivo

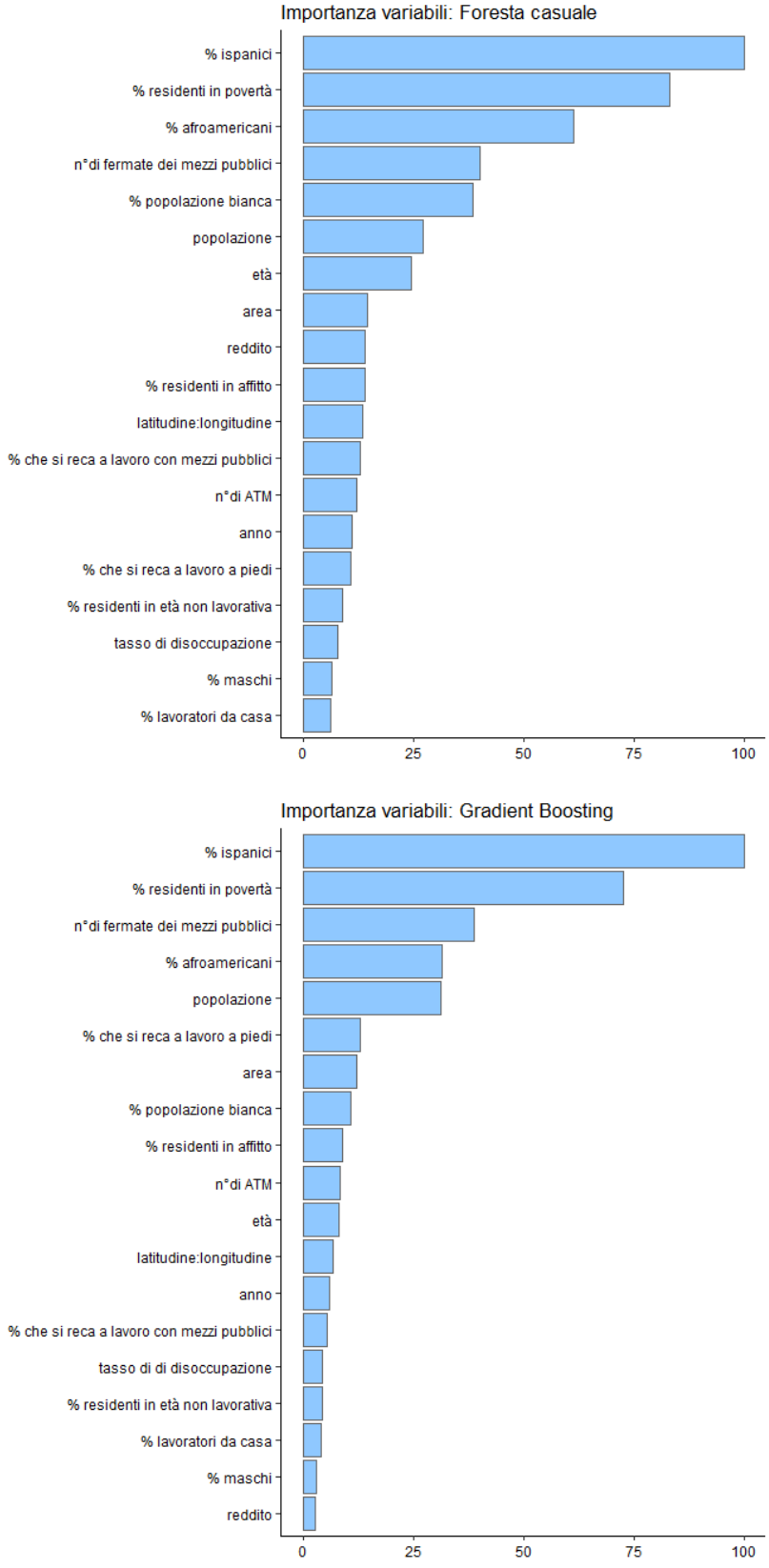


Figura 3.10: Importanza delle variabili per Staten Island

3.4 Discussione dei risultati

3.4.1 Risultati di carattere generale

I modelli adattati hanno permesso di far emergere il ruolo significativo delle variabili "open" nello spiegare il fenomeno. In particolare, tra le variabili di maggior impatto vi sono quelle relative alla situazione economica dei residenti. Infatti tali indicatori (*reddito* e *% residenti in povertà*) compaiono tra le prime cinque variabili per importanza nella maggior parte dei grafici generati da foresta casuale e Gradient Boosting. Inoltre, in generale, è possibile osservare (figure da B.1 a B.8) un incremento nel numero di aggressioni al peggiorare delle condizioni economiche, indipendentemente dal quartiere. Tale risultato è in accordo con quanto discusso nella sezione 2.1 circa la relazione tra criminalità e benessere economico.
L'effetto del tasso di disoccupazione sulla risposta appare invece meno evidente e significativo. In generale dai grafici marginali dei modelli additivi si osserva un andamento crescente del fenomeno al crescere della disoccupazione, tuttavia per molti distretti tale variabile occupa le ultime posizioni per importanza nei grafici generati da foresta casuale e Gradient Boosting. Al contrario, nelle prime posizioni per importanza, si osservano in tutti i distretti le variabili relative ad etnia (*% popolazione bianca* e *% afroamericani*) e origini (*% ispanici*). L'effetto di tali indicatori non è però lo stesso per tutti i quartieri. Ad esempio nel caso di Brooklyn si osserva come l'effetto di questo variabili sia pressoché nullo (figura 3.3), mentre nel Bronx risulta che le aree mono-etniche sono maggiormente interessate dal fenomeno (figura 3.5). Effetti discordanti tra i distretti si osservano anche in relazione alle variabili che indicano l'età media dei residenti e la percentuale di popolazione in età non lavorativa. Si nota infatti che tali variabili assumono importanza diversa a seconda del distretto considerato, apparendo molto importanti per alcuni quartieri e viceversa poco significative per altri. Infine, per quanto riguarda le variabili meno rilevanti, si osserva che la percentuale di individui di sesso maschile, la percentuale di lavoratori da casa e il numero di ATM (ad eccezione di Manhattan) occupano spesso le ultime posizioni per importanza e non sembrano avere un effetto ben definito sul fenomeno.

I risultati discussi in questa sezione sono in accordo con quanto espresso dalla teoria economica circa le motivazioni che inducono gli individui a commettere reati. Tuttavia, in questo elaborato, si propone un possibile approccio differente da quello classico per l'analisi di tali tematiche. In particolare, a differenza di quanto si riscontra negli studi di Brush (2007), Choe (2008), Kelly (2000) e Sachsida et al. (2010), si è scelto di non aggregare le variabili a disposizione per creare indici di disuguaglianza economica, ma di considerare i fattori sociali, economici e demografici nella loro interezza, senza sottoporli ad ulteriori trasformazioni. Ciò permette di limitare l'introduzione di elementi di soggettività, legati alla scelta e costruzione di indici, dando luogo a risultati maggiormente oggettivi basati sui dati stessi. Per gli stessi motivi, si è anche deciso di non avvalersi di variabili strumentali, come si riscontra invece in Lin (2008) e in Raphael e Winter-Ebmer (2001).
Inoltre, discostandosi nuovamente dall'approccio classico, in questo elaborato si è scelto di utilizzare tecniche non parametriche di analisi dei dati. Tale approccio permette anche di affrontare il caso di studio con maggiore oggettività. La possibilità di non vincolarsi ad una struttura parametrica, che richiede assunzioni spesso forti e non sempre verificate sul fenomeno generatore dei dati, consente infatti ai dati stessi di cogliere la complessità del fenomeno in esame, senza che sia necessario introdurre ulteriori vincoli stabiliti a priori.

3.4.2 Adattamento ai dati osservati

A completamento delle analisi svolte si è ritenuto interessante riportare nella scala originaria le previsioni ottenute dai vari modelli, al fine di confrontarle con i valori effettivamente osservati. Per poter offrire una visione d'insieme si è inoltre scelto di rappresentare i risultati sotto forma di mappe, considerando l'intera città di New York e non i soli quartieri singolarmente. Infatti, nonostante i confronti tra modelli possano essere fatti esclusivamente in relazione allo stesso quartiere, si è ritenuto di maggior interesse mostrare il fenomeno nella sua interezza.
La figura 3.11 riporta le mappe relative alle aggressioni osservate e a quelle previste da modello additivo, foresta casuale e Gradient Boosting. Per poter rendere le mappe confrontabili si è scelto di usare la stessa divisione di

categorie adottata per i dati osservati. Inoltre, dal momento che il range di valori all'interno di ciascuna categoria è ampio, si è ritenuto utile riportare l'istogramma relativo al numero di aggressioni. Osservando le mappe si nota infatti non solo che la quasi totalità dei census tracts è mappata nella categoria corretta, ma anche che tutti gli istogrammi sono confrontabili. Ciò permette di affermare che tutti i modelli stimati sono in grado di cogliere in modo appropriato le specificità del fenomeno, producendo un adattamento adeguato. La validità della modellazione non parametrica emerge anche dal confronto con i risultati ottenibili del modello lineare. A completamento delle analisi si è infatti scelto di adattare anche tale modello[3], riportando il confronto tra valori osservati e predetti in figura 3.12. Osservando le due mappe è possibile notare che, rispetto ai casi precedenti, un numero inferiore di census tracts è mappato nella categoria corretta. Inoltre gli istogrammi mostrano distribuzioni non sovrapponibili, rivelando come il modello lineare sottostimi sistematicamente i valori nella coda superiore della distribuzione.
Per quantificare tali scostamenti si sono calcolati gli errori di previsione, utilizzando la stessa metodologia adottata per i modelli non parametrici. La tabella 3.17 riporta tali valori, suddivisi nei cinque distretti. È immediato notare che gli errori ottenuti dalla stima del modello lineare sono considerevolmente superiori a quelli generati da qualunque altro modello adattato (tabelle 3.2, 3.5, 3.8, 3.11, 3.14). Anche in questo caso è però possibile constatare il notevole contributo informativo apportato dagli Open Data. Infatti il rapporto tra gli errori dei modelli senza variabili "open" e quelli dei modelli che le sfruttano è generalmente più che doppio. In conclusione si può affermare che i modelli non parametrici adattati sono in grado di cogliere in modo appropriato le specificità del fenomeno. Al contrario, il modello lineare, a causa della sua rigida struttura, non produce risultati del tutto soddisfacenti.

[3] I coefficienti stimati sono stati riportati nella tabella B.1 in appendice B.

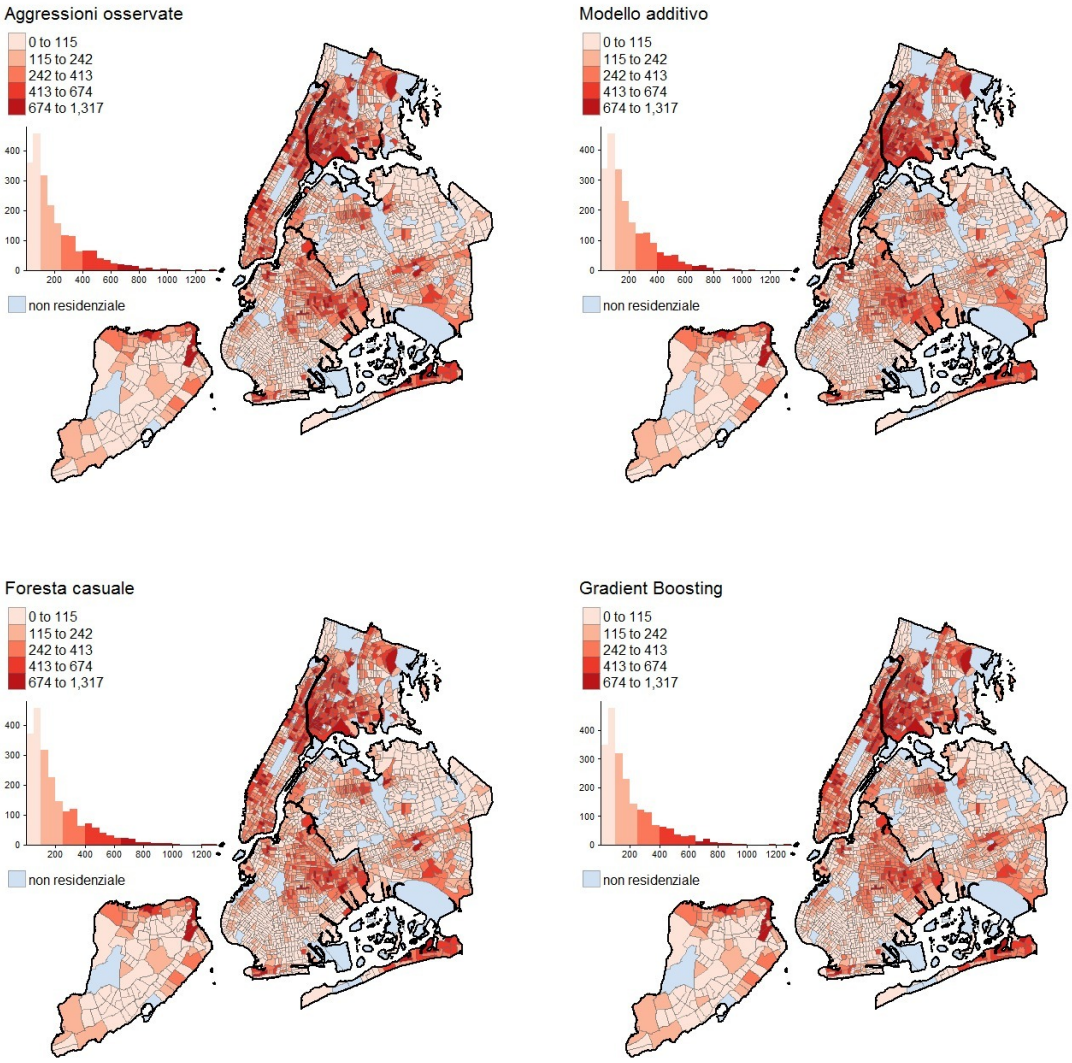


Figura 3.11: Adattamento ai dati osservati: modelli non parametrici

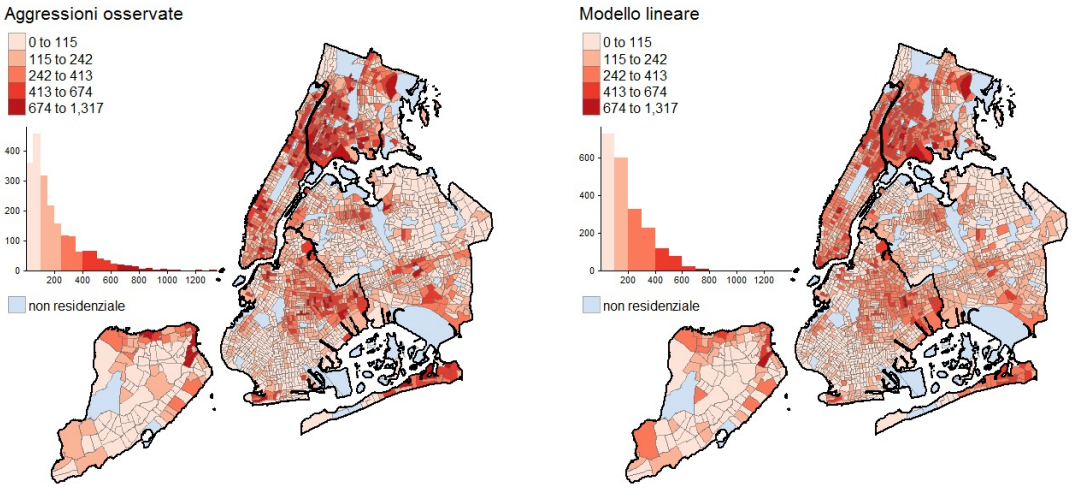


Figura 3.12: Adattamento ai dati osservati: modello di regressione lineare

Tabella 3.17: Errori di previsione per il modello lineare

	MSE modello base	MSE covariate	Rapporto
Manhattan	6.712	3.073	2.184
Brooklyn	4.466	1.646	2.714
Bronx	5.581	2.601	2.146
Queens	3.027	1.592	1.902
Staten Island	4.141	1.338	3.096

Capitolo 4

Conclusioni

L'elaborato si è posto l'obiettivo di mostrare come l'utilizzo di Open Data possa essere d'aiuto nella comprensione di fenomeni di interesse pubblico. A tal proposito, dopo un'accurata ricerca circa la disponibilità di portali di dati aperti e di dati stessi, si è scelto di focalizzarsi sullo studio delle aggressioni nella metropoli di New York. Si è infatti notato che, nonostante gli sforzi degli ultimi anni, in Italia la cultura degli Open Data non si è ancora pienamente affermata. I dati presenti sui portali italiani presentano spesso un grado di aggregazione troppo elevato che li rende non integrabili. Nel capitolo 1 si è infatti sottolineato come l'interoperabilità sia una caratteristica chiave e imprescindibile che gli Open Data devono possedere affinché possano davvero essere riutilizzati.
La costruzione del dataset è stata una fase particolarmente onerosa dell'analisi svolta. È stato infatti necessario aggregare le informazioni provenienti da 16 diversi dataset, disponibili in formati differenti e con diversa granularità. Tra le difficoltà riscontrate in fase di integrazione, vi è stata la mancanza di normalizzazione a livello di codifiche territoriali. Infatti enti diversi usano identificativi diversi per fare riferimento allo stesso oggetto: ad esempio i nomi dei quartieri di New York o gli identificativi dei census tracts. Ciò ha richiesto un ulteriore sforzo nel costruire una chiave che consentisse di identificare univocamente le voci di interesse nei vari dataset.
Ai fini dell'analisi, si è scelto di considerare solo le aree residenziali di New York, escludendo ad esempio parchi, cimiteri ed aeroporti. Questa apparente

semplificazione è giustificata dal fatto che, per il periodo considerato, solo l'1,56% delle aggressioni è avvenuto in queste zone. Inoltre dalle analisi esplorative è emerso che oltre il 50% delle aggressioni si è verificato tra le mura domestiche e quasi il 30% in strada, mentre nei parchi la percentuale è solo dell'1,30%. Si è quindi ritenuto che trascurare le aree non residenziali non costituisse un'eccessiva semplificazione del fenomeno. Inoltre, ad oggi, non è possibile reperire online informazioni socio-demografiche ed economiche su queste aree, in quanto non abitate. Qualora fossero disponibili Open Data relativi ad esempio all'affluenza di persone in queste zone e ad alcune loro caratteristiche, sarebbe possibile costruire indicatori che permettano di includere anche le aree non residenziali nelle analisi.
Dalle analisi esplorative è emersa una significativa eterogeneità tra i quartieri di New York. Per tenere conto di questa realtà diversificata si è ritenuto opportuno, in fase di stima, adattare i modelli separatamente per ciascun quartiere. Inoltre, dal momento che l'obiettivo della tesi è stato utilizzare dati aperti per comprendere meglio il fenomeno di interesse, si è scelto di privilegiare modelli che permettessero una più immediata interpretazione dei risultati. Dalla stima dei modelli additivi è stato infatti possibile valutare l'impatto di ciascuna variabile sulla risposta, mentre dalla stima di foreste casuali e Gradient Boosting se ne è valutata la diversa importanza. In generale si è potuto osservare che le variabili di maggior impatto per il fenomeno sono quelle legate al benessere economico dei residenti. In particolare si è rilevato come all'aggravarsi delle condizioni economiche si abbia un incremento nel numero di aggressioni. Tali osservazioni risultano in accordo con alcune teorie presenti in letteratura (Coccia 2018; Fleisher 1966; Kelly 2000; Lin 2008; Sachsida et al. 2010; Savage, Ellis e Wozniak 2019).
Infine il confronto tra valori predetti e osservati ha evidenziato come i modelli scelti siano in grado di catturare le specificità del fenomeno. Al contrario l'adattamento del modello lineare ha portato a risultati non soddisfacenti. Considerando gli ottimi risultati ottenuti con i modelli non parametrici, non si è quindi ritenuto necessario trattare il fenomeno tramite modelli spaziali. Un approccio spaziale potrebbe però essere una naturale estensione di questa analisi per ulteriori sviluppi. Ad esempio si potrebbe seguire l'approccio descritto in Morris et al. (2019), secondo cui tramite il modello spaziale BYM

(Besag, York e Mollié) si studia la relazione tra fattori sociali, demografici ed economici e il numero di incidenti pedonali nei census tracts di New York.

Come discusso nel capitolo 1, gli Open Data costituiscono una grande risorsa dall'alto potenziale e possono portare ad innumerevoli benefici sia in termini economici sia in termini di benessere sociale. In questo elaborato si è mostrato come l'utilizzo di variabili "open" sia stato fondamentale per i risultati delle analisi. Infatti non era scontato che le variabili scelte fossero di impatto nello studio del fenomeno. I risultati in termini di significatività degli effetti del modello additivo, i grafici relativi all'importanza delle variabili di foreste casuali e Gradient Boosting e il rapporto tra gli errori di previsione forniti dai modelli che utilizzano solo variabili territoriali e da quelli che utilizzano tutte le variabili "open" hanno invece mostrato il notevole impatto degli Open Data nella comprensione del fenomeno. Tali risultati possono anche avere un interesse di tipo pubblico. Infatti proprio una comprensione approfondita del fenomeno può consentire di ricavare utili informazioni per migliorare la qualità della vita dei cittadini in termini di sicurezza personale. Ad esempio alcune iniziative perseguibili potrebbero essere finalizzate alla riqualifica delle aree più povere. Generare nuovi posti di lavoro in queste aree e adeguare i salari ad esempio, potrebbe portare ad un significativo calo nel numero delle aggressioni.

Appendice A

Codice R utilizzato

Codice A.1: Codice per unire due datasets con differente codifica dei census tracts

```
tracts<-readOGR(dsn = "2010 Census Tracts", layer = "geo_export_e62d2353
    -08df-4ac4-a56d-966803b69ef6")
t<-tracts
t<-t[,c(1,6,7,8,2,10)]
t$b<-NA
t$b[t$boro_code %in% c(1)]= 'New York, NY'
t$b[t$boro_code %in% c(2)]= 'Bronx, NY'
t$b[t$boro_code %in% c(3)]= 'Kings, NY'
t$b[t$boro_code %in% c(4)]= 'Queens, NY'
t$b[t$boro_code %in% c(5)]= 'Richmond, NY'

t$b2<-paste("Census Tract" ,t$ctlabel)
t$id<-paste(t$b2 ,t$b, sep=", ")
View(t@data)

t<-t[, c("ntaname","id", "shape_area")]
p2<-as.data.frame(p)

dataset_completo<-merge(t, p2, by.x="id", by.y="name", all.x=T)
```

Codice A.2: Codice per raggruppare il numero di fermate di metro e bus nei census tracts

```
metro<- readOGR("Subway Entrances", "geo_export_39381c4f-5e62-4f2e-b006-
    a95451689003")

metro <- spTransform(metro, CRS(proj4string(p)))
pip <- over(metro, p)

#aggiungo l'informazione sui census tract
metro@data <- cbind(metro@data, pip)

####conto quante fermate ci sono per ciascun census tract####
metro_2 <- aggregate(formula=objectid~geoid, data=metro@data, FUN=length)
names(metro_2) <- c("geoid","n_entrate_metro")
m <- match(x=p@data$geoid, table=metro_2$geoid)
p@data$n_entrate_metro <- metro_2$n_entrate_metro[m]
p$n_entrate_metro[which(is.na(p$n_entrate_metro))]<-0

################################################################

bus<- readOGR("Bus Stop Shelters", "geo_export_5a9c5d64-c635-4c07-9de4-
    eb31053e0032")

bus <- spTransform(bus, CRS(proj4string(p)))
pip <- over(bus, p)

#aggiungo l'informazione sui census tract
bus@data <- cbind(bus@data, pip)

####conto quante fermate ci sono per ciascun census tract####
bus_2 <- aggregate(formula=shelter_id~geoid, data=bus@data, FUN=length)
names(bus_2) <- c("geoid","n_fermate_bus")
m <- match(x=p@data$geoid, table=bus_2$geoid)
p@data$n_fermate_bus <- bus_2$n_fermate_bus[m]
p$n_fermate_bus[which(is.na(p$n_fermate_bus))]<-0
```

Codice A.3: Codice per raggruppare gli ATM nei census tracts

```
library(readr)
library(magrittr)
library(stringr)
library(dplyr)

Bank_Owned_ATM <- read_delim("Bank_Owned_ATM_Locations_in_New_York_State
    (1).csv", ";", escape_double = FALSE, trim_ws = TRUE)

colnames(Bank_Owned_ATM) %<>% str_replace_all("\\s", "_") %>% tolower()

Bank_Owned_ATM$location_1 %<>% str_replace_all("[(]", "__") %>% tolower()
Bank_Owned_ATM$location_1 %<>% str_replace_all("[)]", "__") %>% tolower()

res <- str_match(Bank_Owned_ATM$location_1, "__\\s*(.*?)\\s*__")
Bank_Owned_ATM$cord<-res[,2]

Bank_Owned_ATM$lat<-sub("[,].*", "", Bank_Owned_ATM$cord)
Bank_Owned_ATM$long<-sub(".*[,]", "", Bank_Owned_ATM$cord)

Bank_Owned_ATM<-Bank_Owned_ATM[-which(is.na(Bank_Owned_ATM$long)),]

lats <- Bank_Owned_ATM$lat
lngs <- Bank_Owned_ATM$long

##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
proj4string(points_spdf) <- proj4string(p)

matches <- over(points_spdf, p)
points <- cbind(points, matches)
Bank_Owned_ATM<-cbind(Bank_Owned_ATM, geoid=points[,c(3)],
```

```
                name=points[,c(4)])
Bank_Owned_ATM<-Bank_Owned_ATM[-which(is.na(Bank_Owned_ATM$geoid)),]

d3<-Bank_Owned_ATM %>%
  group_by(geoid, name) %>%
  summarise(count=n())

Bank_Owned_ATM$id<-seq(1:NROW(Bank_Owned_ATM))

####conto quante banche ci sono per ciascun census tract####
count_c <- aggregate(formula=id~geoid, data=Bank_Owned_ATM, FUN=length)
names(count_c) <- c("geoid","n_ATM")
m_c <- match(x=p@data$geoid, table=count_c$geoid)
p$n_ATM <- count_c$n_ATM[m_c]
p$n_ATM[which(is.na(p$n_ATM))]<-0
```

Codice A.4: Codice per ricavare le caratteristiche geografiche dei census tracts

```
library(stringr)
library(dplyr)

c<-coordinates(dataset_completo)
dataset_completo<-cbind(dataset_completo, c)

dataset_completo@data<-dataset_completo@data %>%
  rename(
    long = X1,
    lat = X2
  )

c<-dataset_completo
pjr.ae<-CRS("+proj=aea +lat_0=37.5 +lon_0=-96 +lat_1=29.5 +lat_2=45.5 +x_
    0=0 +y_0=0 +datum=NAD83 +units=m +no_defs") #imposto nuovo riferimento
c<-spTransform(c, pjr.ae)
proj4string(c)
```

```
library(rgeos)
c<-gArea(c, byid=T)

dataset_completo<-cbind(dataset_completo,c)

dataset_completo@data<-dataset_completo@data %>%
  rename(
    area = c
        ..0....231980.046401986...1....177076.749626163...2....172888.840666001..

  )
```

Appendice B

Grafici e tabelle supplementari

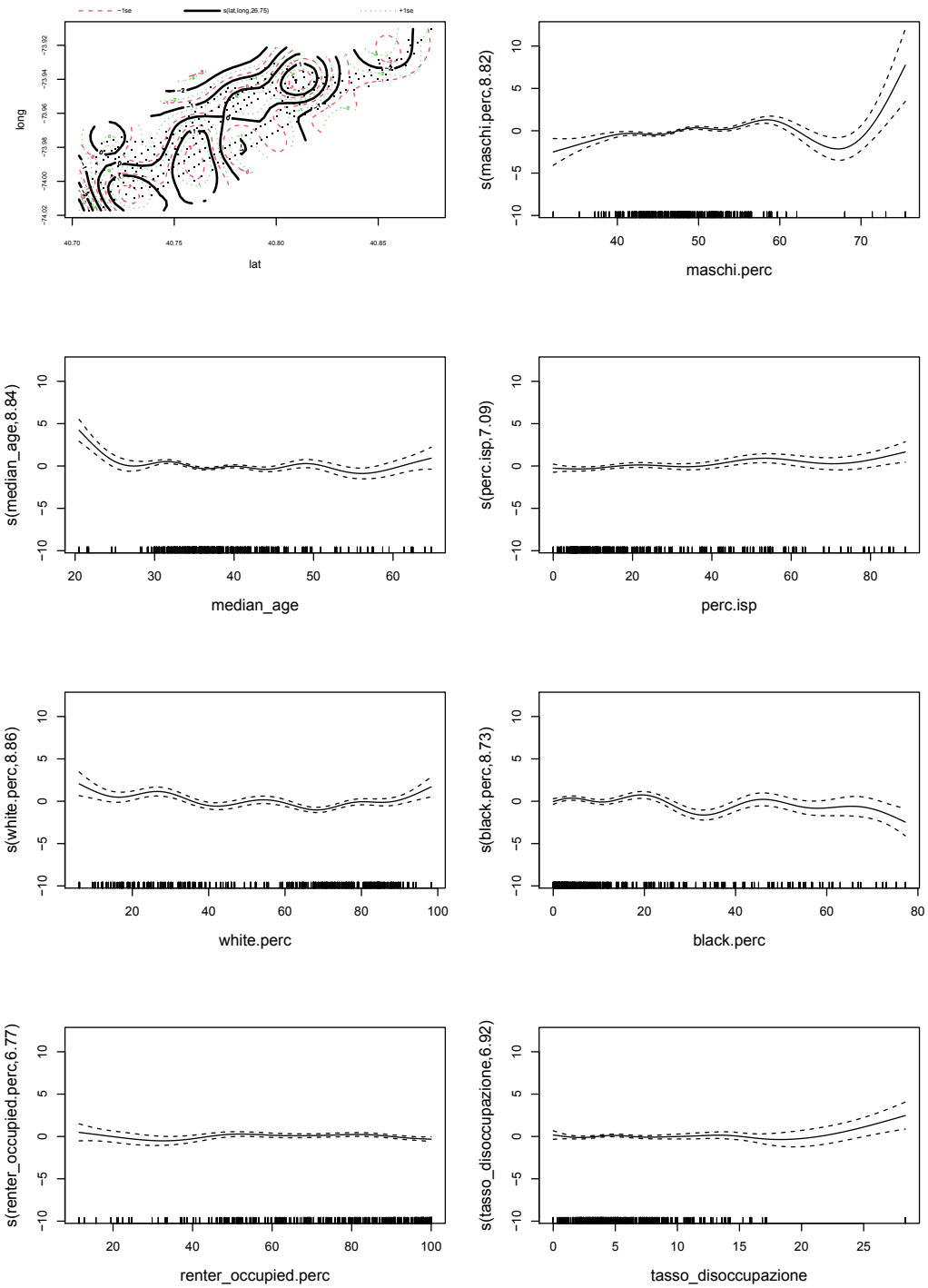


Figura B.1: Modello additivo Manhattan-parte 1

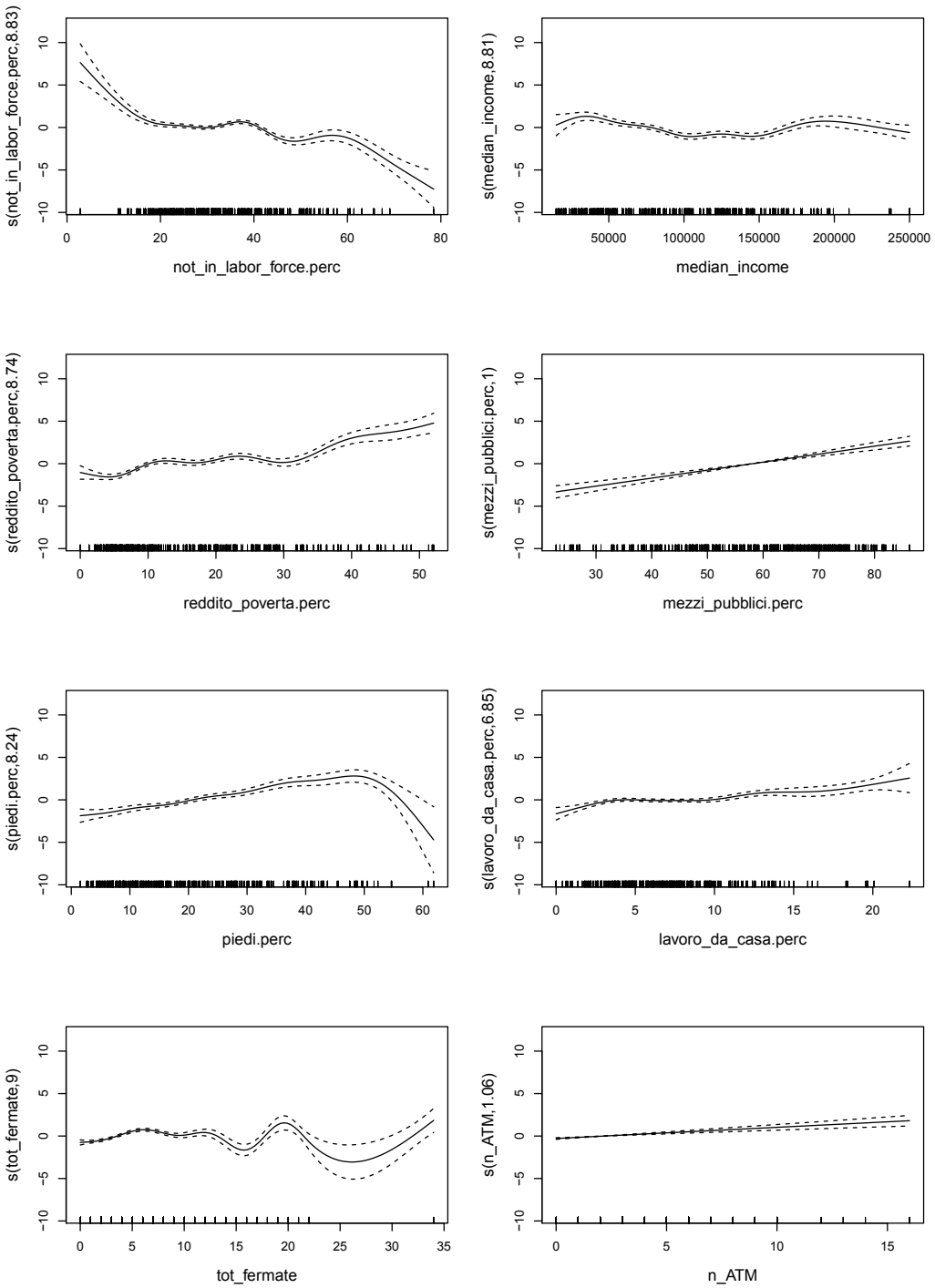


Figura B.2: Modello additivo Manhattan-parte 2

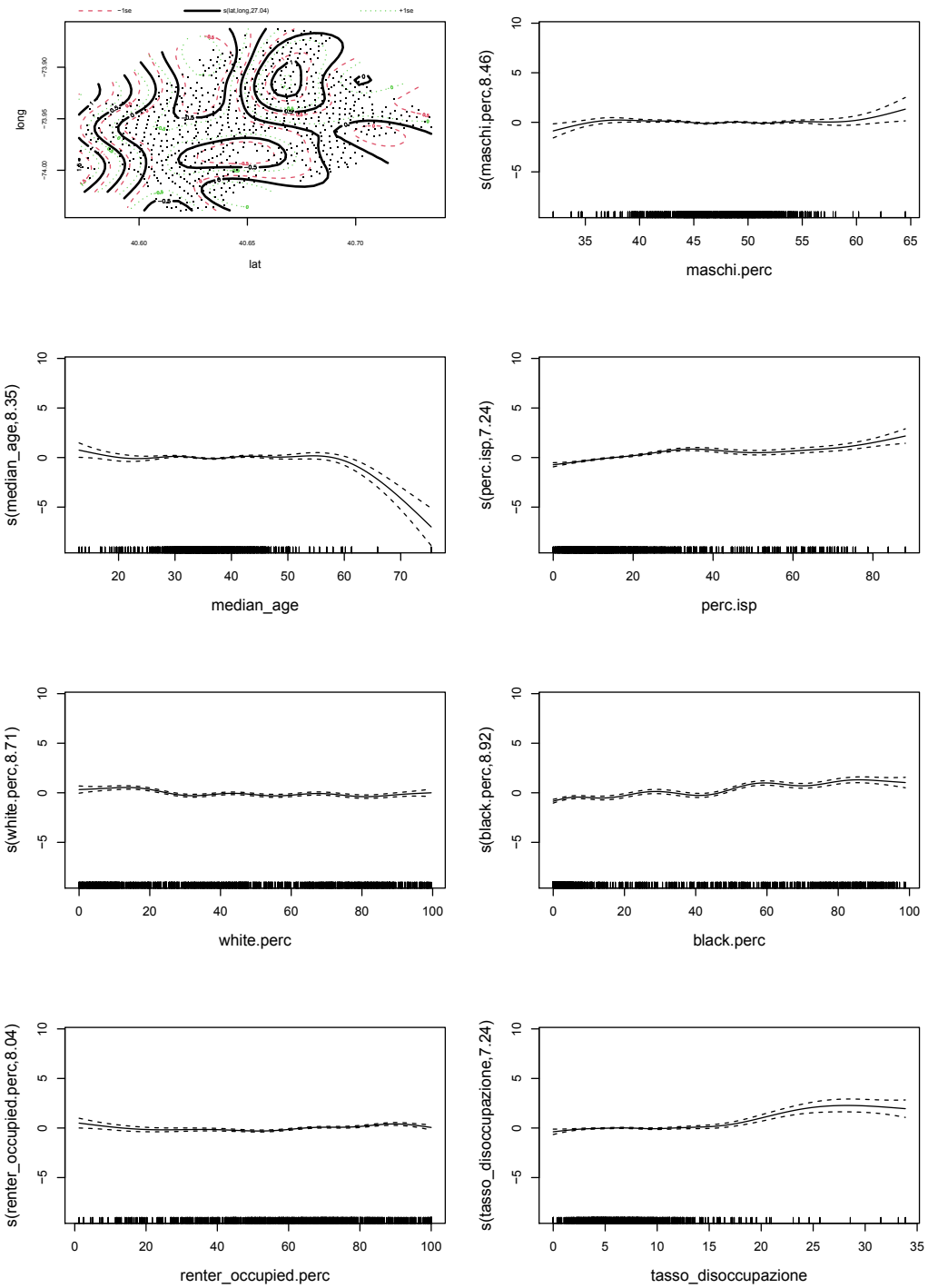


Figura B.3: Modello additivo Brooklyn-parte 1

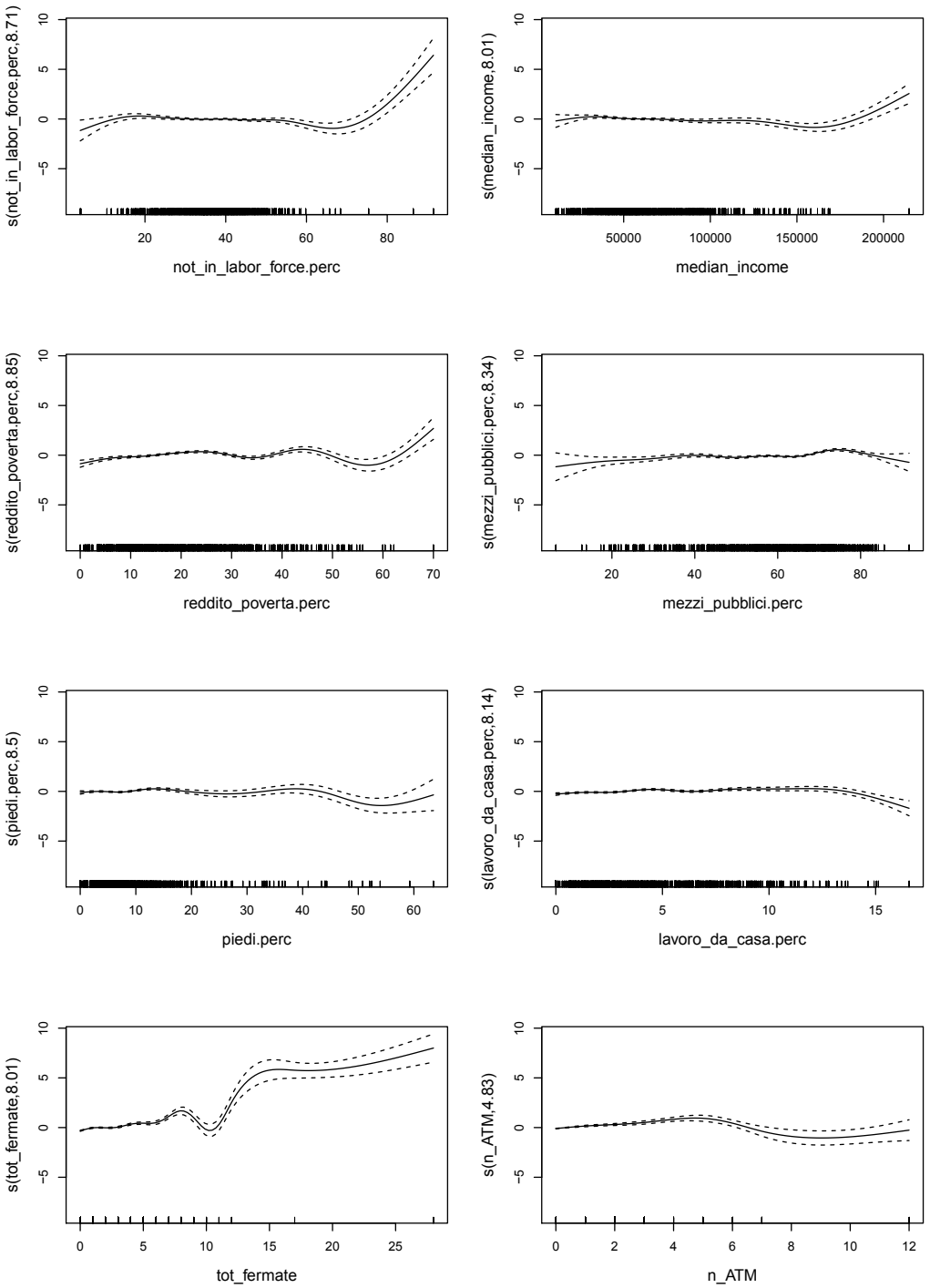


Figura B.4: Modello additivo Brooklyn-parte 2

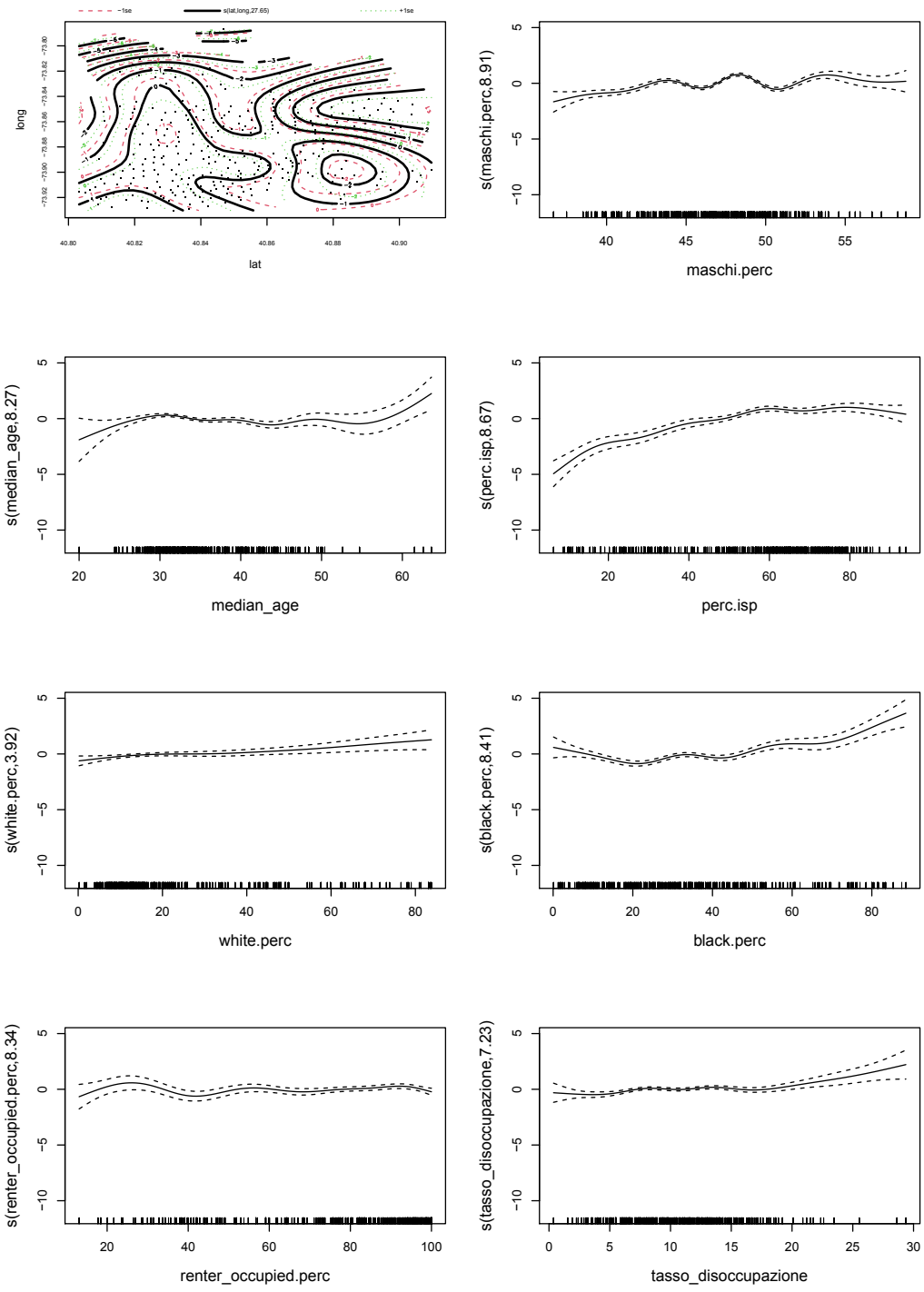


Figura B.5: Modello additivo Bronx-parte 1

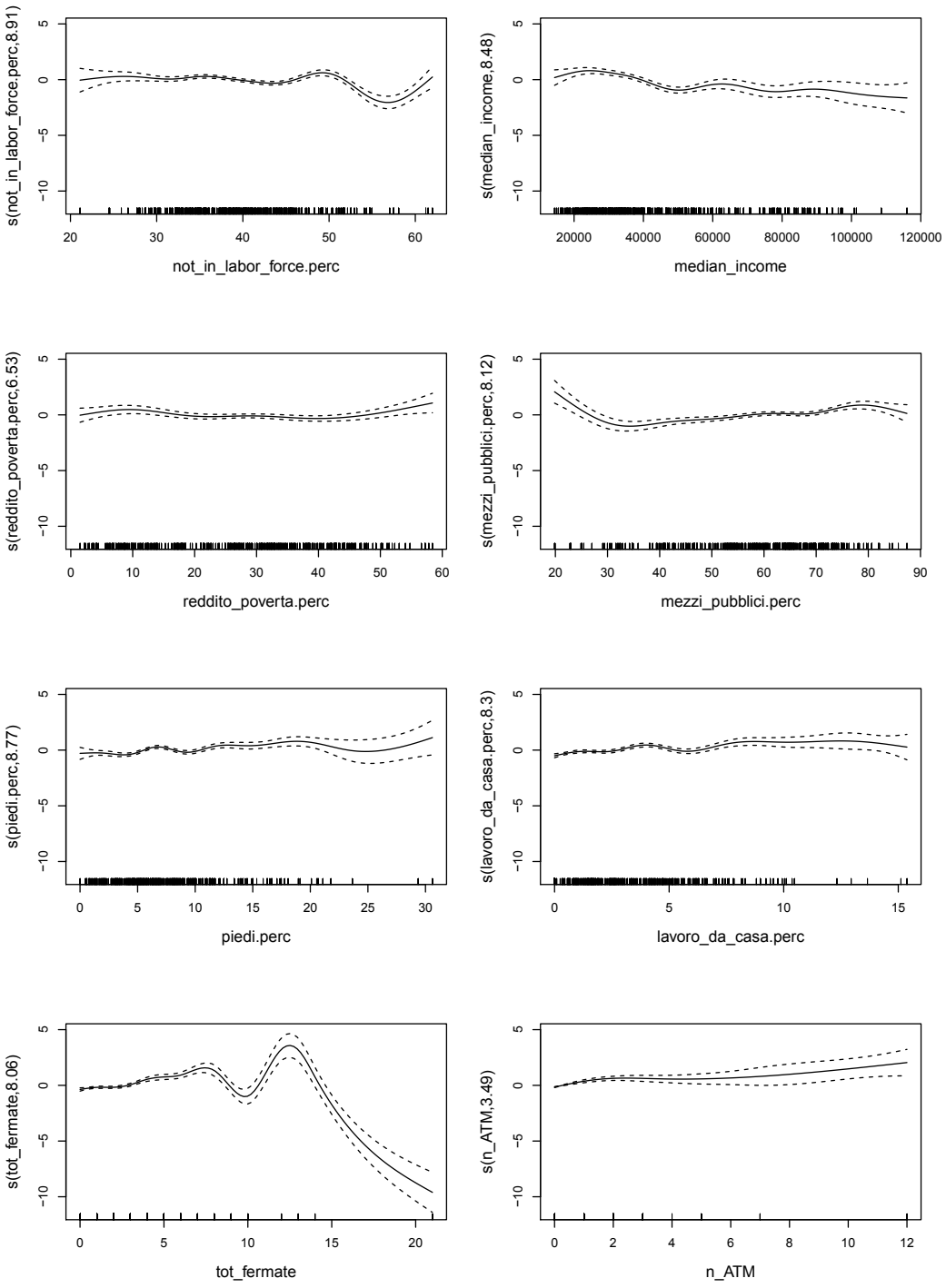


Figura B.6: Modello additivo Bronx-parte 2

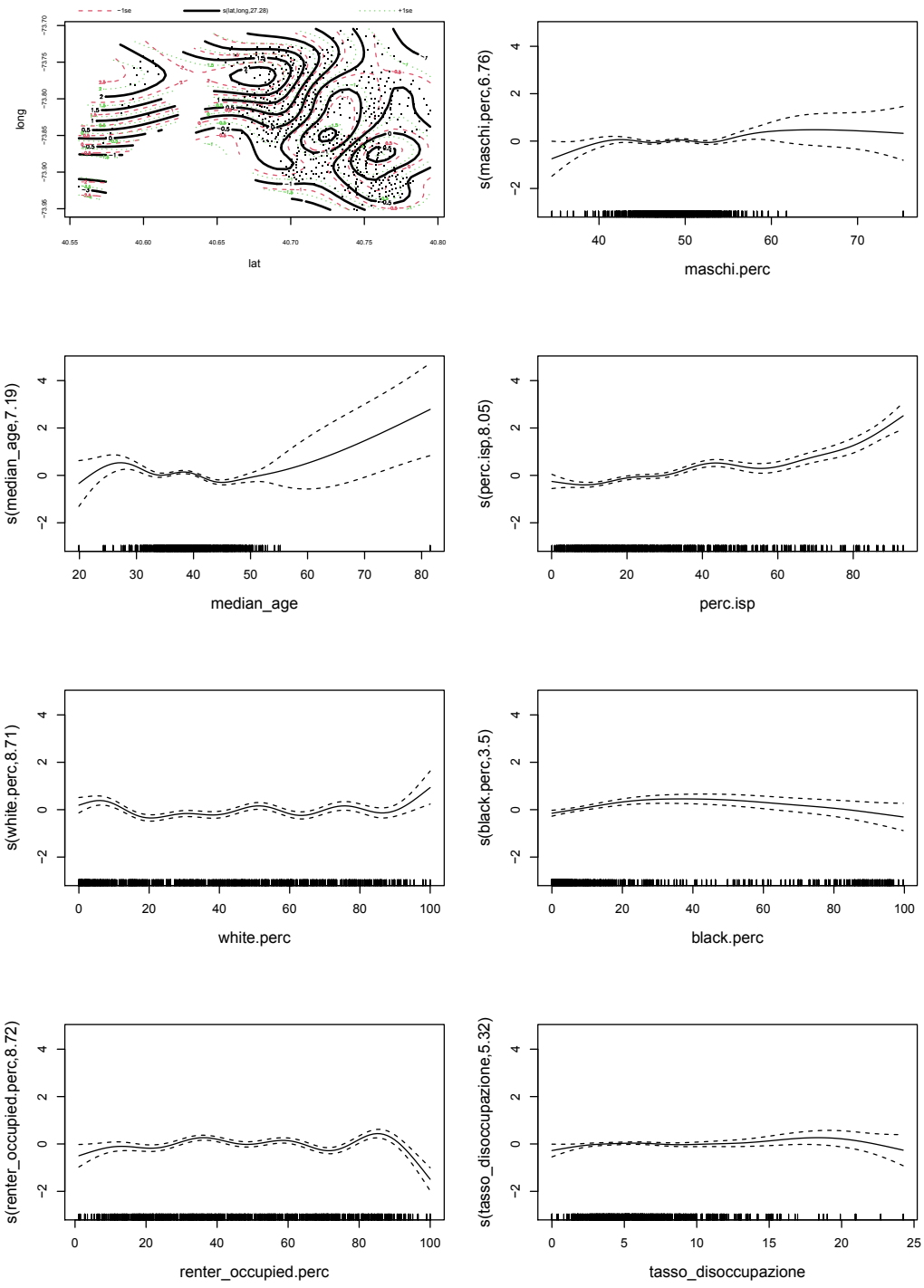


Figura B.7: Modello additivo Queens-parte 1

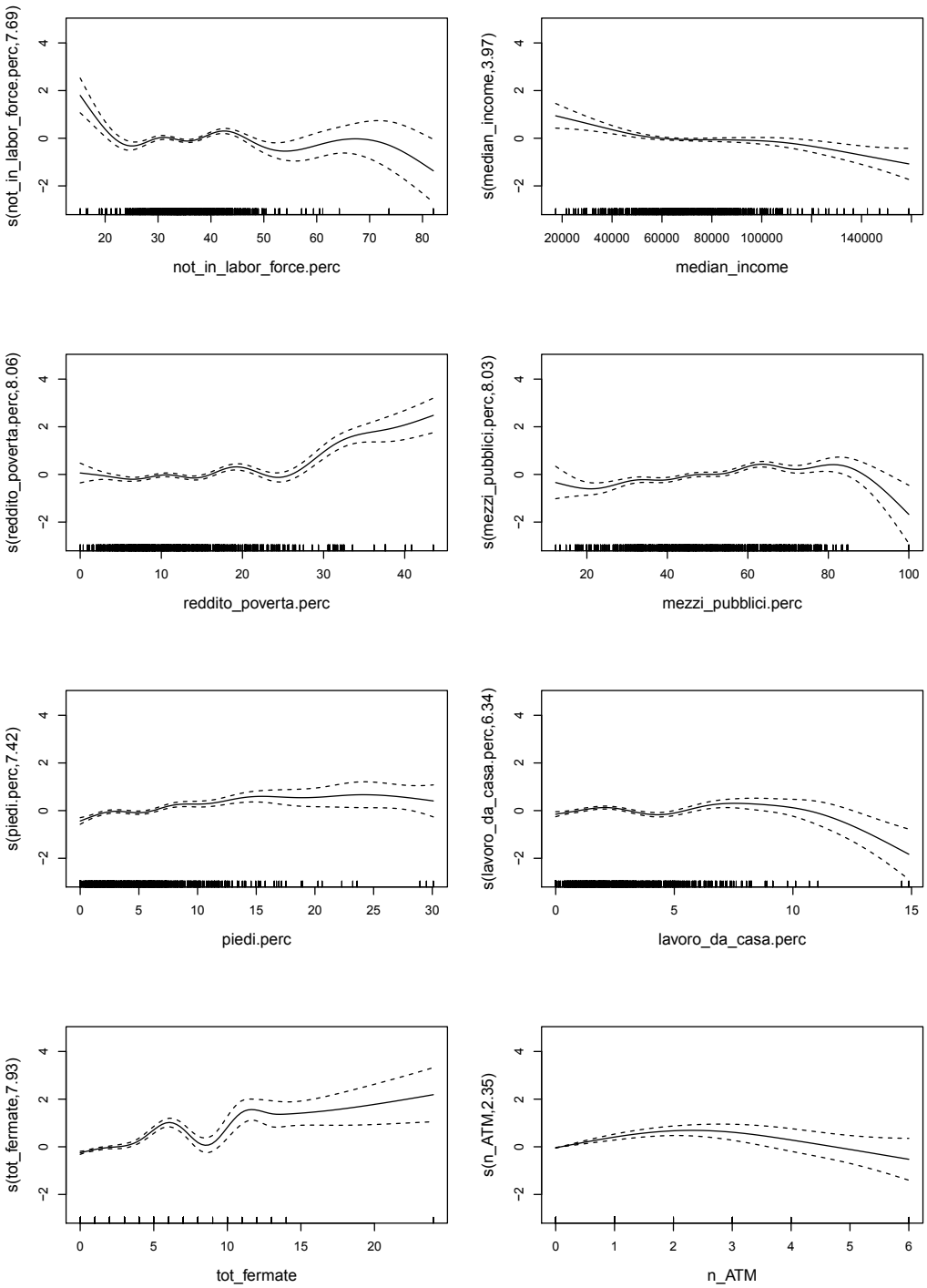


Figura B.8: Modello additivo Queens-parte 2

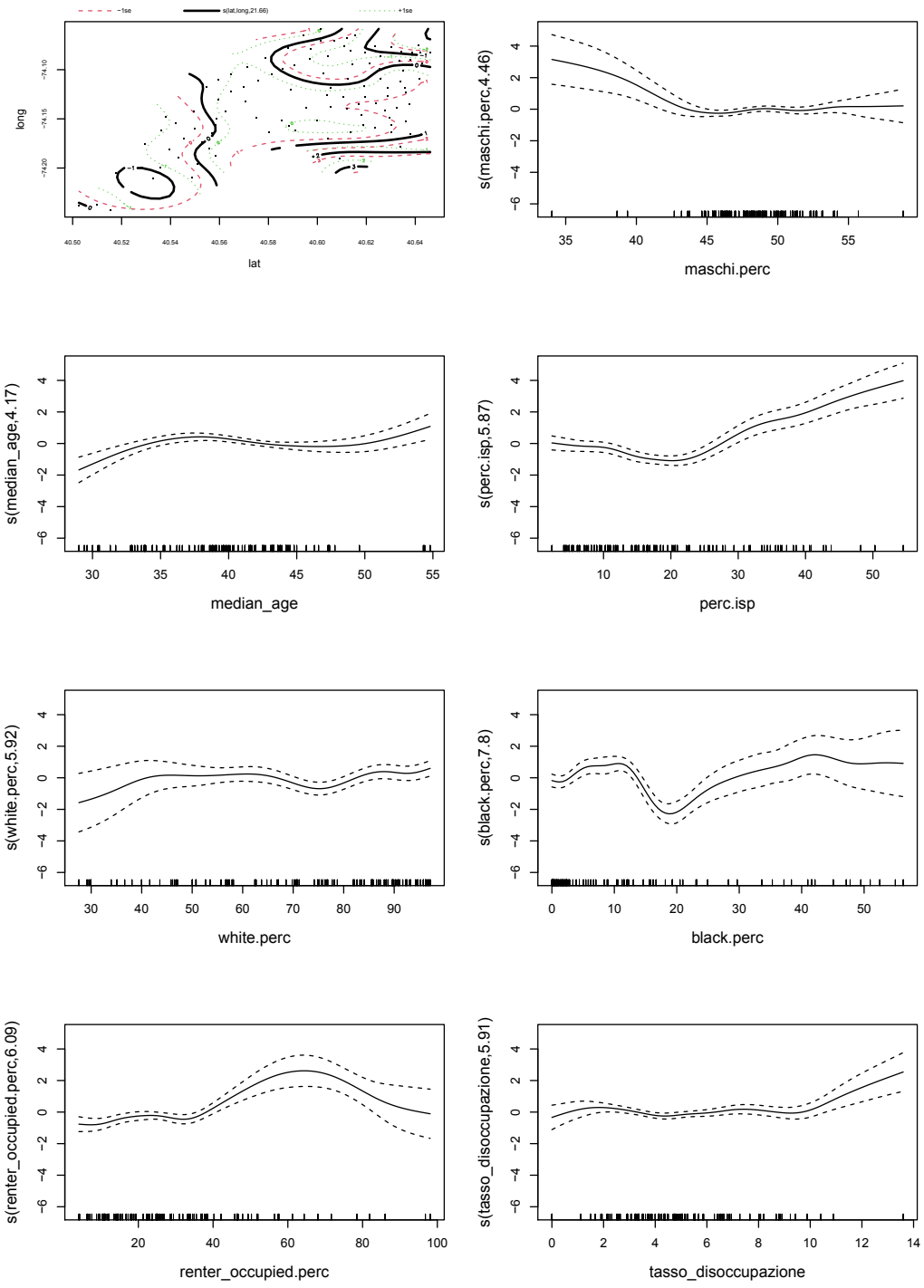


Figura B.9: Modello additivo Staten Island-parte 1

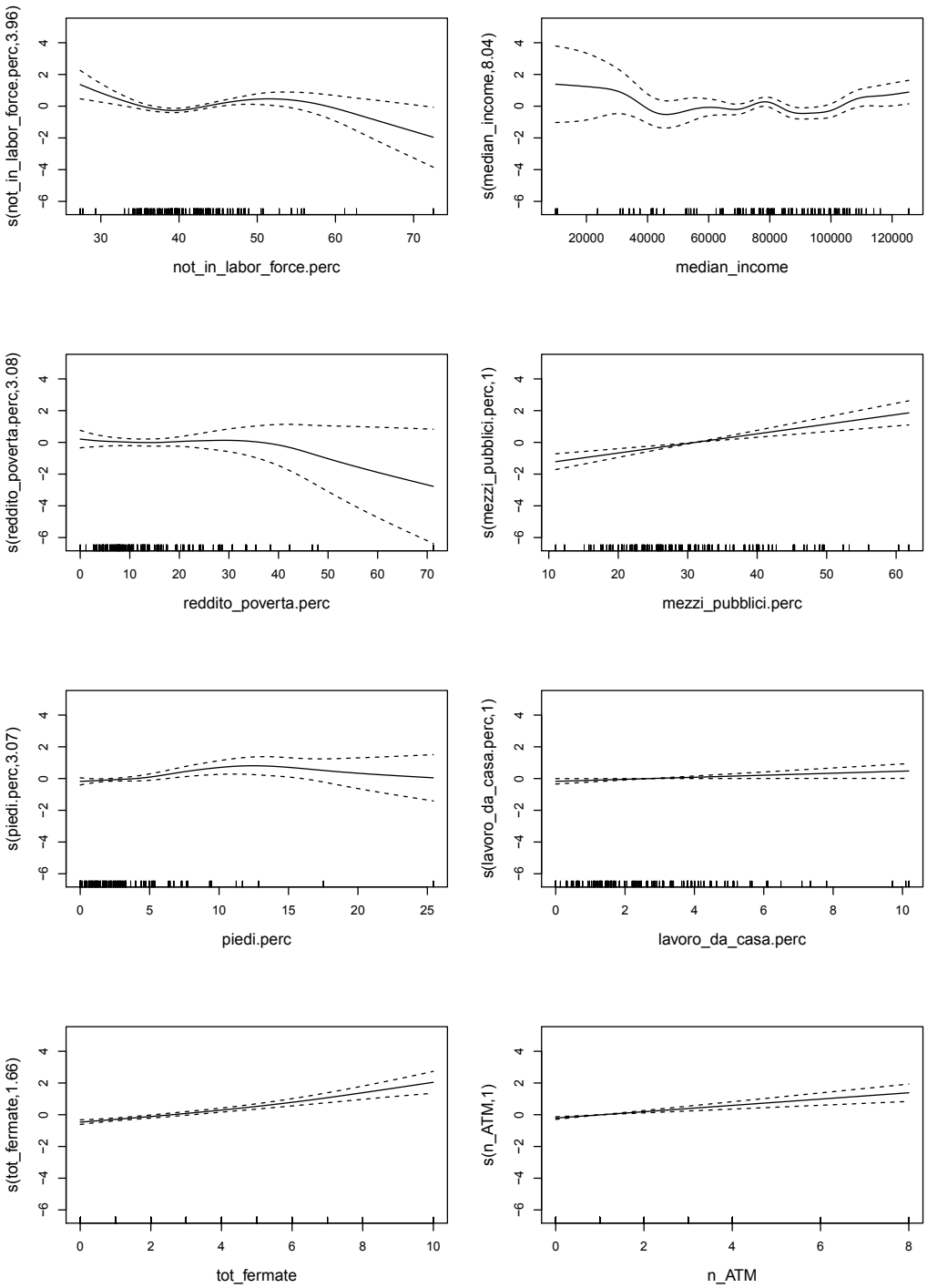


Figura B.10: Modello additivo Staten Island-parte 2

Tabella B.1: Stime del modello lineare per ciascun quartiere

	Manhattan		Brooklyn		Bronx		Queens		Staten Island	
variabile	stima	p-valhe	stima	p-valhe	stima	p-valhe	stima	p-valhe	stima	p-valhe
intercetta	7.213e+02	< 2e-16 ***	-7.511e+01	0.032125 *	2.432e+02	0.000494 ***	1.910e+02	< 2e-16 ***	-1.880e+01	0.868837
popolazione	1.854e-04	< 2e-16 ***	3.916e-04	< 2e-16 ***	4.233e-04	< 2e-16 ***	3.300e-04	< 2e-16 ***	3.684e-04	< 2e-16 ***
area	3.490e-06	4.18e-11 ***	9.234e-07	< 2e-16 ***	3.787e-07	0.021720 *	4.875e-07	1.51e-11 ***	1.255e-07	0.000925 ***
anno 2015	1.297e-01	0.377593	-7.078e-02	0.284207	-3.317e-02	0.787707	-1.131e-01	0.106726	-5.828e-02	0.706992
anno 2016	2.561e-01	0.081592 .	-1.166e-01	0.077838 .	1.067e-01	0.386342	-1.226e-01	0.080355 .	-1.028e-01	0.507363
anno 2017	3.080e-01	0.036288 *	-2.570e-01	0.000102 ***	1.560e-02	0.899236	-1.052e-01	0.133276	-1.655e-01	0.285910
anno 2018	4.191e-01	0.004412 **	-2.565e-01	0.000106 ***	2.625e-01	0.033142 *	-1.159e-01	0.098272 .	-1.215e-01	0.433164
anno 2019	5.159e-01	0.000461 ***	-3.091e-01	2.99e-06 ***	2.049e-01	0.096380 .	-1.268e-03	0.985570	-3.092e-01	0.046467 *
latitudine:longitudine	2.399e-01	< 2e-16 ***	-2.477e-02	0.034062 *	8.120e-02	0.000458 ***	6.292e-02	< 2e-16 ***	-4.537e-03	0.904337
% maschi	6.339e-02	3.76e-13 ***	-2.403e-03	0.670469	7.303e-02	2.58e-12 ***	9.267e-03	0.126019	1.521e-02	0.321327
età	-8.303e-03	0.303312	2.261e-03	0.586954	-4.283e-02	2.52e-05 ***	-3.525e-02	2.00e-11 ***	4.256e-02	0.002679 **
% ispanici	3.058e-02	6.09e-13 ***	2.441e-02	< 2e-16 ***	2.544e-02	1.97e-09 ***	2.375e-02	< 2e-16 ***	8.219e-02	< 2e-16 ***
% popolazione bianca	-9.536e-03	0.037481 *	-2.748e-03	0.075143 .	1.082e-02	0.014360 *	-5.313e-03	3.14e-06 ***	-7.172e-03	0.330598
% afroamericani	3.682e-02	1.54e-13 ***	2.508e-02	< 2e-16 ***	4.399e-02	< 2e-16 ***	2.006e-02	< 2e-16 ***	5.513e-02	7.97e-11 ***
% residenti in affitto	-7.733e-03	0.041143 *	1.269e-02	4.59e-14 ***	-1.513e-03	0.736770	7.758e-03	1.38e-05 ***	-1.170e-03	0.804799
tasso di disoccupazione	5.622e-02	0.000237 ***	5.182e-02	< 2e-16 ***	-1.420e-03	0.881328	2.374e-02	0.000486 ***	2.421e-02	0.294817
% residenti in età non lavorativa	-6.608e-02	< 2e-16 ***	2.192e-03	0.559545	-5.021e-03	0.485837	2.856e-03	0.515172	1.742e-02	0.104762
reddito	-5.530e-06	0.001643 **	-5.785e-06	2.21e-05 ***	-2.501e-05	3.34e-06 ***	-1.010e-05	1.95e-08 ***	1.357e-05	0.004880 **
% residenti in povertà	9.400e-02	< 2e-16 ***	1.968e-02	1.49e-08 ***	3.435e-02	3.06e-07 ***	1.859e-02	5.66e-05 ***	2.315e-03	0.835199
% che si reca a lavoro con mezzi pubblici	3.526e-02	0.000243 ***	2.158e-02	3.57e-16 ***	2.234e-02	6.36e-05 ***	1.443e-02	1.22e-10 ***	2.390e-02	0.001914 **
% che si reca a lavoro a piedi	5.272e-02	5.28e-08 ***	-2.192e-02	1.93e-07 ***	3.943e-02	2.89e-05 ***	5.394e-02	< 2e-16 ***	8.703e-02	3.60e-08 ***
% lavoratori da casa	1.242e-01	3.75e-15 ***	3.116e-02	4.14e-05 ***	9.414e-02	1.02e-10 ***	-4.918e-02	1.13e-06 ***	3.611e-02	0.146685
n° di fermate dei mezzi pubblici	7.422e-02	2.73e-10 ***	2.294e-01	< 2e-16 ***	1.016e-01	1.08e-08 ***	1.326e-01	< 2e-16 ***	2.998e-01	< 2e-16 ***
n° di ATM	1.041e-01	4.04e-07 ***	8.301e-02	6.62e-06 ***	1.385e-01	2.29e-05 ***	2.087e-01	4.35e-06 ***	1.068e-01	0.002830 **

Bibliografia

[1] Simone Aliprandi. *Il fenomeno Open Data: indicazioni e norme per un mondo di dati aperti*. Ledizioni, 2014.

[2] Adelchi Azzalini e Bruno Scarpa. *Data analysis and data mining: An introduction*. OUP USA, 2012.

[3] Marina Bassi. «Open Data: cosa sono, come sfruttarli e stato dell'arte in Italia». In: *Forum PA* (2019).

[4] Gary S. Becker. «Crime and punishment: An economic approach». In: *The economic dimensions of crime*. Springer, 1968, pp. 13–68.

[5] Aura Bertoni. «La scommessa dell'Open Data». In: *Il Sole 24* (2014).

[6] Simonetta Biagio. «Open Data: i Paesi europei verso l'età della maturità». In: *Il Sole 24 Ore Info Data* (2017).

[7] Marit Blank. *Open Data Maturity – Report 2019*. European Data Portal. 2019.

[8] Leo Breiman. «Random forests». In: *Machine learning* 45.1 (2001), pp. 5–32.

[9] Jesse Brush. «Does income inequality lead to more crime? A comparison of cross-sectional and time-series analyses of United States counties». In: *Economics letters* 96.2 (2007), pp. 264–268.

[10] Wendy Carrara, Wae San Chan, Sander Fischer e Eva Van Steenbergen. *Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources*. European Data Portal. 2015.

[11] Chhaya Chauhan e Smriti Sehgal. «A review: Crime analysis using data mining techniques and algorithms». In: *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE. 2017, pp. 21–25.

[12] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin e Michael Chau. «Crime data mining: a general framework and some examples». In: *computer* 37.4 (2004), pp. 50–56.

[13] Jongmook Choe. «Income inequality and crime in the United States». In: *Economics Letters* 101.1 (2008), pp. 31–33.

[14] Mario Coccia. «Violent crime driven by income Inequality between countries». In: *Turkish Economic Review* 5.1 (2018), pp. 33–55.

[15] Cristina Da Rold. «Open science e COVID-19: primo database aperto sugli interventi di ogni paese del mondo». In: *Il Sole 24 Ore Info Data* (2020).

[16] Susana De Juana-Espinosa e Sergio Luján-Mora. «Open government data portals in the European Union: Considerations, development, and expectations». In: *Technological Forecasting and Social Change* 149 (2019), p. 119769.

[17] Michela Finizio. «La miniera degli Open Data pubblici». In: *Il Sole 24* (2015).

[18] Belton M. Fleisher. «The effect of income on delinquency». In: *The American Economic Review* 56.1/2 (1966), pp. 118–137.

[19] *Global Report Fourth Edition*. Open Data Barometer, The World Wide Web Foundation. 2017.

[20] Hossein Hassani, Xu Huang, Emmanuel S. Silva e Mansi Ghodsi. «A review of data mining applications in crime». In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 9.3 (2016), pp. 139–154.

[21] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[22] Esther Huyer e Laura Van Knippenberg. *The Economic Impact of Open Data*. European Data Portal. 2020.

[23] Morgan Kelly. «Inequality and crime». In: *Review of economics and Statistics* 82.4 (2000), pp. 530–539.

[24] Max Kuhn e Kjell Johnson. *Applied predictive modeling*. Vol. 26. Springer, 2013.

[25] Guy Lansley e James Cheshire. *An Introduction to Spatial Data Analysis and Visualisation in R*. Consumer Data Research Centre (CDRC), 2016.

[26] Ming-Jen Lin. «Does unemployment increase crime? Evidence from US data 1974–2000». In: *Journal of Human resources* 43.2 (2008), pp. 413–436.

[27] James Manyika, Michael Chui, Peter Groves, Diana Farrell, Steve Van Kuiken e Elizabeth Almasi Doshi. «Open data: Unlocking innovation and performance with liquid information». In: *McKinsey Global Institute* 21 (2013), p. 116.

[28] Sonia Montegiove. «Il valore dei dati per comprendere la pandemia da Coronavirus». In: *Ingenium* (2020).

[29] Mitzi Morris, Katherine Wheeler-Martin, Dan Simpson, Stephen J. Mooney, Andrew Gelman e Charles DiMaggio. «Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan». In: *Spatial and spatio-temporal epidemiology* 31 (2019), p. 100301.

[30] Maurizio Napolitano. «Ma non basta l'apertura dei dati per aver fatto Open data». In: *Il Sole 24* (2014).

[31] Shyam Varan Nath. «Crime pattern detection using data mining». In: *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*. IEEE. 2006, pp. 41–44.

[32] Luigi Pace e Alessandra Salvan. *Introduzione alla statistica II*. Milano: Cedam, 2001.

[33] Steven Raphael e Rudolf Winter-Ebmer. «Identifying the effect of unemployment on crime». In: *The Journal of Law and Economics* 44.1 (2001), pp. 259–283.

[34] Gianni Rusconi. «L'Europa degli open data: un mercato potenziale da 325 miliardi». In: *Il Sole 24* (2015).

[35] Adolfo Sachsida, Mario Jorge Cardoso de Mendonça, Paulo R.A. Loureiro e Maria Bernadete Sarmiento Gutierrez. «Inequality and criminality revisited: further evidence from Brazil». In: *Empirical Economics* 39.1 (2010), pp. 93–109.

[36] Joanne Savage, Stephanie K. Ellis e Kevin H. Wozniak. «The role of poverty and income in the differential etiology of violence: an empirical test». In: *Journal of poverty* 23.5 (2019), pp. 384–403.

[37] Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel e Kenneth C. Lichtendahl Jr. *Data mining for business analytics: concepts, techniques, and applications in R*. John Wiley & Sons, 2017.

[38] Marco Trabucchi. «L'intelligenza artificiale italiana in campo per elaborare i dati del Covid-19». In: *Il Sole 24* (2020).

[39] Luca Tremolada. «Open data: solo l'11% delle informazioni nel mondo e aperto e accessibile a tutti». In: *Il Sole 24 Ore Info Data* (2014).

[40] Luca Tremolada. «L'Europa dei dati: otto Paesi europei su dieci hanno regole sugli Open data». In: *Il Sole 24 Ore – Info Data* (2017).

[41] Sara Ungaro. «Open Data: un'opportunità che il mercato deve sfruttare». In: *Agenda Digitale* (2015).

[42] *Vademecum Open Data – come rendere aperti i dati delle pubbliche amministrazioni*. Formez PA. 2011.

[43] Simon N. Wood. «Stable and efficient multiple smoothing parameter estimation for generalized additive models». In: *Journal of the American Statistical Association* 99.467 (2004), pp. 673–686.

[44] Simon N. Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

Ringraziamenti

Vorrei dedicare queste ultime righe dell'elaborato a chi in questi mesi mi ha sostenuto e supportato, nonostante il difficile periodo.
Un ringraziamento particolare va alla mia relatrice Dott.ssa Mariangela Guidolin per l'infinita pazienza e disponibilità durante tutto il percorso di tesi. Senza i suoi consigli e incoraggiamenti non sarei mai riuscita a concludere questo elaborato.
Vorrei poi ringraziare i miei genitori che hanno sempre supportato le mie decisioni pur non avendo mai compreso a fondo cosa studiassi.
Un ringraziamento speciale va anche a mia sorella Egle, che a modo suo ha cercato di risollevarmi di morale nei periodi più difficili in cui credevo che non ce l'avrei fatta.
Vorrei anche ringraziare tutti gli amici e parenti che hanno contribuito a rendere un po' più spensierati i giorni di questo anno un po' particolare segnato dal Covid. In particolare un ringraziamento di cuore va a Silvia per le passeggiate pomeridiane e le merende in compagnia. Vorrei ringraziare anche Francesca per le lunghe chiacchierate e per essere sempre stata presente anche se a distanza.
Infine il ringraziamento più grande va al mio ragazzo Simone che è sempre stato presente per sostenermi e incoraggiarmi.